Applied Mathematics and Mathematical Computation 2

Order Stars

A. Iserles
and
S. P. Nørsett

Springer-Science+Business Media, B.V.

Order Stars

APPLIED MATHEMATICS AND MATHEMATICAL COMPUTATION

Editors

R.J. Knops, K.W. Morton

Texts and monographs at graduate and research level covering a wide variety of topics of current research interest in modern and traditional applied mathematics, in numerical analysis, and computation.

1 Introduction to the Thermodynamics of Solids *J.L. Ericksen* (1991)

2 Order Stars *A. Iserles and S.P. Nørsett* (1991)

(Full details concerning this series, and more information on titles in preparation are available from the publisher)

Order Stars

A. ISERLES

Lecturer in Applied Numerical Analysis,
University of Cambridge, UK

and

S.P. NØRSETT

Professor of Numerical Analysis,
Norwegian Institute of Technology

Springer-Science+Business Media, B.V.

First edition 1991

Originally published by Chapman & Hall in 1991.
Softcover reprint of the hardcover 1st edition 1991
Bury St. Edmunds
ISBN 978-0-412-35260-7 ISBN 978-1-4899-3071-2 (eBook)
DOI 10.1007/978-1-4899-3071-2

British Library Cataloguing in Publication Data

Iserles, A.
Order stars: Theory and application.
I. Title II. Nørsett, S.P.
515

Library of Congress Cataloging-in-Publication Data
Available

Contents

Preface

According to Hilbert's dictum, the scaffolding should be invisible in a mathematical edifice. Less kind interpretation of this common principle of writing and presenting mathematics is that we should always strive to do it back-to-front, forever wise after the event. Nobody should be allowed to see the seams in the supposedly seamless robe or compare authors' intentions with the outcome of their endeavour. In particular, the short piece of prose occasionally labelled 'Preface' or 'Forward' ought to be written *after* the main body of the book. And so it is, and we, the authors, can reflect (with much trepidation) on an enterprise that for us is finally over.

Order stars have been originally introduced in the context of numerical solution of ordinary differential equations and, as far as many numerical analysts are concerned, they still belong there. It is our case in this book that the scope of order stars ranges considerably wider and that the cornerstone of the order star theory is a function-theoretic interpretation of complex approximation theory. An application to numerical analysis is a matter of serendipity, not of essence.

Suppose that an analytic function is approximated by another analytic function. As approximation theorists – and as users of approximation theory, inclusive of numerical analysts – we might be interested in several attributes of approximation. The first and foremost is interpolation, other being possibly the location of poles, zeros and other 'exceptional points', boundedness or contractivity in portions of the complex plane (sometimes termed, in deliberately vague terms, 'stability') etc. All these refer to properties of the approximant in various portions of the complex plane that, at first glance, might appear to be unrelated. However, a brief reflection affirms that a connection *does* exist, since the objects of our study are analytic functions.

Unlike, say, splines, analytic functions are devoid of truly local behaviour. Every scintilla of information about an analytic function somewhere in the complex plane (or, to be properly more general, the Riemann sphere) influences its shape everywhere. The situation is similar to certain

oriental mystical theories and, indeed, to that occidental counterpart of mysticism, relativity theory. The universe (in our case, an analytic function on the Riemann sphere) is at one with itself and any minute local disturbance to its 'wave function' emits ripples that reverberate forever and everywhere. Every interpolation condition, every stability requirement, each zero and pole sends 'vibrations' that echo all across the area of interest and influence the approximant there.

The theory of order stars is a means of elucidating and analysing this action-at-distance, by translating the interconnectedness of various approximation-theoretical features into a simple geometrical language. Its relevance to numerical methods for ordinary and partial differential equations is based exclusively upon the 'preprocessing' of underlying numerical-analytic problems into the terminology of approximation theory.

All this sounds somewhat mysterious and opaque, as it indeed should. Had it been possible to explain exactly, with full rigour and ample detail, the minutiae of order stars in a page or two, there would have been no need to write a whole book about it! Thus, the last few paragraphs are just a marker, placing us at the correct spot in the mathematical universe, and a pointer to the next 250 pages.

The theory of order stars has expanded greatly since its introduction in 1978. The process of writing this book was a splendid opportunity to assemble the many available results into a coherent whole. This involved the occasional discovery of lacunae and gaps but – we are glad to say – no cracks that a dollop or two of mathematical mortar cannot put right.

Writing a book involves a great deal of strategic decisions. Every mathematical theory rests on many foundations and it spawns unexpected applications. Thus, how much to assume on part of the reader and how much motivation, explanation and application to provide in the text? There is no absolute answer to such questions since, of course, there is no such thing as a 'standard' or 'average' reader. Each reader has the privilege of bringing her or his mathematical culture and idiosyncrasies to bear on the book, and, as we hasten to admit, so have the authors. It has been our implicit assumption throughout this book that the reader is an intelligent being, with sound mathematical education and readiness to embed new material into a wider mathematical framework. However, we have also assumed that the reader, like the authors, is far too busy (and lazy) to pursue the ideal course of action of studying a whole library shelf each time we mention Riemann surfaces or Toeplitz operators or Pick–Nevanlinna approximation or Runge–Kutta methods. Thus, we tried hard not just to make this book as self-contained as reasonably possible, but also to explain the motivation that underpins the surveyed topics. This provided us with a perfect excuse to depart occasionally from the set course of this book on tangential detours into some fascinating mathematical landscapes.

Throughout the whole book we tried hard to avoid the impression that its subject matter is complete and that all answers are known. Only dead theories are complete – order stars are alive and we expect them to be of important future use. A major motivation behind this book is to familiarize the mathematical community with an analytic tool that is capable of so many applications and we conclude this volume with a list of open problems which might be amenable to analysis with order stars.

We have always found order stars an exciting object to study and describe. This, in a sense, is a spurious statement: of course, *every* mathematician finds her or his current focus of attention, whatever it might be, the most fascinating event since the Big Bang and till (at least) the next week. Unless we have shared a sense of excitement in our topic, we should have better been doing something else! None the less, there is something special in the blend of geometry, approximation theory, numerical analysis and analytic function theory that is called order stars and the theory yields itself to surprising proofs and leads to unexpected insights. It was our ambition to share our enthusiasm and delight in order stars, and it is up to the reader to pronounce on the success of our enterprise.

Even mathematicians cannot stand in the void and, in our experience, the best – and tallest – standing position is on the shoulders of the acknowledged founders of the subject. Order stars were born in the context of numerical analysis of ordinary differential equations and our first and foremost debt is to the generation of numerical mathematicians who have established numerical ODEs as a mathematical discipline. It is this ability, that these days we all take for granted, to answer numerical questions with full mathematical rigour, that has led to order stars. Thus, it is only fair to acknowledge the seminal contribution of John Butcher, Germund Dahlquist, Peter Henrici, Jack Lambert, Hans Stetter and many others, not just to the discipline of numerical ODEs as a whole but also to our personal development as numerical mathematicians.

Standing on tall shoulders is occasionally slippery and one should always hold hands of one's colleagues to maintain the balance. The happy – and relatively small – band of order-starist made this book possible. First and foremost we mention Ernst Hairer and Gerhard Wanner, two-thirds of the original trio that introduced the whole subject, but also Rolf Jeltsch, Olavi Nevanlinna, Mike Powell, Rosie Renaut, Kosie Smit, Klaus-Günther Strack and Gil Strang. It is their work, as much as ours, that made order stars into – we believe – a story worth telling.

Our next acknowledgement is tinged with disappointment. We have learnt by example from books of Gil Strang (himself member of the order star fraternity) how to write mathematics – unstuffy, enthusiastic, devoid of pomposity but, none the less, without yielding an inch of mathematical rigour. Of course, we did not succeed to emulate the master, hence the

disappointment. A skeptical reader might contemplate how worse this book might have been without Gil Strang's influence...

The staff at Chapman and Hall, in particular Elizabeth Johnston, displayed toward us the right mixture of friendly encouragement and firmness to keep us on the straight and narrow. It is much to their credit that a vague idea has been translated into tangible pages of text.

Several of our colleagues and friends read parts of the manuscript, enlightened us with their criticisms and remarks, acted as a sounding board and – most importantly – encouraged us to believe that this has been a worthwhile enterprise. A pride of place belongs to Bill Morton, not just as an editor of the relevant Chapman and Hall series but also as a foremost authority in numerical analysis, whose advice we valued greatly. Brad Baxter, Martin Buhmann and Gustaf Söderlind read various versions of the manuscript and their remarks were always helpful and useful. Bojan Orel helped us with the mysteries of symbolic manipulation.

The Department of Applied Mathematics and Theoretical Physics of Cambridge University and the Institute of Mathematical Sciences of the Norwegian Institute of Technology provided us with a splendid and stimulating working environment and, during our frequent visits to each other's institution, with warm hospitality and mathematical home-away-from-home. We also wish to thank the Norwegian Research Council for Science and Humanities (Project D.02.08.001) for their valuable support.

An excursion in a mathematical garden sounds like a simple and straightforward idea. Soon, alas, the realization dawns that you carry in your backpack a heavy load of administrative, teaching and editorial duties, missed deadlines, pending correspondence, papers to be written and ideas that just *must* be pursued... Hence the constant temptation to pause, meander or even turn back, and hence the need for encouragement and support. We are both fortunate in life-partners who, although blissfully ignorant of all things mathematical and busy with their own careers, never failed with their encouragement. Like stars – and order stars – they were always there when needed. And this is why we are so delighted to dedicate this book to our wives, Dganit and Edith.

Having said all this, we should perhaps sign off with the usual disclaimer. All the help notwithstanding, the buck always stops with the authors. Thus, we wish to exclaim with Casca in Shakespeare's *Julius Cæsar*

> The fault, dear Brutus, is not in our stars,
> But in ourselves...

1

Introduction

He doth elect
The beautiful and fortunate,
And the sons of intellect,
And the souls of ample fate,
Who the Future's gates unbar –
Minions of the Morning Star.

From *Initial, dæmonic and celestial love* by Ralph Waldo Emerson (1803–1882)

Differential equations are undoubtedly the most important and consistently successful means of modelling the physical universe by mathematics. Movement of celestial bodies, the shape of a ship's wake, stress and lift acting on aircraft wings, spread of epidemic through a large population, percolation of crude oil in semi-permeable rock, nuclear processes in the core of stars, transmission of electric pulses down a nerve fibre, the fickle behaviour of stock markets, tear and wear of turbine blades – to all intents and purposes the list is infinite, limited merely by our imagination.

Only the simplest differential equations possess solutions that can be written explicitly in terms of familiar functions: the standard practice of undergraduate lecture courses in differential equations, whereby solutions are always produced explicitly, is misleading! In practice it is imperative to discretize the underlying equations and use approximate solutions. It is important to emphasize here and now that, in real-life situations, a reliable computational solution is likely to be as 'good' as an exact solution, since the data (the geometry of the underlying domain, initial and boundary values, various constants featuring in the equations themselves) usually contain errors and are limited by the imprecision of our measurements. Furthermore, the nature of the application in hand specifies the required finite level of accuracy: it is, after all, futile to design a bridge to a tolerance of a single micron, but modelling a VLSI circuit calls for far greater precision.

It is clear that computational algorithms are central to any practical use of differential equations. Moreover, as long as they are designed and applied correctly, their inexactness has little or no effect on their ultimate

application. Computation and application need not be uneasy bedfellows! In this book we investigate certain mathematical aspects and consequences of the analysis of 'correct' discretizations. Or, to be exact, we present a theory that originated in such analysis. Like many mathematical theories, it eventually acquired life and momentum of its own and has led in unexpected directions. Having stated this, it is only fair to emphasize that differential equations will seldom be far from the main line of argument in this book.

The design of computational algorithms for differential equations pits accuracy against stability. The first and most obvious requirement in this endeavour is high local precision. It is typically ascertained by matching terms in a Taylor expansion and quantized under the concept of order. Alas, this is only part of the story: true, the quality of local approximation increases as the discretization becomes progressively finer – but this is simultaneous with an increase in the number of variables. In other words, pointwise errors decrease, but the overall number of points grows, and, to prevent breakdown, we need theoretical justification to infer from local to global. This adds a crucial condition, usually termed stability, and whose exact nature is strongly problem-dependent.

It is stability in unison with non-trivial order that ensures convergence of the numerical scheme to the exact solution of a well-posed differential equation.

A sound approach to the design of a numerical scheme for a given differential system follows a set pattern. First, we clarify the rules of the game. These are imposed by considerations that originate in the nature of the underlying application and in the computational environment: the computer architecture, the interplay of reliability and cost, considerations of computer speed and storage. These set the stage for further investigation. Next, we narrow the focus (perhaps arbitrarily) to a family of 'candidate' methods, e.g. Runge–Kutta or multistep schemes or finite elements. Finally, we search in that family for the highest-order method, subject to stability. The last stage sometimes requires a degree of simplification and theoretical contortions, since an answer is not always available – indeed, the question is not always well posed – in a general setting. Instead, it might be necessary to restrict the attention to a simplified test equation. Even then, the analysis is far from straightforward and it calls for a hefty measure of intricate mathematical analysis.

Order and stability analysis for a whole range of problems can be transformed into questions in analytic function theory. The precise nature of this transformation is not of our present concern: it will be presented in considerable detail in Chapters 3–7, employing methods from classical, functional and harmonic analysis. The crux of the matter, so to speak, is that both order and stability can be described as features of a certain complex function, which is determined in a well-defined sense by the underlying numerical

method. A few examples will help to clarify matters.

Example 1.1 The ordinary differential system

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \quad \mathbf{y}(t_0) = \mathbf{y}_0, \tag{1.1}$$

is given. Here $\mathbf{y}$ might be a scalar or vector function – its dimensionality makes no difference whatsoever to our discussion! We approximate its solution by a multistep method

$$\sum_{\ell=0}^{N} \alpha_\ell \mathbf{y}_{i+\ell-N} = h \sum_{\ell=0}^{N} \beta_\ell \mathbf{f}(t_{i+\ell-N}, \mathbf{y}_{k+\ell}), \tag{1.2}$$

where $h > 0$ is a time-step and $\mathbf{y}_k$ approximates the exact solution of (1.1) at $t_k = t_0 + kh$. Let

$$\begin{aligned} r(w) &:= \sum_{\ell=0}^{N} \alpha_\ell w^\ell, \\ s(w) &:= \sum_{\ell=0}^{N} \beta_\ell w^\ell. \end{aligned}$$

Then (1.2) is of order p if $r(\exp z) - zs(\exp z) = cz^{p+1} + \mathcal{O}(z^{p+2})$, where the local error constant c is non-zero, and it is zero-stable if all the zeros of the polynomial r are in the closed complex unit disc and all the zeros of unit modulus are simple. According to a celebrated theorem of Dahlquist (1956), (1.2) is convergent if and only if it is consistent (i.e. $p \geq 1$) and zero-stable. Moreover, the method is A-stable if all the N zeros of $r(w) - zs(w)$ are within the open complex unit disc for all $z \in C$ such that $\mathrm{Re}\, z < 0$ (Hairer *et al.*, 1987; Henrici, 1962). A-stability is paramount when we wish to integrate stiff systems of ordinary differential equations, choosing the step length with regard only to local accuracy. ◇

Example 1.2 The system (1.1) is solved with an s-stage Runge–Kutta method

$$\begin{aligned} \mathbf{k}_i &= \mathbf{f}\left(t_k + c_i h, \mathbf{y}_k + h \sum_{j=1}^{s} a_{i,j} \mathbf{k}_j\right), \quad i = 1, \ldots, s, \\ \mathbf{y}_{k+1} &= \mathbf{y}_k + h \sum_{i=1}^{s} b_i \mathbf{k}_i. \end{aligned}$$

Again, $h > 0$ is the step length and $\mathbf{y}_k \approx \mathbf{y}(t_k)$. An application of the method to the scalar linear test equation $y' = \lambda y$ produces the solution sequence $y_{k+1} = R(h\lambda) y_k$, where R is a rational function. If the Runge–Kutta method is of order p then $R(z) = \exp z + \mathcal{O}(z^{q+1})$ for some $q \geq p$

(this is a necessary, but not sufficient, consequence of order p), whereas A-stability is tantamount to $|R(z)| < 1$ for all $z \in \mathcal{C}$ such that $\mathrm{Re}\, z < 0$. $\diamond$

Example 1.3 The space derivative in the advection equation

$$\frac{\partial}{\partial t}u = \frac{\partial}{\partial x}u, \quad u(x,0) = \psi(x), \ -\infty < x < \infty,$$

is discretized by finite differences. This yields a semi-discretization of the form

$$u'_\ell = \frac{1}{\Delta x}\sum_{j=-r}^{s} \beta_j u_{\ell+j}, \quad \ell \in \mathcal{Z}, \tag{1.3}$$

where $u_\ell \approx u(\ell\Delta x, t)$. Let $G(z) := \sum_{j=-r}^{s} \beta_j z^j$. Then (1.3) is of order p if $G(z) = \log z + C(z-1)^{p+1} + \mathcal{O}\,(z^{p+2})$, $C \neq 0$, and it is stable if $\mathrm{Re}\, G(e^{i\theta}) \leq 0$ for all $-\pi \leq \theta \leq \pi$. The situation parallels that for ordinary differential equations: by the Lax equivalence theorem (Richtmyer and Morton, 1967) convergence equals stability and consistency. $\diamond$

Example 1.4 Both space and time are discretized in the advection equation, producing the fully-discretized scheme

$$u_\ell^{k+1} = \sum_{j=-r}^{s} \delta_j(\mu) u_{\ell+j}^k.$$

Here $u_\ell^k \approx u(\ell\Delta x, k\Delta t)$ and μ is the Courant number, $\mu = \frac{\Delta t}{\Delta x}$. Again, both order and stability can be expressed in terms of a single function: we set $H(z,\mu) := \sum_{j=-r}^{s} \delta_j(\mu) z^j$. Then the method is of order p if H approximates z^μ about $z = 1$ to that order, whereas it is stable if $|H(e^{i\theta})| \leq 1$ for all $-\pi \leq \theta \leq \pi$. $\diamond$

A pattern emerges: 'vital statistics' of a numerical method are encapsulated in the behaviour of a meromorphic function. Its order can be expressed by the degree of interpolation to a certain function at some complex point and its stability is equivalent to a condition (e.g. a bound on the modulus or the real part) in a portion of the complex plane. In other words, the relationship between order and stability is determined by the interplay of the behaviour of a meromorphic function in different parts of the complex plane.

Probably the most important property specific to analytic functions is the absence of distinction between local and global behaviour. Specify an analytic function in a neighbourhood (or on a countable set) and you have specified it everywhere in its (connected) domain of analyticity! Require a boundedness condition on the modulus in one portion of the complex plane and the interpolation is affected elsewhere.

The **theory of order stars**, the subject-matter of this book, is a means to elucidate the aforementioned 'action-at-a-distance' of a meromorphic function and its influence on stability and quality of approximation. Deferring the formal introduction of the theory to Chapter 2, we presently provide a simple example of its application.[1]

The importance of rational approximants to the exponential function has already been highlighted in Example 1.2. Let us designate by $\pi_{m/n}$ the set of all rational functions of the form P/Q, where $\deg P \leq m$, $\deg Q \leq n$ and $Q(0) = 1$. Given a function f, analytic at the origin, and integers $m, n \geq 0$, we choose $R_{m/n} \in \pi_{m/n}$ so as to maximize the order at the origin:

$$R_{m/n}(z) - f(z) = \mathcal{O}(z^{p+1})$$

where

$$p = \max\left\{q : \exists R \in \pi_{m/n} \text{ such that } R(z) - f(z) = \mathcal{O}(z^{q+1})\right\}.$$

We call $R_{m/n}$ an $[m/n]$ **Padé approximant** to f. If $p \geq m+n$ then $R_{m/n}$ is unique.[2] It will be proved in Chapter 4 that the Padé approximant to the exponential is of order $m+n$. With no danger of confusion, we reserve in the present chapter the notation $R_{m/n}$ for Padé approximants to the exponential.

A rational function R is A-**acceptable** if $|R(z)| < 1$ for all $z \in \mathcal{C}$ such that $\mathrm{Re}\, z < 0$. It is evident from Example 1.2 that an efficient solution of linear systems of ordinary differential equations by a Runge–Kutta method, without unduly restricting step length to retain stability, is facilitated by A-acceptability of the rational approximant to $\exp z$ that is implicit in the definition of the method.

A-acceptability of $R_{m/n}$ obviously requires that $m \leq n$. Birkhoff and Varga (1965) proved that the $[n/n]$ Padé approximant is A-acceptable, whereas Ehle (1969; 1973) and Nørsett (1975) demonstrated the A-acceptability of the $[(n-1)/n]$ and $[(n-2)/n]$ Padé approximants[3] and showed that $R_{(n-3)/n}$ and $R_{(n-4)/n}$ are not A-acceptable. This, and extensive computer experimentation, justified the **first Ehle conjecture** (Ehle, 1973), to the effect that $[m/n]$ Padé approximants to the exponential are A-acceptable if and only if $n-2 \leq m \leq n$.

We define the **order star** of $R \in \pi_{m/n}$ as the ordered triplet

$$\{\mathcal{A}_+, \mathcal{A}_0, \mathcal{A}_-\},$$

[1] The publication of this result in (Wanner *et al.*, 1978) not only introduced order stars for the first time, but also resolved a long-standing conjecture. It is a tribute to the subsequent success of order stars that we can term it 'simple'.

[2] Some authorities term $R_{m/n}$ a Padé approximant only if $p \geq m+n$ (Baker, 1975).

[3] A-acceptability proofs of $R_{m/n}$ for $n-2 \leq m \leq n$ follow from the Crouzeix–Ruamps theorem, which will be proved with order stars in Chapter 4.

where

$$\begin{aligned} \mathcal{A}_+ &:= \{z \in \mathcal{C} : |e^{-z}R(z)| > 1\}; \\ \mathcal{A}_0 &:= \{z \in \mathcal{C} : |e^{-z}R(z)| = 1\}; \\ \mathcal{A}_- &:= \{z \in \mathcal{C} : |e^{-z}R(z)| < 1\}. \end{aligned}$$

This defines a partition of the complex plane. The following four propositions are special cases of the theory of Chapter 2 and we state them here without proof:

Proposition 1.1 If R is an order-p approximation to the exponential function then the origin is adjoined by $p+1$ sectors of $\mathcal{A}_+$, separated by $p+1$ sectors of $\mathcal{A}_-$. All these sectors approach the origin with the asymptotic angle of $\pi/(p+1)$. □

Proposition 1.2 R is A-acceptable if and only if it is analytic in $\{z \in \mathcal{C} : \mathrm{Re}\, z < 0\}$ and $\mathcal{A}_+ \cap \{i\mathcal{R}\} = \emptyset$. □

Proposition 1.3 The number of zeros (poles) in a bounded connected component of $\mathcal{A}_-$ ($\mathcal{A}_+$) equals the number of interpolation points (i.e. points such that $R(z) = \exp z$) on its oriented boundary, counted with their multiplicity. □

Proposition 1.4 There are exactly two unbounded connected components, one of $\mathcal{A}_+$ and one of $\mathcal{A}_-$. Moreover, for every $\varepsilon > 0$ there exists $r_\varepsilon > 0$ such that for all $r \geq r_\varepsilon$ the segment $\{re^{i\theta} : -\frac{\pi}{2} + \varepsilon \leq \theta \leq \frac{\pi}{2} - \varepsilon\}$ is in $\mathcal{A}_-$, whereas the segment $\{re^{i\theta} : \frac{\pi}{2} + \varepsilon \leq \theta \leq \frac{3\pi}{2} - \varepsilon\}$ is in $\mathcal{A}_+$. □

Any function in $\pi_{m/n}$ possesses three features that are relevant to our investigation: it has a given number of zeros and poles (m and n respectively), it is either A-acceptable or not, and, finally, it has a specific order as an approximation to $\exp z$ at $z = 0$. All these features are reflected in the geometry of the order star. Figure 1.1 displays the order stars[4] for the $[k/(5-k)]$ Padé approximants, $k = 0, 1, \ldots, 5$. The reader is urged to identify the implications of the four propositions in the figure.

Let us suppose that $R \in \pi_{m/n}$ is an A-acceptable approximant of order p. A-acceptability and Proposition 1.2 imply that the pure imaginary axis separates components of $\mathcal{A}_+$. Let ω_- and ω_+ denote the number of sectors of $\mathcal{A}_+$ that approach the origin to the left and to the right of $i\mathcal{R}$, respectively. It follows from Propositions 1.1 and 1.2 that

$$\omega_- + \omega_+ = p + 1 \tag{1.4}$$

and

$$\omega_\pm - 1 \leq \omega_\mp \leq \omega_\pm + 1. \tag{1.5}$$

[4]The set $\mathcal{A}_+$ is shaded. See Table 2.1 for a more comprehensive definition of symbols.

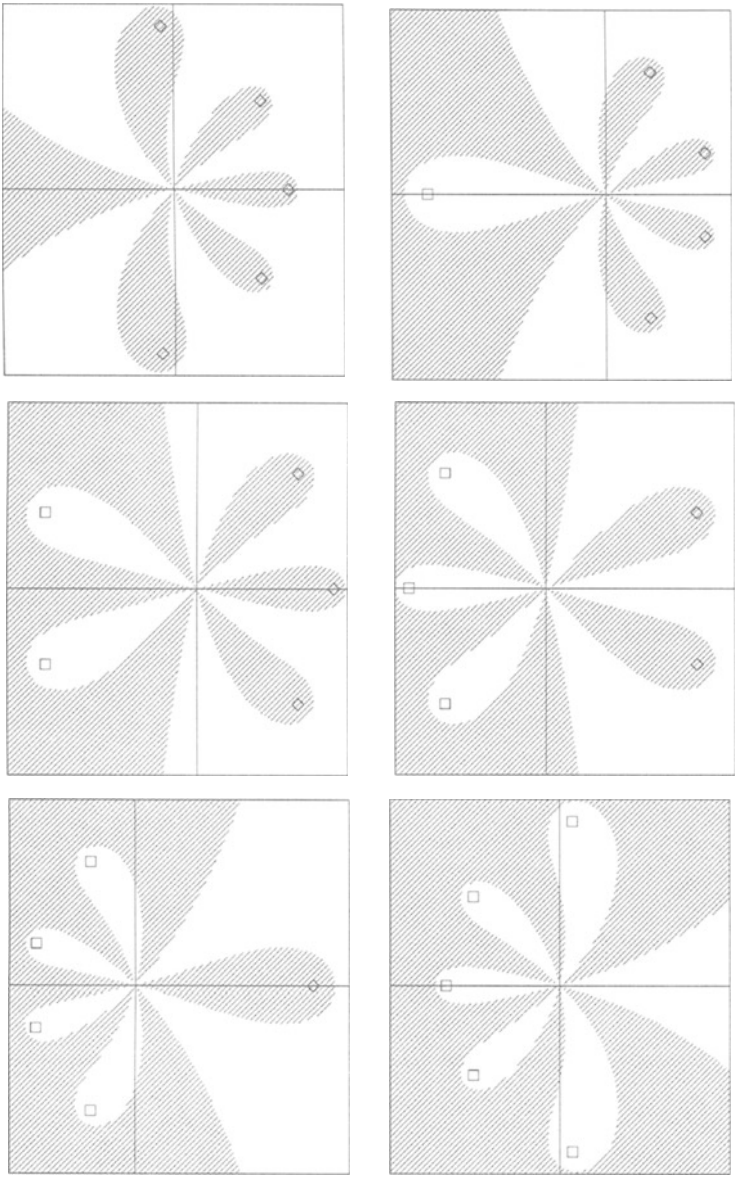


Figure 1.1 Order stars of the $[k/(5-k)]$ Padé approximants to $\exp z$ for $k = 0, 1, \ldots, 5$.

Moreover, by Propositions 1.2 and 1.4, all the sectors of $\mathcal{A}_+$ that approach the origin to the right of the pure imaginary axis belong to bounded components. Since the origin is an interpolation point, it is now a consequence of Proposition 1.3 that $\omega_+ \leq n$: every sector of $\mathcal{A}_+$ there 'costs' a pole. Thus, by (1.5), $\omega_- \leq n+1$ and (1.4) implies that

$$p \leq 2n. \tag{1.6}$$

Next, we turn our attention to the left half-plane. By Proposition 1.2 R is analytic there, hence no poles are allowed. Thus, by Propositions 1.2 and 1.4, all the ω_- sectors of $\mathcal{A}_+$ that adjoin the origin to the left belong to its single unbounded connected component. They separate $\omega_- - 1$ sectors of $\mathcal{A}_-$ that, consequently, belong to bounded connected components of $\mathcal{A}_-$. Each such sector must, by virtue of Proposition 1.3, be 'accounted for' by a zero of R, and the inequality $\omega_- - 1 \leq m$ follows. We now use (1.4) and (1.5) to argue that

$$p \leq 2m+2. \tag{1.7}$$

Given that $R = R_{m/n}$, the Padé approximant, it is true that $p = m+n$. It is now a trivial consequence of the inequalities (1.6–7) that $n-2 \leq m \leq n$, proving the first Ehle conjecture (Wanner *et al.*, 1978).

This example is by way of an *hors d'œuvre*: relatively straightforward, typical of the subsequent fare and, hopefully, whetting the appetite. In Chapter 2 we start with the main course, presenting the general theory of order stars in the complex plane. Chapters 3–5 are concerned with stability barriers that arise in numerical solution of ordinary differential equations, whereas Chapters 6 and 7 consider applications of the theory of order stars to computational schemes for hyperbolic and parabolic partial differential equations, respectively.

Stability barriers motivated the introduction of order stars (Wanner *et al.*, 1978) and much of the subsequent work on the subject. However, as often in mathematics, a new theory is bound to provide surprising insight into unexpected (or, at the very least, unintended) areas. Chapters 8 and 9 are devoted to two such instances. Order stars can be used to examine Padé tableaux of general meromorphic functions and obtain bounds on the size of square blocks therein. Chapter 8 is devoted to this subject-matter, central to the theory of Padé approximation. In Chapter 9 we develop a general theory of contractive approximation in the complex plane, linking order stars with the classical Pick–Nevanlinna interpolation theory.

This book is not intended as an account of an 'open-and-closed' theory. Indeed, it is the belief of the authors that much good is bound to accrue from future work on order stars. Chapter 10 lists open problems and comments on directions for future research.

General Order Stars

'Mais non. Des petites choses dorées qui font rêvasser les fainéants. Mais je suis sérieux, moi! Je n'ai pas le temps de rêvasser.'
'Ah! des étoiles?'
'C'est bien ça. Des étoiles.'

From *Le Petit Prince* by Antoine de Saint-Exupéry (1900–1943)

2.1 Order stars of the first kind

In the present chapter we develop elements of the theory of order stars, the working tool in the remainder of this book. The first two sections, on order stars of the first kind and on the behaviour of order stars near an essential singularity, follow to a large extent the work of Iserles (1985a), whereas the third section, on order stars of the second kind, is a straightforward extension of that paper.

A complex function is said to be **essentially analytic** if it is meromorphic in the closed complex plane $\mathrm{cl}\,\mathcal{C} := \mathcal{C} \cup \{\infty\}$, except perhaps for a finite set of essential singularities. Thus, analyticity may break down only at the aforementioned essential singularities and, possibly, at (at most) a countable set of poles, that are allowed to accumulate exclusively at the essential singularities.

Let f be an essentially analytic function that is being approximated by a rational function $R \in \pi_{m/n}$. Explicitly excluding the case $R \equiv f$, we set

$$\rho(z) := \frac{R(z)}{f(z)}, \quad z \in \mathrm{cl}\,\mathcal{C}.$$

ρ is itself essentially analytic. It inherits all essential singularities of f, its zeros are either the zeros of R or the poles of f, whereas its poles are either the poles of R or the zeros of f – that is, unless zeros, say, of f and R coincide and can be factored out. More importantly, interpolation of f by R is 'encrypted' in ρ, since $R(z_0) = f(z_0)$ is equivalent to $\rho(z_0) = 1$.

The **order star of the first kind** of $\{f, R\}$ is defined as the partition $\{\mathcal{A}_+, \mathcal{A}_0, \mathcal{A}_-\}$ of the closed complex plane, where

$$\mathcal{A}_+ \;:=\; \{z \in \mathrm{cl}\,\mathcal{C} \,:\, |\rho(z)| > 1\};$$

$$\begin{aligned} \mathcal{A}_0 &:= \{z \in \mathrm{cl}\mathcal{C} : |\rho(z)| = 1\}; \\ \mathcal{A}_- &:= \{z \in \mathrm{cl}\mathcal{C} : |\rho(z)| < 1\}. \end{aligned}$$

Excluding the trivial case of $R \equiv f$, it is easy to see that $\mathcal{A}_0$, the common boundary of $\mathcal{A}_+$ and $\mathcal{A}_-$, is a union of closed simple Jordan curves.

Harking back to the example of Chapter 1 – the first Ehle conjecture – and its motivation, we note the following features of approximation that are of interest and that should be traced in the geometry of the order star: (a) interpolation; (b) 'stability'; (c) location of zeros and poles; and (d) behaviour near essential singularities.

The point $z_0 \in \mathrm{cl}\mathcal{C}$ is said to be an **interpolation point** of degree $p \geq 1$ if f is analytic in a neighbourhood of z_0 and

$$\begin{aligned} |z_0| < \infty\text{:} \quad R(z) &= f(z) + C(z - z_0)^p + \mathcal{O}(|z - z_0|^{p+1}); \\ z_0 = \infty\text{:} \quad R(z) &= f(z) + Cz^{-p} + \mathcal{O}(|z|^{-p-1}), \end{aligned}$$

where $C \neq 0$. Clearly, $z_0 \in \mathcal{A}_0$. Moreover, the degree of interpolation can be 'read' from the geometry of the order star, by counting the number of sectors of $\mathcal{A}_+$ and $\mathcal{A}_-$ that approach z_0.

The **index** $\iota(z)$ of a point $z \in \mathcal{A}_0$ is defined, intuitively speaking, as the number of sectors of $\mathcal{A}_-$, say, adjoining z. Unfortunately, the aforementioned 'definition' of the index is incomplete at an essential singularity[1] and we need a somewhat more rigorous approach. Let $z \in \mathcal{A}_0$ be bounded. We define $\wp(z)$ as the set of all non-negative integers k having the property that for every $\varepsilon > 0$ there exists $\delta \in (0, \varepsilon]$ such that there are exactly k arcs of the circle $\{\zeta \in \mathcal{C} : |\zeta - z| = \delta\}$ that belong to $\mathcal{A}_-$. Likewise, if $\infty \in \mathcal{A}_0$ then $\wp(\infty)$ is the set of all non-negative integers having the property that for every $r > 0$ there exists $s \geq r$ such that there are exactly k arcs of the circle $\{z \in \mathcal{C} : |z| = s\}$ that belong to $\mathcal{A}_-$. We define the index of $z \in \mathcal{A}_0$ by

$$\iota(z) := \min\{k : k \in \wp(z)\}. \tag{2.1}$$

Example 2.1 Let

$$f(z) = \frac{1}{\prod_{k=0}^{\infty}(1 - q^k z)}$$

where $q \in \mathcal{C}$, $|q| < 1$. The function is a reciprocal of an entire function, as can be easily argued from the Hadamard factorization (cf. Section 2.2). The Taylor expansion of f about the origin can be produced by employing

[1] It is easy to see that an essential singularity z must belong to $\mathcal{A}_0$: ρ is essentially analytic, hence it is analytic in a punctured neighbourhood $\mathcal{V}$ of z. By a theorem of Weierstrass it comes in $\mathcal{V}$ arbitrarily close to any complex value (Ahlfors, 1966), hence there are points of both $\mathcal{A}_+$ and of $\mathcal{A}_-$ arbitrarily close to z.

a standard technique from the theory of q-hypergeometric functions (Slater, 1966; Gasper and Rahman, 1990).

Let a and q be complex numbers. The q-**factorial coefficient** $(a;q)_n$, where n is either a non-negative integer or ∞, is defined as the product

$$(a;q)_n := \prod_{k=0}^{n-1} \left(1 - aq^k\right)$$

(of course, $(a;q)_0 = 1$). Thus, $f(z) = 1/(z,q)_\infty$. Let

$$g(z) := \sum_{k=0}^{\infty} \frac{z^k}{(q;q)_k}.$$

We assume that $|q| < 1$. It is easy to see, e.g. by the d'Alembert test, that g is analytic in $|z| < 1$. Moreover,

$$g(z) - g(qz) = \sum_{k=0}^{\infty} \frac{1}{(q;q)_k}(1 - q^k)z^k = \sum_{k=1}^{\infty} \frac{z^k}{(q;q)_{k-1}} = zg(z),$$

implying that

$$g(z) = (1 - z)^{-1} g(qz)$$

and it follows readily by induction that

$$g(z) = \left\{ \prod_{k=0}^{n-1} (1 - q^k z)^{-1} \right\} g(q^n z), \quad n \geq 1.$$

We now let $n \to \infty$ and exploit $\lim_{n\to\infty} g(q^n z) = g(0) = 1$ to argue that $g \equiv f$, providing the Taylor expansion of f.

Figure 2.1 displays the order star of f (with $q = \frac{1}{2}$) and $R = R_{3/3}$, its [3/3] Padé approximant. Although it cannot be observed easily from the figure, $+\infty$ is approached by progressively smaller 'bubbles' of $\mathcal{A}_-$, enlosing the zeros of ρ at 2^k, $k = 0, 1, \ldots$, embedded in a 'substrate' of $\mathcal{A}_+$. Evidently, $\wp(\infty) = \{0, 1\}$ (it is, in fact, possible to prove this rigorously by standard methods of analytic function theory) and $\iota(\infty) = 0$.

This example is interesting, since it displays an unavoidable lack of symmetry in our definition: if ρ is analytic at $z \in \mathcal{A}_0$ then it is approached by the same number of sectors of $\mathcal{A}_-$ and $\mathcal{A}_+$ – the choice of $\mathcal{A}_-$ in the definition of $\iota(z)$ is a matter of arbitrariness. The present example demonstrates that this is no longer the case if z is an essential singularity. $\diamond$

Before resuming the orderly course of our exposition, it is appropriate to explain the notation used in Figure 2.1 and elsewhere in this book. As already stated, the shaded area stands for $\mathcal{A}_+$ and the white for $\mathcal{A}_-$. Further symbols that are used to denote various attributes of order stars are listed in Table 2.1.

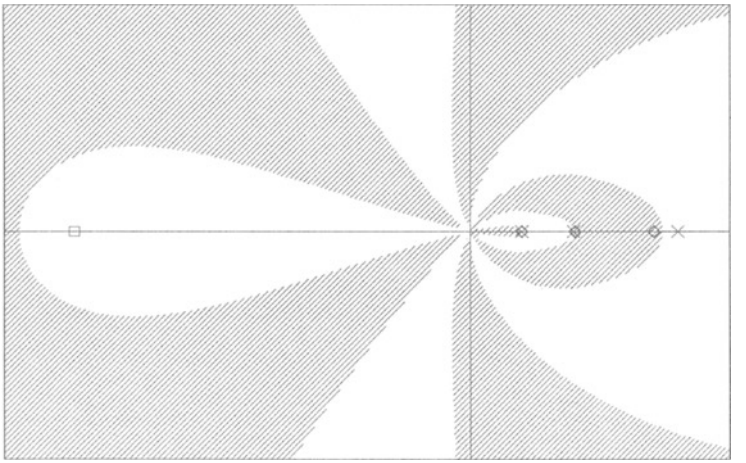


Figure 2.1 Order star of the [3/3] Padé approximant to the theta-like function $\prod_{k=0}^{\infty}(1-2^{-k}z)^{-1}$.

◇	:	Poles of R
□	:	Zeros of R
×	:	Poles of f
■	:	Zeros of f
●	:	Interpolation points

Table 2.1 List of symbols.

To prevent excessive cluttering of figures by a mass of detail, the symbols for interpolation points are used only if desirable from the context. We do not use a special symbol for essential singularities, since their position is invariably obvious.

Let $z \in \mathcal{A}_0$ and $p = \iota(z) > 0$. If ρ is analytic at z and the point is approached by precisely p sectors of $\mathcal{A}_-$ and p sectors of $\mathcal{A}_+$, each of asymptotic angle of π/p, we say that z is **regular**.

Proposition 2.1 Let z_0 be a zero of f of multiplicity $k \geq 0$ (inclusive of the case $k = 0$, i.e. $f(z_0) \neq 0$). If z_0 is an interpolation point of degree $p \geq \max\{1, k+1\}$ then $\iota(z_0) = p - k$ and z_0 is regular.

Proof. We might assume without loss of generality that z is bounded – otherwise we might conformally map $z \mapsto z^{-1}$. Thus

$$f(z) = C_1(z - z_0)^k + \mathcal{O}(|z - z_0|^{k+1}), \quad C_1 \neq 0,$$

and it follows from the definition of ρ that

$$\rho(z) = 1 + \frac{C}{C_1}(z - z_0)^{p-k} + \mathcal{O}(|z - z_0|^{p-k+1}), \quad C \neq 0.$$

Let $s := p - k$, $C_2 := C/C_1 \neq 0$. Letting $z = z_0 + re^{i\theta}$, $r > 0$, we obtain

$$|\rho(z)| = 1 + r^s \mathrm{Re}\{C_2 e^{is\theta}\} + \mathcal{O}(r^{s+1}) := \phi(r, \theta). \tag{2.2}$$

Given any $\varepsilon > 0$ and sufficiently small r, it is clear that $z \in \mathcal{A}_+$ for $\mathrm{Re}\{C_2 e^{is\theta}\} > \varepsilon$ and $z \in \mathcal{A}_-$ for $\mathrm{Re}\{C_2 e^{is\theta}\} < -\varepsilon$. Since $\mathrm{Re}\{C_2 e^{is\theta}\}$ changes sign $2s$ times for equally spaced values of $\theta \in [0, 2\pi)$, the proposition follows, except that $\mathcal{A}_+$ or $\mathcal{A}_-$ might contain some sectors that approach z_0 with asymptotically zero angle (cusps). To complete the proof we need to rule out this possibility.

Cusps may occur only if, as $r \to 0$, some values of θ that obey the equation $\phi(r, \theta) = 1$ tend to coalesce. This implies that a zero of $\mathrm{d}\phi(r, \theta)/\mathrm{d}\theta$ tends to a root of $\phi(r, \theta)$, leading to a contradiction: (2.2), in tandem with the analyticity of ρ in the vicinity of z_0, implies that

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\phi(r, \theta) = -sr^s \mathrm{Im}\{C_2 e^{is\theta}\} + \mathcal{O}(r^{s+1})$$

and $\mathrm{Re}\{C_2 e^{is\theta}\}$, $\mathrm{Im}\{C_2 e^{is\theta}\} = 0$ cannot coexist, since $C_2 \neq 0$. This rules out the existence of cusps. □

Let $\mathcal{V}$ be an open subset of $\mathrm{cl}\, \mathcal{C}$ such that f is analytic in $\mathcal{V}$ and $|f| \equiv 1$ along the boundary of $\mathcal{V}$, except perhaps for a finite number of essential singularities. By the maximal modulus principle (Ahlfors, 1966) an analytic function attains the maximum of its modulus on the boundary. Consequently, $|f(z)| < 1$ for all $z \in \mathcal{V}$ and f maps $\mathrm{cl}\, \mathcal{V}$ onto the closed complex unit disc. We call such a function a $\mathcal{V}^*$-contraction. Moreover, any function that analytically maps $\mathcal{V}$ into the interior of the unit disc is termed a $\mathcal{V}$-contraction. $\mathcal{V}$-contractivity is an important attribute, because of its connection with numerical stability, and frequently it is essential to require its preservation by the approximant R.

Example 2.2 Recall the definition of A-acceptability from Chapter 1: it means nothing but preservation of $\mathcal{C}^-$-contractivity, where $\mathcal{C}^-$ is the left half-plane, by a rational approximant to $\exp z$. Another instance of contractivity is provided by Example 1.4: $H(z, \mu)$ approximates z^μ and, for

the sake of stability, it should retain its contractivity in the unit disc. Note that in both cases f – be it $\exp z$ or z^μ – is $\mathcal{V}^*$-contractive, whereas stability requires only the weaker condition of $\mathcal{V}$-contractivity. ◇

Proposition 2.2 Let f be a $\mathcal{V}^*$-contraction. Then R is a $\mathcal{V}$-contraction if and only if it is analytic in $\mathcal{V}$ and $\mathcal{A}_+ \cap \partial\mathcal{V} = \emptyset$.

Proof. ρ is analytic in any closed domain of $\mathcal{V}$, $|\rho(z)| = |R(z)|$ along $\partial\mathcal{V}$, with the possible exception of a finite set of points, and the proposition is a trivial consequence of the maximal modulus principle. □

We next establish a relationship between the loci of zeros and poles of ρ and the pattern of interpolation. The connected components of $\mathcal{A}_+$ and of $\mathcal{A}_-$ are called $\mathcal{A}_+$-regions and $\mathcal{A}_-$-regions respectively. Such a region is said to be **analytic** if ρ is analytic along its boundary.

Recall that interpolation points lie on $\mathcal{A}_0$, the joint boundary of $\mathcal{A}_-$ and $\mathcal{A}_+$. We say that an $\mathcal{A}_-$-region or an $\mathcal{A}_+$-region is of **multiplicity** L if its directed boundary passes through exactly L interpolation points. Note that these points need not be distinct: the directed boundary of such a region may 'visit' the same point several times.

Proposition 2.3 The multiplicity L of an analytic $\mathcal{A}_+$-region equals the number of poles of ρ, counted with their multiplicity, inside the domain, and $1 \leq L < \infty$. An identical statement, with poles replaced by zeros, is valid for analytic $\mathcal{A}_-$-regions.

Proof. Let $\mathcal{U}$ be an analytic $\mathcal{A}_+$-region of multiplicity L. We parametrize the positively-oriented boundary of $\mathcal{U}$ as $\gamma(t) = \gamma_R(t) + i\gamma_I(t)$, $0 \leq t \leq 1$, where both γ_R and γ_I are real. Since ρ is analytic on $\partial\mathcal{U}$ and $\mathcal{U}$ is a level set of $|\rho|$, γ is analytic in $[0,1]$, except possibly at a finite number of points (that correspond to 'corners' of $\mathcal{U}$).

Let $\mathbf{v}(t) := (\gamma_R(t), \gamma_I(t))$ and $\mathbf{n}(t) := (\gamma_I'(t), -\gamma_R'(t))$ be the tangent and the (outwardly-pointing) normal vector at $\gamma(t)$, respectively. Recall the definition of $\mathcal{A}_+$: as we approach $\partial\mathcal{U}$ from within the domain, $|\rho|$ decreases locally. Thus, $\log|\rho|$ decreases along $\mathbf{n}$ near the boundary.

Representing ρ in polar coordinates,

$$\rho(z) = r(x,y)e^{i\varphi(x,y)}, \quad z = x + iy \in \mathrm{cl}\,\mathcal{U},$$

it follows that

$$\frac{\partial}{\partial \mathbf{n}} \log r(x,y) < 0. \tag{2.3}$$

Analyticity of ρ is equivalent to the satisfaction of the Cauchy–Riemann equations. Expressing these in polar coordinates, we obtain along γ

$$\frac{\partial}{\partial x} \log r \;=\; \frac{\partial}{\partial y}\varphi;$$

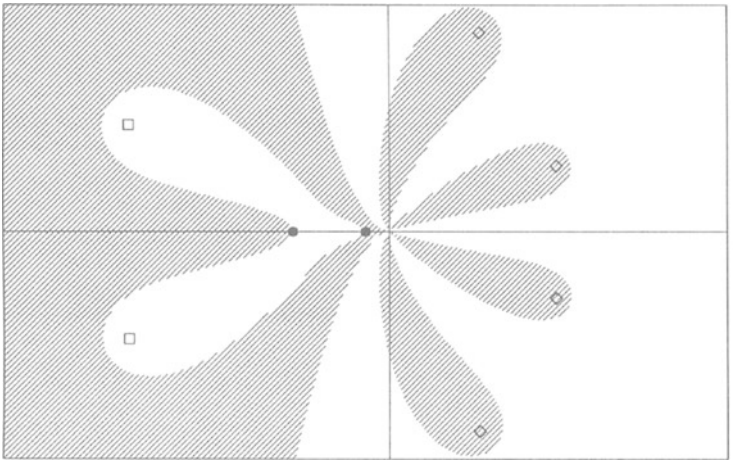


Figure 2.2 Order star of a $\pi_{2/4}$ approximant to $\exp z$ with order 4 at the origin and two negative interpolation points.

$$\frac{\partial}{\partial y}\log r = -\frac{\partial}{\partial x}\varphi.$$

Therefore

$$\begin{aligned}\frac{\partial}{\partial \mathbf{n}}\log r &= \gamma_I'\frac{\partial}{\partial x}\log r - \gamma_R'\frac{\partial}{\partial y}\log r \\ &= \gamma_I'\frac{\partial \varphi}{\partial y} - \gamma_R'\left(-\frac{\partial \varphi}{\partial x}\right) = \frac{\partial \varphi}{\partial \mathbf{v}}.\end{aligned}$$

Thus, by virtue of (2.3), $\arg\rho$ decreases strictly monotonically at all points of analyticity of γ.

The function ρ is meromorphic in $\mathcal{U}$ and γ is a union of closed curves in $\mathrm{cl}\,\mathcal{C}$. Thus, by the argument principle (Ahlfors, 1966), the variation of $\arg\rho$ along γ is $-2\pi L_1$, where L_1 is the number of poles of ρ in $\mathcal{U}$ (since, by the definition of $\mathcal{A}_+$, ρ has no zeros there). Moreover, interpolation points on γ are precisely the points where ρ equals 1; thus $\arg\rho$ is an integer multiple of 2π. Since $\arg\rho$ is monotone on γ, it equals an integer multiple of 2π precisely L_1 times for $t \in [0,1)$, proving that L_1, the number of poles, is equal to the multiplicity L of the $\mathcal{A}_+$-region $\mathcal{U}$.

The inequality $L \geq 1$ is another consequence of the strict monotonicity of $\arg \rho$ almost everywhere along the boundary. Moreover, $L < \infty$, otherwise there will be an infinite number of poles in $\mathcal{U}$ with an accumulation point on the boundary. An accumulation point of poles is an essential singularity, and it is ruled out by the analyticity of ρ along $\partial\mathcal{U}$.

The proof for $\mathcal{A}_-$-regions follows in an identical manner. □

This is perhaps a place to mention the pedigree of the last proposition, the key to the whole theory of order stars. Although it has been stated and proved within the present context in the paper of Wanner *et al.* (1978), the strictly monotonic variation of the argument along level curves of analytic functions and its interplay with the argument principle were observed 33 years earlier, in an important paper of Dame Mary Cartwright (1935) on p-valent analytic functions.

Figure 2.2 displays an order star of a rational approximant to the exponential with three interpolation points: of degree 5 at the origin[2] and of order 1 at two negative points. It should by now be straightforward to identify the interplay of interpolation, A-acceptability and loci of zeros and poles in the geometry of this order star.

2.2 Essential singularities

The greatest diversity of behaviour of an order star is displayed near essential singularities. It is evident from Figure 2.1 that essential singularities need not be regular. Actually, the situation is far more complicated and the determination of the index of an essential singularity is sometimes a non-trivial task.

Let ϑ be an essential singularity of f. It might be an accumulation point of zeros or of poles, but we stipulate that it is not an accumulation point of *both* zeros and poles. The extension of this framework to the realm of essentially analytic functions is usually straightforward.

Clearly, we may decompose

$$\rho(z) = \rho^*(z)\rho^\bullet(z), \tag{2.4}$$

where ρ^* has an essential singularity at ϑ, either ρ^* or $1/\rho^*$ is analytic in $\mathrm{cl}\, C \setminus \{\vartheta\}$ and $\rho^\bullet$ is analytic at ϑ. The behaviour of ρ near ϑ is described solely by the behaviour of ρ^*. Further, the function $z \mapsto 1/(z - \vartheta)$ maps ρ^* conformally to a function which is entire (or has an entire reciprocal). Consequently, it is enough to analyse the behaviour of order stars for entire functions.

Throughout this section, unless stated otherwise, we assume that $1/\rho^*$ is entire (this 'corresponds' to an entire f) and examine the behaviour of

[2]Hence the rational function approximates $\exp z$ to order 4, one less than the degree!

the order star of ρ^* as $|z| \to \infty$. The translation of these results to the realm of ρ is straightforward.

Let g be an entire function. Set

$$M(r) := \max_{|z|=r} \{|g(z)|\}.$$

The order of g is defined (Hille, 1962) as

$$\lambda(g) := \limsup_{r\to\infty} \frac{\log\log M(r)}{\log r}. \tag{2.5}$$

The order is a non-negative real number. Clearly, $g(z) = \exp z^K$, K a non-negative integer, yields $\lambda(g) = K$. However, there is much more to order than merely 'comparison' with the asymptotic behaviour of the exponential. For example, the theta-like function $g(z) = \prod_{k=0}^{\infty}(1 - q^k z)$, where $|q| < 1$, is of order 0, but the essential singularity at infinity is absolutely genuine.

Lemma 2.4 (Hille, 1962) Let $g(z) = \sum_{k=0}^{\infty} g_k z^k$ be an entire function of bounded order. Then

$$\lambda(g) = \limsup_{k\to\infty} \frac{k\log k}{\log |g_k|^{-1}}. \tag{2.6}$$

□

Example 2.3 The **Mittag-Leffler function** (Erdélyi *et al.*, 1953) is defined as

$$E_\alpha(z) := \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + 1)},$$

where $\alpha > 0$ and Γ is the Gamma function. Definition (2.5) is, to all intents and purposes, useless in the present case. Fortunately, the order of E_α can be easily determined from Lemma 2.4. By the Stirling formula (Abramowitz and Stegun, 1965)

$$\Gamma(z) = \sqrt{2\pi} e^{-z} z^{z-\frac{1}{2}} \left\{1 + \mathcal{O}\left(z^{-1}\right)\right\}.$$

Therefore $\log |g_k| = -\alpha k \log k(1 + o(1))$, where

$$E_\alpha(z) = \sum_{k=0}^{\infty} g_k z^k,$$

and (2.6) implies that $\lambda(E_\alpha) = 1/\alpha$. ◇

Further examples will feature in Chapter 8.

The definition of order can be extended to a bounded essential singularity ϑ. Although it can be done formally, replacing (2.5) in an obvious

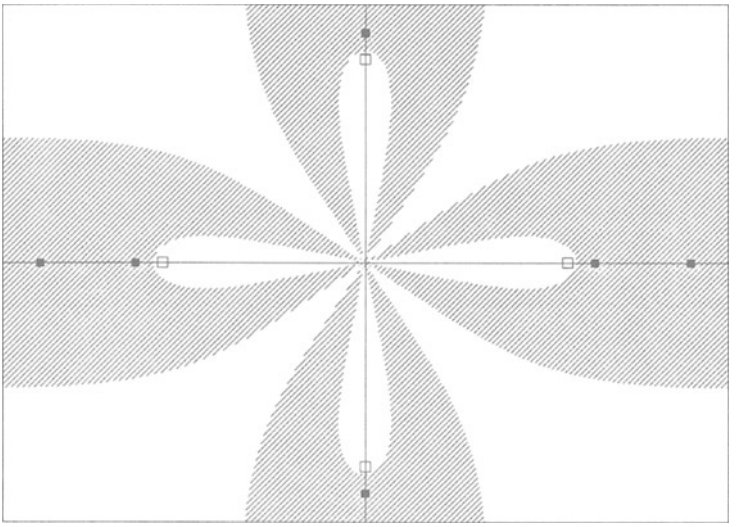


Figure 2.3 Order star for the [4/0] Padé approximant to $z^{-2}\sin z^2$, with $\lambda = 2$ and $\iota(\infty) = 4$. Note that the order of approximation is 7.

manner, probably the simplest approach is by mapping ϑ conformally to infinity. We denote by $\lambda(\rho^*, \vartheta)$ the order of ρ^* at ϑ. Moreover, and without fear of confusion, given an essentially analytic function ρ with an essential singularity at ϑ, we say that the order of ρ at ϑ is the order of the function ρ^* that obeys (2.4) (or of its reciprocal, whichever is analytic in $\mathrm{cl}\,\mathcal{C} \setminus \{\vartheta\}$). This notational device will be employed in the sequel – at present, as already stated, we assume for simplicity that $\vartheta = \infty$.

Proposition 2.5 Let ρ have an essential singularity at ϑ. It is then true that

$$\iota(\vartheta) \leq \max\{1, 2\lambda(\rho, \vartheta)\}. \tag{2.7}$$

Proof. A classical theorem of Ahlfors (Goluzin, 1969) states that an entire, nonconstant function of order λ has at most 2λ finite asymptotic values at ∞. This can be trivially translated to 2λ finite asymptotic values at ϑ in the present framework. Thus, since the asymptotic behaviour of ρ is that of either an entire function or its reciprocal, at most $\max\{1, 2\lambda\}$ distinct

sectors of $\mathcal{A}_-$ may approach ϑ. □

The bound of (2.7) is attainable (without loss of generality, at infinity) for every integer $\lambda \geq 1$ by the function $f(z) = z^{-\lambda} \sin z^\lambda$ (cf. Figure 2.3). However, frequently it is far too generous.

An important type of essential singularity is characterized by the asymptotic behaviour

$$\frac{1}{\rho^*(z)} = az^b e^{cz^K}(1 + o(1)), \quad |z| \gg 0, \tag{2.8}$$

where $a, b, c \in C$, $a, c \neq 0$. In that case we say that the essential singularity is of **exponential type** K. Note that (2.8) implies order K.

Proposition 2.6 If an essential singularity ϑ is of exponential type K then $\iota(\vartheta) = K$ and ϑ is regular.

Proof. Without loss of generality we let $\vartheta = \infty$. Given $z = re^{i\phi}$, (2.8) yields for $r \gg 0$

$$\begin{aligned} -\log|\rho^*(z)| &= \log|a| - \phi \operatorname{Im} b + (\operatorname{Re} b)\log r \\ &\quad + r^K \operatorname{Re}\{ce^{iK\phi}\} + o(1) := \psi(r, \phi). \end{aligned} \tag{2.9}$$

Thus, as r approaches ∞, for every $\varepsilon > 0$, $\operatorname{Re}\{ce^{iK\phi}\} > \varepsilon$ implies that $z \in \mathcal{A}_+$, whereas $\operatorname{Re}\{ce^{iK\phi}\} < -\varepsilon$ gives $z \in \mathcal{A}_-$. This produces K sectors of $\mathcal{A}_+$ and K sectors of $\mathcal{A}_+$, each of an asymptotic angle π/K.

As cusps can be easily excluded by the method of proof of Proposition 2.1, both regularity and $\iota(\infty) = K$ follow. □

The discrepancy between the upper bound of Proposition 2.5 and the sharp – and substantially lower – result of Proposition 2.6 can be attributed to essential singularities that are accumulation points of poles (or zeros or both, if the simplifying restriction of this section is lifted). Let g be an entire function of bounded order λ. By the Hadamard factorization theorem (Hille, 1962) it can be represented as

$$g(z) = e^{q(z)} z^K \prod_{k=1}^{L} E\left(\frac{z}{\kappa_k}, p\right),$$

where q is a polynomial in π_λ, K and p are nonnegative integers, $p \leq \lambda$, L is either a nonnegative integer or ∞, $\kappa_1, \kappa_2, \ldots$ are all the zeros of g and the Weierstrass prime factors E are defined by

$$\begin{aligned} E(z, 0) &:= 1 - z, \\ E(z, p) &:= (1 - z) \exp\left(\sum_{\ell=1}^{p} \frac{z^\ell}{\ell}\right), \quad p \geq 1. \end{aligned}$$

Let us assume that ρ^* has a finite number of poles, $\kappa_1, \ldots, \kappa_L$, say. It can be factorized into

$$\frac{1}{\rho^*(z)} = e^{\tilde{q}(z)} z^K \prod_{k=1}^{L} \left(1 - \frac{z}{\kappa_k}\right),$$

where

$$\tilde{q}(z) = \begin{cases} q(z) & : p = 0, \\ q(z) + \displaystyle\sum_{\ell=1}^{p} \frac{1}{\ell} \left(\sum_{k=1}^{L} \frac{1}{\kappa_k^\ell} \right) z^\ell & : p \geq 1. \end{cases}$$

Consequently $1/\rho^*$ is of exponential type $\deg \tilde{q}$ and the conditions of Proposition 2.6 are satisfied.

An infinite number of poles may sometimes coexist with regularity of the essential singularity. For example, let us assume that all the poles $\kappa_1, \kappa_2, \ldots$ of ρ^* are real, of the form

$$\cdots \leq \kappa_3 \leq \kappa_2 \leq \kappa_1,$$

and that there exist $C > 0$ and $a \in (0,1)$ such that the number of κ_k's in every interval of the form $[-r, \infty)$ is $Cr^a(1+o(1))$ for all $r \gg 0$. We examine the Hadamard factorization of the entire function $1/\rho^*$. By considering the distribution of the κ_k's it is elementary to prove that

$$\sum_{k=1}^{\infty} \sum_{\ell=1}^{p} \frac{1}{\ell} \left(\frac{z}{\kappa_k}\right)^\ell$$

sums up to a pth degree polynomial. Thus, the exponential components in the factors $E(z/\kappa_k, p)$ can be 'added' to the leading exponential term (in other words, we may assume without loss of generality that $p = 0$) and

$$\frac{1}{\rho^*(z)} = e^{-q(z)} \prod_{k=1}^{\infty} \left(1 - \frac{z}{\kappa_k}\right), \tag{2.10}$$

where $q \in \pi_\lambda$, say.

Proposition 2.7 Let ρ^* obey (2.10), such that there are $Cr^a(1 + o(1))$ points κ_k in each interval $[-r, \kappa_1]$ for $r \gg 0$. Then $\lambda \leq \iota(\infty) \leq \lambda + 1$. Moreover, if $0 < a < \frac{1}{2}$ then $\iota(\infty) = \lambda$ and ∞ is regular.

Proof. We exploit a theorem of Pólya and Szegö (Hille, 1962, p. 206): provided the κ_k's are distributed as in the statement of the proposition, it is true that

$$\log \prod_{k=1}^{\infty} \left(1 - \frac{z}{\kappa_k}\right) = \frac{C\pi}{\sin a\pi} z^a (1 + o(1)),$$

uniformly for all z in the set $\{z \in \mathcal{C} : |\arg z| \leq \pi - \varepsilon\}$ for any $\varepsilon > 0$ (both the logarithm and the power function are evaluated on their principal branch). Let $z = re^{i\theta}$, $r > 0$, $0 \leq |\theta| < \pi$. We have

$$-\log|\rho^*(z)| = \operatorname{Re} q(z) + \frac{C\pi}{\sin a\pi} r^a (1 + o(1)). \tag{2.11}$$

This is reminiscent of (2.9) and we may now proceed similarly to the proof of Proposition 2.6, except that (2.11) is invalid on the negative real half-axis. Thus, there are λ sectors of $\mathcal{A}_+$ and λ sectors of $\mathcal{A}_-$ approaching ∞ with the asymptotic angles of π/λ, except that these might be a cusp along the negative half-axis. Consequently, $\iota(\infty) \in \{\lambda, \lambda + 1\}$ and if $\iota(\infty) = \lambda$ then ∞ is regular.

Let us finally stipulate that $0 < a < \frac{1}{2}$. Then (Hille, 1962, p. 207) there exists a sequence of real points x_ℓ such that $\lim_{\ell\to\infty} x_\ell = -\infty$ and

$$\lim_{\ell\to\infty} \prod_{k=1}^{\infty} \left|1 - \frac{x_\ell}{\kappa_k}\right| = \infty.$$

Thus, $\wp(\infty) = \{\lambda, \lambda + 1\}$ and it follows from the definition of index that $\iota(\infty) = \lambda$. □

Proposition 2.8 Let

$$\frac{1}{\rho^*(z)} = \prod_{k=1}^{\infty} \left(1 - \frac{z}{\kappa_k}\right),$$

where $\kappa_k < 0$ for all $k = 1, 2, \ldots$, $\lim_{k\to\infty} \kappa_k = -\infty$ and $\lambda(\rho^{*-1}) = 1$. Then $\iota(\infty) = 1$ and ∞ is regular.

Proof. We let $z = re^{i\theta}$, $r > 0$; thus

$$|\rho(z)^*|^{-2} = \prod_{k=1}^{\infty} \left\{1 - \frac{2r}{\kappa_k}\cos\theta + \left(\frac{r}{\kappa_k}\right)^2\right\}.$$

Let $\varepsilon > 0$ be given. Since $1/\kappa_k \to 0$ and $\kappa_k < 0$, there exists N_ε such that for every $k \geq N_\varepsilon$ it is true that $-\varepsilon \leq r/\kappa_k < 0$. Consequently for all $k \geq N_\varepsilon$

$$\cos\theta > \varepsilon \quad \Rightarrow \quad 1 - \frac{2r}{\kappa_k}\cos\theta + \left(\frac{r}{\kappa_k}\right)^2 > 1 + 3\varepsilon^2 > 1;$$

$$\cos\theta < -\varepsilon \quad \Rightarrow \quad 1 - \frac{2r}{\kappa_k}\cos\theta + \left(\frac{r}{\kappa_k}\right)^2 < 1 - \varepsilon^2 < 1.$$

Thus, $\cos\theta > \varepsilon \Rightarrow z \in \mathcal{A}_+$ and $\cos\theta < -\varepsilon \Rightarrow z \in \mathcal{A}_-$. To prove the proposition we need to rule out the possibility of cusps of $\mathcal{A}_+$ and $\mathcal{A}_-$ fitting into the order star near the pure imaginary axis. Since ρ^* is a real

analytic function, the order star is symmetric with respect to $\mathcal{R}$. In other words, the number of cusps of $\mathcal{A}_-$, say, must be even. Let us assume that it is positive. Then $\iota(\infty) \geq 3$. However, by Proposition 2.5, $\lambda(1/\rho^*) = 1$ implies $\iota(\infty) \leq 2$, hence contradiction. Consequently, no such cusps exist and the proof is complete. □

Let $\rho^*(z) = \prod_{k=1}^{\infty}(1 - z/\kappa_k)^{-1}$, where the κ_k's, in conformity with the conditions of Proposition 2.8, are negative and $\lim_{k\to\infty} \kappa_k = -\infty$. It is easy to verify the remaining condition of the proposition, namely that $\lambda(1/\rho^*) = 1$: this is equivalent to

$$\sum_{k=1}^{\infty} \frac{1}{\kappa_k} < \infty \tag{2.12}$$

and

$$\sum_{k=1}^{\infty} \frac{1}{|\kappa_k|^a} = \infty \tag{2.13}$$

for all $a < 1$ (Hille, 1962).

Example 2.4 Let

$$\rho^*(z) = \prod_{k=2}^{\infty} \left(1 + \frac{z}{k\,(\log k)^b}\right)^{-1}, \quad b > 1.$$

To verify (2.12) and (2.13) it is probably easiest to employ the 2^n test of calculus: the series $\sum c_k$ converges $\Leftrightarrow$ the series $\sum 2^k c_{2^k}$ converges. This ascertains that $\lambda(1/\rho^*) = 1$. Since the remaining conditions of Proposition 2.8 are satisfied, it follows that $\iota(\infty) = 1$ and that ∞ is regular. No order star involving ρ^* has been plotted: the infinite product converges so slowly that even the evaluation of a single value of ρ^* in $|z| \geq 1$ is computationally highly intensive! ◇

At the risk of repeating ourselves, we emphasize that the conditions of the last two propositions need hold only asymptotically for $|z| \gg 0$: we may multiply ρ^* by any function which is analytic and nonzero at ∞ without disturbing the validity of Proposition 2.7 or Proposition 2.8, as the case might be. Thus, ρ^* is allowed zeros, non-real poles and even extra essential singularities, as long as the 'action' is away from infinity. Furthermore, an identical statement is valid for finite essential singularities and the behaviour of the order star (for ρ) in the vicinity of its essential singularities can be synthesized from the results of this section.

2.3 Order stars of the second kind

Essentially analytic functions are allowed to become singular at poles and at essential singularities, but they are devoid of branch cuts. Unfortu-

nately, some functions that are important in stability analysis of evolutionary differential equations possess branch cuts (cf. Examples 1.3 and 1.4 and Chapters 5–7), particular examples being $\log z$ and z^{μ}. Branch cuts can be 'unravelled' by moving from the complex plane to a Riemann surface: this approach is adopted in Chapter 5. In the present section we survey another, more straightforward, technique.

Let f be a complex function, whose precise definition is open for the time being, and R an approximation thereof. We set $\tilde{\rho}(z) := R(z) - f(z)$ for all $z \in \mathrm{cl}\, \mathcal{C}$ where f and R are well defined. The **order star of the second kind** is the triplet $\{\tilde{\mathcal{A}}_+, \tilde{\mathcal{A}}_0, \tilde{\mathcal{A}}_-\}$, where

$$\begin{aligned} \tilde{\mathcal{A}}_+ &:= \{z : \mathrm{Re}\, \tilde{\rho}(z) > 0\}; \\ \tilde{\mathcal{A}}_0 &:= \{z : \mathrm{Re}\, \tilde{\rho}(z) = 0\}; \\ \tilde{\mathcal{A}}_- &:= \{z : \mathrm{Re}\, \tilde{\rho}(z) < 0\}. \end{aligned} \tag{2.14}$$

Definition (2.14) is valid in the whole of $\mathrm{cl}\, \mathcal{C}$ if f is essentially analytic and R is rational (or, with minor amendments to the exposition of Section 2.1, if R is also essentially analytic). However, its main purpose is validity for a more extensive set of functions. We require f to be essentially analytic in $\mathrm{cl}\, \mathcal{C}$, except that it is allowed branch cuts and branch points. Moreover, we stipulate that $\mathrm{Re}\, f$ is continuous across the (open) branch cuts. Such functions do not possess an accepted name and, for the purpose of the present exposition, we call them **slit functions**.

An important collection of slit functions is the class **N**. Following Akhiezer (1965), we say that a function g is in the class **N** if it maps analytically the half-plane $\{z \in \mathcal{C} : \mathrm{Im}\, z > 0\}$ into its closure. Such functions have the representation

$$g(z) = a + bz + \int_{-\infty}^{\infty} \frac{1 + \tau z}{\tau - z} \mathrm{d}\mu(\tau), \quad \mathrm{Im}\, z > 0,$$

where $a, b \in \mathcal{R}$, $b \geq 0$ and μ is a distribution – a right-continuous, non-decreasing function of bounded variation. Although this definition can be extended to the lower half-plane by setting $g(z) = \overline{g(\bar{z})}$, $\mathrm{Im}\, z < 0$, this is usually quite a poor idea: if g is analytic in a real neighbourhood then it should be extended to the lower half-plane by analytic continuation. Either way, it can be proved that $\mathrm{Re}\, g$ is continuous across the real axis (except, possibly, at poles, essential singularities and branch points), while $\mathrm{Im}\, g$ might be discontinuous there. Functions in class **N** are important because of their connection with Riesz resolvents of self-adjoint operators, with the Pick–Nevanlinna interpolation theory (cf. Chapter 9) and with the Hamburger and Stieltjes moment problems (Akhiezer, 1965). A subset of **N** that has been extensively studied is the class of **Stieltjes functions**. Let us assume

that μ is constant for $\tau > 0$, $\int_{-\infty}^{0}(1+\tau^2)\mathrm{d}\mu(\tau) < \infty$ and set

$$\tilde{\mu}(\tau) := \int_{-\infty}^{\tau}(1+\xi^2)\mathrm{d}\mu(\xi), \quad a = \int_{-\infty}^{\infty}\frac{\xi}{1+\xi^2}\mathrm{d}\tilde{\mu}(\xi), \quad b = 0.$$

Then

$$g(z) = \int_{-\infty}^{0}\frac{\mathrm{d}\tilde{\mu}(\tau)}{\tau - z}$$

is said to be a Stieltjes function (Akhiezer, 1965).

Example 2.5 It is easy to verify that

$$\log z = (1+z)\int_{-\infty}^{0}\frac{\mathrm{d}\log(1-\tau)}{\tau - z},$$

therefore $\log z/(1+z)$ is Stieltjes. Typically there is no need to identify the underlying distribution to verify that a given function belongs to the class **N**, merely to check that it maps the open upper half-plane to its closure. An easy example is $g(z) = z^{\mu}$ for $\mu \in (0,1)$. $\diamond$

Note that the reciprocal of an **N** function is also a slit function. Moreover, if g is **N** and h is a real essentially analytic function (i.e. h is real across $\mathcal{R}$) then $g(h(z))$, $h(g(z))$, $h(z)g(z)$ and $h(z)+g(z)$ are all slit functions. Thus, both $\log z$ and z^{μ}, $\mu \in \mathcal{R}$, are allowed. Note that such functions are usually well defined only in $C \setminus \mathcal{R}$. However, definition (2.14) is valid across all of $\mathrm{cl}\,C$, precisely because the real part of any slit function is well defined there.

Order stars of the first and the second kind are linked:

Lemma 2.9 Let $\tilde{f}$ be a given function and $\tilde{R}$ be an approximant thereof. We set

$$f(z) := e^{\tilde{f}(z)}, \quad R(z) := e^{\tilde{R}(z)}$$

and

$$\rho(z) := \frac{R(z)}{f(z)}, \quad \tilde{\rho}(z) := \tilde{R}(z) - \tilde{f}(z).$$

The order star of the first kind of ρ is precisely the order star of the second kind of $\tilde{\rho}$.

Proof. The proof proceeds by comparison of the definitions of order stars of both kinds and $|e^{\zeta}| = e^{\mathrm{Re}\,\zeta}$ for all complex ζ. $\square$

Note that the proof of the lemma tacitly assumes that R is allowed to be essentially analytic, not just rational. As already mentioned, such an extension presents no problems whatsoever.

We extend the definitions from Sections 2.1 and 2.2 as necessary. Lemma 2.9 is instrumental in 'translating' the results on order stars of the first kind, as presented in the these sections, to the present framework:

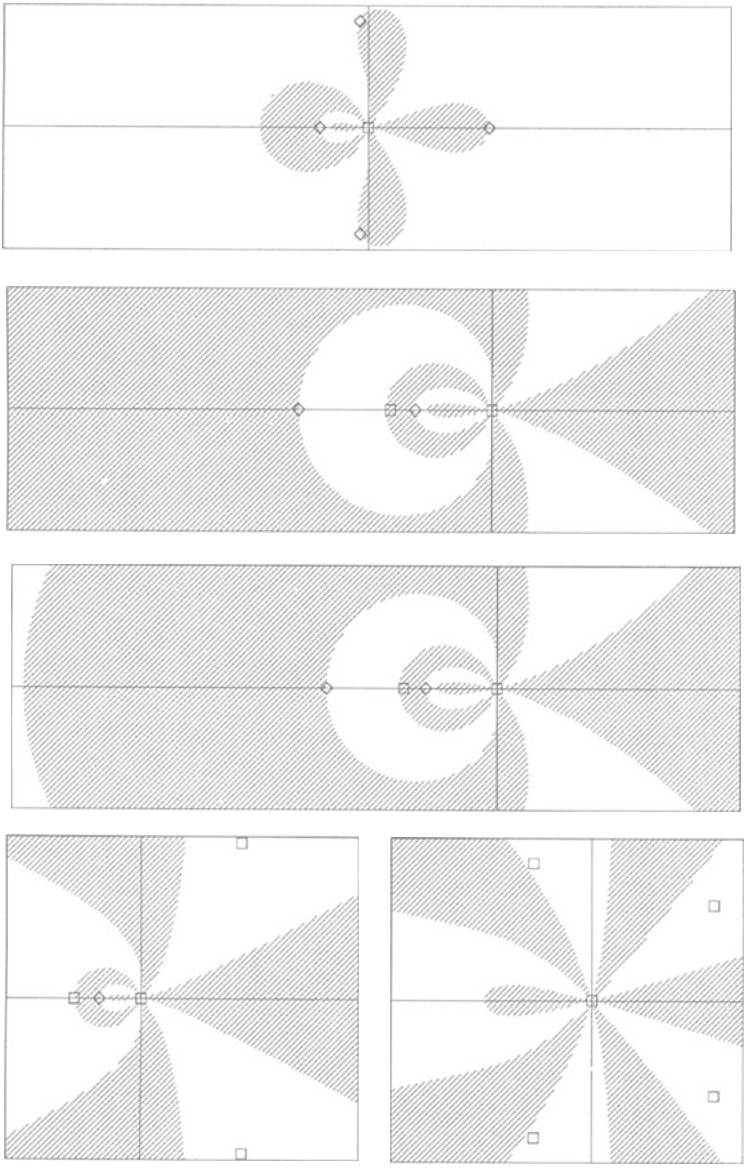


Figure 2.4 Order stars of the second kind for the $[k/(5-k)]$ Padé approximants to $\log(1+z)$ for $k = 1,2,\ldots,5$.

Proposition 2.10 If $z_0 \in \mathrm{cl}\, C$ is a point of analyticity of $\tilde{\rho}$ and an interpolation point of degree p then $z_0 \in \tilde{\mathcal{A}}_0$, $\iota(z_0) = p$ and z_0 is regular.

Proof. This is a straightforward consequence of Proposition 2.1, together with Lemma 2.9. Note that, even if (in the framework of the lemma) z_0 is a zero of $\tilde{f}$, it is true that $f(z_0) \neq 0$, hence the index is not reduced by possibly non-trivial multiplicity of z_0 as a zero of $\tilde{f}$. □

Contractivity is not a feature embedded in the geometry of an order star of the second kind. However, such an order star is particularly suitable when $\mathrm{Re}\, \tilde{f} = 0$ along some complex curve Γ and it is required that $\mathrm{Re}\, \tilde{R}(z) \leq 0$ for all $z \in \Gamma$: trivially, this corresponds to $\Gamma \cap \tilde{\mathcal{A}}_+ = \emptyset$. We will encounter this situation in Chapters 3 and 6.

We say that a closed, positively-oriented, curve of $\tilde{\mathcal{A}}_0$ is an $\mathcal{A}_+$-loop if it is bordered from inside[3], and an $\mathcal{A}_-$-loop if it is bordered from inside by $\tilde{\mathcal{A}}_-$. Note that a closed curve of $\tilde{\mathcal{A}}_0$ may well fail to be either an $\mathcal{A}_+$-loop or an $\mathcal{A}_-$-loop and that portions of an $\mathcal{A}_+$-loop may be also portions of an $\mathcal{A}_-$-loop. We say that a loop is of multiplicity L if it contains precisely L interpolation points.

Proposition 2.11 The multiplicity of an $\mathcal{A}_+$-loop or of an $\mathcal{A}_-$-loop equals the number of singularities of $\tilde{\rho}$ there. Moreover, interpolation points and singularities interlace along a loop.

Proof. It follows from Lemma 2.9 and from the proof of Proposition 2.3 that the imaginary part of $\tilde{\rho}$ is strictly monotone along the oriented $\tilde{\mathcal{A}}_0$, decreasing along $\mathcal{A}_+$-loops and increasing along $\mathcal{A}_-$-loops. The proposition follows, since $\mathrm{Im}\, \tilde{\rho}$ vanishes at an interpolation point and becomes unbounded at a singularity. □

Seemingly we have now four kinds of singularity: poles, essential singularities, branch points and the branch cut. Actually, it follows from the proof of Lemma 2.9 that a pole of $\tilde{\rho}$ corresponds to an essential singularity of exponential type of ρ.

Proposition 2.12 Let z_0 be a pole of $\tilde{\rho}$ of multiplicity K. Then $\iota(z_0) = K$ and z_0 is regular.

Proof. Easy application of Proposition 2.6 and of Lemma 2.9. □

Order stars of the second kind are, in general, insensitive to zeros of $\tilde{f}$ and $\tilde{R}$ – they are not constrained to any particular portion of the order

[3]The closed complex plane is merely a projection of the Riemann sphere by $\tilde{\mathcal{A}}_+$. Thus, if $\infty \in \tilde{\mathcal{A}}_+$ then the boundary of the underlying $\mathcal{A}$-region is an $\mathcal{A}_+$-loop.

star.[4]

Essential singularities of $\tilde{\rho}$ of bounded order are transformed to essential singularities of ρ of infinite order. Thus, many results of Section 2.2 are irrelevant to order stars of the second kind and behaviour near essential singularities calls for a case-by-case analysis – cf. Chapters 3, 6 and 7, as well as the following example.

Example 2.6 Figure 2.4 displays order stars of the second kind for Padé approximants to $\log(1+z)$. Note that the function has a branch cut along $(-\infty,0)$, with a jump of 2π in the imaginary part (moving down across the cut), and branch points at 0 and ∞. Various geometric features can be identified from the preceding propositions. In particular, the branch cut along $(-\infty,-1)$ does not provide the singularity on an $\mathcal{A}_-$-loop that is stipulated in Proposition 2.11, since the jump across the cut cannot nullify the increase of $\operatorname{Im}\tilde{\rho}$ along this loop. The behaviour at the branch points -1 and ∞ is easy to identify by direct techniques: -1 always lies in $\tilde{\mathcal{A}}_+$. If $m > n$ then ∞ is a pole of R of multiplicity $m-n$. This overwhelms the logarithmic singularity there and, by Proposition 2.12, $\infty \in \tilde{\mathcal{A}}_0$ and $\iota(\infty) = m-n$. Finally, if $m \leq n$ then $\infty \in \tilde{\mathcal{D}}$.

It is known from the Padé theory that for all $m \geq n+1$ the $[m/n]$ Padé approximant to $\log(1+z)$, which is of order $m+n$, has n distinct poles on the cut (Baker, 1975). This can be proved quite easily from the order star: we have $\iota(0) = m+n+1$, hence we need to identify $m+n+1$ sectors of $\tilde{\mathcal{A}}_+$ that approach the origin. $m-n$ such sectors can be accounted for by $\iota(\infty) = m-n$ and one by $-1 \in \tilde{\mathcal{A}}_+$. This leaves at least $2n$ sectors that must be 'supported' by a pole of the approximant. Each such pole can support at most a single sector of $\tilde{\mathcal{A}}_+$, unless it lies on $(-\infty,-1)$, at a crossing point with an $\mathcal{A}_-$-loop, since then a single 'ribbon-like' sector of $\tilde{\mathcal{A}}_+$ can approach 0 twice: its 'inner' boundary is the $\mathcal{A}_-$-loop, whereas its 'outer' boundary, an $\mathcal{A}_+$-loop, is catered for, in terms of Proposition 2.11, by crossing the branch cut. The only configuration that provides $2n$ sectors is that of n 'ribbons', implying the existence of precisely n distinct poles of the approximant along the cut. $\diamond$

[4] Except for joint zeros of $\tilde{f}$ and $\tilde{R}$ which are, of course, interpolation points and fall within the frame of reference of Proposition 2.10.

3

Rational Approximants to the Exponential

מוֹנֶה מִסְפָּר לַכּוֹכָבִים לְכֻלָּם שֵׁמוֹת יִקְרָא

From *Psalm 147:4*.

3.1 Introduction

In Example 1.2 we have observed that stability functions of Runge–Kutta methods are rational approximants to the exponential function. In other words, the behaviour of Runge–Kutta for linear problems can be described in terms of properties of certain rational functions – and, of course, 'linear' behaviour is the initial, critical step in nonlinear analysis. It therefore makes perfect sense to construct and analyse general approximants to the exponential. The subject can be introduced in many ways and here we firstly opt for the approach of collocation. Essentially, this means that we approximate locally an ordinary differential equation by a polynomial that satisfies the initial condition and, in addition, obeys the underlying equation at specified collocation points. Obviously, the choice of collocation points is critical to the performance of such a method, for order and stability considerations alike. Moreover, choosing collocation points is tantamount to selecting a specific rational function. This interplay between collocation, rational functions, order and stability led, in fact, to the discovery of order stars in the important work of Wanner *et al.* (1978). This paper has been extended in numerous directions and the body of knowledge on rational approximants to $\exp z$ is substantial. This is not to say, of course, that the theory is cut and dried! Many challenging problems remain and numerous interesting relations still await their discovery. Some of these are collected in Chapter 10.

In the following section we commence by elucidating basic ideas of collocation. Further motivation is provided in Section 3.3, where we survey the Obrechkoff methods. All this leads to the discussion, in the remainder of the present chapter, of general rational approximants. In Chapter 4 we add stability requirements to our brew. And in Chapter 5 we extend the frame-

work from rational to algebraic approximants: these result from numerical methods that reuse past information.

3.2 Collocation

Let us consider the ordinary differential system

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \quad t \in [a, a+h], \quad \mathbf{y}(a) = \mathbf{y}_0. \tag{3.1}$$

The interval of integration extends from a to $a+h$, h is the step size and $\mathbf{y}$ is a function from $\mathcal{R}$ to $\mathcal{R}^d$, which is (at least for the time being) as differentiable as required. There are many methods of constructing an approximant to $\mathbf{y}$. Our approach uses collocation.

Without loss of generality, we assume that $d = 1$ and choose $\mathbf{u}$ in $\pi_n[z]$, the set of polynomials in z of degree not exceeding n, such that $\mathbf{u}(a) = \mathbf{y}_0$. In general, we cannot force a polynomial $\mathbf{u}$ to obey the differential equation for all x. Instead, we can try to satisfy (3.1) at a given set of **collocation points** $x_i = a + hc_i$, $i = 1, \ldots, n$, for some step size $h > 0$, and accept $\mathbf{u}(a+h)$ as our guess of the exact solution $\mathbf{y}(a+h)$. Repeated application of this procedure produces approximants at $a+2h$, $a+3h$ and so on, covering the whole interval of interest. The points c_i, $i = 1, \ldots, n$, are the **collocation parameters** in $\mathcal{R}$ and we assume that they are distinct. We will clarify shortly the considerations that guide their choice.

The n collocation conditions are

$$\mathbf{u}'(a + hc_i) = \mathbf{f}(\mathbf{u}(a + hc_i)), \quad i = 1, \ldots, n. \tag{3.2}$$

Denoting our approximant to $\mathbf{y}(a + ih)$ by $\mathbf{y}_i$, we have

$$\mathbf{y}_1 := \mathbf{u}(a+h). \tag{3.3}$$

Equations (3.2) and (3.3) together define a **collocation method.**

Here the reader should perhaps pause, for a curious question emerges! Our collocation method looks so absolutely plausible and natural, yet the reader will search in vain for its very mention in standard texts on numerical solution of the initial value problem for ordinary differential equations. What's wrong? Not much, really, since the collocation method is nothing more – and nothing less – than a **Runge–Kutta method** in disguise. Thus, let us write $\mathbf{u}$ in terms of the fundamental Lagrangian interpolation polynomials $\ell_i(c)$,

$$\ell_i(c) = \prod_{j \neq i} \frac{c - c_j}{c_i - c_j}, \quad i = 1, \ldots, n.$$

The polynomial u can be written as

$$\mathbf{u}(t) = \mathbf{y}_0 + h \sum_{i=1}^{n} \int_0^{(t-a)/h} \ell_i(c)\mathrm{d}c\, \mathbf{k}_i,$$

where $\mathbf{k}_1, \ldots, \mathbf{k}_n$ are parameters that are determined uniquely by the collocation conditions (3.2). Note that, indeed, $\mathbf{u}(a) = \mathbf{y}_0$. We can now easily deduce the following Runge–Kutta formulation of the collocation method:

$$\mathbf{k}_i = \mathbf{f}(a + c_i h, \mathbf{y}_0 + h \sum_{j=1}^{n} a_{i,j}\mathbf{k}_j), \quad i = 1, \ldots, n, \tag{3.4}$$

and

$$\mathbf{y}_1 = \mathbf{y}_0 + h \sum_{i=1}^{n} b_i \mathbf{k}_i.$$

Here

$$a_{i,j} = \int_0^{c_i} \ell_j(c)\mathrm{d}c, \qquad b_j = \int_0^1 \ell_j(c)\mathrm{d}c$$

for all $i, j = 1, \ldots, n$.

Equation (3.4) has important advantages, as far as computer implementation is concerned. However, the collocation formulation frequently has the edge when we wish to understand the behaviour of the method. For example, its order is fully characterized by the quadrature order of the collocation points:

Theorem 3.1 The order of the collocation method is s if and only if

$$\int_0^1 t^{i-1} N(t)\mathrm{d}t = 0, \quad i = 1, \ldots, s,$$

where $N(t)$ is the collocation polynomial

$$N(t) = \prod_{i=1}^{n} (t - c_i).$$

Proof. See (Hairer *et al.*, 1987). □

Linear stability of a numerical method for ordinary differential equations can be analysed by applying it to the test equation $y' = \lambda y$, where $\lambda \in C$ is a constant. Application of the collocation method to the test equation produces a perturbed differential equation, which is obeyed by the polynomial u,

$$u'(t) = \lambda u(t) + CN((t - a)/h), \quad u(a) = y_0.$$

Here C is a constant, uniquely determined by the requirement that the solution of the differential equation is, in fact, a polynomial. Letting $t =$

$a + hx$ and $v(x) = u(hx + a)$, we obtain the following initial value problem for v

$$v'(x) = zv(x) + KN(x), \quad z := \lambda h, \quad v(0) = y_0,$$

where $K = hC$. The standard variation of constants formula yields its explicit solution,

$$v(x) = e^{zx}y_0 + K\int_0^x e^{z(x-\tau)}N(\tau)\mathrm{d}\tau. \tag{3.5}$$

Straightforward computation gets rid of the integral and affirms that (3.5) is nothing other than

$$v(x) = e^{zx}\left(y_0 + K\sum_{j=0}^{m} N^{(j)}(0)z^{-j-1}\right) - K\sum_{j=0}^{n} N^{(j)}(x)z^{-j-1}y_0.$$

Here, as usual, $N^{(j)} = \mathrm{d}^j N/\mathrm{d}x^j$. Since $v(x)$ is a polynomial, the terms in the brackets must necessarily vanish, producing

$$K = -\frac{z^{n+1}}{\sum_{j=0}^{n} N^{(j)}(0)z^{n-j}}.$$

Thus,

$$u(t) = \frac{S((t-a)/h)}{S(0)}y_0,$$

where the function $S(x)$ has the form

$$S(x) = \sum_{i=0}^{n} z^{n-i}N^{(i)}(x)$$

and

$$C = -\lambda\frac{z^n}{S(0)}.$$

It now follows from (3.2) that

$$y_1 = \frac{S(1)}{S(0)}y_0.$$

Moreover, continuing in this vein to y_2, y_3, etc., affirms that

$$y_i = \left(\frac{S(1)}{S(0)}\right)^i y_0$$

for all i in the domain of interest. The stability function $R(z)$ is therefore

$$R(z) = \frac{P(z)}{Q(z)} \tag{3.6}$$

where the nth degree polynomials $P(z)$ and $Q(z)$ are provided by

$$P(z) \equiv S(1) = \sum_{i=0}^{n} N^{(n-i)}(1) z^i, \tag{3.7}$$

$$Q(z) \equiv S(0) = \sum_{i=0}^{n} N^{(n-i)}(0) z^i. \tag{3.8}$$

$R(z)$ is an approximant to $\exp z$ of order s if

$$R(z) - e^z = e(z) = \mathcal{O}(z^{s+1}).$$

An explicit expression for the error $e(z)$ was presented first by Nørsett (1975), as part and parcel of the theory of C-**polynomials**, introduced therein.

Theorem 3.2 Let $R(z)$ be the rational function introduced in equations (3.5–7). The error $e(z) = R(z) - \exp z$ can be expressed as

$$e(z) = -\frac{\sum_{j=n+1}^{\infty} N_j(1) z^j}{Q(z)}$$

where the quantities $N_j(t)$ are defined recursively,

$$N_1(t) = \int_0^t N(\tau) \mathrm{d}\tau, \quad N_{j+1}(t) = \int_0^t N_j(t), \quad j = 1, 2, \ldots.$$

Proof. The proof follows with very little effort from the explicit representation (3.5). □

3.3 Obrechkoff methods

Collocation methods provide a path from numerical methods for (3.1) to rational approximants. An alternative avenue is provided by **Obrechkoff methods** (Obrechkoff, 1942) that use high derivatives. They were first used in the context of numerically solving (3.1) by Nørsett (1974b).

Let the function $\mathbf{f}$ in (3.1) be analytic.[1] Repeated differentiation yields equations for all derivatives of $\mathbf{y}$ in terms of t and $\mathbf{y}$, the first nontrivial instance being

$$\mathbf{y}'' = \frac{\partial}{\partial t}\mathbf{f}(t, \mathbf{y}) + \left(\frac{\partial}{\partial \mathbf{y}}\mathbf{f}(t, \mathbf{y})\right) \mathbf{f}(t, \mathbf{y}).$$

In general, we write

$$\mathbf{y}^{(k)} = \mathbf{g}_k(t, \mathbf{y}), \quad k = 0, 1, \ldots. \tag{3.9}$$

[1] Hence $\mathbf{f}$ is differentiable of any order and the Taylor series at a has a positive radius of convergence.

Thus, $\mathbf{g}_0(t,\mathbf{y}) = \mathbf{y}$, $\mathbf{g}_1(t,\mathbf{y}) = \mathbf{f}(t,\mathbf{y})$, etc.

Choosing a step length $h > 0$, we seek an approximation to $\mathbf{y}$ at $a+h$, $\mathbf{y}_1$, say. With greater generality, given $\mathbf{y}_i \approx \mathbf{y}(a+ih)$, we wish to time-step to $a+(i+1)h$. This will be accomplished by a scheme of the form

$$\sum_{k=0}^{n} q_k h^k \mathbf{g}_k(a+(i+1)h, \mathbf{y}_{i+1}) = \sum_{k=0}^{n} p_k h^k \mathbf{g}_k(a+ih, \mathbf{y}_i). \tag{3.10}$$

Here n is a nonnegative integer and p_k, q_k, $k = 0,\dots,n$, are fixed parameters, $q_0 = 1$.

To analyse (3.10), we need to ponder for a while on what is meant by a 'solution' of (3.1) and, indeed, by the concept of an ordinary differential equation. Essentially, the task of solving (3.1) (analytically and numerically alike) confronts the problem of finding the unique value at $a+h$, say, of a function that attains a given value at a point a and obeys a specified connection between its values and its derivative at $t \in [a, a+h]$. Let D be the **differential operator**, $Dx(t) := x'(t)$, and let E stand for the **shift operator**, $Ex(t) := x(t+h)$. We can rephrase our problem in an operatorial language: given that $\mathbf{y}$ is known at a and that the action of D on $\mathbf{y}$ is known in $[a, a+h]$, find $E\mathbf{y}(a)$. Now, the connection between D and E (as long as everything in sight is analytic) is provided by the Taylor theorem:

$$\mathbf{y}(a+h) = E\mathbf{y}(a) = \sum_{k=0}^{\infty} \frac{1}{k!} \frac{\mathrm{d}^k \mathbf{y}(a)}{\mathrm{d}t^k} h^k = \left\{ \sum_{k=0}^{\infty} \frac{1}{k!} h^k D^k \right\} \mathbf{y}(a).$$

In other words, formally

$$E = e^{hD}, \tag{3.11}$$

where the exponential of an operator (like that of a scalar or a matrix) is defined via the Taylor expansion.

Let

$$R(z) = \frac{P(z)}{Q(z)} := \frac{\sum_{k=0}^{n} p_k z^k}{\sum_{k=0}^{n} q_k z^k}.$$

We now rephrase (3.10) in an operatorial language, assuming, for the sake of simplicity, that $i = 0$:

$$\sum_{k=0}^{n} q_k h^k D^k E\mathbf{y}(a) \approx \sum_{k=0}^{n} p_k h^k D^k \mathbf{y}(a).$$

where the action of D is expressed via the quantities $\mathbf{g}_k$ from (3.9),

$$D^k \mathbf{y}(a) = \mathbf{g}_k(a, \mathbf{y}(a)).$$

In other words,

$$E \approx R(hD),$$

where the division by $Q(hD)$ is justified by the implicit function theorem, since $q_0 = 1$. Consequently, it follows from (3.11) that R approximates the exponential! More specifically, letting $\mathbf{e}_1 := \mathbf{y}_1 - \mathbf{y}(a+h)$, the error at $a+h$, we have

$$\mathbf{e}_1 = \left(R(hD) - e^{hD}\right)\mathbf{y}(a).$$

Therefore, if

$$R(z) = e^z + ez^{p+1} + \mathcal{O}(z^{p+2})$$

and R is a pth order approximant of the exponential, then

$$\mathbf{e}_1 = eh^{p+1}\frac{\mathrm{d}^{p+1}\mathbf{y}(a)}{\mathrm{d}t^{p+1}} + \mathcal{O}(h^{p+2}).$$

This clearly can be extended to all $a + ih$ in the domain of interest, in the sense that

$$\mathbf{y}_i - \mathbf{y}(a + ih) = \mathcal{O}(h^{p+1}).$$

Theorem 3.3 (Nørsett, 1974b) The Obrechkoff method (3.10) is of order p if and only if R is an approximant to $\exp z$ of order p. Moreover, if (3.10) is applied to the linear equation $y' = \lambda y$, $y(0) = 1$, then the outcome is $y_i = R^i(h\lambda)$, $i = 0, 1, \ldots$.

Proof. The statement on the order follows at once from our analysis, whereas the explicit form of the solution in the linear case can be obtained at once from (3.10), since $g_k(t, y) = \lambda^k y$, $k = 0, 1, \ldots$. □

Rephrasing the second part of the theorem, R is the linear stability function of the Obrechkoff method.

Example 3.1 Letting $n = 1$ and requiring $p \geq 1$ we obtain the **theta method**

$$\mathbf{y}_{i+1} = \mathbf{y}_i + h\left((1-\theta)\mathbf{f}(a + ih, \mathbf{y}_i) + \theta\mathbf{f}(a + (i+1)h, \mathbf{y}_{i+1})\right),$$

where θ is any real constant. In particular, $\theta = 0, \frac{1}{2}, 1$ yield the well-known **forward Euler** method, **trapezoidal rule** and **backward Euler** method respectively. The stability function is

$$R(z) = \frac{1 + (1-\theta)z}{1 - \theta z}.$$

It can be easily verified that $R(z) = \exp z + \mathcal{O}(z^2)$, except for $\theta = \frac{1}{2}$, when the order is boosted by a single unit. ◇

3.4 Examples of approximants

Rational approximants to the exponential appear throughout numerical and approximation-theoretic literature. A fair proportion of them are Padé approximants. Sometimes, however, the structure of the underlying numerical algorithm dictates certain restrictions on the form of the approximant – unlike in the Padé case, not all degrees of freedom are used to maximize order.

In the present section we review these two major examples of rational approximants to $\exp z$.

Example 3.2 Padé approximants.

If N is an appropriate Jacobi polynomial, shifted to the interval $(0,1)$, we recover the $[m/n]$ Padé approximant. Indeed, exploiting the Rodrigues formula (Rainville, 1967), we have (with proper normalization)

$$N(t) = \frac{(-1)^n}{(n+m)!}\frac{\mathrm{d}^m}{\mathrm{d}t^m}\left(t^m(1-t)^n\right).$$

The expressions for $P(z)$ and $Q(z)$ are given explicitly as

$$P(z) = P_{m/n}(z) = \sum_{k=0}^{m}\frac{(m+n-k)!}{(m+n)!}\binom{m}{k}z^k, \tag{3.12}$$

$$Q(z) = Q_{m/n}(z) = P_{n/m}(-z). \tag{3.13}$$

This can be deduced (by a straightforward but tedious analysis) from the aforementioned representation of N. A neater way is provided by the theory of hypergeometric functions. We say that the function f, with the formal power series expansion

$$f(z) = \sum_{\ell=0}^{\infty}\frac{f_\ell}{\ell!}z^\ell$$

is a $\{p,q\}$ **hypergeometric function** if there exists an irreducible function $r_f \in \pi_{p/q}$ such that

$$\frac{f_{\ell+1}}{f_\ell} = r_f(\ell), \quad \ell = 0,1,\ldots.$$

Thus, $r_f(v) \equiv 1$ yields the exponential, $r_f(v) = v - \alpha$ results in $f(z) = (1-z)^\alpha$, and the Jacobi polynomial $P_n^{(\alpha,\beta)}$, where $\alpha,\beta > -1$ and n stands for the degree, is nothing other than

$$\frac{\Gamma(n+\alpha)}{n!\Gamma(\alpha)}f\left((1-z)/2\right),$$

where f is a $\{2,1\}$ hypergeometric function, produced by

$$r_f(v) = (v-n)(v+n+\alpha+\beta+1)/(v+\alpha+1).$$

We require an identity, known as the **Kummer first formula**: let f_1 and f_2 be two $\{1,1\}$ hypergeometric functions, defined by $r_{f_1}(v) = (v+a)/(v+b)$ and $r_{f_2}(v) = (v+b-a)/(v+b)$ respectively. Both a and b are allowed to roam throughout C, except that b must differ from a nonpositive integer (otherwise the r_{f_i}'s will not be well defined for some natural argument v). Then

$$f_1(z) = e^z f_2(-z). \tag{3.14}$$

An important property of any hypergeometric function is that if r_f has a nonpositive integer zero, $r_f(m) = 0$, say, then f is a polynomial of degree m. Suppose that we force f_1 in (3.14) to terminate, choosing $a = -m$. Note that, having done this, we cannot force f_2 to terminate as well, since then b will become a nonpositive integer, and this is not allowed. Anyway, such a 'termination' will be a complete nonsense, since $\exp z$ is not a rational function! We can, however, choose r_{f_2} so that f_2 'almost' terminates, by letting $b = -n-m+\varepsilon$ for some $0 < |\varepsilon| \ll 1$. To emphasize dependence on ε, we let $r_{f_2} \equiv r^\varepsilon$. Splitting the infinite sum

$$\begin{aligned}\sum_{\ell=0}^{\infty} \frac{1}{\ell!} \prod_{j=0}^{\ell-1} r^\varepsilon(j)(-z)^\ell &= \sum_{\ell=0}^{n}(\cdots) + \sum_{\ell=n+1}^{n+m}(\cdots) + \sum_{\ell=n+m+1}^{\infty}(\cdots) \\ &:= I_1^\varepsilon + I_2^\varepsilon + I_3^\varepsilon,\end{aligned}$$

we note first that $\lim_{\varepsilon\to 0} r^\varepsilon(\ell)$ is well posed for $0 \le \ell \le n$, since the denominator does not vanish. Hence, we can let $\varepsilon \to 0$ in I_1^ε. Next, we notice that in the range $\ell \in \{n+1, n+2, \ldots, n+m\}$ we maintain well-posedness and have

$$\lim_{\varepsilon\to 0} r^\varepsilon(\ell) = 0.$$

Therefore, I_2^ε vanishes with ε. Finally, when $\ell \ge n+m+1$ then

$$\begin{aligned}\prod_{j=0}^{\ell-1} r^\varepsilon(j) &= \prod_{j=0}^{\ell-1} \frac{j-n+\varepsilon}{j-n-m+\varepsilon} \\ &= \frac{\prod_{j=-n}^{-1}(j+\varepsilon)\prod_{j=1}^{\ell-n-1}(j+\varepsilon)}{\prod_{j=-n-m}^{-1}(j+\varepsilon)\prod_{j=1}^{\ell-n-m-1}(j+\varepsilon)} \\ &\xrightarrow{\varepsilon\to 0} (-1)^m \frac{n!(\ell-n-1)!}{(n+m)!(\ell-n-m-1)!},\end{aligned}$$

Consequently, $\lim_{\varepsilon\to 0} I_3^\varepsilon$ is a well-defined series. It can be seen at once that it is $\mathcal{O}(z^{n+m+1})$. Likewise, we can let $\varepsilon \to 0$ in f_1 with no ill effects.

Lemma 3.4 The function $R_{m/n} := P_{m/n}/Q_{m/n}$, with $P_{m/n}$ and $Q_{m/n}$ given in (3.12) and (3.13) respectively, is the $[m/n]$ Padé approximant to $\exp z$, with the leading error constant $(-1)^{n+1}n!m!/((n+m)!(n+m+1)!)$.

qq**Proof.** We let ε tend to zero in (3.14). It is easy to see that

$$f_1 \to P_{m/n}, \quad I_1^\varepsilon \to Q_{m/n},$$

hence

$$P_{m/n}(z) = e^z Q_{m/n}(z) + (-1)^{n+1}\frac{n!m!}{(n+m)!}z^{n+m+1} + \mathcal{O}(z^{n+m+1}).$$

But $Q(z) = 1 + \mathcal{O}(z)$, hence division by $Q(z)$ produces

$$R_{m/n}(z) = e^z + (-1)^{n+1}\frac{n!m!}{(n+m)!}z^{n+m+1} + \mathcal{O}(z^{n+m+1}).$$

The proof follows from the definition of Padé approximants in Chapter 1. □

It is usual to arrange Padé approximants (for any function, not necessarily $\exp z$) for all values of $m, n = 0, 1, \ldots$ in an infinite two-dimensional array, called **Padé tableau**. Table 3.1 displays the first few entries of the tableau for $\exp z$.

$n \setminus m$	0	1	2
0	1	$1+z$	$1+z+\frac{1}{2}z^2$
1	$\frac{1}{1-z}$	$\frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}$	$\frac{1+\frac{2}{3}z+\frac{1}{6}z^2}{1-\frac{1}{3}z}$
2	$\frac{1}{1-z+\frac{1}{2}z^2}$	$\frac{1+\frac{1}{3}z}{1-\frac{2}{3}z+\frac{1}{6}z^2}$	$\frac{1+\frac{1}{2}z+\frac{1}{6}z^2}{1-\frac{1}{2}z+\frac{1}{6}z^2}$

Table 3.1 Padé approximants to $\exp z$.

We will encounter Padé approximants again throughout this book. They will be the focus of our attention in Chapter 8. ◇

Example 3.3 Restricted Padé approximants.

This class of approximants was introduced by Nørsett (1974a; 1974b; 1978), motivated by the need to construct rational approximants with a single real pole of multiplicity m: such approximants are linked to specific m-stage Runge–Kutta methods that have important advantages, the Singly-Diagonally Implicit (SDIRK) methods. In the present exposition we depart

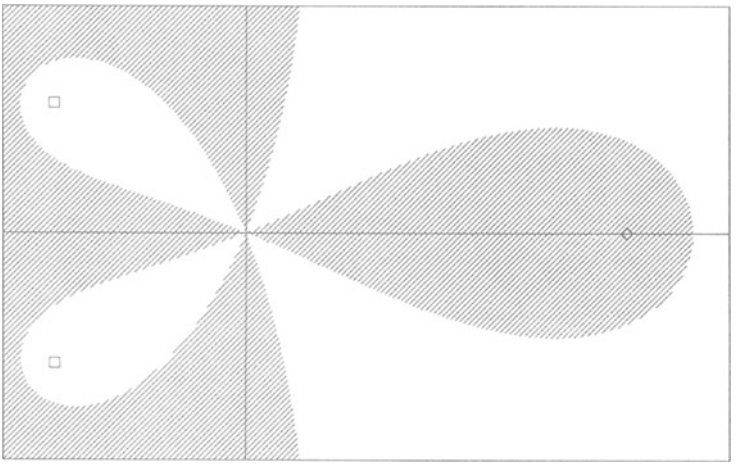


Figure 3.1 Order star of the ${}_1S_{2/2}$ singly p-restricted approximant to $\exp z$.

from this single-pole framework and use the phrase 'restricted' to mean that we are restricting zeros and poles in some way or other, typically by specifying bounds on the number of complex zeros and/or poles.

In line with the Runge–Kutta terminology, we say that the rational function R is **singly p-restricted** if it possesses a single pole, which is real, and **multiply p-restricted** if all its poles are real (but not necessarily distinct). Similar usage pertains to restrictions on zeros, z replacing p. Finally, R is **pz-restricted** (singly or multiply, as the case might be) if it is both p-restricted and z-restricted. Clearly, the multiply p-restricted approximants are of the form

$$R(z) = \frac{P(z)}{\prod_{i=0}^{n}(1 - \gamma_i z)}$$

where $P \in \pi_m[z]$.

Let us commence with the classical case of singly p-restricted approximants. We denote a $\pi_{m/n}$ approximant with the denominator $(1 - \gamma z)^n$ by $S_{m/n}(\cdot, \gamma)$. The underlying collocation polynomial $N(t)$ is allied to Laguerre polynomials: Define N as

$$N(t) = (-1)^n n! \gamma^n L_n(t/\gamma)$$

Here γ is a real number and $L_n(x)$ is the Laguerre polynomial

$$L_m(x) = \sum_{j=0}^{n} \frac{(-1)^j}{j!} \binom{n}{j} x^j$$

(Rainville, 1967). Note as a matter of interest that L_n is a hypergeometric function with $r_{L_n}(v) = (v-n)/(v+1)$. We now use (3.6–8) with this form of the collocation polynomial to derive the diagonal and first subdiagonal **restricted Padé approximants**,

$$S_{n/n}(z) = (-1)^n \frac{\sum_{j=0}^{n} L^{(n-j)}(\gamma^{-1})(z/\gamma)^j}{(1-z/\gamma)^n}.$$

Laguerre polynomials are orthogonal in $(0,\infty)$ (with respect to the weight function $\exp(-x)$). Thus, it follows at once from elementary theory of orthogonal polynomials that all their zeros reside there (Rainville, 1967). We denote them by $\xi_{n,1} < \xi_{n,2} < \cdots < \xi_{n,n}$. Clearly, if γ is a reciprocal of one of these zeros then the coefficient of z^n in the numerator vanishes. This provides the approximant ${}_jS_{(n-1)/n} \equiv S_{(n-1)/n}(\cdot, \xi^{-1}_{n,n+1-j})$, $j = 1,\ldots,n$). Similarly, it is easy to verify that the order of the $\pi_{n/n}$ approximation is boosted up to $n+1$ if γ is a reciprocal of a zero of L'_{n+1}. This, again, is an orthogonal polynomial, hence its zeros are in $(0,\infty)$ (an even easier proof follows at once from the Rolle theorem). Given that these zeros are $\nu_{n,1} < \nu_{n,2} < \cdots < \nu_{n,n}$, we let ${}_jS_{n/n} = S_{n/n}(\cdot, \nu^{-1}_{n,n+1-j})$, $j = 1,\ldots,n$.

Theorem 3.5 (Nørsett, 1974a) The singly p-restricted $[n/n]$ Padé approximant has order n for all real γ, except when $\gamma = \nu^{-1}_{n,j}$ for some $j \in \{1,\ldots,n\}$, when its order equals $n+1$. Moreover, the approximant reduces to an $\pi_{(n-1)/n}$ function exactly when $\gamma = \xi^{-1}_{n,j}$ for some $j \in \{1,\ldots,n\}$. □

It is easy to generalize the last theorem to cater for singly p-restricted $[m/n]$ with any $m \leq n$. Another natural extension is to multiply restricted approximants. However in order to construct these via collocation we need to apply the theory of biorthogonal polynomials (Iserles and Nørsett, 1987; 1988) and that would require another volume, with a very different flavour. However, there is no need for biorthogonality in considering specific examples. The following are borrowed from Orel (1990b).

The A-acceptable[2] singly restricted approximant ${}_1S_{3/3}$ can be slightly perturbed to give the 3/3 fourth-order approximant

$$R(z) = \frac{1 - 2.23046z + 0.773796z^2 + 0.78239z^3}{1 + 3.23046z - 3.46843z^2 + 1.12747z^3}.$$

[2]A rational function R is said to be A-acceptable if $|R(z)| < 1$ for all $\mathrm{Re}\, z < 0$. If, in addition, $\lim_{|z|\to\infty} R(z) = 0$, then it is L-acceptable. Functions like this are highly desirable as stability functions of Runge–Kutta schemes for some 'heavy-duty' ordinary differential equations.

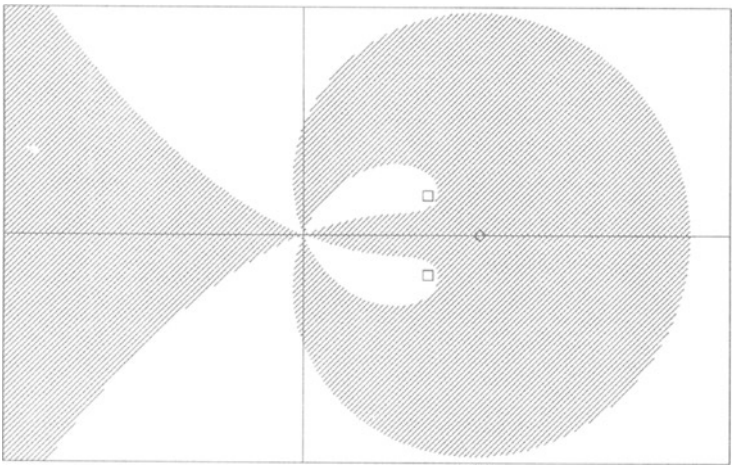


Figure 3.2 Order star of the ${}_3S_{2/3}$ singly p-restricted approximant to $\exp z$.

Based on the L-acceptable singly restricted Padé approximant ${}_2S_{3/2}$, we find after a perturbation the third-order multiple approximant

$$R(z) = \frac{1 - 0.33702z - 0.25143z^2}{1 + 1.33702z - 0.58599z^2 + 0.0837z^3}.$$

Let us remark that both underlying collocation polynomials have zeros outside the interval [0,1]. In practice we seldom use collocation polynomials with that property.

m	${}_1S_{m-1/m}$	${}_1S_{m/m}$
1	$\dfrac{1}{1-z}$	$\dfrac{1-\frac{1}{2}z}{1+\frac{1}{2}z}$
2	$\dfrac{1-(1+\sqrt{2})z}{\left(1-\frac{2+\sqrt{2}}{2}z\right)^2}$	$\dfrac{1-\frac{\sqrt{3}}{3}z-\frac{1+\sqrt{3}}{6}z^2}{\left(1-\frac{3+\sqrt{3}}{6}z\right)^2}$
3	$\frac{1-6.21545z-10.6388z^2}{(1-2.40515z)^3}$	$\frac{1-2.20574z+0.71984z^2+0.76921z^3}{(1-1.06858z)^3}$

Table 3.2 Restricted Padé approximants to $\exp z$.

Some of the first diagonal and subdiagonal restricted Padé approximants are given in Table 3.2 (cf. Nørsett, 1978). ◇

3.5. Order barriers

It is well known that the optimal order of a $\pi_{m/n}$ approximant to the exponential is $m+n$ and that it is attained by the underlying Padé approximant. For restricted approximants the optimal order is not that clear. One could expect the optimal order to be close to $m+n$, at least for a 'good' choice of poles. This, unfortunately, is not the case. The optimal order of a $\pi_{n/n}$ multiply p-restricted approximant is just $n+1$: we cannot do better (at least, as far as order is concerned) with a multiply restricted than with a singly restricted approximant. This result was originally found by Nørsett and Wolfbrandt (1977). Here we use order stars to extend that order bound to general restricted approximants.

Theorem 3.6 (Wanner *et al.*, 1978; Iserles and Nørsett, 1989) Assume that the function $R \in \pi_{m/n}$ has ρ_Z real zeros and ρ_P real poles. The order p (as an approximant to $\exp z$) obeys the bound

$$p \leq m + n - \min\left\{(\rho_Z - 1)_+, (\rho_P - 1)_+\right\},$$

where $(x)_+ := \max\{x, 0\}$ for all $x \in \mathcal{R}$.

Proof. The two cases can be treated in an identical manner and we give the proof for the second case only. The main idea of the proof is in exploiting Proposition 2.3 to link the geometry of zeros and poles with interpolation points. We note first that, in accordance with the discussion in Section 2.2, $\infty \in \mathcal{A}_0$ is regular and that $\iota(\infty) = 1$. Thus, $\mathcal{A}_+$ and $\mathcal{A}_-$ include a single unbounded component each, which we denote by $\mathcal{A}_+^\infty$ and $\mathcal{A}_-^\infty$ respectively. It is easy to verify that $\mathcal{A}_+^\infty$ lies to the left (cf. Figures 3.1 and 3.2).

By Proposition 2.1 there are $\iota(0) = p+1$ sectors of $\mathcal{A}_+$ that adjoin the origin. Of these, κ_∞, say, belong to $\mathcal{A}_+^\infty$, κ_R, say, belong to bounded $\mathcal{A}_+$-regions that contain only real poles, and $\kappa_C := p+1-\kappa_\infty-\kappa_R$ (informally, this means that κ_C sectors at the origin need to be 'supported' by complex poles).

The order star is symmetric with respect to the real axis. Thus, the κ_R bounded $\mathcal{A}_+$-regions that contain real poles necessarily envelop $\kappa_R - 1$ sectors of $\mathcal{A}_-$ that, of course, belong to bounded $\mathcal{A}_-$-regions. By the same token, the κ_∞ sectors of $\mathcal{A}_+^\infty$ enclose $\kappa_\infty - 1$ sectors of $\mathcal{A}_-$. We have

$$\kappa_R + \kappa_\infty \leq m + 2,$$

since Proposition 2.3 implies that we need $\kappa_R + \kappa_\infty - 2$ zeros to account for the multiplicities of the enclosed $\mathcal{A}_-$-regions and at most m zeros are

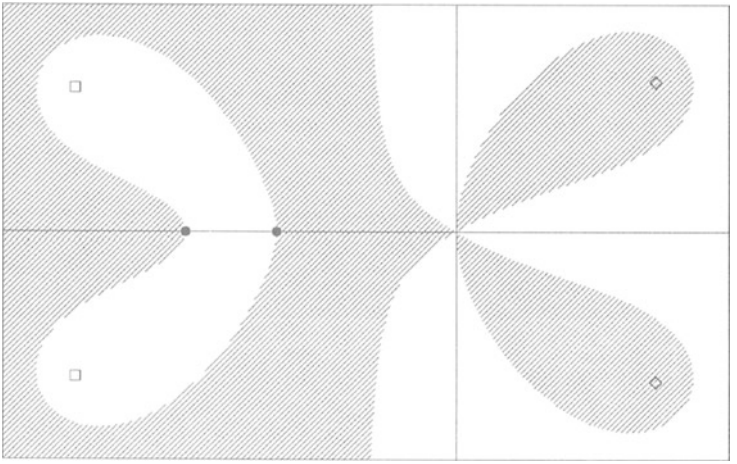


Figure 3.3 Order star of a 2/2 second-order approximant to $\exp z$ with two negative interpolation points.

available. Moreover – and for the same reason, except that now we are counting complex poles – we have

$$\kappa_C \leq n - \rho_P + 1.$$

Since $p = \kappa_R + \kappa_C + \kappa_\infty - 1$, we have

$$p \leq m + n - \rho_P.$$

Finally, if $\rho_P = 0$ then $\kappa_R = 0$ (since no real poles are available!) and the bound is just $p \leq m + n$. This provides the desired bound in terms of ρ_P and completes the proof. □

Figures 3.1 and 3.2 help to comprehend the proof and the reader is urged to identify $\mathcal{A}_+^\infty$, κ_R etc. therein.

3.6 Degree of real interpolation

So far, all the available parameters of a rational function have been used to maximize the order at the origin, subject to possible constraints on zeros and poles. This makes sense, because of the connection with collocation order. However, it is natural to extend the framework by allowing negative

interpolation points. These occasionally make sound computational sense. Consider, for example, the approximation of the linear system

$$\mathbf{y}' = A\mathbf{y}, \quad \mathbf{y}(0) = \mathbf{y}_0,$$

by

$$\mathbf{y}(ih) \approx \mathbf{y}_i := R^i(hA)\mathbf{y}_0,$$

where $h > 0$ and R is a rational approximant to the exponential. One option, advocated by Varga (1962), is to use the best $L_\infty(-\infty, 0]$ $\pi_{n/n}$ approximant, in which case there are $n+m+1$ negative interpolation points (and none at the origin!). Another possibility is to retain a high order of approximation at the origin, but impose exact interpolation at specific negative points (Liniger and Willoughby, 1967).

Iserles (1979) studied the question of how many zeros (which we count with their multiplicity) the error $e(z) = R(z) - \exp z$ can possess in the interval $(-\infty, 0]$. The result, a universal barrier of $m + n + 1$, is termed the **maximal interpolation theorem**, and can be proved by elementary methods. However, order stars provide a far simpler and shorter proof.

Theorem 3.7 (Iserles and Powell, 1981) For any function $R \in \pi_{m/n}$ the number of real zeros of the equation $R(x) = \exp x$, counted with their multiplicities (i.e. the number of real interpolation points, counted with their degrees), cannot exceed $m + n + 1$.

Proof. Let $\zeta \leq 0$ be an interpolation point of degree $q \geq 1$. According to Proposition 2.1, it is a regular member of $\mathcal{A}_0$ and $\iota(\zeta) = q$. Thus, ζ is adjoined by $2q$ sectors of $\mathcal{A}_+$ and $\mathcal{A}_-$.

Consider first the case of both $\mathcal{A}_+^\infty$ and $\mathcal{A}_-^\infty$ adjoining ζ. Thus, at least $q-1$ sectors at ζ belong to bounded regions. Note that there might be at most one such (real) point. In the remaining case ζ is surrounded by an unbounded region, hence all q sectors of 'opposite colour' that adjoin it belong to bounded regions.

Consequently, there are together at least $n + m$ sectors that adjoin interpolation points and belong to bounded $\mathcal{A}_+$ and $\mathcal{A}_-$-regions. We now use Proposition 2.3 to argue that the total multiplicities of bounded $\mathcal{A}_+$-regions may not exceed n, whereas the total multiplicities of bounded $\mathcal{A}_-$-regions are at most m. The proof follows. □

Example 3.4 Clearly, every Padé approximant has by definition an interpolation point of degree $n + m + 1$ (degree of interpolation is counted from zero, while order is counted from one!). Thus, it follows from Theorem 3.6 that $R_{m/n}(x) \neq \exp x$ for $x \in \mathcal{R} \setminus \{0\}$ (cf. Figure 1.1).

A more interesting example was introduced by Ehle and Picel (1975) and Nørsett (1975). It occurs when the rational function is allowed some interpolation points in $(-\infty, 0)$, while retaining a good order of approximation.

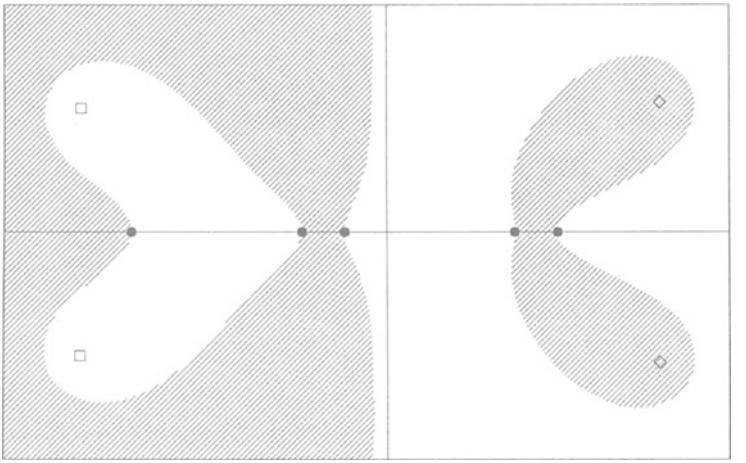


Figure 3.4 Order star of a 2/2 function that interpolates $\exp z$ at $\{-3,-1,-0.5,1.5,2\}$.

Thus, let $m = n = 2$ and require order 2. Then

$$R(z) = \frac{1 + \frac{2-\alpha}{4}z + \frac{1+\beta-\alpha}{8}z^2}{1 - \frac{2+\alpha}{4}z + \frac{1+\beta+\alpha}{8}z^2},$$

where α, β are arbitrary (Nørsett, 1975). It is possible to choose α and β so as to force interpolation at two further points, to bring the sum of interpolation degrees up to the bound of Theorem 3.6. For example, requiring $R(-2) = \exp(-2)$, $R(-1) = \exp(-1)$ yields (in rounded figures)

$$\frac{1 + 0.4501229z + 0.0611269z^2}{1 - 0.5493771z + 0.1110040z^2}$$

(cf. Figure 3.3).

In our final example we go all the way, abandoning order requirements and interpolating at distinct points. Letting again $m = n = 2$, we interpolate at $\{-3,-1,-0.5,1.5,2\}$. This yields (again, in rounded figures)

$$R(z) = \frac{0.9948451 + 0.9758471z + 0.0660616z^2}{1 - 0.5103704z + 0.0799846z^2}$$

and the order star in Figure 3.4. ◇

3.7 Symmetric approximants with real poles

The concept of A-stability forms a backdrop to the present chapter and the focus for the next one. The implicit assumption is that the underlying system of ordinary differential equations is (locally) dissipative and it is perfectly valid for a great many systems of practical importance. However, there exists a class of differential systems that, while being central to modern applied mathematics, displays conservation (as opposed to dissipativity) – the **Hamiltonian equations**. Let $H : \mathcal{R}^d \times \mathcal{R}^d \to \mathcal{R}$ be a C^2 function. The underlying Hamiltonian system is

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = -\frac{\partial H(\mathbf{p},\mathbf{q})}{\partial \mathbf{q}}, \quad \frac{\mathrm{d}\mathbf{q}}{\mathrm{d}t} = \frac{\partial H(\mathbf{p},\mathbf{q})}{\partial \mathbf{p}}. \tag{3.15}$$

This is not the place to debate the importance of (3.15) to classical and quantum mechanics or describe the fascinating mathematics that evolved to analyse its properties and culminating in the Kolmogorov–Arnold–Moser theory. The reader is referred to Arnold (1978) and to Guckenheimer and Holmes (1983) for a thorough exposition. Here we just mention few facts that justify our subsequent investigation into rational approximants of a rather special form.

The cornerstone of any analysis of Hamiltonian equations is the Liouville theorem, stating that the flow of (3.15) (i.e. the solution for all initial values, seen as a mapping of $\mathcal{R}^d \times \mathcal{R}^d$ into itself) is **symplectic**: the differential form $\mathrm{d}\mathbf{p} \wedge \mathrm{d}\mathbf{q}$ is invariant. For readers unfamiliar with symplectic geometry it will be probably the easiest to visualize this state of affairs for $d = 1$, since then the aforementioned differential form is simply the area. Thus, take any 'nice' compact set $\mathcal{U}_0 \subset \mathcal{R} \times \mathcal{R}$ and define

$$\mathcal{U}_t := \{(p(t), q(t)) : (p(0), q(0)) \in \mathcal{U}_0\}, \quad t \geq 0.$$

Symplecticity for $d = 1$ means that the area of $\mathcal{U}_t$ is constant as a function of t.

Conservation of area – or, for that matter, of other quantities – does not imply that the geometry of solution is 'simple'. On the contrary! Within the narrowly-fitting confines of its straightjacket it can (and usually does) evolve in a very complicated fashion. In particular, if $d \geq 2$ then it is possible for the flow to become chaotic.[3] This calls for great care in the numerical solution of (3.15).

Symplecticity of (3.15) can be exploited to prove the existence of limit cycles around elliptic fixed points of the flow in the phase plane. This is

[3]Perhaps counter-intuitively, conservation laws almost invariably spell problems. The chaotic behaviour of (3.15) is one instance of this 'revenge of conservation'. Another, the nonlinear hyperbolic conservation law $\partial u/\partial t + \partial f(u)/\partial x = 0$, is mentioned in Chapter 6. There the price of conservation (of energy) is spontaneously-arising discontinuities.

a major feature of Hamiltonian systems, a feature that we wish numerical methods to replicate.

We say that a numerical method is **symplectic** (or **canonical**) if it maintains the symplectic structure of the flow of (3.15). Although specific symplectic methods have been known for a decade or so, a major breakthrough occurred relatively recently, when Lasagni (1988), Sanz-Serna (1988) and Suris (1989) simultaneously produced a characterization of all symplectic Runge–Kutta methods

$$\begin{aligned} \mathbf{k}_\ell &= \mathbf{f}\left(t_n + c_\ell h, \mathbf{y}_n + h\sum_{j=1}^{s} a_{\ell,j}\mathbf{k}_j\right), \quad \ell = 1,\ldots,s, \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + h\sum_{\ell=1}^{s} b_\ell \mathbf{k}_\ell, \end{aligned} \tag{3.16}$$

by means of a simple algebraic test, namely

$$b_k a_{k,\ell} + b_\ell a_{\ell,k} = b_k b_\ell, \quad k,\ell = 1,\ldots,s. \tag{3.17}$$

The most striking feature of (3.17) is not its simplicity – although it helps! – but the fact that some of the most thoroughly studied methods (3.16) obey it. This is, in particular, the case with the **Gauss–Legendre** schemes, where $c_1,\ldots,c_s$ are the zeros of the sth Legendre polynomial, shifted to $[0,1]$ (Hairer *et al.*, 1987). Gauss–Legendre methods represent, on the face of it, the best of all worlds: they are of order $2s$, the largest attainable, they are A-stable[4] and symplectic. They are also very expensive to implement.

The main reason can be traced to the fact that the linear stability function of the s-stage Gauss–Legendre method is the $[s/s]$ Padé approximant to $\exp z$. In general, given a Runge–Kutta method (3.16) with the linear stability function $R \in \pi_{s/s}$, the expense of implementing is intimately connected with the number of real poles of R: if all the poles are real the stages can be, effectively, decoupled and solved one at a time (Butcher, 1988), while complex conjugate poles present a considerably more formidable computational problem. Alas, a diagonal Padé approximant possesses either a single real pole or none, depending on the parity of s,[5] all other poles forming complex conjugate pairs. It makes perfect sense to trade order off for the sake of real poles, motivating renewed consideration of the multiply p-restricted approximants.

We are back to the theme of Example 3.3, but the rules of the game are new and different from those in the remainder of this and the next chapters. We are no longer interested in A-acceptability. On the other hand, we restrict the field to **symmetric** approximants: functions $R(z) = Q(-z)/Q(z)$,

[4] Actually, they obey a stronger stability requirement, namely **algebraic stability**...

[5] The proof – by order stars, what else? – is trivial and left for the reader.

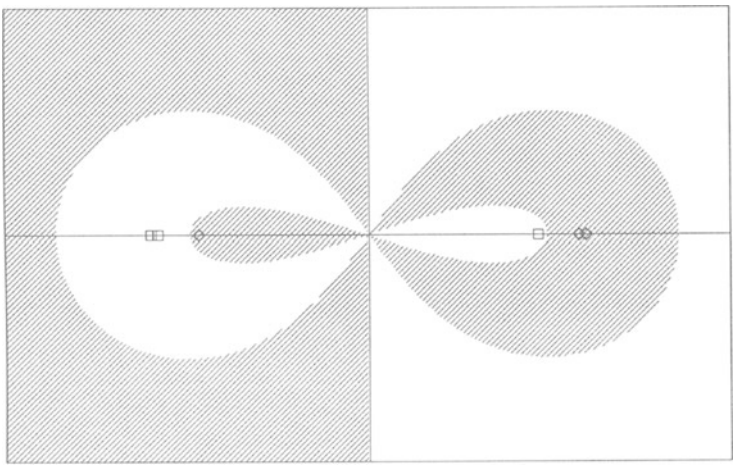


Figure 3.5 Order star of a symmetric multiply pz-restricted fourth-order approximant.

$\deg Q = s$, that obey $|R(it)| \equiv 1$, $t \in \mathcal{R}$ – otherwise symplecticity is lost already for the linear Hamiltonian equation with $H(p,q) = \frac{1}{2}(p^2+q^2)$. Hence, our multiply p-restricted approximants are really multiply pz-restricted. Moreover, as long as R is irreducible, it is anti-symmetric with respect to the imaginary axis: z^* is a pole if and only if $-z^*$ is a zero.

It is a consequence of Theorem 3.6 that the order of R cannot exceed $s+1$. Moreover, since

$$Q(-z)e^{\frac{z}{2}} - Q(z)e^{-\frac{z}{2}}$$

is an odd function, it follows that the order is even. Thus, order $s+1$ can be realized only if s is odd. We devote the remainder of this section to the investigation of symmetric multiply pz-restricted approximants of order $s+1$, s odd. Our analysis follows that in (Iserles, 1990).

Figure 3.5 displays the order star of the first kind for

$$\begin{aligned} R(z) &= \frac{1+\frac{1}{2}z-\frac{239}{344}z^2-\frac{803}{2064}z^3}{1-\frac{1}{2}z-\frac{239}{344}z^2+\frac{803}{2064}z^3} \\ &= \frac{(z+1.5039524\ldots)(z+1.4558135\ldots)(z-1.1739627\ldots)}{(z-1.5039524\ldots)(z-1.4558135\ldots)(z+1.1739627\ldots)}. \end{aligned}$$

It is of order 4 and possesse two positive poles and a single negative pole.

We denote by μ_- and μ_+ the number of sectors of $\mathcal{A}_+$ that adjoin the origin from within $\mathcal{C}^-$ and $\mathcal{C}^+$ respectively. Here, in line with Example 2.2, $\mathcal{C}^-$ is the complex left half-plane, whereas $\mathcal{C}^+$ is the right half-plane. The numbers μ_- and μ_+ are well-defined, since $i\mathcal{R} \subseteq \mathcal{A}_0$, and they obey, in line with Proposition 2.1,

$$\mu_- + \mu_+ = s + 1, \quad |\mu_+ - \mu_-| \leq 1. \tag{3.18}$$

Next we deduce that at most two sectors of $\mathcal{A}_+$ that adjoin the origin in $\mathcal{C}^-$ may lie in $\mathcal{A}_+^\infty$ (moreover, obviously, $\mathcal{A}_+^\infty \cap \mathcal{C}^+ = \emptyset$): For suppose that more than two such sectors of $\mathcal{A}_+$ are in $\mathcal{A}_+^\infty$. By symmetry of the order star with respect to the real axis, they must enclose at least two bounded $\mathcal{A}_-$-regions that do not intersect $\mathcal{R}$. According to Proposition 2.3, each such region must contain a zero of ρ, and this contradicts our stipulation that R is multiply pz-resticted.

Because of (3.18), it transpires that at least s sectors of $\mathcal{A}_+$ at the origin belong to bounded $\mathcal{A}_+$-regions. However, it follows from Proposition 2.3 that the total multiplicity of all bounded $\mathcal{A}_+$-regions is precisely s. Hence there are precisely s such sectors of $\mathcal{A}_+$. Let $\tilde{\mathcal{A}}_+$ be such a region and suppose that it contains $\varsigma \geq 1$ sectors that adjoin the origin – hence, by Proposition 2.3, at least ς poles. Hence, $\tilde{\mathcal{A}}_+$ envelops $\varsigma - 1$ sectors of $\mathcal{A}_-$ at the origin and all such sectors lie in bounded $\mathcal{A}_-$-regions. Using again Proposition 2.3, it follows that $\tilde{\mathcal{A}}_+$ surrounds at least $\varsigma - 1$ zeros. The approximant being multiply pz-restricted, all these zeros are real. It follows readily from the symmetry of the order star that $\varsigma \leq 2$. Moreover, *either* $\varsigma = 1$ and $\tilde{\mathcal{A}}_+$ adjoins the origin via the real axis *or* $\varsigma = 2$ and $\tilde{\mathcal{A}}_+$ surrounds a portion of the real axis that belongs to $\mathcal{A}_-$. Because of the anti-symmetry with respect to the imaginary axis, exactly one bounded $\mathcal{A}_+$-region has $\varsigma = 1$ and a further $(s-1)/2$ bounded $\mathcal{A}_+$-regions satisfy $\varsigma = 2$.

Theorem 3.8 (Iserles, 1990) Given a symmetric multiply pz-restricted sth degree approximant R of order $s+1$, it is necessarily true that

(a) The integer s is odd;

(b) R has exactly $(s-1)/2$ poles in $\mathcal{C}^-$ and $(s+1)/2$ poles in $\mathcal{C}^+$; and

(c) There are at least $(s+1)/2$ distinct poles; at least $[(s+3)/4]$ in $\mathcal{C}^-$ and at least $[(s+1)/4]$ in $\mathcal{C}^+$.

Proof. We have already proved that order $s+1$ is inconsistent with even s, hence (a). To prove (b), we recall that the origin is adjoined by at most two sectors from $\mathcal{A}_+^\infty$, both from within $\mathcal{C}^-$. The remaining sectors of $\mathcal{A}_+$ belong to bounded regions, hence, because of Proposition 2.3, we need at least $\mu_- - 2$ poles of R in $\mathcal{C}^-$ and μ_+ poles in $\mathcal{C}^+$. The only configuration

that is consistent with odd s, (3.18) and anti-symmetry is

$$\mu_- = \frac{s+1}{2} + 1, \quad \mu_+ = \frac{s+1}{2}. \tag{3.19}$$

Thus, we need at least $(s-1)/2$ poles to the left and $(s+1)/2$ poles to the right of $i\mathcal{R}$. Since s poles are available, all the 'at most' and 'at least' statements of this paragraph can be replaced by 'exactly': *exactly* two sectors of $\mathcal{A}_+^\infty$ in $\mathcal{C}^-$, *exactly* $\mu_- - 2$ poles of R there etc. This proves (b).

Recall that there exists precisely one bounded $\mathcal{A}_+$-region with $\varsigma = 1$, otherwise the directed boundary of every bounded $\mathcal{A}_+$-region crosses the origin twice. Consequently, the order star consists of exactly $(s+1)/2$ bounded $\mathcal{A}_+$-regions and must have at least $(s+1)/2$ distinct poles (each of multiplicity either 1 or 2).

We deduce from (3.19) that

$$\begin{aligned} s \bmod 4 = 1 \quad &\Longrightarrow \quad \text{the unique } \mathcal{A}_+\text{-region with } \varsigma \text{ lies in } \mathcal{C}^+, \\ s \bmod 4 = 3 \quad &\Longrightarrow \quad \text{the unique } \mathcal{A}_+\text{-region with } \varsigma \text{ lies in } \mathcal{C}^- \end{aligned}$$

The proof can now be easily completed by counting multiplicities of bounded $\mathcal{A}_+$-regions. □

Theorem 3.8 can be used to some effect in the derivation of symplectic Runge–Kutta methods (Iserles, 1990). In particular, it emphasizes that p-restricted (as distinguished from multiply p-restricted) methods are of little benefit when $s \geq 3$.

It is a simple matter to use order stars to analyse symmetric approximants to $\exp z$ that possess q poles in $\mathcal{C} \setminus \mathcal{R}$ ($q = 0$ in Theorem 3.8). We leave this as an exercise for the reader.

4

A-acceptability barriers

That might control
The starry pole,
And fallen, fallen light renew!

From *Hear the Voice* by William Blake (1757–1827)

4.1 Motivation

In the last chapter, having motivated our study of rational approximants to $\exp z$, we examined barriers on their order or on combined degrees of interpolation. The emphasis on order is important, because of its correspondence with the order of some numerical schemes. However, the main interest in rational approximants to $\exp z$ focuses on their stability properties. We say that a rational function R is **A-acceptable** if $|R(z)| < 1$ for all $z \in C$, $\mathrm{Re}\, z < 0$.[1] The main virtue of A-acceptability is that it causes the underlying numerical method (implemented with a constant step length) to solve stably any stable linear equation. This has ramifications well outside linear analysis. For suppose that $\hat{\mathbf{y}}$ is a strongly attractive equilibrium of (3.1), where we require that $\mathbf{f} = \mathbf{f}(\mathbf{y})$, independent of t (an autonomous system). In the vicinity of $\hat{\mathbf{y}}$ the behaviour of the ordinary differential system can be explained in terms of the variational equation

$$\mathbf{y}' = J(\mathbf{y} - \hat{\mathbf{y}}), \quad J := \left.\frac{\partial \mathbf{f}}{\partial \mathbf{y}}\right|_{\mathbf{y}=\hat{\mathbf{y}}}.$$

Because of strong attractivity, all the eigenvalues of the Jacobian matrix J are in the complex left half-plane. Consequently, subject to A-acceptability and possible restriction on the step length (so that we remain 'in the vicinity'), attractivity of the original equilibrium is inherited by the numerical method (Iserles *et al.*, 1990).

[1] We have already defined A-acceptability in Chapter 1 and mentioned the definition again in Chapter 3. However, it is so central to the exposition in the present chapter that it certainly bears repeating.

The original stimulus behind the development of order stars was the well-known **first Ehle conjecture**, that the only A-acceptable Padé approximants to $\exp z$ occur when $0 \leq m \leq n \leq m+2$. We have already seen the proof in Chapter 1. In the present chapter we venture well beyond the Ehle conjecture.

4.2 Maximal approximants

Theorem 3.7 states that the maximal number of real interpolation points (counted with their degree) of any $R \in \pi_{m/n}$ is $m+n+1$. Any function that attains this barrier with nonnegative interpolation points is called a **maximal approximant**. This section is concerned with characterizing all A-acceptable maximal approximants.

We say that $R \in \pi_{m/n}$ is in the class $\mathcal{E}_{m/n,p}$ if it has order p. Note that, order p being the same as interpolation of degree $p+1$ at the origin, such a function has exactly $m+n-p$ negative and no positive interpolation points. Of course, every maximal approximant belongs to $\mathcal{E}_{m/n,p}$ for some $p \in \{-1, 0, \ldots, m+n\}$.[2] In particular, the Padé approximant $R_{m/n}$ is the unique (up to normalization) member of $\mathcal{E}_{m/n,m+n+1}$.

Among the earliest examples of maximal approximants that depart from Padé are $\mathcal{E}_{m/n,m+n}$ and $\mathcal{E}_{m/n,m+n-1}$, independently studied by Ehle and Picel (1975) and by Nørsett (1975). Even earlier, Varga (1962) contemplated the use of best $L_\infty(-\infty, 0]$ approximants from $\pi_{m/n}$ (see also (Cody *et al.*, 1969)). It follows at once from standard results in approximation theory (Cheney, 1966; Powell, 1981) that such functions lie in $\mathcal{E}_{m/n,0}$. The **second Ehle conjecture** (Ehle, 1976) claims that the A-acceptable maximal approximants are precisely all members of $\mathcal{E}_{m/n,p}$ that obey $p \geq 2n-2$. Since it is always true that $p \leq m+n$, the first Ehle conjecture is a special case.

Let us closely examine the second Ehle conjecture. Clearly, it consists of two statements:

(a) that all members of $\mathcal{E}_{m/n,p}$ for $p \geq 2n-2$ (thus, necessarily, $0 \leq m \leq n \leq m+2$) are A-acceptable; and

(b) that, whenever either m, n or p ventures outside these bounds, A-acceptability is lost.

Statement (a) was proved in three developments of increasing generality: Birkhoff and Varga (1965) showed that $\mathcal{E}_{n/n,2n+1}$ is A-acceptable, Ehle (1973) verified that this is the case with all of $\mathcal{E}_{m/n,m+n+1}$, $m \leq n \leq m+2$ and, finally, Ehle and Picel (1975) and Nørsett (1975) proved that the 'affirmative' part of the conjecture is true. It required several steps, as well, to affirm that statement (b) is valid: Clearly, no $\pi_{m/n}$ function may be A-acceptable when $m > n$. Nørsett (1975) showed that all functions in

[2] $p = -1$ means that the origin is not an interpolation point.

$\mathcal{E}_{m/n,m+n+1}$ are A-unacceptable when $m+3 \leq n \leq m+4$. Wanner *et al.* (1978) proved the first Ehle conjecture, namely that $m+3 \leq n$ leads to A-unacceptability in $\mathcal{E}_{m/n,m+n+1}$. This was the first paper to use (and, indeed, to introduce) order stars, the seed from which sprung the whole subject-matter of this book. Finally, the full statement – and, therefore, the second Ehle conjecture in its full glory – was proved by Iserles and Powell (1981), by extending the order star technique.

Fortunately, we do not require to repeat all these steps, since the whole conjecture can be proved quite easily by order stars. The secret of our success is in the exponential function itself being strictly bounded by unity in the left half-plane and $|\exp z| \equiv 1$ for $z \in i\mathcal{R}$. Thus, letting $\mathcal{V} = \{z \in \mathcal{C} : \mathrm{Re}\, z < 0\}$, the left half-plane, A-acceptability of a rational function R is nothing other than the requirement that R is a $\mathcal{V}^*$-contraction, as defined in Section 2.1. The following theorem is a variation on a result of Crouzeix and Ruamps (1977).

Theorem 4.1 If $R = P/Q \in \pi_{m/n}$ is a p-th order approximant to $\exp z$, $p \geq 2n-2$, such that $\lim_{|z|\to\infty} |R(z)| \leq 1$ and the coefficients of Q alternate in sign then R is A-acceptable.

Proof. Because of the maximal modulus theorem, we just need to prove that R is analytic in $\mathrm{cl}\,\mathcal{V}$ (i.e. that all the zeros of Q have positive real part) and $|R(it)| \leq 1$ for all $t \in \mathcal{R}$. Let

$$E(t) := |Q(it)|^2 - |P(it)|^2$$

be the **E-polynomial** (Nørsett, 1975) of the function R. Order p means that $R(z) = \exp z + \mathcal{O}(z^{p+1})$, hence

$$|R(z)| = \exp(\mathrm{Re}\, z) + \mathcal{O}(z^{p+1}).$$

Along $i\mathcal{R}$ this implies $|Q(it)| - |P(it)| = \mathcal{O}(t^{p+1})$. But

$$E(t) = (|Q(it)| - |P(it)|)(|Q(it)| + |P(it)|),$$

consequently $E(t) = \mathcal{O}(t^{p+1})$. Because we stipulate A-acceptability, necessarily $m \leq n$. Thus, $E \in \pi_{2n}[t]$. Moreover, it is an even polynomial. We now exploit the inequality $p \geq 2n-2$ to argue that there exists a constant $C \in \mathcal{R}$ such that $E(t) = Ct^{2n}$. Finally, $\lim_{|t|\to\infty} |R(it)| \leq 1$ implies that $C \geq 0$. But

$$|R(it)|^2 = 1 - C\frac{t^{2n}}{|Q(it)|^2},$$

therefore $|R(it)| \leq 1$ for all $t \in i\mathcal{R}$. All that remains to prove A-acceptability is that all the zeros of Q are in the open right half-plane.

We examine the order star of the first kind of $\{R, \exp z\}$. Since, by Proposition 2.1, $\iota(0) = p+1$, the origin is adjoined by at least $[(p+1)/2]$

sectors of $\mathcal{A}_+$ to the right of $i\mathcal{R}$. Since $|R(it)| \leq 1$, it is clear that $\mathcal{A}_+ \cap i\mathcal{R} = \emptyset$. Thus, none of these sectors belongs to $\mathcal{A}_+^\infty$ (which, as we have already seen in the proof of Theorem 3.6, approaches ∞ to the left of $i\mathcal{R}$) and all must belong to bounded $\mathcal{A}_+$-regions. Thus, in line with Proposition 2.3, we require at least $[(p+1)/2]$ poles of R (thus, zeros of Q) in the right half-plane.

Finally, we note that $[(p+1)/2] \geq n-1$. Thus, there might be at most one zero of Q to the left. Coefficients of Q being real, this zero must be negative. This is ruled out by the Descartes rule of signs (Pólya and Szegő, 1979), since the signs of the coefficients of Q alternate. Hence all the poles of R are to the right and the proof is complete. □

Corollary Padé approximants to $\exp z$ are A-acceptable for $0 \leq m \leq n \leq m+2$.

Proof. The order is $n+m \geq 2n-2$. Moreover, (3.12) and (3.13) implies that $\lim_{|z|\to\infty} |R(z)| \leq 1$ and the coefficients of Q alternate in sign. Thus, all three conditions of the last theorem are satisfied. □

We wish to prove that all members of $\mathcal{E}_{m/n,p}$, $p \geq 2n-2$, not just Padé approximants, are A-acceptable. First we need two technical results, of little virtue *per se*.

Lemma 4.2 Given any $n, m \geq 1$, it is true that

$$P_{m/n}(z) = P_{(m-1)/n}(z) + czP_{(m-1)/(n-1)}(z), \tag{4.1}$$

$$Q_{m/n}(z) = Q_{(m-1)/n}(z) + czQ_{(m-1)/(n-1)}(z), \tag{4.2}$$

where

$$c = \frac{n}{(m+n-1)(m+n)}.$$

Proof. The lemma is proved by straightforward substitution into (3.12) and (3.13). □

Lemma 4.3 (Ehle and Picel, 1975; Nørsett, 1975) For every $x_1, x_2 < 0$, $x_1 \neq x_2$, it is possible to find $R^{[1]} \in \mathcal{E}_{n/n,2n-1}$, $R^{[2]} \in \mathcal{E}_{(n-1)/n,2n-2}$ and $R^{[3]} \in \mathcal{E}_{n/n,2n-2}$ such that

$$R^{[1]}(x_1) = R^{[2]}(x_1) = R^{[3]}(x_1) = e^{x_1}, \quad R^{[3]}(x_2) = e^{x_2}.$$

Proof. Given any rational approximant $R = P/Q$ to $\exp z$, we denote the **modified error function** by

$$\psi(z) := P(z) - e^z Q(z).$$

In particular, we let $\psi_{m/n}(z) = P_{m/n}(z) - e^z Q_{m/n}(z)$. To prove the statement on $R^{[1]}$, we let

$$R^{[1]}(z) = \frac{\alpha P_{n/n}(z) + (1-\alpha)P_{(n-1)/n}(z)}{\alpha Q_{n/n}(z) + (1-\alpha)Q_{(n-1)/n}(z)}$$

and prove that there exists $\alpha \in \mathcal{R}$ such that $R^{[1]}(x_1) = \exp x_1$. We have

$$\psi(z) = \alpha\psi_{n/n}(z) + (1-\alpha)\psi_{(n-1)/n}(z) = \mathcal{O}(z^{2n}),$$

hence $R^{[1]}$ is, indeed, an order-$(2n-1)$ approximant. Moreover, $\psi(x_1) = 0$, the interpolation condition, is equivalent to

$$\left\{\psi_{n/n}(x_1) - \psi_{(n-1)/n}(x_1)\right\}\alpha = -\psi_{(n-1)/n}(x_1),$$

hence the 'right' α exists, unless

$$\psi_{n/n}(x_1) - \psi_{(n-1)/n}(x_1) = 0.$$

According to (4.1) and (4.2), the last equality reduces to

$$\frac{1}{2(2n-1)} x_1 \psi_{(n-1)/(n-1)}(x_1) = 0.$$

This is impossible according to Theorem 3.7, because $x_1 < 0$.

The proof for $R^{[2]}$ is almost identical: we are letting

$$R^{[2]}(z) = \frac{\alpha P_{(n-1)/n}(z) + (1-\alpha)P_{(n-2)/n}(z)}{\alpha Q_{(n-1)/n}(z) + (1-\alpha)Q_{(n-2)/n}(z)}.$$

It is left to the reader.[3]

Unsurprisingly, the more difficult part of the proof pertains to the double interpolation by $R^{[3]}$. We assume that $n \geq 2$ (the case $n = 1$ is both easier and, frankly, not very interesting, since the order is 0) and let

$$R^{[3]}(z) = \frac{\alpha P_{n/n}(z) + (1-\alpha-\beta)P_{(n-1)/n}(z) + \beta P_{(n-2)/n}(z)}{\alpha Q_{n/n}(z) + (1-\alpha-\beta)Q_{(n-1)/n}(z) + \beta Q_{(n-2)/n}(z)}.$$

Finding α and β to satisfy interpolation conditions is equivalent to solving the linear system

$$\begin{aligned} \left\{\psi_{n/n}(x_\ell) - \psi_{\underline{n-1}/n}(x_\ell)\right\}\alpha \; &- \; \left\{\psi_{\underline{n-1}/n}(x_\ell) - \psi_{\underline{n-2}/n}(x_\ell)\right\}\beta \\ &= \; -\psi_{\underline{n-1}/n}(x_\ell), \quad \ell = 1, 2. \end{aligned}$$

[3]Our representation of $R^{[2]}$ excludes the case $n = 1$. It can be solved in a straightforward manner and is left to the reader.

We now exploit (4.1) and (4.2) to argue that the determinant of this system is

$$-\frac{1}{4(2n-1)^2}\det\begin{bmatrix}\psi_{(n-1)/(n-1)}(x_1) & \psi_{(n-2)/(n-1)}(x_1)\\ \psi_{(n-1)/(n-1)}(x_2) & \psi_{(n-2)/(n-1)}(x_2)\end{bmatrix}.$$

Both x_1 and x_2 being negative, it may vanish only if

$$\frac{\psi_{(n-2)/(n-1)}(x_1)}{\psi_{(n-1)/(n-1)}(x_1)} = \frac{\psi_{(n-2)/(n-1)}(x_2)}{\psi_{(n-1)/(n-1)}(x_2)}. \tag{4.3}$$

Note first that it is impossible for $\psi_{(n-1)/(n-1)}$ and $\psi_{(n-2)/(n-1)}$ to vanish simultaneously, at $x_2 < 0$, say. Otherwise (4.1) and (4.2) will imply that $\psi_{(n-2)/(n-2)}(x_2) = 0$ and this is ruled out by Theorem 3.7. Thus, the function

$$\hat{R}(z) := \frac{\psi_{\frac{n-2}{n}}(x_2)P_{\frac{n-1}{n-1}}(z) - \psi_{\frac{n-1}{n-1}}(x_2)P_{\frac{n-2}{n}}(z)}{\psi_{\frac{n-2}{n}}(x_2)Q_{\frac{n-1}{n-1}}(z) - \psi_{\frac{n-1}{n-1}}(x_2)Q_{\frac{n-2}{n}}(z)}$$

is well defined. It resides in $\pi_{(n-1)/(n-1)}$ and approximates $\exp z$ of order $2n-3$. Moreover, it is trivial to verify that

$$\hat{R}(x_2) = e^{x_2}.$$

Therefore, $\hat{R} \in \mathcal{E}_{(n-1)/(n-1),2n-3}$, with an interpolation point at $x_2 < 0$, and Theorem 3.7 imply that $\hat{R}(x) \neq 0$ for all $x < 0$, $x \neq x_2$. It follows that (4.3) cannot be satisfied for $x_1 \neq x_2$, the determinant may not vanish and our proof is complete. □

Theorem 4.4 (Ehle and Picel, 1975; Nørsett, 1975) Every member of $\mathcal{E}_{m/n,p}$, $p \geq 2n-2$, is A-acceptable.

Proof. Our point of departure is Theorem 4.1. It is possible to prove that the constants α, β and $\alpha+\beta$ in the proof of Lemma 4.3 all reside in $[0,1]$. Hence, by expressions (3.12) and (3.13), all the conditions of Theorem 4.1 are satisfied. In this proof we pursue a different approach, based solely on order stars.

Let $R \in \mathcal{E}_{n/n,2n-1}$ interpolate at $x_1 < 0$ and suppose that it is not A-acceptable. According to the proof of Theorem 4.1, $|R(it)| \leq 1$ for all $t \in \mathcal{R}$, thus, necessarily, R must have a pole in the left half-plane. Consider now the approximants $R_\tau \in \mathcal{E}_{n/n,2n-1}$, such that

$$R_\tau(\tau x_1) = e^{\tau x_1}, \quad \tau \in [0,1].$$

Thus, $R_0 \equiv R_{n/n}$ is A-acceptable and has all its poles to the right of $i\mathcal{R}$, whereas $R_1 \equiv R$ has a pole to the left. Let us trace the poles as the parameter τ increases from 0 to 1. Poles are continuous in τ and they cannot 'jump' from the right to the left of $i\mathcal{R}$ via infinity, because then

they coincide there with a zero – and this is ruled out by Theorem 3.7. Consequently, as τ increases, they must cross $i\mathcal{R}$. Before the onset of this event, $\mathcal{A}_0$ must cross $i\mathcal{R} \setminus \{0\}$. This is impossible according to the proof of Theorem 4.1, since $E(t) = \mathcal{O}(t^{2n})$. Consequently, R is A-acceptable.

The proof for $\mathcal{E}_{(n-1)/n,2n-2}$ and $\mathcal{E}_{n/n,2n-2}$ is identical, except that, in the latter case, Lemma 4.3 cannot be used when $x_1 = x_2$ (i.e. $R(x) = \exp x_1 + \mathcal{O}(|x - x_1|^2)$). This can be fixed quite easily by a limiting argument and the details are left to the reader. □

Theorems 4.1 and 4.4 are somewhat exceptional: we have so far seen the technique of order stars as a 'negative' tool, handy in proving nonexistence, whereas these two theorems present it in a more 'positive' context. We revert back to form in our next theorem.

Theorem 4.5 (Iserles and Powell, 1981) A maximal approximant in $\mathcal{E}_{m/n,p}$ is A-acceptable only if $p \geq 2n - 2$.

Proof. We repeat, with minor modifications, the proof of the first Ehle conjecture from Chapter 1: because of Propositions 2.1 and 2.2, there are at least $[(p-1)/2]$ sectors of $\mathcal{A}_-$ that adjoin the origin from within $\mathcal{A}_+^\infty$, and they all belong to bounded $\mathcal{A}_-$-regions. Moreover, there are a further $p - m - n$ negative interpolation points, which are adjoined by bounded $\mathcal{A}_-$-regions. We count the sum of multiplicities of bounded $\mathcal{A}_-$-regions: because of Proposition 2.3, it is true that

$$\left[\frac{p-1}{2}\right] + (m + n - p) \leq m,$$

the number of available zeros. Thus, $[(p-1)/2] + n \leq p$, and this implies $p \geq 2n - 2$. □

We have already seen a few figures of order stars for maximal approximants, in particular Figures 3.3 and 3.4. Note that the first displays an A-acceptable approximant and the second depicts a non-A-acceptable one, exactly in line with our results.

Corollary The best $L_\infty(-\infty, 0]$ approximant from $\pi_{n/n}$ is not A-acceptable for all $n \geq 1$.

Proof. Since such an approximant possesses $2n + 1$ interpolation points in $(-\infty, 0)$, its A-unacceptability follows at once from Theorem 4.5. □

Some authors have advocated using best L_∞ approximants, mainly in the context of numerical solution of semi-discretized parabolic differential equations (Varga, 1962). It frequently makes perfect sense to relax acceptability requirements. However, the corollary calls for the exercise of

caution: although it does not sink L_∞ approximants, it nevertheless sends a shot across their bows.

4.3 Hairer's representation of rational approximants

The treatment of rational approximants in Section 4.2 was motivated by interpolation – we relaxed order to fit the exponential at specific negative points. An alternative approach is to abandon the connection to interpolation altogether and consider general n/n approximants of arbitrary order. As long as the order exceeds $n-1$, this can be done by using the theory of C-**polynomials** (Nørsett, 1975), which was mentioned in Section 3.2. We adopt here an alternative approach, expanding on the original succint account of Hairer (1982).

Let $M_n \equiv P_{n/n}$, the numerator of the n/n Padé approximant to the exponential. Thus, according to Lemma 3.4,

$$R_{n/n}(z) = \frac{M_n(z)}{M_n(-z)} = e^z + C_n z^{2n+1} + \mathcal{O}(z^{2n+2}),$$

where

$$C_n = (-1)^{n-1} \frac{(n!)^2}{(2n)!(2n+1)!}.$$

For purely technical reasons we define

$$M_{-1}(z) := -\frac{2}{z}.$$

The function $R = P/Q \in \pi_{n/n}[z]$ is said to be given in the **Hairer representation** if it is of the form

$$R(z) = \frac{f(z)M_{n-k}(z) + z^2 g(z) M_{n-k-1}(z)}{f(z)M_{n-k}(-z) + z^2 g(z) M_{n-k-1}(-z)}, \tag{4.4}$$

where $k \in \{0, 1, \ldots, n\}$, f and g are polynomials, $\deg f \leq k$, $\deg g \leq k-1$ and $f(0) = 1$ ($g \equiv 0$ if $k = 0$).

Theorem 4.6 (Hairer, 1982) The function R approximates the exponential to order $p \geq 2(n-k)$ if and only if it can be represented in the form (4.4).

Proof. Let us suppose that R obeys (4.4). Thus,

$$\begin{aligned} P(z) - e^z Q(z) &= f(z)\psi_{(n-k)/(n-k)}(z) + z^2 g(z) \psi_{(n-k-1)/(n-k-1)}(z) \\ &= \mathcal{O}(z^{2(n-k)+1}) \end{aligned}$$

and R is indeed of order $2(n-k)$ at least.

To establish the proof in the other direction we assume that $R = P/Q$, $Q(0) = 1$, is of the requisite order and observe that, according to Lemma 3.4 (and, if $m = 0$, our 'strange' definition of M_{-1}),

$$\begin{aligned} D_m(z) &:= M_m(z)M_{m-1}(-z) - M_m(-z)M_{m-1}(z) \\ &= M_{m-1}(-z)M_m(-z)\left\{\left(R_{m/m}(z) - e^z\right) - \left(R_{\frac{m-1}{m-1}}(z) - e^z\right)\right\} \\ &= -M_{m-1}(-z)M_m(-z)\left\{C_{m-1}z^{2m-1} + \mathcal{O}(z^{2m})\right\} \\ &= -C_{m-1}z^{2m-1} + \mathcal{O}(z^{2m}), \quad m = 0, 1, \ldots. \end{aligned}$$

But D_m belongs to $\pi_{2m-1}[z]$, hence it follows that

$$D_m(z) = -C_{m-1}z^{2m-1}. \tag{4.5}$$

Similarily, we can show that

$$\begin{aligned} P(z)M_{n-k-1}(-z) - Q(z)M_{n-k-1}(z) &= \mathcal{O}(z^{2(n-k)-1}), \\ P(z)M_{n-k}(-z) - Q(z)M_{n-k}(z) &= \mathcal{O}(z^{2(n-k)+1}). \end{aligned}$$

Thus, solving the linear algebraic system

$$\begin{bmatrix} M_{n-k}(z) & z^2M_{n-k-1}(z) \\ M_{n-k}(-z) & z^2M_{n-k-1}(-z) \end{bmatrix} \begin{bmatrix} f(z) \\ g(z) \end{bmatrix} = \begin{bmatrix} P(z) \\ Q(z) \end{bmatrix}$$

explicitly with Cramer's rule we show that f and g are polynomials of 'correct' degree. Letting $z = 0$ affirms that $f(0) = 1$ and completes the proof. □

Simple calculation affirms that $p \geq 2(m-k)+1$ if and only if

$$g(0) = \frac{1}{4(4(m-k)^2 - 1)}.$$

Although it is perfectly possible to obtain conditions for order $p \geq 2(m - k + 1)$, it is a much more sensible course of action in that case to use an appropriately smaller value of k.

Equation (4.4) lends itself well to the evaluation of the E-polynomial (cf. page 52). Using (4.5), we have

$$\begin{aligned} E(t) &= \left|M_{n-k}(-it)f(it) - t^2M_{n-k-1}(-it)g(it)\right|^2 \\ &\quad - \left|M_{n-k}(it)f(it) - t^2M_{n-k-1}(it)g(it)\right|^2 \\ &= t^2D_{n-k}(it)\{f(it)g(-it) - f(-it)g(it)\} \\ &= |C_{n-k}|t^{2(n-k)+1}\mathrm{Im}\,\{f(-it)g(it)\}. \end{aligned}$$

Lemma 4.7 The approximant R satisfies the inequality

$$|R(it)|^2 \leq 1, \quad t \in \mathcal{R}, \tag{4.6}$$

if and only if $\mathrm{Im}\{f(-it)g(it)\} \geq 0$ for all $t \geq 0$.

Proof. The proof follows at once from the definition of the E-polynomial. □

Example 4.1 (Hairer, 1982) Let us consider the set of all n/n approximants of order $p \geq 2n-6$. Thus, $k = 3$ and, given

$$\begin{aligned} f(z) &= 1 + f_1 z + f_2 z^2 + f_3 z^3, \\ g(z) &= g_0 + g_1 z + g_2 z^2, \end{aligned}$$

we are within the conditions of Lemma 4.7 if and only if $g_1 \geq g_0 f_1$, $g_2 f_3 \leq 0$ and

$$g_0 f_3 - g_1 f_2 + g_2 f_1 < 0 \quad \Rightarrow \quad (g_0 f_3 - g_1 f_2 + g_2 f_1)^2 \leq 4 g_2 f_3 (g_1 - g_0 f_1).$$

◇

To study the A-acceptability of R, subject to the satisfaction of (4.6), we examine the order star of the first kind with respect to

$$\rho(z) := \frac{R(z)}{R_{(n-k)/(n-k)}(z)}.$$

Thus, we consider R as an 'approximant' to a diagonal Padé approximant. Order stars of this type are sometimes called **relative order stars** (Hairer, 1982; Jeltsch and Nevanlinna, 1981; Wanner *et al.*, 1978). Note that the 'order' of ρ need not coincide with p. Indeed, simple calculation affirms that

$$\rho(z) = 1 - \frac{z^2 g(z) D_{n-k}(z)}{Q(z) M_n(-z)} = 1 + C_{n-k-1} \frac{z^{2(n-k)+1} g(z)}{Q(z) M_n(-z)}. \tag{4.7}$$

Hence the 'order' of ρ exceeds $2(n-k)$ if $g(0) = 0$.

The relative order star has some remarkable features: its zeros are precisely the zeros of R and the poles of $R_{(n-k)/(n-k)}$ and a symmetric statement holds for poles. The function ρ is rational, hence all regions, whether bounded or not, are analytic.

We stipulate in the remainder of the present section that $k \geq 1$ and that R is irreducible (that is, P and Q have no common zeros). Moreover, we disregard the case $g \equiv 0$, since it corresponds to a diagonal Padé approximant which, as we know from Theorem 4.4, is A-acceptable.

Theorem 4.8 (Hairer, 1982) An approximant R is A-acceptable if and only if the following conditions are satisfied:

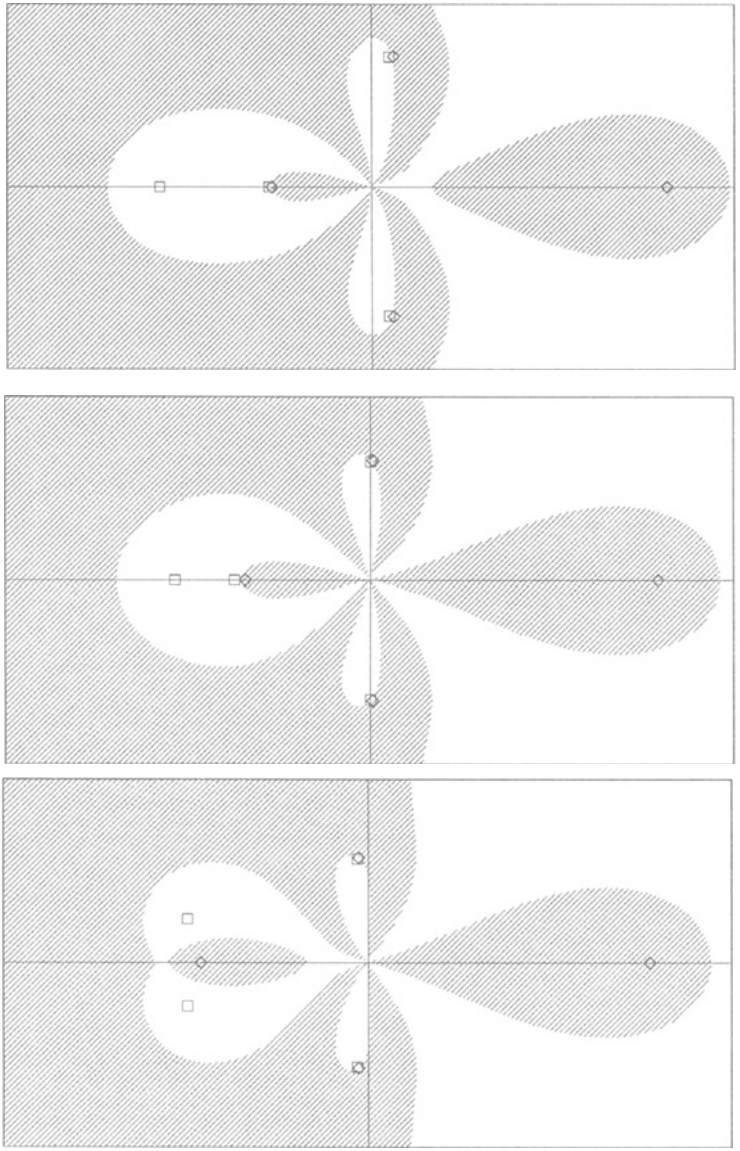


Figure 4.1 Relative order stars dying breed for $n = 4$, $K = 2$, $f(z) = 1 + z + z^2$ and $g(z) = \alpha + z$, $\alpha \in \left\{-\frac{1}{2}, 0, \frac{1}{2}\right\}$.

(a) R obeys (4.6);

(b) All the zeros of g lie in $\{z \in C : \mathrm{Re}\, z \geq 0,\ z \neq 0\}$;

(c) If g has a zero on $i\mathcal{R}$, $g(it_0) = 0$, say, then $it_0 g'(it_0) f(-it_0) > 0$;

(d) If $\deg g \leq \min\{\deg f, k\} - 2$ then $\lim_{|z|\to\infty} z^2 g(z) f(\bar{z}) > 0$.

Proof. A-acceptability is equivalent to condition (a) (that is, to $\mathcal{A}_+ \cap i\mathcal{R} = \emptyset$), in tandem with zeros of Q residing in the open left half-plane.

Suppose that (a) is satisfied and denote by ω_- and ω_+ the number of sectors of $\mathcal{A}_+$ that approach the origin to the left and to the right of $i\mathcal{R}$ respectively. Note that, because of (a), the imaginary axis does not intersect $\mathcal{A}_+$, hence these numbers are well defined.

Let us first assume that $g(0) \neq 0$. Thus, by Proposition 2.1, $\iota(0) = \omega_- + \omega_+ = 2(m-k)+1$ and

$$m - k \leq \omega_-, \omega_+ \leq m - k + 1.$$

If $g(0) > 0$ and $m-k$ is even then $C_{m-k-1} > 0$, thus, by (4.7) the interval $(0, x)$ for $0 < x \ll 1$ lies in $\mathcal{A}_+$. Consequently, ω_+ must be odd and we deduce that $\omega_- = m - k$. Moreover, $g(0) > 0$ and $m-k$ odd imply that $C_{m-k-1} < 0$, thus $(0, x) \in \mathcal{A}_-$ for $0 < x \ll 1$. Hence ω_+ is even – again, it equals $m-k+1$. On the other hand, if $g(0) < 0$ then an identical argument implies that $\omega_- = m - k + 1$.

If $g(0)$ vanishes then $\iota(0) > 2(m-k)+1$ and it follows at once from (a) that $\omega_- \geq m - k + 1$.

We conclude that $g(0) \leq 0$ implies $\omega_- \geq m - k + 1$, whereas $g(0) > 0$ means that $\omega_- = m - k$. If R is A-acceptable then there are at most $m-k$ poles of ρ to the left of $i\mathcal{R}$ – precisely the zeros of M_{m-k}, except for those that also annihilate P (the latter is possible only if $g(0) = 0$). Since all $\mathcal{A}_\infty$-regions are analytic, we need $\omega_- \leq m - k$. Therefore, A-acceptability cannot coexist with $g(0) \leq 0$.

Figure 4.1 displays relative order stars that illustrate the aforementioned cases of negative, zero and positive $g(0)$ respectively.

Suppose that $g(0) > 0$. Thus, $\omega_- = m - k$, just right to cater for the zeros of M_{m-k} in the left half-plane. In that case R has no poles there and is A-acceptable if and only if the following three conditions are obeyed:

(i) There are no interpolation points to the left of $i\mathcal{R}$;

(ii) No interpolation points on $i\mathcal{R} \setminus \{0\}$ may be 'supported' by poles of ρ in the left half-plane;

(iii) If $\rho(\infty) = 1$ (note that (a) implies that $|\rho(\infty)| \leq 1$) then it may not be approached from within the left half-plane by points of $\mathcal{A}_+$.

Observe that, according to (4.7), interpolation points away from the origin are precisely the zeros of g (inclusive of 'zero at infinity', which corresponds to reduction in degree). Hence, (i) is equivalent to condition (b). Suppose

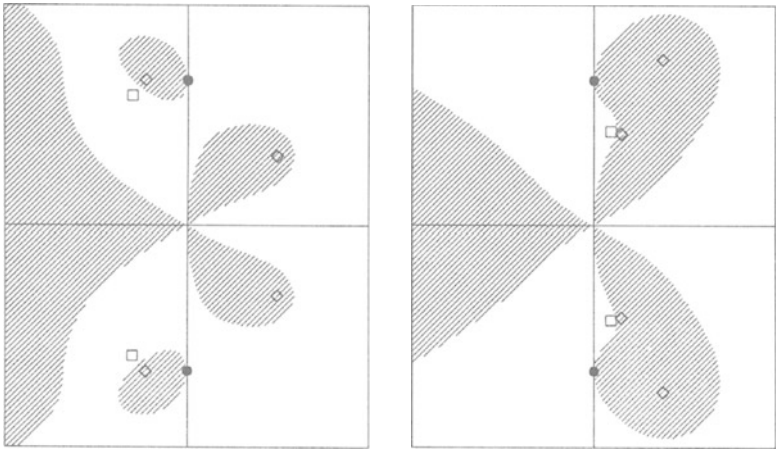


Figure 4.2 Relative order stars for $n = 4$, $K = 3$, $f(z) = 1 - z + f_2 z^2 - z^3$, $f_2 \in \{0, 2\}$, and $g(z) = 1 + z^2$.

that $g(it_0) = 0$, $f(it_0) \neq 0$, for some $t_0 \neq 0$. Then

$$M_{n-k}(it_0)Q(it_0) = f(it_0)|M_{n-k}(it_0)|^2 \neq 0$$

and, for small $\varepsilon > 0$,

$$\rho(it_0 + \varepsilon e^{i\tau}) \approx 1 + i\varepsilon \frac{|C_{n-k-1}|t_0^{2(n-k)+1}}{|M_{n-k}(it_0)|^2} e^{i\tau} \frac{g'(it_0)}{f(it_0)}.$$

Requiring $|\rho(it_0 + \varepsilon e^{i\tau})| < 1$ for $|\tau - \pi| < \pi/2$ demonstrates that (ii) and (c) are equivalent. Similar calculation affirms that (iii) and (d) are the same. This completes the proof. □

Example 4.2 (Hairer, 1982) Let $n = 4$, $k = 3$, $f(z) = 1 - z + f_2 z^2 - z^3$, $g(z) = 1 + z^2$. Thus,

$$\text{Im}\{g(it)f(-it)\} = t(1 - t^2) \geq 0, \quad t \geq 0,$$

hence, according to Lemma 4.7, condition (a) is satisfied irrespectively of the value of f_2. It is easy to check that (b) and (d) are obeyed as well. However, both zeros of g reside on the imaginary axis , hence (c) cannot be taken for granted. It is easy to verify that it is valid if and only if $f_2 > 1$, and

this is the necessary and sufficient condition for A-acceptability. Figure 4.2 displays the relative orders stars for $f_2 = 0$ and $f_2 = 2$ respectively and illustrated an important point in the proof of Theorem 4.8. $\diamond$

It is perfectly possible to rephrase Theorem 4.8 in the 'language' of zeros of f, rather than zeros of g and the proof is very similar indeed.

In section 4.6 we exploit a special case of Hairer's representation in our investigation of complex interpolation.

4.4 Restricted approximants

The first discussion of A-acceptability for p-restricted approximants (cf. Example 3.3) was presented by Nørsett (1974a; 1974b; 1978) and continued by Wolfbrandt (1977). The 'order-star paper' of Wanner *et al.* (1978) fully characterized all A-acceptable singly p-restricted Padé approximants. The latest and most definitive work in this subject is that of Orel (1990, 1991). Our exposition focuses on singly p-restricted approximants with $m, n \geq 1$ and it proceeds along the lines of Wanner *et al.* and Orel.

This is the place to recall Theorem 3.5: the singly p-restricted $[n/n]$ Padé approximant $S_{n/n}(\cdot, \gamma)$ (as introduced in Example 3.3) has order $n+1$ whenever $\gamma = \nu_{n,j}^{-1}$ for some $j \in \{1, \ldots, n\}$, where $\nu_{n,1}, \nu_{n,2}, \ldots, \nu_{n,n}$ are the zeros of L'_{n+1}. All other values of γ yield order n. The order $(n+1)$ approximants are also denoted by ${}_jS_{n/n} \equiv S_{n/n}(\cdot, \nu_{n,j}^{-1})$.

Theorem 3.5 can be generalized to cater for all $S_{m/n}$ such that $m \leq n$ (Nørsett, 1978).[4] Straightforward expansion into the series of

$$\sum_{k=0}^{m} p_k z^k = (1 - \gamma z)^n e^z + \mathcal{O}(z^{m+1})$$

verifies that order m is equivalent to

$$p_k = \sum_{\ell=0}^{k} (-1)^\ell \binom{n}{\ell} \frac{\gamma^\ell}{(k-\ell)!}, \quad k = 0, 1, \ldots, m.$$

Moreover, to attain order $m+1$ the parameter γ must obey

$$\sum_{\ell=0}^{\min\{m+1,n\}} (-1)^\ell \binom{n}{\ell} \frac{\gamma^\ell}{(m+1-\ell)!} = 0.$$

It is easy to identify the p_k's and the order $(m+1)$ condition explicitly in terms of special functions:

$$p_k = (-\gamma)^k L_k^{(n-k)}(\gamma^{-1}), \quad k = 0, \ldots, m \tag{4.8}$$

[4] There is nothing to prevent the contemplation of $S_{m/n}$ for $m > n$, except that, of course, A-acceptability is ruled out.

and

$$L^{(|n-m-1|)}_{\min\{n,m+1\}}(\gamma^{-1}) = 0 \tag{4.9}$$

respectively. Here $L^{(\alpha)}_n$ is the nth degree **generalized Laguerre polynomial**, which is orthogonal in the interval $[0,\infty)$ with respect to the weight function $x^\alpha \exp x$ (Rainville, 1967). Every nth degree orthogonal polynomial has n distinct zeros, all in the support of the underlying weight function. Thus, it is a consequence of (4.9) that for every such m and n there exist m distinct positive numbers, $\gamma_1^{(m/n)} < \gamma_2^{(m/n)} < \cdots < \gamma_m^{(m/n)}$, say, such that ${}_jS_{m/n} := S_{m/n}(\cdot,\gamma_j^{(m/n)})$ is of order $m+1$ for all $j = 1,\ldots,m$. Moreover, $S_{m/n}(\cdot,\gamma)$ is of order m for all $\gamma \in \mathcal{R} \setminus \{\gamma_1^{(m/n)},\ldots,\gamma_m^{(m/n)}\}$. Note, incidentally, that the $\gamma_\ell^{(m/n)}$ are ordered in the *opposite order* to that of the zeros of Laguerre polynomials.

Example 4.3 The approximant $S_{2/2}(\cdot,\gamma)$ can be written easily down explicitly:

$$S_{2/2}(z,\gamma) = \frac{1 + (1-2\gamma)z + \left(\frac{1}{2} - 2\gamma + \gamma^2\right) z^2}{(1-\gamma z)^2}.$$

As γ increases from 0, $S_{2/2}$ has an interesting 'life history'. At $\frac{1}{2} - \frac{\sqrt{3}}{6} = \gamma_1^{(2/2)}$ the order momentarily increases to 3. Further, at $1 - \frac{\sqrt{2}}{2} = \gamma_1^{(1/2)}$, the numerator is linear, hence we have an $S_{1/2}$ approximant of maximal order 2. Even further, at $\frac{1}{2} = \gamma_1^{(1/1)}$, a zero and a pole coalesce and the function is reducible to the $[1/1]$ Padé approximant, which is coincidentally an $S_{1/1}$ function of maximal order 2. Finally, at $\frac{1}{2} + \frac{\sqrt{3}}{6} = \gamma_2^{(2/2)}$ the order is yet again boosted to 3. We prove in the remainder ot this section that this pattern is universal.

It is elementary to verify that A-acceptability is equivalent to $\gamma \geq \frac{1}{4}$. In particular, it follows that ${}_1S_{1/2}$, ${}_1S_{1/1} \equiv R_{1/1}$ and ${}_2S_{2/2}$ – but not ${}_1S_{2/2}$ – are A-acceptable. $\diamond$

Recall from Chapter 3 that there are just two unbounded components in the underlying order star, which we have denoted by $\mathcal{A}_+^\infty$ and $\mathcal{A}_-^\infty$. Moreover, S having a single pole, there exists (by Proposition 2.3) exactly one bounded $\mathcal{A}_+$-region, which we denote by $\mathcal{A}_+^\circ$.

Given an open set $\mathcal{U} \subset \mathcal{C}$ with a Jordan boundary, we define its **external boundary** as the boundary of the unique connected component of the closed set $\mathrm{cl}\mathcal{C} \setminus \mathcal{U}$ that contains ∞. Intuitively speaking, the external boundary is the boundary of the set that is obtained by filling in all the 'holes' that may exist inside $\mathcal{U}$. We denote by Γ_L and Γ_R the external boundaries of $\mathcal{A}_+^\infty$ and $\mathcal{A}_+^\circ$ respectively. Moreover, we designate by the symbols $\Theta_L(\gamma)$ and $\Theta_R(\gamma)$ the number of zeros of $S_{m/n}(\cdot,\gamma)$ to the left of Γ_L and inside

Γ_R respectively.[5]

Lemma 4.9 Given a singly p-restricted approximant $S_{m/n}(\cdot,\gamma)$ of order $p \geq m$, it is true that

$$\Theta_L(\gamma) + \Theta_R(\gamma) \begin{cases} \geq m & : p = m, \\ = m+1 & : p = m+1. \end{cases} \tag{4.10}$$

Proof. All the sectors of $\mathcal{A}_+$ approaching the origin may belong either to $\mathcal{A}_+^\infty$ or to $\mathcal{A}_+^\circ$. Therefore, there are at most two[6] sectors of $\mathcal{A}_-^\infty$ that adjoin the origin. According to Proposition 2.1, the origin is adjoined by $p+1$ sectors of $\mathcal{A}_-$, hence, by Proposition 2.3, at least $p-1$ zeros of $S_{m/n}$ must reside in $\mathcal{A}_-$-regions that are surrounded either by $\mathcal{A}_+^\infty$ or by $\mathcal{A}_+^\circ$. The lemma follows from the definition of external boundary. □

A central consideration to A-acceptability analysis of singly p-restricted approximants is the variation of Θ_L and Θ_R as a function of γ. We already know that it is constrained by (4.10).

Lemma 4.10 Let p_ℓ, $\ell = 0,\ldots,m$, be polynomials in a parameter β and $p_0 \equiv 1$. Define the polynomial P by

$$P(z,\beta) := \sum_{\ell=0}^{n} p_\ell(\beta) z^\ell$$

and let $r(\beta)$ be a real root of $P(z,\beta) = 0$. As a function of β, r moves from one complex half-plane to the other (i.e. from $\mathcal{C}^\pm$ to $\mathcal{C}^\mp$) if and only if β is a zero of p_m of odd multiplicity. Moreover, suppose that $p_m(\beta^*) = 0$, $p_m'(\beta^*) \neq 0$ and

$$C := \frac{p_{m-1}(\beta^*)}{p_m'(\beta^*)} < 0.$$

Then, as β increases through β^*, a real zero $r(\beta)$ jumps from $\mathcal{C}^-$ to $\mathcal{C}^+$.

Proof. The product of the roots being $p^*(\beta) := -1/p_m(\beta) \neq 0$, no real zero can 'jump' a half-plane through the origin. Movement from $\mathcal{C}^\pm$ to $\mathcal{C}^\mp$ can occur either via $i\mathcal{R}\setminus\{0\}$ or through ∞. The first option is ruled out for real zeros, whereas for the second $p^*(\beta)$ becomes unbounded, hence $p_m(\beta)$ must vanish. It is elementary that β must be a zero of odd multiplicity, otherwise $r(\beta)$ 'bounces' back from ∞ into the original half-plane.

[5]The definition of Θ_R is open to certain ambiguity, since it may happen – only to the right of Γ_R – that a zero and a pole coincide. In that case we may replace the numerator and the denominator by relatively prime lower-degree polynomials. To remove the ambiguity we forbid this procedure and count a zero even if it is 'removable'.

[6]In fact, exactly two.

Suppose now that β^* is a simple zero of p_m and $C < 0$. Since

$$\frac{P(z,\beta)}{p_m(\beta)} \approx z^m + \frac{C}{\beta - \beta^*} z^{m-1} + \cdots, \qquad |\beta - \beta^*| \ll 1,$$

it is true that

$$r(\beta) \approx \frac{C}{\beta^* - \beta}$$

and a zero jumps at β^* from left to right. □

As soon as we identify P with the numerator of $S_{m/n}$ and β with γ, we can note that, as a consequence of (4.8), all the zeros of $p_m(\gamma)$ are simple, hence of odd multiplicity. Thus, each passage of γ^{-1} through a zero of the 'right' Laguerre polynomial corresponds to a jump of a zero of $S_{m/n}(\cdot,\gamma)$ from $\mathcal{C}^{\pm}$ to $\mathcal{C}^{\mp}$. However, $p_m(\gamma) = 0$ implies that $S_{m/n}(\cdot,\gamma)$ is an $(m-1)/n$ function. In other words, $\gamma = \gamma_j^{((m-1)/n)}$ for some $j \in \{1,\ldots,m-1\}$.

Considering $S_{m/n}$ and its order star as a function of the parameter γ, three types of event are of interest:

(a) The aforementioned jumps of a zero from a half-plane to a half-plane, $\gamma = \gamma_j^{((m-1)/n)}$ for some j;

(b) Values $\gamma = \gamma_\ell^{(m/n)}$, whereby the order is increased to $m+1$; and

(c) Values $\gamma = \gamma_k^{((m-1)/(n-1))}$, whereby $S_{m/n}$ reduces to $S_{(m-1)/(n-1)}$.

According to the Markov theorem (Szegő, 1939), zeros of $L_n^{(\alpha)}$ are monotone in α. It follows that the events (a) and (b) interlace. Moreover, elementary manipulation of mixed recurrences for Laguerre polynomials (Rainville, 1967) affirms that there is a $\gamma_k^{((m-1)/(n-1))}$ between any $\gamma_j^{((m-1)/n)}$ and $\gamma_\ell^{(m/n)}$. Consequently,

$$\begin{aligned} \gamma_1^{(m/n)} &< \gamma_1^{((m-1)/n)} < \gamma_1^{((m-1)/(n-1))} < \gamma_2^{(m/n)} < \gamma_2^{((m-1)/n)} < \cdots \\ &< \cdots < \gamma_{m-1}^{((m-1)/n)} < \gamma_{m-1}^{((m-1)/(n-1))} < \gamma_m^{(m/n)}. \end{aligned} \tag{4.11}$$

Figure 4.3 displays an 'order-star movie' for $m = n = 3$: γ passes through the following eight values:

	γ	approximant	Θ_L	Θ_R
1	0.12		3	0
2	0.128886...	${}_1S_{3/3}$	3	0
3	0.135		3	0
4	0.158984...	${}_1S_{2/3}$	2	0
5	0.19		2	0
6	0.211325...	${}_1S_{2/2}$	2	1
7	0.27		2	1
8	0.302535...	${}_2S_{3/3}$	2	1

Clearly, the 'movie' is consistent with (4.11). Moreover, we observe that both Θ_L and Θ_R are weakly monotone: the first decreasing and the second increasing.

Lemma 4.11 The integer function Θ_L is monotonically decreasing from m to 0 as γ traverses the interval $(0, \infty)$.

Proof. Upon $\gamma = 0$ the rational function $S_{m/n}$ reduces to an mth degree polynomial. There are no bounded $\mathcal{A}_+$-regions in the underlying order star. The order being m, there are $m+1$ sectors of $\mathcal{A}_-$ at the origin. Because of the absence of bounded $\mathcal{A}_+$-regions, just one such sector may belong to $\mathcal{A}_+^\infty$ and the rest lie in bounded $\mathcal{A}_-$-regions. These regions are surrounded by the unique $\mathcal{A}_+$-region, namely $\mathcal{A}_+^\infty$, and it follows that $\Theta_L(0) = m$.

The desired result follows by demonstrating that the quantities C from Lemma 4.10 are negative for all $\gamma = \gamma_\ell^{(m/n)}$. Using (4.8) we can express them in terms of Laguerre polynomials and make use of the identity

$$\frac{\mathrm{d}L_n^{(\alpha)}(z)}{\mathrm{d}z} = -L_{n-1}^{(\alpha)}(z)$$

(Rainville, 1967). Thus, for $m \leq n$ we have

$$\begin{aligned} p_{m-1}(\gamma) &= (-\gamma)^{m-1} L_{m-1}^{(n-m+1)}(\gamma^{-1}), \\ p_m(\gamma) &= (-\gamma)^m L_m^{(n-m)}(\gamma^{-1}) \end{aligned}$$

consequently

$$p_m'(\gamma) = -m(-\gamma)^{m-1} L_m^{(n-m)}(\gamma^{-1}) + (-\gamma)^{m-2} L_{m-1}^{(n-m+1)}(\gamma^{-1})$$

and

$$C = C(\gamma_\ell^{(m/n)}) = -\gamma_\ell^{(m/n)} < 0, \quad \ell = 1, \ldots, m.$$

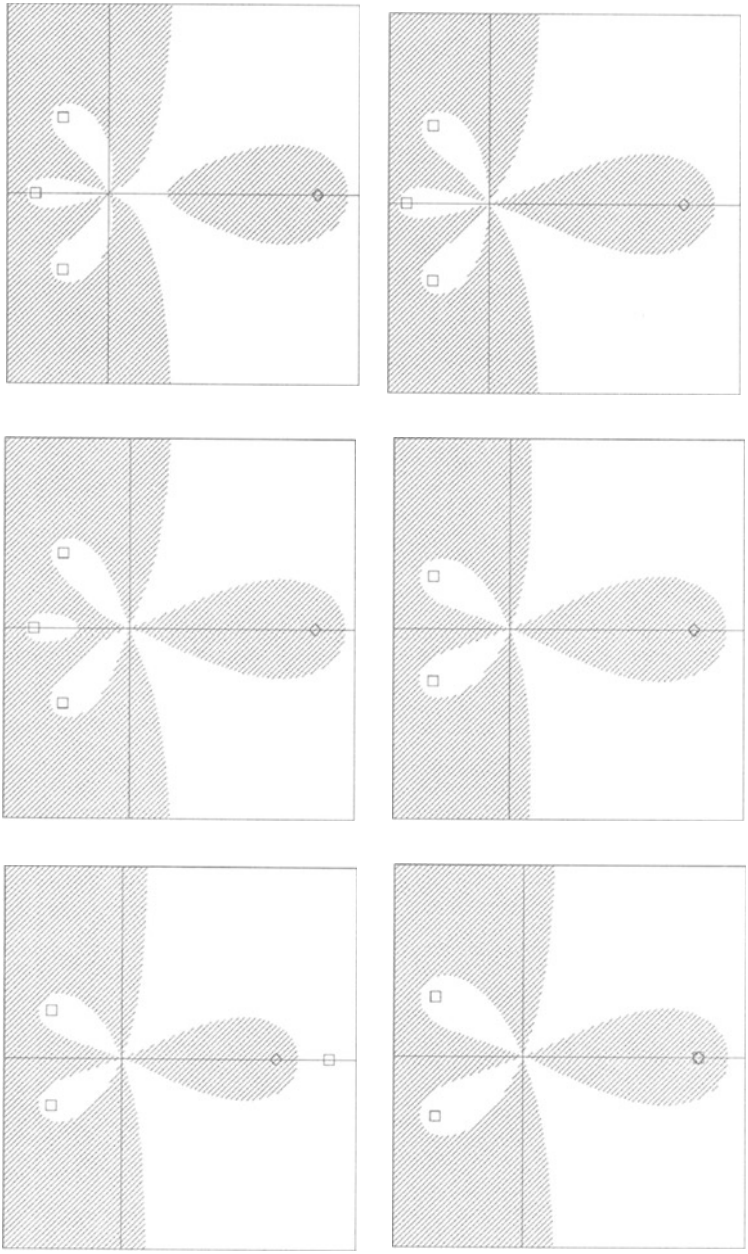
A similar proof can be used for $m = n+1$ and the lemma is true. □

Theorem 4.12 The singly p-restricted approximant ${}_\ell S_{m/n}$, $0 \leq m \leq n+1$, of order $m+1$ can be A-acceptable only when

$$\left[\frac{m}{2}\right] + 1 \leq \ell \leq \left[\frac{m+1}{2}\right] + 1. \tag{4.12}$$

Proof. Order $m+1$ and Proposition 2.1 imply that the origin is approached by exactly $\Theta_L + 1$ sectors of $\mathcal{A}_+$ from within $\mathcal{A}_+^\infty$ and $\Theta_R + 1$ sectors of $\mathcal{A}_+$ that belong to $\mathcal{A}_+^\circ$. All these sectors are equiangular and, by A-acceptability and Proposition 2.2, they must stay clear of the imaginary axis. This implies that

$$\Theta_L, \Theta_R \leq \left[\frac{m+3}{2}\right]. \tag{4.13}$$



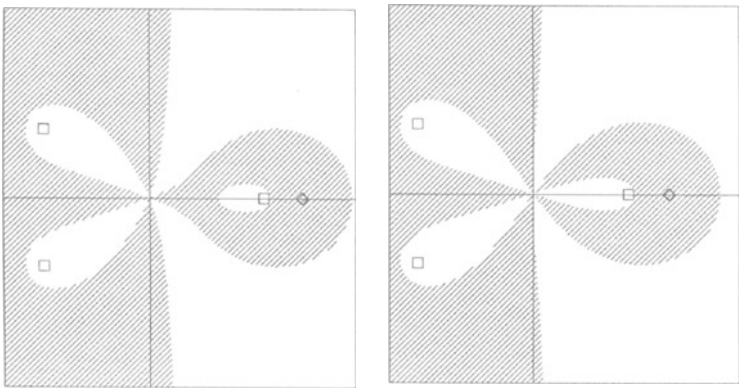


Figure 4.3 Order star 'movie' of $S_{3/3}(\cdot,\gamma)$ for eight different values of γ.

We presently exploit Lemmas 4.9 and 4.11 to deduce that $\gamma = \gamma_\ell^{(m/n)}$ implies $\Theta_L = m + 1 - \ell$, $\Theta_R = \ell - 1$. Substitution into (4.13) completes the proof. □

It is important to emphasize that inequality (4.12) is necessary for A-acceptability, but it is by no means sufficient. Herewith we find – with no help from order stars – all the A-acceptable approximants for the special case $m = n$. Nørsett (1974a) verified that ${}_1S_{1/1}$, ${}_2S_{2/2}$, ${}_3S_{3/3}$ and ${}_4S_{5/5}$ are A-acceptable (they are all, of course, consistent with (4.12)). Amazingly enough, these are all the A-acceptable approximants of that form:

Theorem 4.13 (Wanner *et al.*, 1978) The A-acceptable singly p-restricted approximants ${}_\ell S_{n/n}$ are precisely ${}_1S_{1/1}$, ${}_2S_{2/2}$, ${}_3S_{3/3}$ and ${}_4S_{5/5}$.

Proof. The proof is technical – here we just outline its main steps and ideas, leaving the details to the more masochistic among our readers. It follows from (4.8) that

$$\lim_{|z|\to\infty} \left|{}_\ell S_{n/n}(z)\right| = \left|L_n\left(1/\gamma_\ell^{(n/n)}\right)\right|.$$

Hence, it is necessary for stability that $|L_n(1/\gamma_\ell^{(n/n)})| \leq 1$. Laguerre polynomials for large values of n can be represented by the Fejér–Perron asymptotic formula (Szegő, 1939):

$$L_n^{(\alpha)}(x) = \pi^{-\frac{1}{2}} x^{-\frac{1}{2}\alpha-\frac{1}{4}} n^{\frac{1}{2}\alpha-\frac{1}{4}} e^{\frac{1}{2}x} \cos\left(2\sqrt{nx} - \frac{(2\alpha+1)\pi}{4}\right) + \mathcal{O}\left(n^{\frac{1}{2}\alpha-\frac{3}{4}}\right)$$

uniformly for all x in a compact positive interval. Thus, the ℓth extremum point of $L_{n+1}^{(1)}$ is

$$\frac{1}{\gamma_{n+1-\ell}^{(n/n)}} \approx \frac{\pi^2\left(\ell+\frac{1}{4}\right)}{4(n+1)}.$$

Laguerre polynomials obey the recurrence

$$(n+1+\alpha)L_n^{(\alpha)}(x) = (n+1)L_{n+1}^{(\alpha)}(x) - nL_n^{(\alpha+1)}(x)$$

(Rainville, 1967). By definition

$$L_n^{(1)}\left(1/\gamma_{n+1-\ell}^{(n/n)}\right) = 0,$$

therefore

$$L_n\left(1/\gamma_{n+1-\ell}^{(n/n)}\right) = L_{n+1}\left(1/\gamma_{n+1-\ell}^{(n/n)}\right).$$

It follows after elementary manipulation that

$$\left|L_n\left(1/\gamma_{n+1-\ell}^{(n/n)}\right)\right| \le 1 \quad \Leftrightarrow \quad n+1 \ge \frac{\pi^2\left(\ell+\frac{1}{4}\right)^2}{2\log\frac{\pi^4\left(\ell+\frac{1}{4}\right)^2}{4}}.$$

This yields the A-acceptability condition

$$n \ge \begin{cases} 6(n-\ell)-4 & : \ell = 1,2,\ldots,n-3, \\ 9 & : \ell = n-2, \\ 5 & : \ell = n-1, \\ 1 & : \ell = n. \end{cases}$$

The last inequality, in tandem with (4.12), reduces the field to a finite number of candidates. These are checked one by one and only the four approximants from the statement of the theorem are shown to be A-acceptable. □

No characterization of A-acceptable approximants, along the lines of Theorem 4.13, is available for $m < n$. The main problem is that, the degree of the numerator being strictly smaller than the degree of the denominator, we have $\lim_{|z|\to\infty} S_{m/n}(z,\gamma) = 0$. Thus, the technique that by excluding most candidates for A-acceptability was central to the proof of Theorem 4.13 is of little help in the general case. Orel (1990) computed the values shown in Table 4.10.

		m										
		0	1	2	3	4	5	6	7	8	9	10
	0	–	–	–	–	–	–	–	–	–	–	–
	1	1	1	–	–	–	–	–	–	–	–	–
	2	1	1,2	2	–	–	–	–	–	–	–	–
	3	1	1,2	2	3	–	–	–	–	–	–	–
	4	1	1,2	2	3	–	–	–	–	–	–	–
	5	1	1,2	2	3	3	4	–	–	–	–	–
n	6	1	1,2	2	3	3	4	–	–	–	–	–
	7	1	1,2	2	2,3	3	4	–	–	–	–	–
	8	1	1,2	2	2,3	3	4	4	5	–	–	–
	9	1	1,2	2	2,3	3	4	4	5	–	–	–
	10	1	1,2	2	2,3	3	4	4	5	–	–	–
	11	1	1,2	2	2,3	3	4	4	5	–	6	–
	12	1	1,2	2	2,3	3	4	4	5	5	6	–

Table 4.1 Values of ℓ for all the A-acceptable singly p-restricted approximants ${}_{\ell}S_{m/n}$, $m = 0, 1, \ldots, 10$, $n = 0, 1, \ldots, 12$.

Orel advances two interesting conjectures on A-acceptability of singly p-restricted approximants. They are presented in Chapter 10.

4.5 Interpolation with p-restricted approximants

Section 4.2 was devoted to interpolation of $\exp z$ by 'relaxing' Padé approximants $R_{m/n}$, whereas the theme of section 4.4 is A-acceptability – or otherwise – of restricted Padé approximants $S_{m/n}$. It is natural to pose the question how well we can do with the raw materials of Section 4.4, p-restricted approximants, to solve the major problem of section 4.2, namely A-acceptable interpolation in $(-\infty, 0]$.

Recall that

$$S_{n/n}(z;\gamma) = \frac{\sum_{k=0}^{n}(-1)^k L_k^{(n-k)}(1/\gamma)(\gamma z)^k}{(1-\gamma z)^n} \tag{4.14}$$

and let

$$E_n(z;\gamma) := S_{n/n}(z;\gamma) - e^z$$

be the approximation error. As we already know from Section 3.4,

$$E_n(z;\gamma) = \mathcal{O}(z^{p+1}),$$

where $p = n$, unless $L_n'(1/\gamma) = 0$, when $p = n+1$. We again denote the special values of γ by $\gamma_j^{(n/n)}$, $j = 1, \ldots, n$, ordered as in (4.11).

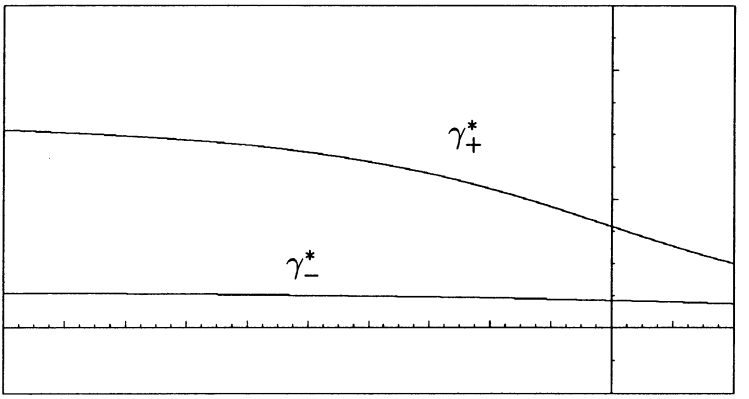


Figure 4.4 A graph of $\gamma_\pm^*$, the interpolating parameters from Example 4.4.

Given $x^* \in (-\infty, 0)$, we wish to determine γ^* such that $E_n(x^*;\gamma^*) = 0$. Nørsett and Trickett (1984) have discussed this problem for general $S_{m/n}$, $m \leq n$, and here we follow in their footsteps.

Example 4.4 Letting $E_2(x^*;\gamma) = 0$ produces a quadratic in γ, with the solutions

$$\gamma_\pm^*(x^*) = \frac{1 + x^* - e^{x^*} \pm \sqrt{x^*(1+e^{x^*})\left(\frac{1}{2}x^* - \tanh\frac{1}{2}x^*\right)}}{x^*(1-e^{x^*})}.$$

Figure 4.4 displays the graphs of the solutions for $x^* \in (-10, 2)$ and we observe that they are both real. Hence, interpolation is possible for $n = 2$ and we trade the third order of $S_{2/2}$ off, in exchange for second order *and* interpolation. ◇

A similar result is valid for all $n \geq 1$: all the n roots of the polynomial equation $E_n(x^*;\gamma) = 0$ are real. To prove this statement, we require a few classical identities on Laguerre polynomials, namely:

$$L_n^{(\alpha)}(x) = L_{n-1}^{(\alpha)}(x) + L_n^{(\alpha-1)}(x) \tag{4.15}$$

$$xL_{n-1}^{(\alpha+1)}(x) = (\alpha+n)L_{n-1}^{(\alpha)}(x) - nL_n^{(\alpha)}(x) \tag{4.16}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}L_n^{(\alpha)}(x) = -L_{n-1}^{(\alpha+1)}(x) \tag{4.17}$$

(Rainville, 1967).

Lemma 4.14 It is true that

$$\frac{\partial}{\partial\gamma}E_n(z;\gamma) = \frac{n}{\gamma^2}\left(\frac{-\gamma z}{1-\gamma z}\right)^{n+1} L_{n-1}^{(1)}(1/\gamma). \tag{4.18}$$

Proof. The partial derivatives of E_n and $S_{n/n}$ with respect to γ coincide, hence we can apply $\partial/\partial\gamma$ to (4.14). The identity (4.17) yields

$$\begin{aligned}\frac{\partial}{\partial\gamma}E_n(z;\gamma) &= (1-\gamma z)^{-n-1}\left\{nz\sum_{k=0}^{n}(-\gamma z)^k L_k^{(n-k)}(1/\gamma)\right.\\ &\quad + \frac{1-\gamma z}{\gamma}\sum_{k=1}^{n}(-\gamma z)^k L_{k-1}^{(n-k+1)}(1/\gamma)\\ &\quad \left. + \frac{1-\gamma z}{\gamma}\sum_{k=0}^{n}(-\gamma z)^k k L_k^{(n-k)}(1/\gamma)\right\}.\end{aligned}$$

We next use (4.15) and (4.16). This produces

$$\begin{aligned}\frac{\partial}{\partial\gamma}E_n(z;\gamma) &= \frac{n}{(1-\gamma z)^{n+1}}\left\{z\sum_{k=0}^{n}(-\gamma z)^k L_k^{(n-k)}(1/\gamma)\right.\\ &\quad \left. + \frac{1-\gamma z}{\gamma}\sum_{k=1}^{n}(-\gamma z)^k L_{k-1}^{(n-k)}(1/\gamma)\right\}\end{aligned}$$

and further trivial manipulation completes the proof. □

Recall that the order increases by one when $\gamma = \gamma_k^{(n/n)}$. A reformulation of the last sentence is that $S_{n/n}(\,\cdot\,,\gamma_k^{(n/n)})$ possesses $n+2$ interpolation points at the origin. As γ moves away from $\gamma_k^{(n/n)}$, $n+1$ interpolation points stay at 0 and it is natural to suspect that the remaining one moves 'away'. A perusal of Figure 4.3 reinforces this suspicion, as does the discussion preceding Lemma 4.11. The next lemma affirms it as true.

Lemma 4.15 (Nørsett and Trickett, 1984) Given any fixed $x^* \in (-\infty,0)$, there exist precisely n distinct real parameters

$$\gamma_1^*(x^*) < \gamma_2^*(x^*) < \cdots < \gamma_n^*(x^*)$$

such that $E_n(x^*;\gamma_k^*) = 0$ for $k = 1,2,\ldots,n$. Moreover,

$$\gamma_k^*(x^*) \in \left(\gamma_k^{(n/n)}, \gamma_k^{((n-1)/n)}\right)$$

and

$$\lim_{x^* \uparrow 0} \gamma_k^*(x^*) = \gamma_k^{(n/n)}, \quad \lim_{x^* \downarrow -\infty} \gamma_k^*(x^*) = \gamma_k^{((n-1)/n)}$$

for all $k = 1, 2, \ldots, n$.

Proof. Since $E_n(x^*, \gamma^*) = 0$ boils down to an nth degree polynomial equation in γ^*, the existence of n complex parameters follows at once. We need just to prove that they are all real, reside in the stated intervals and possess the right asymptotic behaviour when x^* approaches the end-points of $(-\infty, 0)$.

It is possible to furnish a proof by conventional means, representing

$$E_n(x;\gamma) = e^z \int_0^{-x} \left(\frac{\gamma\tau}{1+\gamma\tau}\right)^{n+1} \left(L_n(1/\gamma) - \frac{1}{\gamma\tau} L_n^{(1)}(1/\gamma)\right) \mathrm{d}\tau,$$

carefully monitoring the sign of E_n as x travels from the origin to $-\infty$ and exploiting the continuity of the γ_k^*'s. An alternative approach, left to the reader, is to use order stars, in line with the analysis in section 4.4. □

So far we have always 'moved' from the interpolation space to the parameter space. It is interesting to note that the converse of Lemma 4.15 is true and we are allowed to 'move' in the opposite direction: for every parameter value in the 'right' range there exists an interpolation point.

Lemma 4.16 (Nørsett and Trickett, 1984) Let

$$\mathcal{I}_n := \bigcup_{k=1}^{n} (\gamma_k^{(n/n)}, \gamma_k^{((n-1)/n)}).$$

For every $\gamma \in \mathcal{I}_n$ there exists a unique $x^* \in (-\infty, 0)$ such that $E_n(x^*, \gamma) = 0$.

Proof. Compare the estimate

$$E_n(z;\gamma) = -\gamma^{-1} L_n^{(1)}(1/\gamma)(-\gamma z)^{n+1} + \mathcal{O}(|z|^{n+2}), \quad z \to 0,$$

(Nørsett, 1978) with

$$E_n(z;\gamma) \approx L_n(1/\gamma), \quad |z| \to \infty,$$

(the latter follows at once from (4.14)). These are precisely the cases (b) and (a) respectively from the discussion preceding Lemma 4.11 and it follows at once from the definition of the $\gamma_k^{(m/n)}$'s as reciprocals of zeros of Laguerre polynomials that

$$(-1)^n L_n^{(1)}(1/\gamma) L_n(1/\gamma) < 0 \quad \gamma \in \mathcal{I}_n.$$

Thus, E_n changes sign in $(-\infty, 0)$, proving the existence of a negative interpolation point. Supposing that more than one such point exists, we

contradict the known structure of the order star that has been used in section 4.4 – this can be shown readily by a standard zero-counting argument and is left to the reader. □

The analysis so far has been more in the spirit of Chapter 3 – the relationship between parameters and interpolation. We now add to our discussion the extra ingredient of A-acceptability. Bearing in mind Theorem 4.13, it is clear that there is very little hope for A-acceptability when $n \geq 6$ or $n = 4$. Not being particularly fond of lost causes, we simply reproduce a table, as computed by Burrage (1988), of intervals Γ_n of A-acceptable γ's for $n = 1, 2, 3$ (see Table 4.2). It is instructive to compare these sets with the $\mathcal{I}_n$'s.

$n \setminus k$	Γ_n	$\mathcal{I}_n$
1	$(0.5000, \infty)$	$(0.5000, 1.0000)$
2	$(0.2500, \infty)$	$(0.1667, 0.2113) \cup (0.5000, 0.78868)$
3	$(0.3333, 1.0686)$	$(0.1090, 0.1289) \cup (0.2319, 0.3025) \cup (0.6590, 1.0686)$

Table 4.2 A-acceptability intervals, compared with the sets $\mathcal{I}_n$, for $n = 1, 2, 3$. All numbers are presented correct to five significant digits.

It emerges that Γ_n, at least for $n \leq 3$, includes only the rightmost subinterval $(\gamma_n^{(n/n)}, \gamma_n^{((n-1)/n)}) \subseteq \mathcal{I}_n$.

Interpolation is frequently a means to an end: minimizing the L_∞ norm of $E_n(\,\cdot\,;\gamma)$ in $(-\infty, 0)$. Other things being equal, this represents a sensible use of the parameter γ, since it is arguably worth trading one 'unit' of order off, in exchange for better overall error across the whole negative half-axis. Thus, we set

$$F_n(\gamma) = \sup_{x \in (-\infty, 0)} |E_n(x;\gamma)|$$

and seek $\gamma \geq 0$ that minimizes F_n. Direct differentiation of (4.14) readily verifies that

$$\frac{\mathrm{d}}{\mathrm{d}x} S_{n/n}(x;\gamma) = S_{(n-1)/(n+1)}(x;\gamma),$$

hence the maxima of $|E_n(x;\gamma)|$ are *precisely* either $-\infty$ or the interpolation points of $S_{(n-1)/(n+1)}(\,\cdot\,;\gamma)$. Nørsett and Trickett (1984) prove results similar to Lemmas 4.15 and 4.16 for arbitrary $S_{m/n}$, $m \geq n$. In particular, $S_{(n-1)/(n+1)}$ may interpolate in $(-\infty, 0)$ only when γ belongs to an interval of the form $(\gamma_k^{((n-1)/(n+1))}, \gamma_k^{(n/(n+1))})$ for some $k \in \{1, 2, \ldots, n\}$.

Theorem 4.17 (Nørsett and Trickett, 1986) The function $F_n(\gamma)$ has a local minimum in each of the intervals $(\gamma_k^{(n/n)}, \gamma_k^{((n-1)/n)})$, $k = 1, 2, \ldots, n$. The global minimum of F_n is the least of these minima.

Proof. The proof follows from the aforementioned analysis, formula (4.18) (which provides an explicit expression for the stationary points of E_n as a function of γ), the interlace inequalities (4.11) and the identity

$$\lim_{x \to -\infty} |E_n(x;\gamma)| = |L_n(1/\gamma)|$$

after short manipulation. □

Minima for $n \leq 3$ were computed by Nørsett and Trickett (1986) and are displayed in Table 4.3.

$n \setminus k$	Interval 1	Interval 2	Interval 3
1	1.1390		
2	0.6691	3.5220	
3	0.4753	2.3709	6.3690

Table 4.3 Local minima of F_n in sub-intervals of $\mathcal{I}_n$, $n = 1,2,3$. All numbers are presented correct to five significant digits.

Needless to say, as long as $n \leq 3$, it frequently makes sense to choose the local minimum in the rightmost interval, in preference to the global minimum, to safeguard A-acceptability.

4.6 Complex fitting

So far in this chapter we have discussed interpolation of the exponential by rationals at real points. This restriction to *real* interpolation is, up to a point, perfectly sensible. Firstly, the order corresponds to real interpolation (at the origin, of course) of sufficiently high degree. Secondly, in many cases (semi-discretization of parabolic partial differential equations being just one example) the eigenvalues of the Jacobian matrix of the ordinary differential system[7] are real. Thirdly, some ordinary differential equations are stiff because of one or two 'parasitic' components (usually corresponding to very fast modes), which can be supressed by means of real interpolation (Liniger and Willoughby, 1970). Fourthly, whereas any rational function has a finite bound on the total multiplicity of real interpolation points (cf. Theorem 3.7), it also has an *infinity* of complex interpolation points. The reason is that the exponential is periodic with period $2\pi i$ – matters are already clear in the trivial case $R(z) \equiv 1$. Thus, complex approximation is different in kind from its real brethren.

[7] We might be deliberating here on rational approximants to the exponential. However, it should never be forgetten that all this is done to an end: numerically solving ordinary differential equations!

Having said all this, we do not wish to argue that complex interpolation[8] is without any merit or interest. To the contrary! The need to interpolate at complex conjugate points arises frequently in the context of **highly oscillatory equations**: ordinary differential systems that display very fast oscillations, imposed on a slower-varying signal. These frequently arise within the context of electrocardiography, seismology, sonar, speech recognition and elsewhere. Such oscillations sometimes correspond to just few modes, which can be conviniently filtered out by complex fitting.

The first to discuss complex fitting were Liniger and Willoughby (1970). They have studied the case of 2/2, second-order approximants R with two complex conjugate interpolation points, ζ and $\bar{\zeta}$, say. The natural question to contemplate is: 'what is the portion $\mathcal{F}_2$ of the complex plane such that $\zeta, \bar{\zeta} \in \mathcal{F}_2 \Leftrightarrow R$ is A-stable?' and the Liniger–Willoughby answer, an outcome of a computer search, was the domain bordered on the right by a wavy line, which is displayed in Figure 4.5. More recently, Iserles and Nørsett (1983), as well as Hairer *et al.* (1985) further explored complex fitting, for general n/n functions and interpolation at a single complex conjugate pair of points. This work led to characterization of A-acceptable approximants of that form, *inter alia* identifying the 'wavy' lines in Figure 4.5. The presentation in this section is modelled after Hairer *et al.* 1985). To be absolutely fair to the reader, we should perhaps emphasize that order stars play only a secondary role in this work. Our two justifications for the inclusion of this material are completeness of exposition and an intriguing connection, unveiled in the sequel, between complex fitting and zeros of Bessel functions.

Example 4.5 The approximant

$$R_1(z) = \frac{1+\gamma z}{1+\delta z}$$

is not in the class of approximants that are treated subsequently in this section. Nonetheless, it is highly instructive and highlights many aspects of our analysis. Let $\zeta = x + iy$, $y \neq 0$. Requiring $R_1(x \pm iy) = \exp(x \pm iy)$ yields

$$\begin{aligned} \gamma &= \frac{ye^x - (x \sin y + y \cos y)}{(x^2+y^2)\sin y} \\ \delta &= \frac{-ye^{-x} - x \sin y + y \cos y}{(x^2+y^2)\sin y} \end{aligned}$$

Neither γ nor δ exist when y is an integer multiple of π. This includes not only the case $y = 0$, which we have ruled out, but also many bona fide

[8]Strictly speaking, of course, real is a special case of complex. Thus, for the benefit of the purists, by 'complex fitting' we mean 'interpolation to $\exp z$ at complex conjugate points with nonzero imaginary parts'.

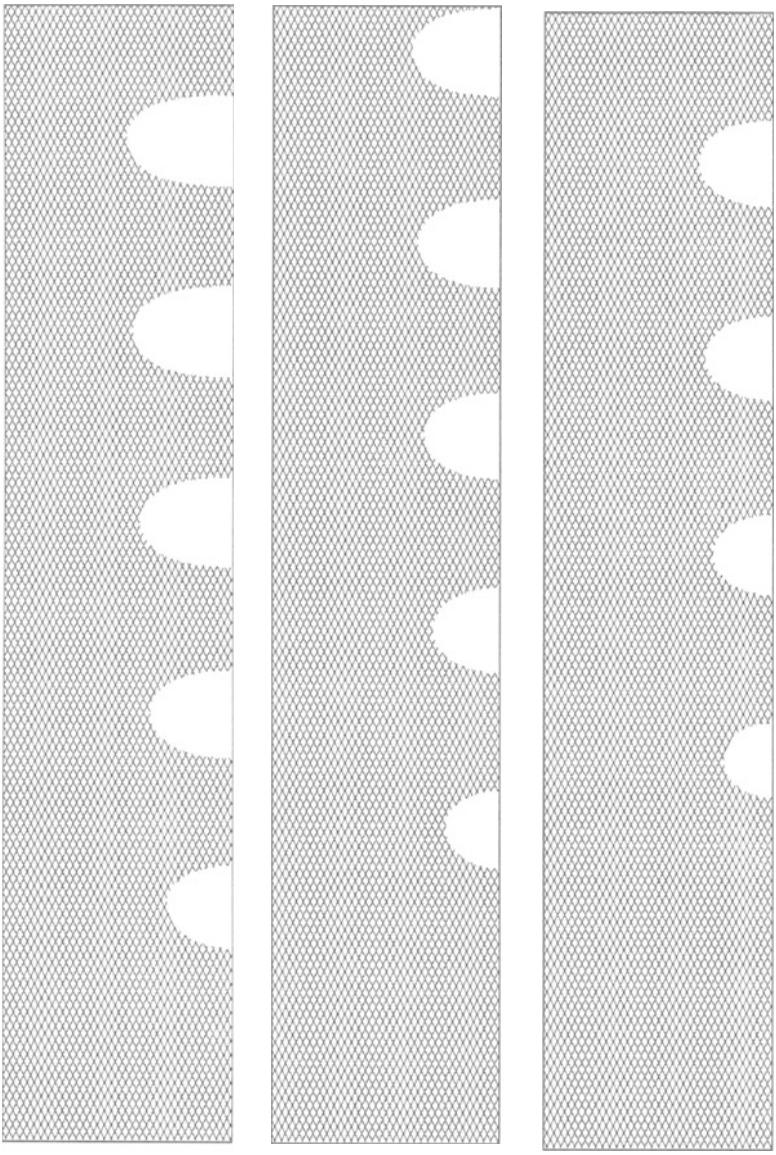


Figure 4.5 The A-acceptability sets $\mathcal{F}_n$ for $n = 2, 3, 4$, and $-7.5 \leq \mathrm{Re}\,\zeta$. The $\mathcal{F}_n$'s are cross-hatched.

values where complex fitting fails. However, as long as $\sin y \neq 0$, we have

$$R_1(z; x+iy) = \frac{(x^2+y^2)\sin y - (x\sin y + y\cos y - ye^x)\,z}{(x^2+y^2)\sin y - (x\sin y - y\cos y + ye^{-x})\,z}.$$

The special case $x = 0$ produces[9]

$$R_1(z; iy) = \frac{y + z\tan\frac{y}{2}}{y - z\tan\frac{y}{2}}.$$

The approximant is symmetric. Consequently, it is A-acceptable if and only if the pole resides in the right half-plane – the condition for that is $\tan(y/2) > 0$. It follows that the pure imaginary axis of the ζ-plane, our parameter space, can be partitioned into alternating intervals of A-acceptability and non-A-acceptability, separated by points where neither γ nor δ is bounded. $\diamond$

Recall the definiton

$$\psi_{n/n}(z) := P_{n/n}(z) - e^z Q_{n/n}(z).$$

The function $\psi_{n/n}$ obeys the following identities, which will be used without further ado in the sequel:

$$\psi_{n/n}(z) - \psi_{(n-1)/(n-1)}(z) = \frac{z^2}{4(2n-3)(2n-1)}\psi_{(n-2)/(n-2)}(z) \tag{4.19}$$

$$\psi'_{n/n}(z) = \frac{1}{2}\psi_{n/n}(z) - \frac{z}{4(2n-1)}\psi_{(n-1)/(n-1)}(z) \tag{4.20}$$

Their verification is straightforward, since they are obeyed with ψ's replaced with M's.

Given a complex point $\zeta \in \mathcal{C} \setminus \mathcal{R}$, we investigate a function $R_n(z) \equiv R_n(z;\zeta) \in \pi_{n/n}$ that obeys

$$R_n(z) = e^z + \mathcal{O}(z^{2n-1}), \quad R_n(\zeta) = e^\zeta,\ R_n(\bar{\zeta}) = e^{\bar{\zeta}}.$$

We assume that $n \geq 2$ and write R_n in the **Hairer representation** (4.4). The order being $p \geq 2n-2$, it follows from Theorem 4.6 that f is linear, whereas g is a constant: $f(z) = 1 - f_1 z$ and $g(z) \equiv g_0$. Note that f_1 and g_0 depend on n and the point ζ. It follows that

$$R_n(z;\zeta) = \frac{P_n(z)}{Q_n(z)} = \frac{(1-f_1 z)M_{n-1}(z) + g_0 z^2 M_{n-2}(z)}{(1-f_1 z)M_{n-1}(-z) + g_0 z^2 M_{n-2}(-z)}, \tag{4.21}$$

[9]Complex fitting with $\mathrm{Re}\,\zeta = 0$ is sometimes called **frequency fitting** (Iserles and Nørsett, 1983).

where $M_m(z) = P_{m/m}(z)$, $m = 0, 1, \ldots$.

Imposing complex interpolation on R, it is easy to deduce that (4.19) holds if and only if (f_1, g_0) are a solution of the linear system

$$\begin{bmatrix} \zeta\psi_{\frac{n-1}{n-1}}(\zeta) & -\zeta^2\psi_{\frac{n-2}{n-2}}(\zeta) \\ \bar{\zeta}\psi_{\frac{n-1}{n-1}}(\bar{\zeta}) & -\bar{\zeta}^2\psi_{\frac{n-2}{n-2}}(\bar{\zeta}) \end{bmatrix} \begin{bmatrix} f_1 \\ g_0 \end{bmatrix} = \begin{bmatrix} \psi_{\frac{n-1}{n-1}}(\zeta) \\ \psi_{\frac{n-1}{n-1}}(\bar{\zeta}) \end{bmatrix}. \tag{4.22}$$

We need to investigate nonsingularity of this linear system for $\operatorname{Im}\zeta \neq 0$. Its determinant is

$$\Delta_n := -2i|\zeta|^2 \operatorname{Im}\left\{\zeta\psi_{(n-2)/(n-2)}(\zeta)\psi_{(n-1)/(n-1)}(\bar{\zeta})\right\}, \tag{4.23}$$

which can be evaluated explicitly: Substitution of (4.19) into (4.23) yields

$$\begin{aligned} \Delta_n &= -2i|\zeta|^2 \operatorname{Im}\left\{\zeta\psi_{(n-2)/(n-2)}(\zeta)\left(\psi_{(n-2)/(n-2)}(\bar{\zeta})\right.\right. \\ &\qquad \left.\left. - \frac{\bar{\zeta}^2}{2(2n-5)(2n-3)}\psi_{(n-3)/(n-3)}(\bar{\zeta})\right)\right\} \\ &= -2i\left|\zeta\psi_{(n-2)/(n-2)}(\zeta)\right|^2 \operatorname{Im}\zeta - \frac{|\zeta|^2}{4(2n-5)(2n-3)}\Delta_{n-1}. \end{aligned}$$

Moreover,

$$\Delta_2 = -\frac{1}{6}i|\zeta|^2\left\{(1 - e^x\cos y)\,y + \left(x - \frac{|\zeta|^2}{2}\right)e^x \sin y\right\},$$

where $\zeta = x + iy$. By elementary induction it follows that $\operatorname{Im}\zeta = 0$ implies $\Delta_n = 0$ for all $n \geq 2$ – no great surprise there! However (and we have already seen something similar in Example 4.5) there are further values of $\zeta \in \mathcal{C}$ where complex fitting fails. We devote much of the remainder of this section to investigating the loci of such points and their connection with A-acceptability.

Lemma 4.18 (Iserles and Nørsett, 1983) For every $n \geq 2$ and $T \in \mathcal{R}$ it is true that

$$\psi_{n/n}(iT) = -i\sqrt{2\pi}\frac{m!}{(2n)!}e^{\frac{1}{2}iT}(T/2)^{n+\frac{1}{2}}J_{n+1/2}\left(\frac{T}{2}\right) \tag{4.24}$$

where $J_{n+\frac{1}{2}}$ is the spherical Bessel function of the first kind (Rainville, 1967).

Proof: We split M_n into a sum of even and odd polynomials,

$$M_n(z) = E_n(z^2) + zU_n(z^2).$$

Substitution into the definition of $\psi_{n/n}$ leads to

$$\begin{aligned}\psi_{n/n}(iT) &= \left(E_n(-T^2) - iTU_n(-T^2)\right)e^{iT} - \left(E_n(-T^2) + iTU(-T^2)\right) \\ &= 2ie^{\frac{1}{2}iT}\left\{E_n(-T^2)\sin\left(\frac{T}{2}\right) - TU_n(-T^2)\cos\left(T/2\right)\right\}.\end{aligned}$$

Let

$$r_n(T) = E_n(-T^2)\sin\left(\frac{T}{2}\right) - TU_n(-T^2)\cos\left(\frac{T}{2}\right).$$

We employ (4.19) to derive the three-term recurrence relation

$$r_n(T) = r_{n-1}(T) - \frac{1}{4(2n-3)(2n-1)}T^2 r_{n-2}(T).$$

This, in conjunction with

$$\begin{aligned}r_0(T) &= \sin\left(\frac{T}{2}\right) = \sum_{k=0}^{\infty}\frac{(-1)^k}{(2k+1)!}\left(\frac{T}{2}\right)^{2k+1} \\ r_1(T) &= \sin\left(\frac{T}{2}\right) - \frac{T}{2}\cos\left(\frac{T}{2}\right) = -2\sum_{k=1}^{\infty}\frac{(-1)^k k}{(2k+1)!}\left(\frac{T}{2}\right)^{2k+1}\end{aligned}$$

determines the r_n's for every $n \geq 0$. A simple induction argument proves that

$$r_n(T) = \frac{(-1)^n n! 2^{2n}}{(2n)!}\sum_{k=n}^{\infty}\frac{(-1)^k k!}{(2k+1)!(k-n)!}\left(\frac{T}{2}\right)^{2k+1}.$$

Compare this to the Taylor expansion of a Bessel function J_ν, where $\nu \in C$ is neither zero nor a negative integer,

$$J_\nu(z) = z^\nu \sum_{k=0}^{\infty}\frac{1}{k!\Gamma(k+\nu+1)}\left(-\frac{z^2}{4}\right)^k$$

(Rainville, 1967). Letting $\nu = \frac{1}{2}$, it is an easy matter to verify that (4.24) is true. □

True connoisseurs of Bessel functions – admittedly, a dying breed – will view (4.24) and its proof with little surprise and immediately see its relationship with Neumann (Rainville, 1967), Lommel (Chihara, 1978) and Bessel (Krall and Frink, 1949) polynomials.

Lemma 4.19 All the roots of $\psi_{n/n}(z) = 0$ are imaginary. Except for $z = 0$, they are of the form iT, where T is a zero of $J_{n+\frac{1}{2}}(T/2)$.

Proof: The zeros of $\psi_{n/n}(z)$ are precisely the points where the diagonal function $R_{n/n}(z) \equiv M_n(z)/M_n(-z)$ interpolates the exponential. We can deduce at once from the order star that all roots of $\psi_{n/n}(z) = 0$ lie on the imaginary axis and the form of the zeros of $\psi_{n/n}$ is now a consequence of (4.24). We conclude by recalling that all zeros of $J_{n+\frac{1}{2}}$ are real (Rainville, 1967). □

We have already looked into the coefficients f_1 and g_0 in (4.21). Now we discuss them further, providing their explicit representation. Let

$$s_n(\zeta) := \omega_{n-1}|\zeta|^2 \mathrm{Im}\left\{\zeta \frac{\psi_{(n-2)/(n-2)}(\zeta)}{\psi_{(n-1)/(n-1)}(\zeta)}\right\}, \quad n \geq 2,$$

where $\omega_n = 1/(4(4n^2 - 1))$.

Lemma 4.20 Suppose that $\mathrm{Im}\,\zeta \neq 0$. Then

$$g_0 = \frac{\mathrm{Im}\,\zeta}{s_n(\zeta)}, \tag{4.25}$$

$$f_1 = 4(2n-3)(2n-1)\frac{\mathrm{Im}\,(\psi_{n/n}(\zeta))/\psi_{(n-1)/(n-1)}(\zeta))}{s_n(\zeta)}. \tag{4.26}$$

In particular, both f_1 and g_0 are real and well defined for all $\zeta \in \mathcal{C} \setminus \mathcal{R}$ such that $s_n(\zeta) \neq 0$.

Proof: Divide the first equation in (4.22) by $\psi_{(n-1)/(n-1)}$ and multiply by $\bar{\zeta}$. We now have

$$\bar{\zeta} - f_1|\zeta|^2 + g_0\omega_{n-1}|\zeta|^2\zeta\frac{\psi_{(n-2)/(n-2)}(\zeta)}{\psi_{(n-1)/(n-1)}(\zeta)} = 0.$$

Similar operations on the second equation yield a complex conjugate of the last formula. In both cases f_1 is multiplied by the same real coefficient and disappears upon subtraction. This yields (4.25). Likewise, adding both formulae and substituting the explicit value of g_0 produces, after short manipulation and using the identity (4.19), the expression (4.26). □

Having discussed the existence of n/n complex fittings of order at least $2n - 2$ we turn our attention to A-acceptability, defining

$$\mathcal{F}_n := \{\zeta \in \mathcal{C} : R_n(\,\cdot\,;\zeta) \text{ exists and is } A\text{-acceptable}\}.$$

The purpose of the remainder of this section is to characterize completely the set $\mathcal{F}_n$.

Theorem 4.21 Let the function R_n be given by formula (4.21). Then
(a) $|R_n(it)| \leq 1$ for all $t \in \mathcal{R}$ if and only if $f_1 g_0 \geq 0$;
(b) R_n is A-acceptable if and only if both $f_1 \geq 0$ and $g_0 \geq 0$.

Proof: The desired result follows at once from Lemma 4.7 and Theorem 4.8, except for the case $g_0 = 0$, which reduces to a diagonal Padé approximant (which is A-acceptable).

An alternative method of proving (b) is by tracking the movement of the poles as f_1 and g_0 vary, $f_1 g_0 \geq 0$. It does not require order stars altogether and may appeal to some readers. In principle, there are three possible ways for a pole to cross $i\mathcal{R}$: via the origin, through the punctured line $i\mathcal{R} \setminus \{0\}$ and by 'jumping' at infinity.

The possibility of a pole travelling via the origin is ruled out at once for bounded f_1 and g_0, since $Q_n(0) = 1$.

Let us suppose that a pole moves through the punctured imaginary axis. Since R_n is a real function, its complex conjugate moves through the same axis and, because $f_1 g_0 \geq 0$, both these poles coalesce with zeros. It follows that $R_n \in \pi_{(n-2)/(n-2)}$, but this is contradicted by R_n being an order-$2(n-1)$ approximation. Consequently, no movement of poles through $i\mathcal{R} \setminus \{0\}$ is possible.

No pole can jump thorough infinity (with a single exception), but for an entirely different reason: the coefficient of z^n in Q_n vanishes and a pole becomes unbounded only when $f_1 + (2n-1)g_0 = 0$, and this cannot happen for $f_1 g_0 \geq 0$, except when $f_1 = g_0 = 0$ and R_n reduces to $R_{(n-1)/(n-1)}$.

We conclude that the number of poles of R_n on the left of the imaginary axis is constant as long as f_1, g_0 stay in the same quadrant, subject to $f_1 g_0 \geq 0$. It is straightforward to verify that

$$f_1 = \frac{1}{2(2n-1)} > 0, \quad g_0 = \frac{1}{4(2n-3)(2n-1)} > 0$$

produces the $(n-1)/n$ Padé approximant, with n poles in the right half-plane (cf. Lemma 3.4). Thus, all the poles of Q_n stay to the right of $i\mathcal{R}$ and R_n is A-acceptable for all $f_1, g_0 \geq 0$.

Finally, to rule out A-acceptability for $f_1, g_0 < 0$, we observe that, as $g_0 \uparrow 0$, a pole and a zero coalesce at $1/f_1 < 0$. Thus, because of continuity of poles with respect to g_0, a pole stays to the left of the imaginary axis as g_0 decreases away from zero, preventing A-acceptability. This concludes the proof. □

Theorem 4.21 is crucial to our understanding of complex fitting. However, it is not entirely adequate, since we wish to characterize the set $\mathcal{F}_n$ in terms of the complex interpolation points, rather than the parameters f_1 and g_0. Rephrasing its main result by using (4.25) and (4.26), R_n is

A-acceptable if and only if

$$s_n(\zeta) \neq 0, \quad s_n(\zeta)\mathrm{Im}\,\zeta \geq 0, \quad s_n(\zeta)\mathrm{Im}\,\{\psi_{n/n}(\zeta)\psi_{(n-1)/(n-1)}(\bar{\zeta})\} \geq 0.$$

Lemma 4.22 Excluding the points where $\psi_{(n-1)/(n-1)}(\zeta) = 0$, we have

$$\mathrm{Im}\left(\frac{\psi_{n/n}(\zeta)}{\psi_{(n-1)/(n-1)}(\zeta)}\right)\begin{cases} = 0 & : (\mathrm{Re}\,\zeta)(\mathrm{Im}\,\zeta) = 0 \\ > 0 & : (\mathrm{Re}\,\zeta)(\mathrm{Im}\,\zeta) < 0 \\ < 0 & : (\mathrm{Re}\,\zeta)(\mathrm{Im}\,\zeta) > 0 \end{cases} \tag{4.27}$$

Proof. Let

$$\Psi(\zeta) := \frac{\psi_{n/n}(\zeta)}{\psi_{(n-1)/(n-1)}(\zeta)}.$$

Suppose first that $\mathrm{Im}\,\zeta = 0$. Then both $\psi_{n/n}$ and $\psi_{(n-1)/(n-1)}$ are real and $\mathrm{Im}\,\Psi(\zeta) = 0$. If $\mathrm{Re}\,\zeta = 0$ then

$$\psi_{m/m}(it) = -e^{it}\psi_{m/m}(-it), \quad m = 0, 1, \ldots$$

implies that for $\zeta = it, \in \mathcal{R}$,

$$\Psi(it) = \frac{\psi_{n/n}(it)}{\psi_{(n-1)/(n-1)}(it)} = \frac{\psi_{n/n}(-it)}{\psi_{(n-1)/(n-1)}(-it)} = \overline{\Psi(it)}.$$

Thus, $\mathrm{Im}\,\Psi(it) = 0$.

Next, we examine the asymptotic behaviour of Ψ for large $|\zeta|$. Letting $\zeta = re^{i\tau}$, $r \gg 1$, we obtain

$$\Psi(\zeta) \approx \begin{cases} -\frac{re^{i\tau}}{2(2n-1)} & : -\frac{\pi}{2} < \tau < \frac{\pi}{2} \\ \frac{re^{i\tau}}{2(2n-1)} & : \frac{\pi}{2} < \tau < \frac{3\pi}{2} \end{cases}.$$

Thus, far away from the origin

$$\mathrm{Im}\,\Psi(re^{i\tau}) \approx \begin{cases} -\frac{r\sin\tau}{2(2n-1)} & : -\frac{\pi}{2} < \tau < \frac{\pi}{2} \\ \frac{r\sin\tau}{2(2n-1)} & : \frac{\pi}{2} < \tau < \frac{3\pi}{2} \end{cases}, \tag{4.28}$$

consistently with (4.27).

Finally, we examine Ψ near its poles. It is straightforward to deduce from the order star theory that all the zeros of $\psi_{m/m}$, regardless of m, reside on the pure imaginary axis: each such zero is an interpolation point of the underlying Padé approximant and it must necessarily lie along the joint portion of $\mathcal{A}_+^\infty$ and $\mathcal{A}_-^\infty$, otherwise there are not enough zeros or poles to 'support' it. However, in the case of diagonal Padé we have $\partial\mathcal{A}_+^\infty \cap \partial\mathcal{A}_i^\infty = i\mathcal{R}$ and our assertion is true.

Thus, let it^* be a zero of $\psi_{(n-1)/(n-1)}$. We may assume that $t^* \neq 0$, since the origin is an analytic point of Ψ. Given $\zeta = it^* + \varepsilon e^{i\tau}$, $0 < \varepsilon \ll 1$, we have

$$\Psi(\zeta) \approx \frac{\psi_{n/n}(it^*)}{\varepsilon e^{it^*}\psi'_{(n-1)/(n-1)}(it^*)}.$$

We express $\psi_{n/n}$ in the numerator and $\psi'_{(n-1)/(n-1)}$ in the denominator by using (4.19) and (4.20) respectively. This and $\psi_{(n-1)/(n-1)}(it^*) = 0$ yield

$$\Psi(\zeta) \approx -\frac{it^* e^{-it^*}}{(2n-1)\varepsilon},$$

therefore

$$\operatorname{Im}\Psi(\zeta) \approx -\frac{t^* \cos\tau}{(2n-1)\varepsilon}. \tag{4.29}$$

Given $0 < \varepsilon \ll 1$ we let

$$D_\varepsilon := \left\{\zeta \in \mathcal{C} : |\zeta| < \frac{1}{\varepsilon},\ |\zeta - it^*| > \varepsilon \text{ for every } t^* \text{ such that } \psi_{\frac{n-1}{n-1}}(it^*) = 0\right\}$$

and consider first

$$D_\varepsilon^{(++)} := \{\zeta \in \mathcal{C} : \operatorname{Re}\zeta, \operatorname{Im}\zeta > 0\} \cap D_\varepsilon.$$

We have already seen that $\operatorname{Im}\Psi(\zeta)$ vanishes on the part of $\partial D_\varepsilon^{(++)}$ that coincides with the axes. Moreover, according to (4.28) and (4.29), it is negative at the remaining portion of the boundary. According to the maximum principle for subharmonic functions (Ahlfors, 1966), the imaginary part of an analytic function attains its maximum on a boundary. Letting $\varepsilon \downarrow 0$ affirms that $\operatorname{Im}\Psi(\zeta) < 0$ for $\operatorname{Re}\zeta, \operatorname{Im}\zeta > 0$.

A similar technique, extended to the three other quadrants, completes the proof of the lemma. □

Theorem 4.23 The approximant R_n is A-acceptable if and only if $\operatorname{Re}\zeta \leq 0$ and $s_n(\zeta)\operatorname{Im}\zeta > 0$.

Proof. We already know, as a consequence of Theorem 4.21, that A-acceptability is equivalent to $s_n(\zeta) \neq 0$ (otherwise R_n is not well defined) and to ζ, $s_n(\zeta)$ and $\Psi(\zeta)$ being all of the same sign.

If $s_n(\zeta) > 0$ then necessarily $\operatorname{Im}\zeta > 0$. Moreover, $\operatorname{Im}\Psi(\zeta) \geq 0$ and Lemma 4.22 imply $\operatorname{Re}\zeta \leq 0$. On the other hand, if $s_n(\zeta) < 0$ then also $\operatorname{Im}\zeta < 0$ and, again, Lemma 4.22 is used to argue that $\operatorname{Re}\zeta \leq 0$. This proves the necessity of the given conditions. To prove their sufficiency we simply travel in the opposite direction, demonstrating that they are consistent with Theorem 4.21 and Lemma 4.22. □

Theorem 4.23 provides a firm theoretical framework to characterize the set $\mathcal{F}_n \subset \mathcal{C}$ of parameters ζ that produce A-acceptable approximants. The crux is to determine when $s_n(\zeta)\mathrm{Im}\,\zeta > 0$ and the next lemma is helpful in this context.

Lemma 4.24 The function s_n possesses the following properties for all $\zeta = x + iy \in \mathcal{C}$.

$$s_n(x+iy) = s_n(-x+iy); \tag{4.30}$$

$$\lim_{|x|\to\infty} y s_n(x+iy) > 0, \quad y \neq 0; \tag{4.31}$$

$$s_n(iy) = \omega_{n-1} y^3 \frac{\psi_{(n-2)/(n-2)}(iy)}{\psi_{(n-1)/(n-1)}(iy)} \in \mathcal{R}; \tag{4.32}$$

$\forall$ fixed $y \neq 0$, $s_n(x+iy)$ has at most one *negative* root. (4.33)

The constant $\omega_n = 1/(4(4n^2-1))$ has been already defined in the preamble to Lemma 4.20.

Proof. To verify (4.31) we note that, according to the definition of $\psi_{n/n}$,

$$\psi_{n/n}(-x+iy) = -e^{-x+iy}\overline{\psi_{n/n}(x+iy)}. \tag{4.34}$$

The desired identity follows at once from the definition of s_n.

Expansion of $\psi_{(n-2)/(n-2)}(\zeta)/\psi_{(n-1)/(n-1)}(\zeta)$ in powers of ζ^{-1} for $|\zeta| \to \infty$ affirms that

$$\zeta\frac{\psi_{(n-2)/(n-2)}(\zeta)}{\psi_{(n-1)/(n-1)}(\zeta)} = 2(2n-3)\mathrm{sign}\,x - 4(n-1)(2n-3)\frac{1}{\zeta} + \frac{C}{\zeta^2}\mathrm{sign}\,x + \mathcal{O}\left(\frac{1}{\zeta^3}\right),$$

where C is a real constant. By taking the imaginary part we find that

$$\mathrm{Im}\,\zeta\frac{\psi_{(n-2)/(n-2)}(\zeta)}{\psi_{(n-1)/(n-1)}(\zeta)} \approx 4(n-1)(2n-3)\frac{1}{y}, \quad |x| \to \infty$$

and (4.31) follows.

As an immediate consequence of the definition of s_n, we know that

$$s_n(iy) = \omega_{n-1} y^3 \mathrm{Re}\,\frac{\psi_{(n-2)/(n-2)}(iy)}{\psi_{(n-1)/(n-1)}(iy)}.$$

However, because of (4.34), $\psi_{(n-2)/(n-2)}(iy)/\psi_{(n-1)/(n-1)}(iy)$ is real and (4.32) is true.

Let $y \neq 0$ be fixed and consider $s_n(x+iy)$ as a function of x. We define an auxiliary function

$$g(\zeta) := \mathrm{Im}\left\{\zeta\frac{\psi_{(n-2)/(n-2)}(\zeta)}{\psi_{(n-1)/(n-1)}(\zeta)}\right\}.$$

Since $y \neq 0$, clearly $s_n(\tilde{x}+iy) = 0$ implies $g(\tilde{x}+iy) = 0$. Thus, it is enough to prove the statement for g, instead of s_n.

Suppose that $g(x_1+iy) = g(x_2+iy) = 0$ and that $g(x+iy) \neq 0$ for all $x \in (x_1, x_2)$. Since $g(\cdot + iy)$ can only have isolated zeros, it follows that

$$g'(x_1+iy)g'(x_2+iy) \leq 0.$$

Thus, to prove (4.33), it is enough to demonstrate that $\tilde{x} < 0$ and $g(\tilde{x}+iy) = 0$ imply that $yg'(\tilde{x}+iy) < 0$. Straightforward, albeit quite tedious, calculation, with a little help from (4.19) and (4.20), yields

$$\begin{aligned} g'(\zeta) &= (2n-1)\mathrm{Im}\left\{\frac{\psi_{(n-2)/(n-2)}(\zeta)}{\psi_{(n-1)/(n-1)}(\zeta)}\right\} \\ &\quad + \frac{1}{4(2n-3)}\mathrm{Im}\left\{\left(\zeta\frac{\psi_{(n-2)/(n-2)}(\zeta)}{\psi_{(n-1)/(n-1)}(\zeta)}\right)^2\right\}. \end{aligned}$$

However, the second term vanishes when $\zeta = \tilde{x}+iy$. We can now employ Lemma 4.22 to argue that $yg'(\tilde{x}+iy) < 0$, concluding our proof. □

We have already mentioned that Liniger and Willoughby were the first to discuss complex fitting in a heuristic setting. General characterization of $\mathcal{F}_n$ has been presented by Hairer *et al.* (1985). The following lemma, given originally by Iserles and Nørsett (1982), establishes an all-important connection between the geometry of $\mathcal{F}_n$ and Lemma 4.19. Thus, recall that all the zeros of $\psi_{n/n}$ are imaginary and denote them by $\{iy_k^{[n]}\}_{k=0}^{\infty}$, where

$$0 = y_0^{[n]} < y_1^{[n]} < y_2^{[n]} < \cdots. \tag{4.35}$$

Theorem 4.25 (Iserles and Nørsett, 1982) The frequency-fitted rational approximant $R_n(\cdot\,; iy)$, $n \geq 2$, $y > 0$, is A-acceptable, provided that

$$y_{k-1}^{[n-1]} < \mathrm{Im}\,\zeta < y_k^{[n-2]} \tag{4.36}$$

for some $k \geq 1$.

Proof. We observe first that, according to Lemma 4.19, the y_k^m's are zeros of spherical Bessel functions. In particular, zeros of $\psi_{(n-2)/(n-2)}$ and $\psi_{(n-1)/(n-1)}$ interlace along the upper imaginary half-axis,

$$0 = y_0^{[n-1]} < y_1^{[n-2]} < y_1^{[n-1]} < y_2^{[n-2]} < y_2^{[n-1]} < \cdots$$

(Abramowitz and Stegun, 1965, p. 370).

Recall that

$$\psi_{m/m}(z) = C_m z^{2m+1} + \mathcal{O}(z^{2m+2}), \quad (-1)^{m+1}C_m > 0,$$

for all $m \geq 0$. Thus, provided that $y \to 0$,

$$s_n(iy) = C_n^* y + \mathcal{O}(y^2), \quad C_n^* = \omega_{n-1}\frac{C_{n-2}}{C_{n-1}} > 0.$$

According to (4.32) and Lemma 4.19, the function s_n is real along the upper imaginary half-axis and it possesses there simple poles $\{iy_k^{[n-1]}\}_{k=1}^{\infty}$ and simple zeros $\{iy_k^{[n-2]}\}_{k=1}^{\infty}$. These are exactly the points where it changes sign. Since it is positive for $0 < y \ll 1$, positivity persists throughout $(0 = y_0^{[n-1]}, y_1^{[n-2]})$ and is lost at the right end-point, only to be recovered at $y_1^{[n-1]}$. We can now easily produce an inductive proof that s_n is positive precisely in the intervals of the form (4.36). Since $y > 0$, the statement of the theorem is an immediate conclusion of Theorem 4.23. □

What happens away from the imaginary axis? Of course, the open right half-plane is not very interesting, since Theorem 4.23 affirms that no A-acceptability is possible there. Let us suppose without loss of generality that $\operatorname{Im}\zeta > 0$ and examine the left half-plane (or, to be more exact, the open upper-left quadrant) of the ζ-plane. By Theorem 4.23, A-acceptability there is equivalent to $s_n(\zeta) > 0$. The two crucial items of information are provided by (4.31) and (4.33). Thus, if $\operatorname{Re}\zeta$ is sufficiently small then $s_n(\zeta) > 0$ and $R_n(\,\cdot\,;\zeta)$ is A-acceptable.

Let us examine A-acceptability as ζ travels leftwards along the line $\{x \in (-\infty, 0],\ y \text{ constant}\}$. We know from (4.33) that s_n cannot change sign more than once. Thus, *either* it is positive throughout *or* a single sign-change occurs and there exists $x^* = x^*(y)$ such that

$$s_n(x+iy)\begin{cases} < 0 & : x^* < x < 0 \\ > 0 & : x < x^* \end{cases}$$

In other words, if y belongs to one of the intervals (4.36), where $R_n(\,\cdot\,;iy)$ is A-acceptable, then A-acceptability persists along the whole line, otherwise it is achieved after a while and never lost again.

The 'wavy' boundary of $\mathcal{F}_n$ in Figure 4.5, which looks somewhat like Gruyère cheese after a mice invasion, is now fully explained. Disregarding the axes, we see that s_n vanishes precisely on a family of simple, closed curves $\mathcal{D}_k$, $k = 1, 2, \ldots$, such that

$$\mathcal{D}_k \cap i\mathcal{R} = \{y_{k-1}^{[n-1]}, y_k^{[n-2]}\},$$

and their reflections $\mathcal{D}_k$ in the lower half-plane. Each $\mathcal{D}_k$ is symmetric with respect to the imaginary axis, as can be seen at once from (4.30).

Theorem 4.26 The interior of the set $\mathcal{F}_n$ is precisely the portion of the closed left half-plane that lies outside the curves $\mathcal{D}_{\pm k}$ for all $k = 1, 2, \ldots$.
□

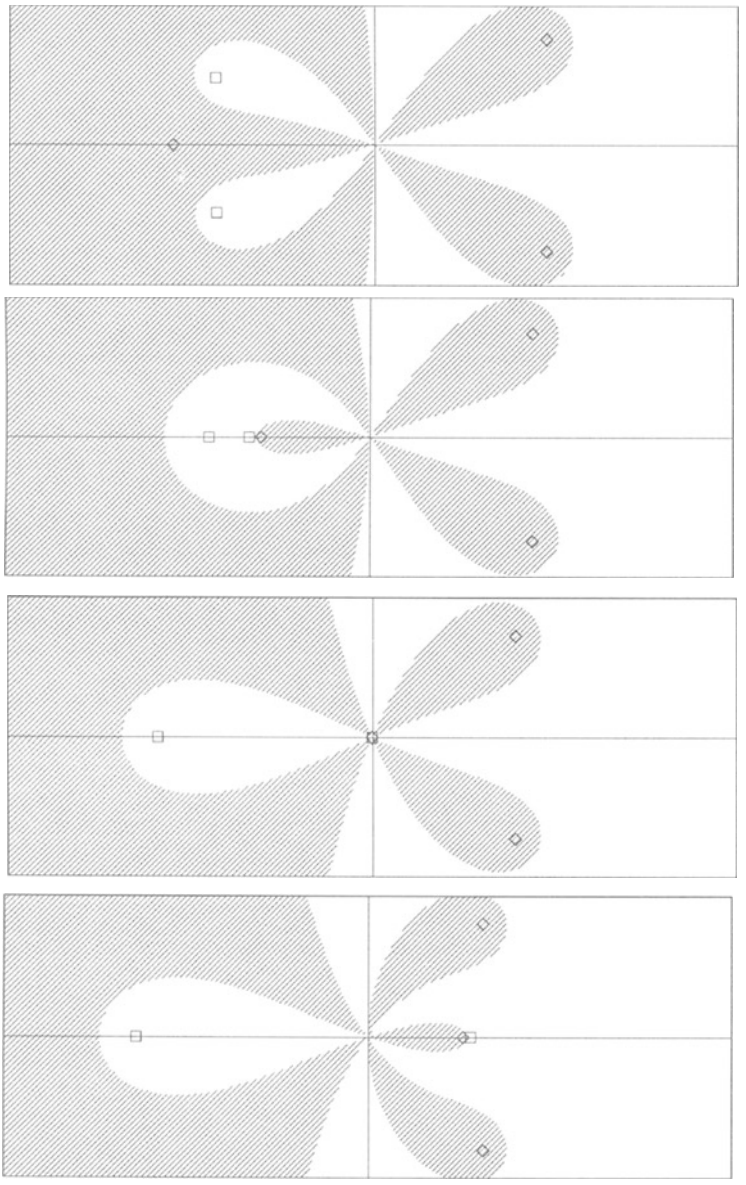


Figure 4.6 Order stars for $R_3(\,\cdot\,;\zeta)$, where $\operatorname{Im}\zeta = -10$ and $\operatorname{Re}\zeta \in \{-1, -1.3, -1.6914577\ldots, -2\}$.

Having embarked on the task of characterizing $\mathcal{F}_n$, we have produced, in Theorem 4.26, a characterization of $\operatorname{int}\mathcal{F}_n$. In mathematical texts, as in insurance contracts, the reader should beware of the small print! The boundary of $\mathcal{F}_n$ is composed of two kinds of 'object': portions of $\partial\mathcal{D}_k$ for some $k \in \mathcal{Z} \setminus \{0\}$ and subintervals of $i\mathcal{R}$. Because of Theorem 4.23, points of

$$\partial\mathcal{F}_n \setminus \bigcup_{|k|\geq 1} \mathcal{D}_k$$

do not produce A-acceptable approximants, hence they lie outside $\mathcal{F}_n$. This leaves out three kinds of point:

(a) $\zeta = \pm iy_{k-1}^{[n-1]}$: In this case $\psi_{(n-1)/(n-1)}(\zeta) = 0$ implies $g_0 = 0$ and R_n reduces to the Padé approximant $R_{(n-1)/(n-1)}$, which is A-acceptable. Therefore ζ belongs to $\mathcal{F}_n$.

(b) $\zeta = \pm iy_k^{[n-2]}$: Since $\psi_{(n-2)/(n-2)}(\zeta)$ vanishes, $|g_0|$ becomes unbounded and $f_1/g_0 = 0$. Consequently, R_n reduces to $R_{(n-2)/(n-2)}$, an A-acceptable function, and $\zeta \in \mathcal{F}_n$.

(c) $\zeta \in \partial\mathcal{F}_n$, $\operatorname{Re}\zeta < 0$: We have $s_n(\zeta) = 0$ and both f_1 and g_0 become unbounded. However, their ratio stays bounded and short calculation yields

$$\varpi := \frac{f_1}{g_0} = \frac{1}{\operatorname{Im}\zeta} \times \operatorname{Im}\left\{\frac{\psi_{n/n}(\zeta)}{\psi_{(n-1)/(n-1)}(\zeta)}\right\},$$

an expression that is neither zero nor unbounded. Using (4.19) (with $n-1$ replacing n), we can write

$$R_n(z;\zeta) = \frac{M_{n-2}(z)(1-\varpi z) + z^2 M_{n-3}(z)}{M_{n-2}(-z)(1-\varpi z) + z^2 M_{n-3}(z)}.$$

It now follows at once from Theorem 4.8 that R_n is A-acceptable.

Theorem 4.27 For every $n \geq 2$, the set $\mathcal{F}_n$ consists of the open set that has been already defined in the statement of Theorem 4.26, as well as all points ζ of $\mathcal{D}_k$ for $k \in \mathcal{Z} \setminus \{0\}$ such that $\operatorname{Re}\zeta \leq 0$. □

Corollary Let $\operatorname{Re}\zeta$ be negative. Then $\zeta \notin \mathcal{F}_n$ implies $\zeta \in \mathcal{F}_{n-1}, \mathcal{F}_{n+1}$.

Proof. In the open left half-plane each $\mathcal{D}_k$ lies inside the band

$$\left\{\zeta \in \mathcal{C} : y_{k-1}^{[n-1]} < \operatorname{Im}\zeta < y_k^{[n-2]}\right\}.$$

This follows at once from our analysis and from (4.33). The statement of the lemma is now a consequence of the inequalities (4.35). □

To conclude this chapter, we present in Figure 4.6 four order stars of the first kind, with

$$\rho(z) = e^{-z} R_n(z;\zeta).$$

Although they have no role in our discourse, they demonstrate quite vividly the evolution of R_n as $\mathrm{Re}\,\zeta$ decreases, while $\mathrm{Im}\,\zeta$ is kept constant. Letting $n = 3$, we commence with $\zeta = -1 + 10i$, in the non-A-acceptable set (cf. Figure 4.5). It is clear that the reason for the breakdown of A-acceptability is in a negative pole. As $\mathrm{Re}\,\zeta$ decreases and the interpolation point moves toward $\mathcal{F}_3$, that pole is 'chased' by a zero and this is demonstrated in the second order star, for $\zeta = -1.3 + 10i$. At $\zeta = -1.6914577\ldots + 10i$ we reach the boundary of $\mathcal{F}_3$ and the zero catches up with the pole at the origin. The underlying approximant is third-order and A-acceptable. Finally, in the bottom figure, $\zeta = -2 + 10i$ is well within the A-acceptability set. The pole and the zero have re-emerged in the right half-plane, but now it is the zero that leads the way...

5

Multistep methods

Yes, the old lost stars wheel back, dear lass,
That blaze in the velvet blue.
They're all old friends, on the old trail, our own trail, the out trail,
They're God's own guides on the Long Trail—the trail that is always new.

From *L'Envoi* by Rudyard Kipling (1865–1936).

5.1 The first Dahlquist barrier

Even the more casual students of numerical mathematics are aware that, as far as ordinary differential equations are concerned, there is much besides Runge–Kutta and multiderivative methods in the computational toolbox. Most important among the numerical techniques that have so far been conspicuous by their absence in this book are the **multistep methods**, which exploit information from a whole range of time steps to enhance the quality of the solution.

Similarly to the notation of Chapter 3, we denote by $\mathbf{y}_i$ an approximation to the solution at $a + ih$, where $h > 0$ and $i = 0, 1, \ldots$. Moreover, we let $\mathbf{f}_i := \mathbf{f}(a + ih, \mathbf{y}_i)$. A general multistep method has the form

$$\sum_{\ell=0}^{N} \alpha_\ell \mathbf{y}_{i+\ell-N} = h \sum_{\ell=0}^{N} \beta_\ell \mathbf{f}_{i+\ell-N}, \quad \alpha_N = 1. \tag{5.1}$$

Note that, if $N \geq 2$, we require $N - 1$ **starting values** $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{N-1}$, in addition to the initial value $\mathbf{y}_0$. The provision of starting values is of no concern to this exposition. We say that a multistep method is **explicit** if $\beta_N = 0$, otherwise it is called **implicit** and its practical evaluation requires the solution of a nonlinear algebraic system.

Clearly, the method (5.1) is fully characterized by its coefficients. In other words, letting

$$r(w) := \sum_{\ell=0}^{N} \alpha_\ell w^\ell, \qquad s(w) := \sum_{\ell=0}^{N} \beta_\ell w^\ell,$$

produces a natural isomorphism from pairs of Nth degree polynomials to N-step methods (5.1).[1] We assume that r and s are relatively prime – the method (5.1) is in that case said to be **irreducible**. This does not bring about any loss of generality, while somewhat simplifying the exposition.

We have seen in the last two chapters three important features of numerical schemes – convergence, order and A-stability – and it is of interest to examine how to express them in terms of r and s. A central place in this discussion must be reserved for **convergence** – recall that a method is convergent if, as $h \downarrow 0$, the local error in a compact interval is uniformly bounded by the imperfections in the initial data and in the solution of the nonlinear algebraic equations. Thus, unless a method is convergent, we must not use it, since, even as the time step becomes infinitesimally small, it might produce completely the wrong solution. It is important to grasp the difference between A-stability and convergence: A-stability is all about practicability and it can be relaxed, for differing intents and purposes, by weaker definitions, whereas convergence is an imperative that should never be traded off.

Deferring an examination of A-stability and its relationship with order to section 5.3, we commence our exposition by stating a famous theorem, due to Dahlquist (1956). There are solid grounds to the argument that the publication of the **Dahlquist equivalence theorem** augured the coming of age of numerical analysis as a mathematical discipline.

A polynomial p is said to obey the **root condition** if all its zeros reside in the closed complex unit disc and the zeros with unit modulus are simple.

Theorem 5.1 (Dahlquist, 1956) The method (5.1) is convergent if and only if it is of order $p \geq 1$ and the polynomial r obeys the root condition. □

The pattern of

Convergence	$\Longleftrightarrow$	$p \geq 1$	+	An algebraic condition

is critical to numerical analysis of evolutionary differential equations and will be encountered again in Chapter 6 in a different context.[2]

It is easy to evaluate the order of (5.1), and we need nothing more than the polynomials r and s. Let us define a **multistep operator** $\mathcal{M}_h$, that

[1] It is usual to denote these polynomials by ρ and σ, rather than r and s respectively. However, since ρ is reserved in this book to the 'order star function', we opt for this nonstandard notation.

[2] We anticipate here the **Lax equivalence theorem** for evolutionary linear partial differential equations. Of course, the 'algebraic condition' therein is no longer the root condition.

acts on elements of arbitrary sequences $\vec{\mathbf{x}} := \{\mathbf{x}_i\}_{i=0}^{\infty}$:

$$(\mathcal{M}_h(\vec{\mathbf{x}}))_i := \sum_{\ell=0}^{N} \alpha_\ell \mathbf{x}_{i+\ell-N} - h \sum_{\ell=0}^{N} \beta_\ell \mathbf{f}(a + (i+\ell-N)h, \mathbf{x}_{i+\ell-N}), \quad i \geq N.$$

In other words, $\vec{\mathbf{y}}$ is the solution of (5.1) if

$$(\mathcal{M}_h(\vec{\mathbf{y}}))_i = 0, \quad i = N, N+1, \ldots. \tag{5.2}$$

Set $\mathbf{z}_i := \mathbf{y}(ih)$, $i = 0, 1, \ldots$ and recall the **shift operator** E and the **differential operator** D from section 3.3. We have

$$\begin{aligned}
(\mathcal{M}_h(\vec{\mathbf{z}}))_i &= \sum_{\ell=0}^{N} \alpha_\ell \mathbf{y}(a + (i+\ell-N)h) - h \sum_{\ell=0}^{N} \beta_\ell \mathbf{y}'(a + (i+\ell-N)h) \\
&= \left(\sum_{\ell=0}^{N} \alpha_\ell E^\ell \right) \mathbf{y}(a + (i-N)h) - hD \sum_{\ell=0}^{N} \beta_\ell \mathbf{y}(a + (i-N)h) \\
&= (r(E) - s(E) \log E)\, \mathbf{y}(a + (i-N)h) \\
&= M(\log E, E) \mathbf{y}(a + (i-N)h),
\end{aligned}$$

where

$$M(z, w) := r(w) - zs(w).$$

Suppose that

$$M(\log w, w) = \mathcal{O}(|w-1|^{p+1})$$

for some natural number p. Since $E = I + \mathcal{O}(h)$, it follows that

$$(\mathcal{M}_h(\vec{\mathbf{z}}))_i = \mathcal{O}(h^{p+1}), \quad i = N, N+1, \ldots. \tag{5.3}$$

Let $\mathbf{e}_i := \mathbf{y}_i - \mathbf{z}_i$ be the numerical error. Subtracting (5.3) from (5.2) we obtain

$$(\mathcal{M}_h(\vec{\mathbf{e}}))_i = \mathcal{O}(h^{p+1}), \quad i = N, N+1, \ldots,$$

hence

$$\mathbf{e}_i - h\beta_N \mathbf{f}(a + Nh, \mathbf{e}_N) = \tag{5.4}$$

$$- \sum_{\ell=0}^{N-1} \alpha_\ell \mathbf{e}_{i+\ell-N} + h \sum_{\ell=0}^{N-1} \beta_\ell \mathbf{f}(a + (i+\ell-N)h, \mathbf{e}_{i+\ell-N}) + \mathcal{O}(h^{p+1}).$$

We suppose that $\mathbf{e}_i = \mathcal{O}(h^{p+1})$ for $i = 0, \ldots, N-1$ and employ induction to argue that this estimate persists for all $i \geq 0$, at least for sufficiently small $h > 0$. For suppose that it is true up to $i-1$. Thus, by the Taylor theorem,

$$\mathbf{f}(a + (i+\ell-N)h, \mathbf{e}_{i+\ell-N}) = \mathcal{O}(h^{p+1})$$

and the inductive proof follows from the implicit function theorem upon substitution in (5.4). Recall that $\mathbf{e}_i$ is the numerical error to realize that we have just derived an order condition.

Proposition 5.2 The multistep method (5.1) is of order p if and only if $M(\log w, w) = \mathcal{O}(|w-1|^{p+1})$. □

Example 5.1 The natural instinct is to exploit all the available degrees of freedom to maximize the order. Since there are $2N+1$ free coefficients in r and s (recall that $\alpha_N = 1$), we may expect order $2N$. It is clear from irreducibility and Proposition 5.2 that

$$\text{order } p \quad \Leftrightarrow \quad \frac{r(w)}{s(w)} = \log w + \mathcal{O}(|w-1|^{p+1}),$$

hence the maximal order occurs when r and s are the numerator and the denominator, respectively, of the N/N Padé approximant to the logarithm at $w = 1$.

Such approximants can be computed easily for modest values of N. Thus, $N = 1$ gives

$$\frac{w-1}{\frac{1}{2}(1+w)},$$

hence $r(w) = w - 1$, $s(w) = \frac{1}{2}(1+w)$ and we recover the second-order **trapezoidal rule**

$$\mathbf{y}_i - \mathbf{y}_{i-1} = \frac{1}{2}h(\mathbf{f}_i + \mathbf{f}_{i-1}).$$

Note that r satisfies the root condition, consequently the trapezoidal rule is convergent. Letting $N = 2$ produces

$$\frac{-1+w^2}{\frac{1}{3}+\frac{4}{3}w+\frac{1}{3}w^2}$$

and the fourth-order method

$$\mathbf{y}_i - \mathbf{y}_{i-2} = \frac{1}{3}h(\mathbf{f}_i + 4\mathbf{f}_{i-1} + \mathbf{f}_{i-2}).$$

Again, the method is convergent: although both zeros of r reside on the unit circle, they are simple.

Slightly more substantial calculation is required in the case $N = 3$. Now the Padé approximant is

$$\frac{-1-\frac{27}{11}w+\frac{27}{11}w^2+w^3}{\frac{3}{11}(1+9w+9w^2+w^3)}.$$

In particular, the zeros of r are 1 and $\frac{1}{11}\left(-19 \pm 4\sqrt{15}\right)$. It is easy to see that one zero exceeds unity in modulus: the method is not convergent! The

implication is that, as long as we desire a meaningful numerical solution, we cannot exploit all the coefficients in a three-step method to maximize order.

The coefficients for an arbitrary N can be written down explicitly (Dahlquist, 1963),

$$\alpha_k = \chi_N^{-1}(\chi_k - \chi_{N-k})\binom{N}{k}^2, \quad \beta_k = \frac{1}{2}\chi_N^{-1}\binom{N}{k}^2, \quad k = 0,1,\ldots,N,$$

where

$$\chi_0 := 0, \qquad \chi_m := \sum_{j=1}^{m} \frac{1}{j}, \quad m = 1,2,\ldots.$$

One way of producing the above expression exploits the aforementioned fact that r/s is a Padé approximant to $\log w$ at $w = 1$, with a known closed form (Baker, 1975).

It is easy to observe that $z^N r(z^{-1}) = -r(z)$, therefore if $r(\omega) = 0$, say, then also $r(\bar{\omega}^{-1}) = 0$. In other words, *either* all the zeros reside on the unit circle *or* the root condition is violated.

Suppose that the root condition holds and that $N \geq 2$. Then, in particular, the sum of moduli of the zeros cannot exceed N. However, the sum of the zeros equals $-r_{N-1}$, therefore the root condition implies that $|r_{N-1}| \leq N$. Substituting the explicit form of r_{N-1}, we obtain after elementary manipulation

$$\kappa_N := \frac{1}{N-1}\left(1 + \frac{1}{N}\right) - \sum_{j=1}^{N} \frac{1}{j} \geq 0.$$

Note that $\kappa_2 = 0$, hence all is right (as it should be!) for $N = 2$. However, for every $N \geq 2$

$$\kappa_{N+1} = \kappa_N - \frac{N^2+3}{(N^2-1)N} < 0,$$

hence the inequality is violated, r fails the root condition and the highest-order method cannot be convergent for $N \geq 3$. ◇

The lesson of the last example is that order and the root condition compete. Thus, it is essential to determine for every $N \geq 1$ the maximal order that is consistent with the root condition, hence with convergence.

Given r and order $p \geq N + 1$, the polynomial s is determined uniquely by Proposition 5.1: $M(\log w, w) = \mathcal{O}(|w-1|^{p+1})$ implies that

$$s(w) = \frac{r(w)}{\log w} + \mathcal{O}(|w-1|^p), \tag{5.5}$$

hence s is the truncated Taylor expansion of $r(w)/\log w$ about $w = 1$. In particular, it follows that order $p = N + 1$ is always consistent with the root condition, because we can always choose arbitrary r that obeys the condition and derive s from (5.5). Moreover, if $p \geq N + 1$, we have just one polynomial, r, to 'play' with. We are within the framework where order stars feature as an ideal tool: two features of a polynomial, each determined by a different geometric aspect in the complex plane, are in competition.

The question of maximal order consistent with the root condition has been determined by Dahlquist (1956), with no help from order stars. It is widely known as the **first Dahlquist barrier**. However, a proof based on order stars exists (Iserles and Nørsett, 1984)[3] and is presented in the remainder of this section.

We have already seen order stars of the second kind that depict Padé approximants to the logarithm in Figure 2.4. Unfortunately, the role of zeros is obscured there, hence they are of little use to the task in hand. Instead, we consider order stars of the second kind that are generated by

$$\tilde{\rho}(z) = \frac{s(e^z)}{r(e^z)} - \frac{1}{z}.$$

This 'inversion' means that zeros of r translate to poles of $\tilde{\rho}$, hence are points of nonanalyticity of the order star.

Figure 5.1 displays four order stars, corresponding to the methods

$$\mathbf{y}_i - \mathbf{y}_{i-1} = \frac{1}{720}h(251\mathbf{f}_i + 646\mathbf{f}_{i-1} - 264\mathbf{f}_{i-2} + 106\mathbf{f}_{i-3} - 19\mathbf{f}_{i-4})$$

(the fifth-order four-step **Adams–Moulton scheme** (Henrici, 1962)),

$$\mathbf{y}_i - \mathbf{y}_{i-2} = \frac{1}{3}h(\mathbf{f}_i + 4\mathbf{f}_{i-1} + \mathbf{f}_{i-2})$$

(the fourth-order two-step **Milne scheme**, cf. (Hairer *et al.*, 1987) and Example 5.1),

$$\mathbf{y}_i + \frac{27}{11}\mathbf{y}_{i-1} - \frac{27}{11}\mathbf{y}_{i-2} - \mathbf{y}_{i-3} = \frac{3}{11}(\mathbf{f}_i + 9\mathbf{f}_{i-1} + 9\mathbf{f}_{i-2} + \mathbf{f}_{i-3})$$

(sixth-order, three-step, cf. Example 5.1) and

$$\mathbf{y}_i - \frac{980}{363}\mathbf{y}_{i-1} + \frac{490}{121}\mathbf{y}_{i-2} - \frac{4900}{1089}\mathbf{y}_{i-3} + \frac{1225}{363}\mathbf{y}_{i-4} - \frac{196}{121}\mathbf{y}_{i-5} + \frac{490}{1089}\mathbf{y}_{i-6} - \frac{60}{1089}\mathbf{y}_{i-7} = \frac{140}{363}h\mathbf{f}_i$$

[3] Appropriately enough, this order star proof was originally published in a volume that celebrated Germund Dahlquist's sixtieth birthday.

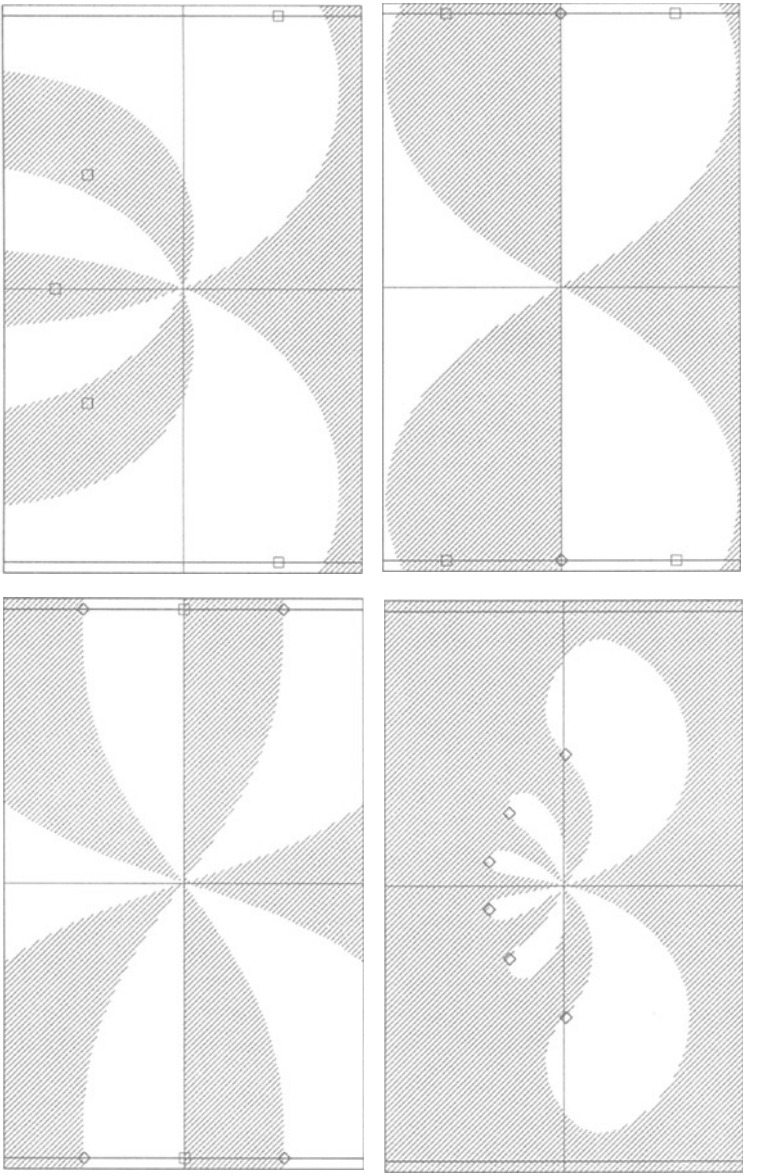


Figure 5.1 Order stars of the second kind for $s(e^z)/r(e^z) - 1/z$ and four multistep methods.

(the seventh-order, seven-step **backward differentiation formula** (Hairer *et al.*, 1987)). Note that the underlying polynomials r in the first two methods obey, and in the remaining two violate, the root condition.

Proposition 5.3 Satisfaction of the root condition is equivalent to all the poles of $\tilde{\rho}$ residing in the closed left half-plane and to the poles along $i\mathcal{R}$ being simple.

Proof. The proof follows at once from the definition of the root condition, since $z \mapsto \log z$ maps the unit disc onto the left half-plane and the unit circle onto the imaginary axis. □

Proposition 5.4 If the multistep method (5.1) is of order $p \geq 2$ then $\iota(0) = p - 1$ and the origin is a regular point of the order star.

Proof. Proposition 5.2 implies that

$$\tilde{\rho}(z) = 1 + \mathcal{O}(|z|^{p-1})$$

and the desired result follows at once from Proposition 2.10. □

The function $\tilde{\rho}$ involves $\exp z$, which is periodic in the complex plane. Hence, both zeros and poles are replicated by multiples of $2\pi i$, and this creates obvious difficulties to the zero-and-pole counting arguments that are central to the order star theory. The remedy is to restrict our attention to the strip

$$\mathcal{S} := \{z \in \mathcal{C} : |\mathrm{Im}\, z| \leq \pi\}.$$

Let

$$\mathcal{S}^+ := \mathrm{int}\,\mathcal{S} \cap \{\mathrm{Re}\, z > 0\} \qquad \text{and} \qquad \mathcal{S}^- := \mathrm{int}\,\mathcal{S} \cap \{\mathrm{Re}\, z < 0\}.$$

Lemma 5.5 There exists $\kappa \in \mathcal{R}$ such that $\{z \in \mathcal{C} : \mathrm{Re}\, z \geq \kappa\} \cap \mathcal{S}$ wholly belongs to one of the sets $\tilde{\mathcal{A}}_+$ or $\tilde{\mathcal{A}}_-$: if $\beta_N > 0$ then it belongs to $\tilde{\mathcal{A}}_+$, otherwise it lies in $\tilde{\mathcal{A}}_-$.

Proof. If $\beta_N \neq 0$ then $\lim_{\mathrm{Re}\, z \to \infty} \tilde{\rho}(z) = \beta_N$, whereas if β_N vanishes and $\mathrm{Re}\, z \to \infty$ then $\mathrm{Re}\{s(e^z)/r(e^z)\}$ decays faster than $\mathrm{Re}\{-1/z\}$. In either case the lemma follows from the definition of the order star. □

Both poles of $\tilde{\rho}$ and roots of the equation $\tilde{\rho}(z) = 0$ lie along $\tilde{\mathcal{A}}_0$ and it is important to our analysis to determine their relative positions. Recall that a **loop** is a closed curve in $\tilde{\mathcal{A}}_0$.

Lemma 5.6 Let γ be a loop such that $\gamma \cap \mathcal{S} \neq \emptyset$ and $\gamma \cap \partial\mathcal{S} = \emptyset$. Then there is on γ exactly one pole of $\tilde{\rho}$ between any two roots of $\tilde{\rho}(z) = 0$. Moreover, every pole of $\tilde{\rho}$ in $\mathrm{int}\,\mathcal{S}$ is a regular point of index equal to its multiplicity.

Proof. The first part of the lemma follows at once from Proposition 2.11 and its remainder from Proposition 2.12. □

Let $\mathcal{U}$ be either a bounded $\mathcal{A}_+$-region or a bounded $\mathcal{A}_-$-region and suppose that $\mathrm{cl}\,\mathcal{U} \cap \{\mathcal{R} + \pi i\} \neq \emptyset$. We set

$$\begin{aligned} x_- &:= \min\{x \in \mathcal{R} : x + \pi i \in \mathrm{cl}\,\mathcal{U}\} > -\infty, \\ x_+ &:= \max\{x \in \mathcal{R} : x + \pi i \in \mathrm{cl}\,\mathcal{U}\} < \infty. \end{aligned}$$

Lemma 5.7 Let $z_0 \in \partial\mathcal{U} \cap \mathrm{int}\,\mathcal{S}$ be a zero of $\tilde{\rho}$. Then

(a) If $\mathcal{U}$ is an $\mathcal{A}_+$-region then either $x_+ + \pi i$ is a pole of $\tilde{\rho}$ or there is a pole along the positively oriented portion of $\partial\mathcal{U}$ extending from z_0 to $x_+ + \pi i$;

(b) If $\mathcal{U}$ is an $\mathcal{A}_-$-region then either $x_- + \pi i$ is a pole of $\tilde{\rho}$ or there is a pole along the positively oriented portion of $\partial\mathcal{U}$ extending from $x_- + \pi i$ to z_0.

Furthermore, a similar statement is valid if $\mathcal{R} + \pi i$ is replaced by $\mathcal{R} - \pi i$ in the definition of $\mathcal{U}$.

Proof. Recall that $\mathrm{Im}\,\tilde{\rho}$ is monotone along $\partial\mathcal{U}$. Moreover, by direct calculation,

$$\mathrm{Im}\,\tilde{\rho}(x \pm \pi i) = \pm(x^2 + \pi^2)^{-1}, \quad x \in \mathcal{R}.$$

The lemma follows, because poles are the only points of nonanalyticity of $\tilde{\rho}$ along $\partial\mathcal{U}$. □

Let

$$F(t) := \mathrm{Re}\,\{s(e^{it})r(e^{-it})\} = |r(e^{it})|^2 \mathrm{Re}\,\tilde{\rho}(it), \quad t \in \mathcal{R}.$$

Clearly, F is an Nth degree polynomial in $(1 - \cos t)$. Note that, consequently, it is an even function. Moreover, Proposition 5.2 implies

$$r(e^{-it}) = -its(e^{-it}) + \mathcal{O}(t^{p+1})$$

and substitution into the definition of F affirms that

$$F(t) = \mathcal{O}(t^{2[(p+2)/2]}) = \mathcal{O}((1 - \cos t)^{[p+1]/2}).$$

Thus, there are exactly two possibilities: either F is identically zero or it may have at most $N - [(p+2)/2]$ zeros in $(0, \pi)$.

We stipulate that the method (5.1) obeys the root condition and consider first the case of $\beta_N > 0$.

If $F \equiv 0$ then $i\mathcal{R} \subset \tilde{\mathcal{A}}_0$ and it follows from Proposition 5.4 that exactly $p - 1$ sectors of $\tilde{\mathcal{A}}_+$ and $p - 1$ sectors of $\tilde{\mathcal{A}}_+$ adjoin the origin from within $\mathcal{S}^+$ (cf. Figure 5.1). Since the pure imaginary axis lies on $\tilde{\mathcal{A}}_0$, no loop can

cross into $\mathcal{S}^-$. Moreover, according to Lemma 5.5 there is just a single unbounded region in $\mathcal{S}^+$. Therefore, there must be at least $p-2$ bounded loops in $\mathcal{S}^+$. According to Lemmas 5.6 and 5.7 each loop must 'account' for a pole of $\tilde{\rho}$. However, the root condition and Proposition 5.3 imply that there are no poles in $\mathcal{S}^+$ except along the imaginary axis – and the latter are simple. Thus, each bounded loop must contain a portion of $i\mathcal{R}$. Moreover, by Proposition 2.12 and simplicity of the poles there, no pole may 'support' more than a single loop.

Only $N-1$ poles are available to 'support' loops, since $r(1)=0$ has migrated to the origin. However, if N is even then we can count a pole twice, provided that it lies at πi (and so, by periodicity, at $-\pi$). Hence $p-2 \leq N$. Note that if N is odd then $r(-1) \neq 0$ and $\pm\pi i$ is not a pole – otherwise $F \equiv 0$ would have implied that it is of even multiplicity, contradicting the root condition. Therefore, $p-2 \leq N-1$. We obtain the inequality

$$p \leq 2\left[\frac{N+2}{2}\right]. \tag{5.6}$$

The case $F \not\equiv 0$ is more delicate, since $i\mathcal{R}$ no longer belongs wholly to $\tilde{\mathcal{A}}_0$. All we can deduce now from Proposition 5.4 is that between $[(p+1)/2]-1$ and $[(p+1)/2]$ $\mathcal{A}_+$-regions adjoin the origin from $\mathcal{S}^+$. Since $\beta_N > 0$, Lemma 5.5 implies that none may extend to $+\infty$. Thus, all the corresponding loops must be 'supported' by poles of $\tilde{\rho}$ which, by Proposition 5.3, are either on or to the left of $i\mathcal{R}$.

There are the following possibilities for an $\mathcal{A}_-$-region $\mathcal{U}$ in $\mathcal{S}^+$ to 'use' a pole from $\mathcal{C} \setminus \mathcal{S}^+$. In the following, $\gamma = \partial\mathcal{U}$, the $\mathcal{A}_+$- loop formed by the boundary of $\mathcal{U}$:

(1) $\mathcal{U}$ adjoins the origin wholly within $\mathcal{S}^+$, γ crosses $i\mathcal{R}$ and it contains a pole from $\operatorname{int}\mathcal{S}^-$;

(2) $\mathcal{U}$ adjoins the origin wholly within $\mathcal{S}^+$, γ crosses $i\mathcal{R}$ and it contains a pole on the line $\{\operatorname{Re} z < 0,\ |\operatorname{Im} z| = \pi\}$. It is easy to verify that this pole is a zero of F of even multiplicity;

(3) $\mathcal{U}$ adjoins the origin wholly within $\mathcal{S}^+$ and γ contains a zero on $i\mathcal{R}$;

(4) $\mathcal{U}$ adjoins the origin wholly within $\mathcal{S}^+$, crosses into $\mathcal{S}^-$ and becomes unbounded there (tending to $-\infty$). In that case necessarily $r(0)=0$ and, counting the multiplicity, for each such region we have one less pole to 'support' the remaining loops;

(5) $\mathcal{U}$ adjoins the origin along $i\mathcal{R}$ (in other words, in an arbitrarily small neighbourhood of the origin it contains points from both $\mathcal{S}^+$ and $\mathcal{S}^-$), γ crosses $i\mathcal{R}$ and it contains a pole from either $\operatorname{int}\mathcal{S}^-$ or $i\mathcal{R}$;

(6) $\mathcal{U}$ adjoins the origin along $i\mathcal{R}$ and γ contains a zero on the line $\{\operatorname{Re} z < 0,\ |\operatorname{Im} z| = \pi\}$.

Note that at most two $\mathcal{A}_-$-regions may be of type (5) or (6).

Careful examination of all possibilities affirms that there are at least $2[(p+1)/2]-4$ points of $\tilde{\mathcal{A}}_0$ along the line segment $\{it\,;\,0<t<\pi\}$. Each such point is, by definition, a zero of F (inclusive of poles of $\tilde{\rho}$ along the segment). As the number of zeros of F there may not exceed $2(N-[(p+2)/2])$, it follows that

$$p \le N+1. \tag{5.7}$$

A point of minor interest is that, unlike in the case $F \equiv 0$, simplicity of poles along $i\mathcal{R}$ is not used in the proof.

Next, consider the remaining case, $\beta_N \le 0$. Note that, by Lemma 5.7, $+\infty$ is now approached by an $\mathcal{A}_-$-region. The analysis is similar to the previous case, with few important differences.

If $F \equiv 0$ then the number of sectors of $\tilde{\mathcal{A}}_-$ adjoining the origin is odd, hence, by Proposition 5.4, p must be even. If the origin is adjoined from the right along the real axis by $\tilde{\mathcal{A}}_-$ then it follows easily that $p \le N$, otherwise, by symmetry, $p-2 \le N-3$, thus $p \le N-1$. We derive the bound

$$p \le N, \qquad p \text{ even}. \tag{5.8}$$

Finally, if $F \not\equiv 0$ we denote by μ the number of sectors of $\tilde{\mathcal{A}}_-$ that adjoin the origin in $\mathcal{S}^+$ and approach $+\infty$. They envelop $\mu-1$ sectors of $\tilde{\mathcal{A}}_+$ which must, as a consequence of the root condition and our analysis, either cross into $\operatorname{int}\mathcal{S}^-$ or adjoin poles along the imaginary axis. We now count zeros of F due to *both* bounded $\mathcal{A}_-$-regions and $\mathcal{A}_+$-regions crossing to the left in search of poles. This yields again the bound (5.8).

The bounds (5.6–8), taken together, produce the desired restriction on order, subject to stability – the **first Dahlquist barrier.**

Theorem 5.8 (Dahlquist, 1956) Subject to convergence, the order of a multistep method (5.1) is bounded by $2[(N+2)/2]$. The order exceeds $N+1$ only if N is even and all the zeros of r lie on the unit circle. If $\beta_N \le 0$ (and, in particular, if the method is explicit) the order is bounded by N. □

The first Dahlquist barrier (which, of course, was proved by Germund Dahlquist using an entirely different technique, essentially by exploiting the Riesz–Herglotz representation of some asymptotic series) was an early triumph of theoretical numerical mathematics. By highlighting the subtle influence of 'stability' of the attainable order in one particular case, it had contributed threefold to our understanding: firstly by solving the underlying problem, secondly by implicitly raising the whole issue of interplay between order and stability in numerical solution of differential equations and, finally, by demonstrating that the aforementioned issue is tractable by complex-analytic means.

5.2 Order stars on Riemann surfaces

Order stars have been defined in Chapter 2 for essentially analytic functions and so far this framework proved itself amply adequate for our purposes. However, multistep methods confront us with functions that fail in a most significant way to be essentially analytic. Moreover, although these functions are complex, the complex plane is not the right medium for their analysis.

We solve the classical test equation for linear stability,

$$y' = \lambda y, \quad y(0) = 1 \text{ with } \lambda \in C$$

by using the multistep method (5.1) with a fixed step. This procedure yields the linear difference equation

$$\sum_{\ell=0}^{N} (\alpha_\ell - z\beta_\ell) y_{i+\ell-N} = 0, \quad i = 0, 1, \ldots, \tag{5.9}$$

where we are letting $z := h\lambda$. Thus, the sequence $\{y_i\}_{i=0}^{\infty}$ tends to zero and z lies in the **linear stability set** of the method if and only if all the N zeros $w_1(z), w_2(z), \ldots, w_N(z)$ of the underlying characteristic polynomial

$$M(z; w) := \sum_{\ell=0}^{N} (\alpha_\ell - z\beta_\ell) = r(w) - zs(w)$$

lie inside the complex unit disc. In other words, the method (5.1) is **A-stable** (that is, it solves stably all stable linear differential equations) if

$$z \in C, \ \mathrm{Re}\, z < 0 \quad \Longrightarrow \quad |w_\ell(z)| < 1 \text{ for all } \ell = 1, 2, \ldots, N.$$

In the one-step case we have just one root w_1 to reckon with and it is, of course, a rational function. This is no longer true when $N \geq 2$. The stability function is now multivalued: for every relevant value of z we need N distinct quantities to be bounded by unity at one and the same time. To rephrase this important observation, *although we have more degrees of freedom than in the one-step case, we need to keep more quantities under control.* This is not just a source of considerable added conceptual difficulty but also explains why – as will be amply clear in the sequel – stability properties of multistep methods are often disappointing.

Example 5.2 The two step, second order BDF method is given by

$$y_{n+2} - \frac{4}{3} y_{n+1} + \frac{1}{3} y_n = \frac{2}{3} f_{n+2}$$

and its application to the linear test equation yields

$$M(w, z) = \left(1 - \frac{2}{3} z\right) w^2 - \frac{4}{3} w + \frac{1}{3}.$$

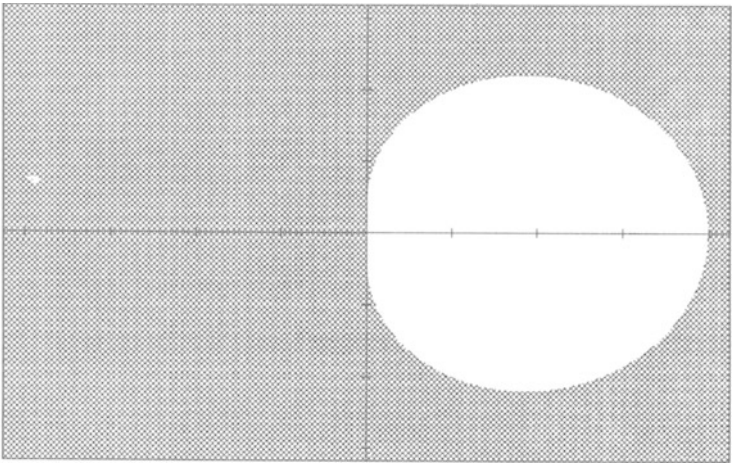


Figure 5.2 Stability domain (the cross-hatched set) of the two-step BDF method.

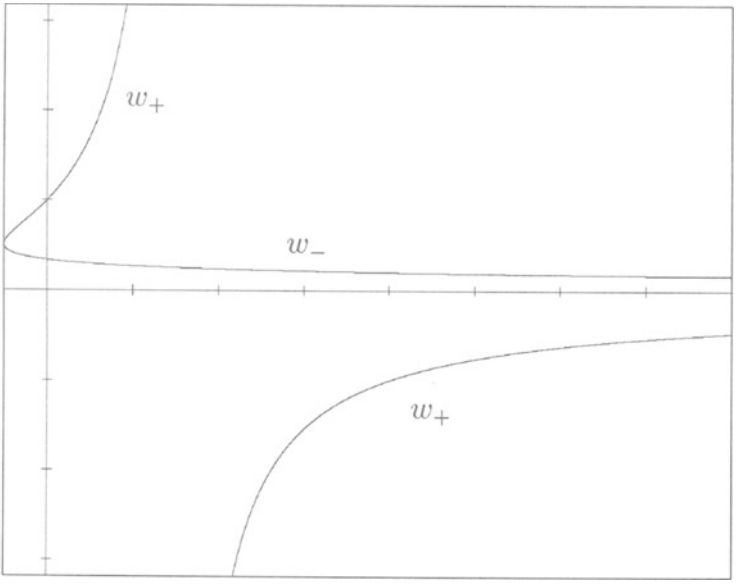


Figure 5.3 The 'solutions' w_+ and w_- of the two-step BDF method for $-\frac{1}{2} \leq x \leq 8$.

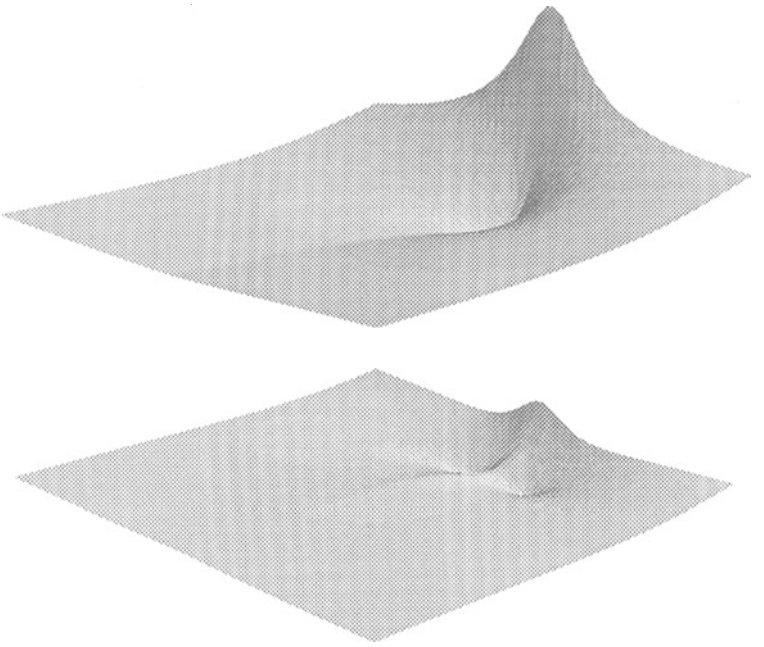


Figure 5.4 The level curves for $w_{\pm}$ for the two-step BDF method.

To examine A-stability we use the **Cohn–Schur** criterion (Lambert, 1973): both zeros of the quadratic

$$aw^2 + bw + c = 0, \quad a, b, c \in C, \; a \neq 0,$$

reside in the open unit disc provided that

$$|a|^2 > |c|^2, \quad \left(|a|^2 - |b|^2\right) > |a\bar{b} - b\bar{c}|^2 . \tag{5.10}$$

Letting $a = 3 - 2z$, $b = 4$ and $c = 1$, (5.10) produces

$$\begin{aligned} 2 - 3\mathrm{Re}\, z + |z|^2 &> 0, \\ |z|^4 + 9(\mathrm{Re}\, z)^2 &> \left(4 + 6|z|^2\right) \mathrm{Re}\, z. \end{aligned}$$

Since both inequalities are satisfied when $\mathrm{Re}\, z < 0$, it follows that the method is A-stable. Actually, the set of stable points ventures well outside the left half plane, as can be seen in Figure 5.2.

Of course, it is trivial to write the zeros explicitly:

$$w_{\pm}(z) = \frac{1}{3-2z}\left(2 \pm \sqrt{1+2z}\right).$$

Note that

$$w_{+}(z) = e^{z} + \mathcal{O}(|z|^{3}).$$

This reflects the fact that we are dealing with a second-order method. However, the function w_{-} does not approximate any meaningful quantity. This root is purely extraneous to accuracy considerations although, of course, it can spoil stability just as much as its more germane relative. Figure 5.3 displays w_{+} and w_{-} for real values of z whereas, in Figure 5.4, we present the level curves $|w_{+}|$ and $|w_{-}|$ in a portion of the complex plane.

Since w_{+} approximates the exponential, it is a natural idea to examine the order star with respect to $\rho(z) = e^{-z}w_{+}(z)$. It is displayed in Figure 5.5, where it can be seen that $\iota(0) = 3$, as predicted by Proposition 2.1. However, two aspects of the underlying approximant render this approach inadequate. Firstly, stability can be damaged by w_{-}, as well as by w_{+}, and the first function is not reflected at all in the order star. Secondly, we can distinguish between w_{+} and w_{-} only within the radius of convergence of the Taylor series (of both functions) about the origin. This radius equals $\frac{1}{2}$, since they possess a branch point at $z = -\frac{1}{2}$.[4] $\diamond$

The last example presents in stark relief the difficulties that are implicit in stability analysis of multistep methods. The right approach is to abandon the treatment of $w_{\pm}$ as complex functions altogether and consider instead the multivalued function

$$\mathbf{w}(z) = \begin{bmatrix} w_{+}(z) \\ w_{-}(z) \end{bmatrix},$$

defined on a Riemann surface.

Before we describe the general theoretical framework of Riemann surfaces, it makes sense to broaden our canvas to multistep multiderivative methods, that combine the multistep form (5.1) and the multiderivative form (3.10).[5] Recall the function $\mathbf{g}_j$ from section 3.3, that stands for the

[4] The order star in Figure 5.4 has been drawn, of course, for values within the 'set of unambiguity'.

[5] This is not the most general 'melange' of methods that is tractable by the technique of this section. It is perfectly possible to combine multistep, multiderivative and Runge–Kutta methods into **general linear methods** (Butcher, 1987). This leads again to algebraic functions in A-stability analysis, the sole difference being that the order conditions of the principal solution are no longer sufficient for the order of accuracy of the scheme.

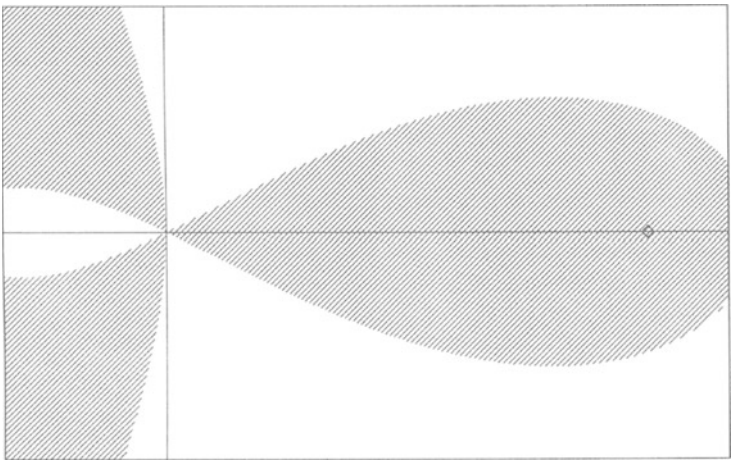


Figure 5.5 Order star of the principal solution w_+ of the two-step BDF method, for the range $-\frac{1}{2} \leq \mathrm{Re}\, z \leq \frac{7}{4}$.

jth derivative of $\mathbf{y}$, as given by formal differentiation of the ordinary differential equation (3.1). A **multistep multiderivative method** reads

$$\sum_{\ell=0}^{N} \sum_{j=0}^{n} q_{\ell,j} h^j \mathbf{g}_j(a + (i + \ell - N), \mathbf{y}_{i+\ell-N}) = 0. \tag{5.11}$$

We stipulate the normalization $q_{N,0} = 1$.

Application of (5.11) to the linear test equation gives

$$\sum_{\ell=0}^{N} \sum_{j=0}^{n} q_{\ell,j} z^j y_{i+\ell-N} = 0,$$

where, again $z = h\lambda$. Therefore, $y_i \to 0$ and the method is stable for the underlying value of z if all the zeros of

$$M(z; w) := \sum_{\ell=0}^{N} Q_\ell(z) w^\ell$$

are in the open unit disc. Here

$$Q_\ell(z) = \sum_{j=0}^{n} q_{\ell,j} z^j \in \pi_n[z], \quad \ell = 0, 1, \ldots, N.$$

We assume forthwith that the function M is **irreducible**: it cannot be expressed as a product

$$M(z;w) = M_1(z;w)M_2(z;w),$$

where both M_1 and M_2 are non-trivial polynomials in z and w.[6] Furthermore, we stipulate that $\partial M(1,0)/\partial w \neq 0$ – this is a necessary condition for convergence. In fact, satisfaction of the **root condition** (as defined in section 5.1 in the context of one-derivative methods) by the polynomial Q_n is necessary and, in tandem with consistency, sufficient for convergence of (5.11) (Stetter, 1973). Another (and more immediate) benefit of $\partial M(1,0)/\partial w \neq 0$ is that the origin is not a branch point.

The N-tuple

$$\mathbf{w}(z) := \begin{bmatrix} w_1(z) \\ w_2(z) \\ \cdots \\ w_N(z) \end{bmatrix} \tag{5.12}$$

'lives' on a Riemann surface (Ahlfors, 1966), that is herewith described more formally. Note that one of the components – without loss of generality we can assume that it is w_1 – approximates $\exp z$ to an order equal to the order of accuracy of the method (5.11). We call it the **principal solution.**

An analytic function f, specified in a region $\mathcal{X} \subset \mathcal{C}$, defines a **function element**, which we denote by $(f,\mathcal{X})$. A **general analytic function** is a non-void, countable collection W of function elements, such that for every pair $(f_\alpha,\mathcal{X}),(f_\beta,\mathcal{X}) \in W$ there exists a chain $\{(f_{\gamma_i},\mathcal{X})\}_{i=1}^{M}$ such that $\gamma_1 = \alpha$, $\gamma_M = \beta$ and for each $i \in \{1,2,\ldots,M-1\}$ the elements $(f_{\gamma_i},\mathcal{X})$ and $(f_{\gamma_{i+1}},\mathcal{X})$ are analytic continuations of each other. A **complete analytic function** is a general analytic function that contains all analytic continuations of any of its elements (Ahlfors, 1966).

Example 5.3 The two solutions w_+ and w_- from Example 5.2 are function elements of the general analytic function

$$w(z) = \frac{2+\sqrt{2z-1}}{3-2z}$$

(where, of course, the square-root function is two-valued), with $\mathcal{X} = \mathcal{C}\setminus\{\frac{3}{2}\}$. In general, the N components of the function $\mathbf{w}$ from (5.12) are functional elements that combine into a general analytic function. The set $\mathcal{X}$ encompasses $\mathcal{C}$, punctured at the zeros of Q_N. It is easy to prove that the function is complete. $\diamond$

[6] If M is reducible then an identical solution sequence (for a linear problem) will be produced by the method which is defined by the function M_1, say. Thus, our assumption entails no loss of generality.

We say that a complete analytic function W is an **algebraic function** if there exists a polynomial $S(\cdot;\cdot)$ such that all the function elements $(f,\mathcal{X})$ obey the relation $S(z,f(z)) = 0$ (Ahlfors, 1966). This is precisely the case with the function whose elements feature in (5.12).

Proposition 5.9 (Ahlfors, 1966) Let $z^* \in \mathcal{C}$ be neither a zero of Q_N nor such that $M(z^*;w) = 0$ possesses a multiple zero. Then the equation $M(z^*;w) = 0$ has N distinct solutions $w_1^*, w_2^*, \ldots, w_N^*$, say, and there exist an open neighbourhood $\mathcal{Y}$ such that $z^* \in \mathcal{Y}$ and N function elements $(w_k,\mathcal{Y})$, $k = 1,2,\ldots,N$ such that

(a) $M(z;w_k(z)) \equiv 0$ for $z \in \mathcal{Y}$ and $k = 1,2,\ldots,N$;

(b) $w_k(z^*) = w_k^*$, $k = 1,2,\ldots,N$;

(c) If $M(z,w) = 0$ for some $z \in \mathcal{Y}$ then there exists $k \in \{1,2,\ldots,N\}$ such that $w = w_k(z)$.

□

As Ahlfors (1966) concludes, the above reasoning allows for the construction of an algebraic function W. Specifically, we choose W as the complete analytic function that is determined by the element $(w_1,\mathcal{X})$ at any $z \in \mathcal{X}$, except for a finite number of **excluded points**: zeros of Q_N (thus, poles of W) and multiple zeros of M (i.e. branch points). We can then continue W to its other $N-1$ **branches** through the branch points.[7] In other words, for every branch and every non-excluded point z there corresponds a unique function value $w(z)$ and, w being analytic, we can talk about derivatives, expansions and more sophisticated features of analytic function theory.

We define the **Riemann surface** $\mathcal{M}$ in the following manner: For each non-excluded $z \in \mathcal{X}$, we add to $\mathcal{M}$ all the values of the branches $(w_k(z),\mathcal{X})$, $k = 1,2,\ldots,N$. The outcome, a set of N sheets, punctured at the branch points, is extended to these by continuation. It is, of course, helpful to visualize them as collections of points (w,z) such that $M(z;w) = 0$. The whole point of considering Riemann surfaces is that the multivalued function W becomes single-valued on $\mathcal{M}$. Since $\zeta \in \mathcal{M}$ is a collection of values $(w_k(z),\mathcal{X})$, we can define the **natural projection** $\pi : \mathcal{M} \to \mathcal{X}$ by $\pi(\zeta) = z$. The inverse π^{-1} is well defined away from branch points and it assigns to z an N-tuple of values, each on a different branch.

Before we define formally order stars on Riemann surfaces, we just mention for further reference that both order and stability of (5.11) express themselves in the geometry of $\mathcal{M}$. We have already mentioned the principal solution w_1, which reflects the order of accuracy. In the vicinity of the origin it lies on a distinct branch, which we term the **principal branch**. By virtue of analyticity, we are back to the old framework of approximants to $\exp z$, albeit only as long as we do not 'hit' a branch point. Moreover, a

[7] Irreducibility implies that all the branches can be 'connected' in this fashion.

method is A-stable if for every $z \in \mathcal{C}$ such that $\mathrm{Re}\, z < 0$ it is true that all components of $\pi^{-1}(z)$ are strictly bounded by 1 in modulus.

We let $\mathcal{X}$ be the whole closed complex plane (thus, inclusive of ∞), except for the zeros of Q_N, and define an order star on $\mathcal{M}$ as

$$\begin{aligned} \mathcal{A}_+^R &:= \{\zeta \in \mathcal{M} : |\rho(\zeta)| > 1\}, \\ \mathcal{A}_0^R &:= \{\zeta \in \mathcal{M} : |\rho(\zeta)| = 1\}, \\ \mathcal{A}_-^R &:= \{\zeta \in \mathcal{M} : |\rho(\zeta)| < 1\}, \end{aligned}$$

where

$$\rho(\zeta) := e^{-\pi(\zeta)} w(\zeta), \quad \zeta \in \mathcal{M}.$$

Of course, it is impossible to present pictures of order stars on Riemann surfaces in a two-dimensional book. However, any neighbourhood on any sheet (away from branch points) is locally diffeomorphic to a complex neighboorhood. In other words, as long as we stay away from branch points, we might inspect 'flattened-out' portions of the order star, for example in Figure 5.5. Moreover, limited insight into the geometry of the order star is provided by its two-dimensional projection, detaching the sheets along the branch cuts. This 'surgery' involves, needless to say, an element of arbitrariness.

We have introduced in section 2.1 the concept of an interpolation point of degree p. This can be extended easily from $\mathcal{C}$ to $\mathcal{M}$ and the present context of approximating $\exp z$. Thus, $\zeta^* \in \mathcal{M}$ is an **interpolation point** of degree $p \geq 1$ if

$$w(\zeta) = e^{\pi(\zeta)} + C(\zeta - \zeta^*)^p + \mathcal{O}(|\zeta - \zeta^*|^{p+1}), \quad C \neq 0, \ \zeta \to \zeta^*,$$

when $|\zeta^*|$ is bounded or

$$w(\zeta) = C\zeta^{-p} + \mathcal{O}(|\zeta|^{-p-1}), \quad C \neq 0, \ |\zeta| \to \infty,$$

for $\zeta^* = \infty$.

The phrase 'interpolation' is perhaps misleading. The function w might well interpolate the exponential, but, in general, the values on the remaining branches are totally meaningless. 'Interpolation' assumes implicitly that we are approximating alike with alike, and this is not the case here – except in a strictly local sense.

Much of the theory in Chapter 2 can be transplanted *in toto* to Riemann surfaces. Thus, instead of repeating, point by boring point, all its statements and proofs, it suffices to highlight the main results and the subtle differences in the present framework.

Since we are interested in modelling order of the method (5.11), rather than more general interpolation features, we restrict the definition of **index**

to two locations: the principal branch and infinity (the latter being the only essential singularity of the order star).

Proposition 5.10 Let w_1 represent the principal solution of a method (5.11) of order p. Then $\iota(0) = p+1$ and the origin is a regular point of $\mathcal{A}_0^R$. Moreover, $\iota(\infty) = 1$ along all the branches, ∞ is regular and for every $\varepsilon > 0$ there exists $r > 0$ such that the arcs

$$\left\{re^{i\theta} : -\frac{1}{2}\pi + \varepsilon \leq \theta \leq \frac{1}{2}\pi - \varepsilon\right\} \quad \text{and} \quad \left\{re^{i\theta} : \frac{1}{2}\pi + \varepsilon \leq \theta \leq \frac{3}{2}\pi - \varepsilon\right\}$$

belong to $\mathcal{A}_-^R$ and $\mathcal{A}_+^R$ respectively.

Proof. The first part follows trivially from Proposition 2.1, since w_1 is analytic in the vicinity of the origin. The statement on the behaviour at infinity is a consequence of the growth (or decay, as the case might be) of $\exp(\mathrm{Re}\, z)$ ultimately overwhelming $|w_k(z)|$ as z approaches ∞ along any ray emanating from the origin, away from the imaginary axis. □

Transplanting Proposition 2.3, that links the multiplicity of an $\mathcal{A}_\pm$-region with the number of zeros or poles therein, requires some subtlety. The definition of $\mathcal{A}_\pm^R$-regions and their multiplicity is a straightforward generalization and, as in section 2.1, we say that a region is **analytic** if its boundary does not cross an essential singularity – in the present case it means, of course, that the region is bounded. Moreover, we say that an analytic $\mathcal{A}_\pm^R$-region is **strictly analytic** if it does not contain any branch points. In other words, a strictly analytic region is bounded and confined to a single sheet of $\mathcal{M}$.

Proposition 5.11 The multiplicity of an analytic $\mathcal{A}_+^R$ region $\mathcal{U}$ cannot exceed the number of zeros of Q_N inside $\mathcal{U}$, counted with their multiplicity. The multiplicity of a strictly analytic $\mathcal{A}_-^R$-region equals the number of zeros of Q_0 within the region.

Proof. We note first that zeros of Q_N and of Q_0 correspond to poles and zeros of ρ, respectively. An assertion that the multiplicity of a strictly analytic $\mathcal{A}_+^R$-region ($\mathcal{A}_-^R$-region) $\mathcal{U}$ equals the number of zeros of Q_N (Q_0) in $\mathcal{U}$ follows at once from Proposition 2.3, because ρ is meromorphic there and the standard argument principle of analytic function theory remains valid. This is no longer the case if there are branch points inside the region, since every branch point contributes an increase in the argument (of magnitude, a fraction of 2π, that depends on multiplicity, cf. (Ahlfors, 1966)). In other words, branch points act, for the purposes of the argument principle, as zeros of fractional multiplicity. This means that extra zeros of Q_N might be required to 'compensate' for branch points in non-strictly analytic $\mathcal{A}_+^R$-regions and, *mutatis mutandis*, the multiplicity of non-strictly analytic $\mathcal{A}_-^R$-regions may exceed the number of zeros of Q_0 there. □

The definition of $\mathcal{V}^*$-contractivity and $\mathcal{V}$-contractivity can be extended effortlessly to Riemann surfaces by using the natural projection. This is true not just for the Riemann surface under consideration (which is 'generated' by an algebraic function). Thus, let $\mathcal{V}$ be an open subset of $\mathcal{M}$, with a Jordan boundary, and suppose that a function $f : \mathcal{M} \to \mathcal{C}$ is analytic in $\mathcal{V}$ and $|f| \equiv 1$ along $\partial\mathcal{V}$ (except, possibly, for a finite number of points). The maximal modulus theorem survives the transition from the complex plane to Riemann surfaces and can be used to argue that $|f(\zeta)| < 1$ for all $\zeta \in \mathcal{M}$. Accordingly, we say that f is a **$\mathcal{V}^*$-contraction** if it maps $\mathrm{cl}\,\mathcal{V}$ onto the closed complex unit disc, and we call it a **$\mathcal{V}$-contraction** if it maps $\mathcal{V}$ into the open unit disc.

Proposition 5.12 The algebraic function w is A-acceptable and the underlying method (5.11) is A-stable if w is a $\pi^{-1}(\mathcal{C}^-)$-contraction, where $\mathcal{C}^-$ denotes the open left half plane. This is true if and only if no zero of Q_N resides in $\mathcal{C}^-$ and

$$\mathcal{A}_+^R \cap \pi^{-1}(i\mathcal{R}) = \emptyset.$$

□

We have developed in this section a generalization of order stars from the complex plane to Riemann surfaces. The purpose of this exposition should be clear in the next section, where the focus is on A-stability barriers for multistep multiderivative methods.

5.3 The Daniel–Moore conjecture and its solution

What is the highest attainable order of an A-stable method? This, quite obviously, is a question of an overwhelming practical importance: other things being equal, solving stiff ordinary differential systems we should opt for the A-stable method of highest order! It is implicit in the discussion in section 5.2 that the problem is non-trivial. As the number of steps increases, so does not just the number of degrees of freedom (hence greater scope for high order), but also the number of branches in the underlying Riemann surface – and, of course, all these branches must be 'controlled'. The situation is similar to the first Dahlquist barrier from section 5.1, except that A-stability is a much more severe requirement than convergence. Consequently, we can anticipate the answer to our question to be considerably smaller than that of Theorem 5.8.

Prior to presenting the answer, we further sharpen the question. For, simply finding the highest order of an A-stable N-step, n-derivative method for given N and n still leaves possibly an infinity of candidates for the 'optimal' method. To single out the best in that crowd, we need to have a

look at the error constant.

Given that w is the algebraic function corresponding to a pth order convergent method (5.11), we say that c is the **error constant** if

$$w_1(z) = e^z + cz^{p+1} + \mathcal{O}(|z|^{p+2}), \quad z \to 0,$$

where w_1 is the principal solution.

Theorem 5.13 (Dahlquist, 1963) The highest order of an A-stable multistep method (5.1) is two. Moreover, the error constant of an A-stable second-order method obeys the inequality $|c| \geq \frac{1}{12}$. □

Let us ponder briefly on the implications of this theorem, known as the **second Dahlquist barrier**. Firstly – and these are good news – the question has a meaningful and clear answer, at the very least for one-derivative methods. Secondly, adding extra steps cannot increase the order of an A-stable method. Thirdly – to compund further the rout of multistep methods – the least error constant among second-order A-stable methods 'belongs' to the trapezoidal rule which, needless to say, is a one-step method!

Restriction to second order of acuracy is quite severe and Dahlquist's result has prompted much work on circumventing the barrier – either by relaxing stability requirements or by using different methods.[8] Moreover, it has been conjectured by Daniel and Moore (1970) that the order of an A-stable multiderivative method (5.11) is restricted by a similar barrier. To get the right 'feel' for such a barrier, let us recall the Obrechkoff methods from section 3.3. Each scheme (3.10) is a one-step, $\max\{m,n\}$-derivative method of order $m+n$, whose stability function is the m/n Padé approximant. According to Lemma 3.4, the error constant is $(-1)^{n+1}n!m!/((n+m)!(n+m+1)!)$. Because of the connection with Padé approximants, the corollary to Theorem 4.1 implies that the Obrechkoff method with $m=n$ is A-stable. In other words, for each $n \geq 1$ we have an example of a multistep, n-derivative, A-stable method of order $2n$ and with an error constant $(-1)^{n+1}(n!)^2/((2n)!(2n+1)!)$. The **Daniel–Moore conjecture** states that this is the best that can be achieved by any A-stable n-derivative method.

Some attempts had been undertaken to solve the Daniel–Moore conjecture, most notably by Genin (1974) and Jeltsch (1976). The complete proof was provided by Wanner, Hairer and Nørsett (1978) in the original paper that introduced order stars, and it relies heavily on the material in the previous section.

Theorem 5.14 (Wanner *et. al.*, 1978) The method (5.11) of order p is A-stable only if $p \leq 2n$. Moreover, for every A-stable method of order $2n$, the

[8] There exists an A-stable s-stage Runge–Kutta method of order $2s$. As this is the maximal order of any s-stage method, evidently Runge–Kutta methods are not a subject to stability barriers *à la* multistep schemes.

sign of the error constant is $(-1)^{n+1}$.

Proof. Order p, in tandem with Proposition 5.10, imply that, on the principal branch of w, at least $[(p+1)/2]$ $\mathcal{A}_+^R$-regions approach the origin to the right of $i\mathcal{R}$. The method being A-stable, according to Proposition 5.12 no $\mathcal{A}_+^R$-region is allowed to cross $\pi^{-1}(i\mathcal{R})$. Proposition 5.10 implies that these regions are analytic and we can now use Proposition 5.11 to argue that there must be at least $[(p+1)/2]$ poles of w, counted with their multiplicities, in $\pi^{-1}(\mathbf{C}^+)$, the image of the open right half-plane under the inverse of the natural map. The poles of w are precisely the zeros of the nth degree polynomial Q_N and it follows that $[(p+1)/2] \leq n$. This yields the desired inequality $p \leq 2n$.

Let us consider an A-stable method of maximal order $2n$.[9] It follows from our analysis that *exactly* n $\mathcal{A}_+^R$-regions adjoin the origin from the right on the principal branch. Moreover, $\iota(0)$ being $2n+1$, exactly $n+1$ $\mathcal{A}_-^R$-regions approach the origin there. Thus, if n is even then the origin is approached along the positive semi-axis by an $\mathcal{A}_-^R$-region, whereas in the case of an odd n, it is approached there by an $\mathcal{A}_+^R$-region. The assertion on the sign of the error constant follows at once from the definition of the order star. □

The proof of the first part of the Daniel–Moore conjecture is complete. The prove the remaining part, namely that the error constant c of an A-stable n-derivative method of order $2n$ obeys the inequality

$$|c| \leq \frac{(n!)^2}{(2n)!(2n+1)!}, \tag{5.13}$$

we use **relative order stars** (cf. section 4.3). Thus, we redefine ρ as

$$\rho(\zeta) := \frac{w(\zeta)}{R_{n/n}(\pi(\zeta))}, \quad \zeta \in \mathcal{M},$$

with $\mathcal{A}_+^R$, $\mathcal{A}_-^R$ and $\mathcal{A}_0^R$ amended in a similar vain. Here, as elsewhere, $R_{n/n}$ is the n/n Padé approximant to $\exp z$. Since $|R_{n/n}(it)| \equiv 1$ for all $t \in \mathcal{R}$, Proposition 5.12 remains true, as does Proposition 5.11. As far as Proposition 5.10 is concerned, its first part is valid: $\iota(0) = 2n+1$ on the principal branch. However, infinity is no longer an essential singularity, hence all $\mathcal{A}_\pm^R$-regions are analytic.

The Padé approximant $R_{n/n}$ is A-acceptable, hence all its poles are in the right half-plane. Furthermore, it is symmetric with respect to $i\mathcal{R}$, hence all its zeros reside in $\mathcal{C}^-$. It follows that all the poles of ρ in $\pi^{-1}(\mathcal{C}^+)$ are just the zeros of Q_N and they must 'account' for all the sectors of $\mathcal{A}_+^R$ that

[9] We are not discussing an empty set! According to a previous remark, the n/n Obrechkoff methods are A-stable and of order $2n$.

adjoin the origin. Note that some of these sectors may belong to unbounded $\mathcal{A}_+^R$-regions, but, ∞ not being an essential singularity, this does not interfere with our count.

Recall that $(-1)^{n+1}c > 0$. On the other hand, revisiting the proof of Theorem 5.14, it is clear that replacing the exponential with $R_{n/n}$ in the definition of ρ does not interfere with the conclusion that, for sufficiently small ε,

$$(0,\varepsilon) \in \begin{cases} \mathcal{A}_-^R & :n \text{ is even}, \\ \mathcal{A}_+^R & :n \text{ is odd}. \end{cases}$$

Therefore, in the case of even n, $c < 0$ and, for $x \in (0,\varepsilon)$,

$$w_1(x) < R_{n/n}(x).$$

But

$$\begin{aligned} w_1(x) &= e^x + cx^{2n+1} + \mathcal{O}(|x|^{2n+2}), \\ R_{n/n}(x) &= e^x - \frac{(n!)^2}{(2n)!(2n+1)!}x^{2n+1} + \mathcal{O}(|x|^{2n+2}). \end{aligned}$$

Consequently

$$c \le -\frac{(n!)^2}{(2n)!(2n+1)!},$$

with equality possible *only* when (5.11) is itself the Obrechkoff method corresponding to the n/n Padé approximant.[10]

Similar reasoning is valid for odd n and we obtain

$$\frac{(n!)^2}{(2n)!(2n+1)!} \le c,$$

with sharp inequality unless (5.11) is an Obrechkoff method.

Theorem 5.15 (Wanner *et. al.*, 1978) The error constant c of any A-stable n-derivative method (5.11) of order $2n$ obeys the inequality (5.13), with equality possible only in the case of the n/n Obrechkoff method. □

This completes the proof of the Daniel–Moore conjecture. We conclude that, as far as the order and the size of the error constant are concerned, multistep multiderivative methods do not confer any advantages over the one-step Obrechkoff methods (3.10) within the realm of A-stable schemes. Having said that, it should be emphasized that there is much more to the numerical solution of stiff differential equations than just A-stability. Multistep methods, even with inferior stability characteristics, are of great use and feature prominently in many state-of-the art numerical packages.

[10] In that case the order star degenerates and $\mathrm{cl}\,\mathcal{C} = \mathcal{A}_0^R$ – but this is precisely the one case when order stars are not required!

Alternatives to A-stability abound and they all circumvent the second Dahlquist barrier and its generalization. A particularly useful concept is that of $A(\alpha)$-stability: we say that the method (5.1) (or, for that matter, (5.11)) is $A(\alpha)$-**stable** if it solves stably the linear test equation $y' = \lambda y$, $y(0) = 1$, with unit step length, for all λ in the wedge-shaped domain

$$\{z \in \mathcal{C} : z \neq 0, \ |\arg(-z)| < \alpha\}.$$

Of course, $A(\pi/2)$-stability is nothing else but A-stability, but decreasing $\alpha \geq 0$ allows for relaxed stability requirements. These are amply sufficient for the numerical integration of stiff differential systems in the absence of oscillating components.

Widlund (1967) proved that for every $\alpha \in [0, \pi/2)$ there exists an $A(\alpha)$-stable method (5.1) of order N for $N \leq 4$. This, quite clearly, is an improvement over the barrier of Theorem 5.14 and methods of this form can be useful. Unfortunately, as $\alpha \uparrow \pi/2$, the coefficients of these methods become unbounded. Further improvement to the ratio of order *versus* the number of steps is barred by another result of Widlund (1967), namely that no $A(0)$-stable method (5.1), except for the trapezoidal rule, may possess an order exceeding N.

$A(\alpha)$-stability (or, consistently with our usage, $A(\alpha)$-acceptability) can be also used in the context of m/n Padé approximants to $\exp z$. Although, according to the corollary to Theorem 4.1, these are A-acceptable only for $n - 2 \leq m \leq n$ (the first Ehle barrier), it is easy to exploit their explicit form (3.12–13) to deduce $A(0)$-acceptability for all $m \leq n$. Consequently, for each $m \leq n - 3$ there exists a maximal $\alpha_{m/n} < \frac{1}{2}\pi$ such that the m/n Padé approximant to $\exp z$ is $A(\alpha_{m/n})$-acceptable. We return to this point in Chapter 10, where we describe a conjecture on the value of $\alpha_{m/n}$.

5.4 The Jeltsch–Nevanlinna comparison theorem

How to compare linear stability domains of explicit methods? Can we do better stability-wise with some methods, not with others? To set the stage we must decide on the rules of the game, and it is quite obvious, with little reflection, that simply comparing stability domains is wrong. An n-stage method 'costs' n function evaluations per step, and it is necessary to offset this by scaling the set $\mathcal{S}$.

The simplest comparison theorem is due, like all other results in that area, to the collaborative effort of Jeltsch and Nevanlinna (1981, 1982): Let $\mathcal{S}$ be the stability domain of an explicit, n-stage, single-derivative, (possibly) multistep, irreducible method. Then $\mathcal{S}$ may not properly include the disc

$$\mathcal{S}_n^E := \left\{ z \in \mathcal{C} : \left| 1 + \frac{z}{n} \right| < 1 \right\}.$$

The disc $\mathcal{S}_n^E$ is obtained when the forward Euler method is applied s times with the step h/n. The latter can be written as an n-stage explicit Runge–Kutta scheme

$$\begin{aligned}
\mathbf{k}_1 &= \mathbf{f}(a, \mathbf{y}_0), \\
\mathbf{k}_i &= \mathbf{f}\left(a + \frac{(i-1)h}{n}, \mathbf{y}_0 + \frac{h}{n}\sum_{j=1}^{i-1} \mathbf{k}_j\right), \quad i = 2, 3, \ldots, n, \\
\mathbf{y}_1 &= \mathbf{y}_0 + h\mathbf{k}_n.
\end{aligned}$$

In other words, as long as the goal is to maximize the radius of a circle (symmetric with respect to the real axis and passing through the origin) inside $\mathcal{S}$, the best choice is also the simplest – forward Euler.

Insofar as the last paragraph seems to imply that the forward Euler scheme optimizes the stability/work ratio, this impression is grossly misleading, since a much more interesting state of affairs prevails: no method is wholly 'better' then any other method. Specifically, consider any two distinct explicit methods (multistep, multiderivative, multistage *etc.*) of non-trivial orders of accuracy, that produce an algebraic stability functions (cf. section 5.2) $w^{(i)}$, $i = 1, 2$, with

$$w^{(i)}(z) = c_i z^n + \mathcal{O}(|z|^{n-1}), \quad c_i \neq 0, \; i = 1, 2 \quad |z| \to \infty. \tag{5.14}$$

The number s is, of course, related to the number of stages or derivatives in the underlying schemes. We denote by $\mathcal{S}^{(i)}$, $i = 1, 2$, the linear stability domains and note that, the scheme being explicit, they are bounded. By its definition, the boundary of $\mathcal{S}^{(i)}$ is characterized by

$$\max\{|w_j^{(i)}(z)| \; : \; j = 1, 2, \ldots, n\} \equiv 1.$$

We assume that this is determined by just one branch. In other words, there exist branches $w_e^{(i)}$, $i = 1, 2$, such that

$$z \in \partial\mathcal{S}^{(i)} \quad \Longrightarrow \quad |w_e^{(i)}(z)| = 1.$$

We stipulate that each $w_e^{(i)}$ is analytic (that is, has no branch points) in the open set $\mathcal{Q}^{(i)} := \mathcal{C} \setminus \mathrm{cl}\,\mathcal{S}^{(i)}$, $i = 1, 2$, and that it has an nth-fold pole at infinity. A method that obeys these conditions is said to satisfy **property** C.

Proposition 5.16 The branch $w_e^{(i)}$ can be identified in the vicinity of the origin with the principal branch.

Proof. An immediate consequence of $0 \in \partial\mathcal{S}^{(i)}$. □

We are fully justified to abandon the notation $w_e^{(i)}$ for $w_1^{(i)}$ and note that property C is all about the principal branch.

Proposition 5.17 For $i = 1,2$ and all $z \in \mathcal{Q}^{(i)}$ it is true that $|w_1(z)| > 1$ and $|w_j(z)| < 1$ for $j = 2,3,\ldots,n$.

Proof. First we note that the analyticity of $w_1^{(i)}$ implies that it is disconnected from the remaining branches in $\mathcal{Q}^{(i)}$. In particular, by (5.14), $w_2^{(i)},\ldots,w_n^{(i)}$ are disconnected from the pole at ∞. It follows at once from the maximum principle on Riemann surfaces and from the definition of $\mathcal{Q}^{(i)}$ that $|w_j| < 1$, $j = 2,3,\ldots,n$. This, in turn, implies that $|w_1| > 1$ in $\mathcal{Q}^{(i)}$: For suppose that $|w_1(z)| \leq 1$ for some $z \in \mathcal{Q}^{(i)}$. Then the moduli of all the branches are bounded by unity there and $z \in \mathrm{cl}\,\mathcal{S}^{(i)}$. Since w_1 cannot be constant in $\mathcal{Q}^{(i)}$ (recall the pole at ∞), this leads to a contradiction. □

The importance of the last two propositions goes well beyond the technical. Essentially, it means that the geometry of $\mathcal{S}^{(i)}$ is determined by just one branch. Of course, inside the stability domain the distinction between the branches is ultimately lost, because irreducibility implies that no sheet of the Riemann surface can be disconnected throughout $\mathrm{cl}\,\mathcal{C}$, but this has no serious consequences to the task in hand.

Theorem 5.18 (Jeltsch and Nevanlinna, 1981) The stability domains $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$ are mutually non-inclusive, i.e.

$$\mathcal{S}^{(1)} \not\subset \mathcal{S}^{(2)}, \qquad \mathcal{S}^{(2)} \not\subset \mathcal{S}^{(1)}.$$

Proof. We plan to prove that the assumption $\mathcal{S}^{(2)} \subset \mathcal{S}^{(1)}$ leads to a contradiction. To this end we construct the relative order star with respect to

$$\rho(\zeta) = \frac{w^{(1)}(\zeta)}{w^{(2)}(\zeta)}, \quad \zeta \in \mathcal{M}.$$

Given $z \in \partial\mathcal{S}^{(2)}$, it follows at once that $\pi^{-1}(z) \in \mathcal{A}_-^R$ and it is easy to verify that

$$\mathcal{U} := \mathcal{S}^{(1)} \setminus \mathcal{S}^{(2)} \subset \mathcal{A}_-^R$$

and that

$$\zeta \in \mathcal{U}_0 := \partial\mathcal{S}^{(1)} \cap \partial\mathcal{U} \qquad \Longrightarrow \qquad \zeta \in \mathcal{A}_-^R \cap \mathcal{A}_0^R.$$

Hence, for all ζ above, along the branch $w_1^{(1)}$ we have $|w^{(2)}| \geq 1$ and it follows from Proposition 5.17 that the principal branches coincide for both methods. Moreover, since $|w_1^{(2)}| > 1$, we can deduce that, for sufficiently small $\varepsilon > 0$, there exists an open set $\mathcal{U}_\varepsilon$ such that

$$\mathcal{U}_0 \subset \mathcal{U}_\varepsilon, \qquad \mathrm{dist}\,(\mathcal{U}_0, \partial\mathcal{U}_\varepsilon) > \varepsilon, \qquad \mathcal{U}_\varepsilon \subset \mathcal{A}_-^R. \tag{5.15}$$

Both methods being consistent, it is necessarily true that 0 lies on $\partial\mathcal{S}^{(1)} \cap \partial\mathcal{S}^{(2)}$ and that for every sufficiently small $\varepsilon > 0$ there exists $r_\varepsilon > 0$ such that

$$r_\varepsilon e^{i\theta} \in \mathcal{C} \setminus \mathcal{S}^{(1)}, \quad |\theta| < \frac{\pi}{2} - \varepsilon,$$

and $\lim_{\varepsilon \downarrow 0} r_\varepsilon = 0$. Consequently, by virtue of (5.15), it is true that this arc belongs to $\mathcal{A}_-^R$ for sufficiently small ε. In other words, the origin is adjoined by just a single sector of $\mathcal{A}_-^R$ and no sectors of $\mathcal{A}_+^R$ to the right of $i\mathcal{R}$ along the principal branch. In other words, $\iota(0) \leq 1$ and it follows from Proposition 5.10 that at least one of the methods cannot be of non-trivial order. This contradiction to our original assumptions affirms that it is impossible for $\mathcal{S}^{(2)}$ to be included wholly in $\mathcal{S}^{(1)}$, or the other way round. □

Like children in a progressive classroom, no method is 'better' or 'worse' than any other method. Each possesses a unique set of excellence – its very own stability domain.

Propert C is not unduly restrictive – Jeltsch and Nevanlinna (1981) prove that it is obeyed by most methods of interest. However, interesting questions remain and we return to this topic in Chapter 10, in our exposition of open problems.

6

The advection equation

But when the vigilant patrol
Of stars walks round about the pole,
Their leaves, that to the stalks are curl'd,
Seem to theirs staves the ensigns furl'd.

From *A Garden* by Andrew Marvell (1621–1678).

6.1 Order and Stability Conditions

The **advection equation** reads

$$\frac{\partial}{\partial t}u = \frac{\partial}{\partial x}u, \tag{6.1}$$

where $u = u(x,t)$ is given either along the whole real line $-\infty < x < \infty$, $t = 0$ (the **Cauchy problem**) or in an interval (a,b), $b < \infty$, and along the line $x = b$, $t \geq 0$ (the **initial-boundary-value problem**). It acts as a convenient model for a whole range of important nonlinear hyperbolic equations: the **conservation laws**

$$\frac{\partial}{\partial t}u = \operatorname{div} f(u),$$

where $u = u(\mathbf{x},t)$ is defined for appropriate initial and boundary data (note that $\mathbf{x}$ can be a vector) (Lax, 1973). Conservation laws are common in mathematical modelling of compressible flow, e.g. the inviscid Burgers' equation, the Euler equations of compressible inviscid flow etc., as well as in the study of solitons. Hence the practical importance of being able to approximate the solution of (6.1) well by a computational procedure.

The advection equation is seemingly the simplest partial differential equation imaginable, since it corresponds to a unilateral shift: thus, the exact solution of the Cauchy problem is $u(x,t) = u(x+t,0)$. This simplicity is highly deceptive from the numerical point of view. To underpin the last sentence we should note that the conservation of certain invariants, as implied in the name 'conservation laws', is not an unreserved blessing

and is achieved at a price: discontinuities and rarefaction fans (Lax, 1973). Specifically – and this is already apparent from the unilateral shift associated with (6.1) – the characteristic curves traverse the (x,t) plane diagonally and the solution favours a particular direction. Hence it makes sense that the numerical solution should have a built-in asymmetry between 'left' and 'right'.

The discontinuities that arise in the solution of nonlinear conservation laws are **shocks**, i.e. they are impenetrable to the flow of information. Obviously, this ought to be expressed in a numerical procedure: a discretized equation to determine the value at a point $(\ell\Delta x,(n+1)\Delta t)$ should not rely on information from grid points $(j\Delta x,n\Delta t)$ that are separated by a shock. This, again, forces a directional slant (occasionally termed **upwinding**).

The theme of the present chapter is the interplay between order, stability and upwinding. We demonstrate that stability, in conjunction with high order, leads to upwinding of specific and predictable form. *Mutatis mutandis*, stability and given 'slant' restrict the order.

Numerical methods for partial differential equations of evolution come in two basic varieties: **semi-discretizations** and **full discretizations**. In a semi-discretized scheme we first replace the spatial derivatives by finite differences.[1] This yields a system of ordinary differential equations (ODEs) which can be solved by any of a large choice of well-understood methods. On the other hand, full discretization (as the name implies) replaces simultaneously both space and time derivatives. Note that, of course, each combination of semi-discretization and an ODE solver results in a full discretization.[2] None the less, semi-discretizations merit special attention, since they typically produce schemes that are easier to analyse.

We commence by analysing the Cauchy problem for (6.1) by semi-discretized finite differences (Example 1.3 already anticipated, without proving, some of our results). We assume that the initial condition $u(x,0)$ is an $L_2(-\infty,\infty)$ function and that it is sufficiently differentiable for our purposes. Let $u_\ell(t)$ be an approximant to $u(\ell\Delta x,t)$, $\ell\in\mathcal{Z}$, where $\Delta x>0$ is the space discretization parameter. We endeavour to approximate the operator $\partial/\partial x$ at $(\ell\Delta x,t)$ by using the u_j's. A finite difference scheme of this form and of some generality reads

$$\sum_{j=-\bar{r}}^{\bar{s}}\alpha_j u'_{\ell+j}=\frac{1}{\Delta x}\sum_{j=-r}^{s}\beta_j u_{\ell+j},\quad \ell\in\mathcal{Z}. \tag{6.2}$$

Note that the method (1.3) corresponds to $\bar{r}=\bar{s}=0$ – an **explicit** scheme.

[1] Or by a Galerkin-type finite element approximation, a pseudospectral scheme etc.

[2] The converse is false: not every full discretization can be 'factorized' meaningfully into a semi-discretization and an application of an ODE solver. An example will be provided in Chapter 7.

Given any sequence $\mathbf{g} = \{g_\ell\}_{\ell=-\infty}^{\infty}$ which is ℓ_2 (that is, $\sum_{\ell=-\infty}^{\infty} |g_\ell|^2 < \infty$), we define the **Fourier transform** $\mathcal{F}_{\Delta x}$ by the standard formula

$$\hat{g}(\theta) = \{\mathcal{F}_{\Delta x}\mathbf{g}\}(\theta) := \sum_{\ell=-\infty}^{\infty} e^{i\ell\theta\Delta x} g_\ell, \quad |\theta| \le \frac{\pi}{\Delta x}. \tag{6.3}$$

Thus, $\mathcal{F}_{\Delta x}$ maps bi-infinite ℓ_2 sequences into functions acting on the interval $(-\pi/\Delta x, \pi/\Delta x)$. Moreover, defining the 2-norms in the underlying spaces in the usual way,

$$\begin{aligned} \|\mathbf{g}\| &:= \left\{ \sum_{\ell=-\infty}^{\infty} |g_\ell|^2 \right\}^{\frac{1}{2}}, \\ |\hat{g}|_{\Delta x} &:= \left\{ \frac{\Delta x}{2\pi} \int_{-\pi/\Delta x}^{\pi/\Delta x} |\hat{g}(\theta)|^2 \mathrm{d}\theta \right\}^{\frac{1}{2}}, \end{aligned}$$

we readily verify that the Fourier transform is an isometry:

$$\|\mathbf{g}\| = |\mathcal{F}_{\Delta x}\mathbf{g}|_{\Delta x}. \tag{6.4}$$

We reserve the notation $\hat{u}$, $\hat{U}$ and $\hat{e}$ for the Fourier transforms of the sequences $\{u(\ell\Delta x, t)\}$, $\{u_\ell(t)\}$ and $\{u_\ell(t) - u(\ell\Delta x, t)\}$ respectively. All three are, of course, functions of t. It is a trivial consequence of (6.1), via integration by parts, that

$$\frac{\partial}{\partial t}\hat{u} = -i\theta\hat{u}, \tag{6.5}$$

whereas multiplying (6.2) by $e^{i\ell\theta\Delta x}$ for each $\ell \in \mathcal{Z}$, summing up and exchanging the order of summation speedily produces

$$\frac{\partial}{\partial t}\hat{U} = \frac{1}{\Delta x} h(e^{-i\theta\Delta x})\hat{U}, \tag{6.6}$$

where the **characteristic function** h is defined by

$$h(z) := \frac{\sum_{j=-r}^{s} \beta_j z^j}{\sum_{j=-\bar{r}}^{\bar{s}} \alpha_j z^j}.$$

Subtracting (6.5) from (6.6) and conjugation result in the ordinary differential equation

$$\frac{\partial}{\partial t}\hat{e} = \frac{1}{\Delta x} h(e^{i\theta\Delta x})\hat{e} + \frac{1}{\Delta x}\left(h(e^{i\theta\Delta x}) - i\theta\Delta x\right)\hat{u}, \tag{6.7}$$

which is valid for every $|\theta| \le \pi/\Delta x$ with the initial condition $\hat{e}(\theta, 0) = 0$ (since $u_\ell(0) = u(\ell\Delta x, 0)$, the given initial condition). Solving (6.7) yields

$$\hat{e} = \frac{1}{\Delta x}\left(h(e^{i\theta\Delta x}) - i\theta\Delta x\right) \int_0^t \exp\left(\frac{t-\tau}{\Delta x} h(e^{i\theta\Delta x})\right) \hat{u}(\theta, \tau)\mathrm{d}\tau. \tag{6.8}$$

We say that the characteristic function h is **good** if it is analytic on the complex unit circle. In other words, the denominator of the rational function does not vanish there. Clearly, we need to stipulate that h is good, otherwise (6.7) becomes meaningless for some values of θ.

The **approximation order** of h is the integer p such that

$$h(z) = \log z + c(z-1)^{p+1} + \mathcal{O}(|z-1|^{p+2}), \quad c \neq 0, \quad z \to 1.$$

Since h is good, we always have $p \geq -1$. Herewith we assume that the order is non-trivial, i.e. that $p \geq 1$. Since

$$\frac{h(e^{i\theta\Delta x}) - i\theta\Delta x}{\Delta x} = O((\Delta x)^p),$$

the equation (6.8) implies that

$$\hat{e} = O(t(\Delta x)^p).$$

We now exploit the isometry property (6.4) of the Fourier transform to argue that

$$\|e(t)\| = |\hat{e}(\cdot, t)|_{\Delta x} = O(t(\Delta x)^p).$$

Thus, p is also the **order of accuracy** of the numerical scheme (6.2).

Fourier transforms, which were so helpful in analysing the order of the semi-discretization (6.2), are equally central to the study of its stability. Let $\mathbf{u}$ be the bi-infinite vector of the u_ℓ's. Since $p \geq 1$ and consequent on the celebrated theorem of Lax (Richtmyer and Morton, 1967), convergence of (6.2) (for the Cauchy problem) to the correct solution of (6.1) is equivalent to the uniform boundedness of $\|\mathbf{u}\|$ for all t in a compact interval as $\Delta x \downarrow 0$. This provides the right definition of **stability** in the present framework.

Exploiting again the isometry property (6.4), we deduce that stability is equivalent to the uniform boundedness of $|\hat{U}|_{\Delta x}$ for the requisite range of t and Δx.

Lemma 6.1 The semi-discretization (6.2) is of order of accuracy p if and only if h approximates $\log z$ at $z = 1$ up to order p, and it is stable for the Cauchy problem if and only if $\operatorname{Re} h(e^{i\theta}) \leq 0$ for all $-\pi \leq \theta \leq \pi$.

Proof. It remains to verify the given condition for stability. This can be easily accomplished by deriving the explicit form of $\hat{U}$ from the ordinary differential equation (6.6): we have

$$\hat{U}(\theta, t) = \exp\left(\frac{t}{\Delta x} h(e^{-i\theta\Delta x})\right) \hat{U}(\theta, 0)$$

and, clearly, uniform boundedness for all $\Delta x \downarrow 0$ is equivalent to $\operatorname{Re} h(z) \leq 0$ for all $|z| = 1$. □

Example 6.1 The choice $\bar{r} = \bar{s} = 0$ yields explicit schemes. Since h is good, we may set, without loss of generality, $\alpha_0 = 1$. We choose the coefficients $\beta_{-r}, \ldots, \beta_s$ so as to maximize the order of accuracy. By Lemma 6.1 this is equivalent to maximizing p in

$$h(z) = \sum_{j=-r}^{s} \beta_j z^j = \log z + \mathcal{O}(|z-1|^{p+1}), \quad z \to 1.$$

We assert that, for fixed r and s, the maximum is attained for $p = r + s$ and the coefficients

$$\begin{aligned} \beta_j &= = \frac{(-1)^{j+1}}{j} \frac{r!s!}{(r+j)!(s-j)!}, \quad j = -r, \ldots, s, \; j \neq 0; \\ \beta_0 &= \begin{cases} -\sum_{i=r+1}^{s} \frac{1}{i} & : s \geq r+1, \\ 0 & : s = r, \\ \sum_{i=s+1}^{r} \frac{1}{i} & : r \geq s+1 \end{cases} \end{aligned} \tag{6.9}$$

The proof is quite straightforward and is based on the original meaning of the function h: let the function values $f_{\ell+j}$ be given at $(\ell + j)\Delta x$ for $j = -r, \ldots, s$. They can be interpolated in a unique manner by an $(r+s)$-degree polynomial with the Lagrange formula

$$q(x) = \sum_{j=-r}^{s} f_{\ell+j} \prod_{\substack{i=-r \\ i \neq j}}^{s} \frac{x - (\ell+i)\Delta x}{(j-i)\Delta x}.$$

Thus, the best approximation to the derivative at $\ell \Delta x$ is

$$q'(\ell \Delta x) = \sum_{j=-r}^{s} \beta_j f_{\ell+j}.$$

The values (6.9) follow easily, as does the order of approximation. Finally, the uniqueness of Lagrange interpolants proves that order $p = r + s$ cannot be exceeded by any choice of the coefficients. In our terminology, the choice (6.9) leads to **interpolatory methods** $h^{(r,s)}$ (Iserles, 1982).

We examine the stability (for Cauchy problems) of $h^{(r,s)}$. Clearly, by symmetry, $\operatorname{Re} h^{(r,r)}(e^{i\theta}) \equiv 0$ and, by Lemma 6.1, the method is stable. To examine some other instances of r and s we first note that (6.9) implies

$$h^{(r,s+1)}(z) = h^{(r,s)}(z) + (-1)^{r+1} \frac{r!s!}{(r+s+1)!} z^{-r}(1-z)^{r+s+1}. \tag{6.10}$$

Since $e^{-ir\theta}(1-e^{i\theta})^{2r} = (-1)^r|1-e^{i\theta}|^{2r} = (-1)^r 2^r(1-\cos\theta)^r$, we have

$$\begin{aligned}\operatorname{Re} h^{(r,r+1)}(e^{i\theta}) &= \operatorname{Re} h^{(r,r)}(e^{i\theta}) \\ &\quad + (-1)^{r+1}\frac{(r!)^2}{(2r+1)!}\operatorname{Re} e^{-ir\theta}(1-e^{i\theta})^{2r+1} \\ &= -\frac{(r!)^2}{(2r+1)!}2^r(1-\cos\theta)^{r+1} \le 0\end{aligned}$$

for all $|\theta| \le \pi$. Stability follows by Lemma 6.1. Likewise, exploiting (6.10) produces

$$\begin{aligned}\operatorname{Re} h^{(r,r+2)}(e^{i\theta}) &= \operatorname{Re} h^{(r,r+1)}(e^{i\theta}) \\ &\quad + (-1)^{r+1}\frac{r!(r+1)!}{(2r+2)!}\operatorname{Re} e^{-ir\theta}(1-e^{i\theta})^{2r+2} \\ &= -\frac{(r!)^2}{(2r+1)!}2^r(1-\cos\theta)^{r+2}.\end{aligned}$$

Hence stability for $s = r+2$. It will be proved in Section 6.3 that the conditon $r \le s \le r+2$ characterizes all the stable choices of r and s. ◇

Analysis of full discretizations is very similar to the study of semi-discretized schemes, except that ordinary differential equations are replaced by algebraic equations and instead of time t we have a **Courant number** $\mu := \Delta t/\Delta x$. A general scheme reads

$$\sum_{j=-\bar{r}}^{\bar{s}} \gamma_j(\mu) u_{\ell+j}^{k+1} = \sum_{j=-r}^{s} \delta_j(\mu) u_{\ell+j}^{k}, \tag{6.11}$$

where $u_\ell^j \approx u(\ell\Delta x, k\Delta t)$. We set the characteristic function

$$H(z,\mu) := \frac{\sum_{j=-r}^{s}\delta_j(\mu)z^j}{\sum_{j=-\bar{r}}^{\bar{s}}\gamma_j(\mu)z^j}.$$

Fourier analysis leads to an approximation-theoretical characterization of order and stability conditions. Again, we need to stipulate that the characteristic function is good, i.e. that it has no poles on the complex unit circle.

Lemma 6.2 Let a Courant number $\mu > 0$ be given. The full discretization (6.11) is of order of accuracy p if and only if $H(z,\mu) = z^\mu + \mathcal{O}(|z-1|^{p+1})$ for $z \to 1$. Moreover, it is stable for the Cauchy problem if and only if $|H(e^{i\theta},\mu)| \le 1$ for all $-\pi \le \theta \le \pi$. □

Example 6.2 Let $r = \bar{r} = \bar{s} = 0$, $s = 1$ and

$$u_\ell^{k+1} = (1-\mu)u_\ell^k + \mu u_{\ell+1}^k. \tag{6.12}$$

It is easy to obtain this method by discretizing with forward Euler (Richtmyer and Morton, 1967) in both space and time. We have

$$H(z,\mu) = 1 - \mu + \mu z = z^{\mu} + \frac{1}{2}\mu(1-\mu)(1-z)^2 + \cdots, \quad z \to 1,$$

hence $p = 1$. Moreover,

$$|H(e^{i\theta},\mu)|^2 = 1 - 2\mu(1-\mu)(1-\cos\theta)$$

and Lemma 6.2 implies stability for all $\mu \in (0,1)$ and instability for $\mu > 1$.

The last result seems quite trivial – indeed, it *is* quite trivial! However, it serves to emphasize that considerable care is needed in stability analysis. Many students of numerical mathematics gain an impression that the 'correct' approach to stability is by deriving (or estimating) the spectral radius of the underlying iteration matrix. This happens to be entirely right as long as the underlying matrix is symmetric (or, in general, normal, i.e. it commutes with its adjoint matrix A^T or A^*), but is false otherwise.[3] The iteration matrix of (6.12) is bi-infinite, of the form $A = (a_{i,j})_{i,j=-\infty}^{\infty}$, where $a_{i,i} = 1-\mu$, $a_{i,i+1} = \mu$ and $a_{i,j} = 0$ for $j \neq i, i+1$. A hand-waving argument might read: '...the eigenvalues of any finite-dimensional section of A are all at $1-\mu$. Hence stability is equivalent to $|1-\mu| \leq 1$, thus $0 < \mu \leq 2$...' This, of course, is a wrong result, based on a wrong argument. Firstly, the infinite-dimensional operator A has no eigenvalues at all and $\sigma(A)$ consists solely of a continuous spectrum. Secondly, it is the ℓ_2 norm of A and its powers that plays a crucial role in stability analysis (the formal definition of Lax stability being that $\|A\|^k$ is uniformly bounded for all k and $\Delta x \downarrow 0$), not its spectral radius. Were A normal, the norm and the spectral radius would have coincided. This, of course, is not the case with (6.12). $\diamond$

Example 6.3 Similarly to Example 6.1, we consider explicit full discretizations. Thus, without loss of generality, $\gamma_0(\mu) \equiv 1$ and

$$H(z,\mu) = \sum_{j=-r}^{s} \delta_j(\mu) z^j.$$

Multiplying both sides by z^r and shifting the index of summation yields

$$\sum_{j=0}^{r+s} \delta_{j-r}(\mu) = z^{\mu+r} + \mathcal{O}(|z-1|^{p+1}), \tag{6.13}$$

[3] Unfortunately, the staple fare of undergraduate lecture courses in numerical solution of partial differential equations of evolution typically consists of the diffusion and the wave equations. Both lead to normal iteration matrices and eigenvalue analysis is justified. This might lead to a woefully wrong impression!

p being the order of accuracy of the method. We presently require the concept of the **factorial symbol** (also known as the Pochhammer symbol): $(x)_0 := 1$ and $(x)_m = x(x+1)\cdots(x+m-1) = (x)_{m-1}(x+m-1)$ for $m \geq 1$. The Taylor expansion of the binomial function is

$$(1-\zeta)^\kappa = \sum_{j=0}^{\infty} \frac{(-\kappa)_j}{j!}\zeta^j.$$

Comparison with (6.13) demonstrates that, in general, $p \leq r+s$ (except for specific integer Courant numbers) and that the bound is attained when

$$\sum_{j=0}^{r+s} \delta_{j-r}(\mu) z^j = \sum_{j=0}^{r+s} \frac{(-\mu-r)_j}{j!}(1-z)^j.$$

Differentiating the last expression k times yields

$$\sum_{j=k}^{r+s} \frac{j!}{(j-k)!}\delta_{j-r}(\mu) z^{j-k} = (-1)^k \sum_{j=k}^{r+s} \frac{(-\mu-r)_j}{(j-k)!}(1-z)^{j-k}.$$

Substitution of $z = 0$ produces

$$\begin{aligned}\delta_{k-r}(\mu) &= \frac{(-1)^k}{k!} \sum_{j=0}^{r+s-k} \frac{(-\mu-r)_{j+k}}{j!} \\ &= \frac{(-1)^k(-\mu-r)_k}{k!} \sum_{j=0}^{r+s-k} \frac{(-\mu-r+k)_j}{j!}.\end{aligned}$$

It is a trivial matter to prove (for example, by induction) that

$$\sum_{j=0}^{m} \frac{(\kappa)_j}{j!} = \frac{(\kappa+1)_m}{m!}, \quad \kappa \in \mathcal{C}.$$

This, with a little further manipulation, ascertains that

$$\delta_j(\mu) = \frac{(-1)^{r+1+j}}{\mu-j} \frac{(-\mu-r)_{r+s+1}}{(r+j)!(s-j)!}, \quad j = -r,\ldots,s. \tag{6.14}$$

Formula (6.14) has been derived by Iserles and Strang (1983) by different means.

We denote the highest-order method by $H^{(r,s)}$. Note, as a matter of interest, that if μ equals an integer L in $\{-r, -r+1, \ldots, s\}$ then $H^{(r,s)}(z, L) = z^L$. It will be demonstrated in Section 6.4 that $H^{(r,r)}$, $H^{(r,r+1)}$ and $H^{(r,r+2)}$ are stable for all $\mu \in (0,1)$ – this parallels the results on the explicit semi-discretizations $h^{(r,s)}$ in Example 6.1. Moreover, in Section 6.5 we prove

that these are all the choices of r and s that lead to stability for any μ in $(0,1)$. $\diamond$

The similarity between the stability results for $H^{(r,s)}$ and $h^{(r,s)}$ is not a matter of coincidence. We say that a full discretization is **canonical** if its characteristic function, which we denote by H, is smoothly differentiable with respect to μ for $|\mu| \downarrow 0$ and $H(z,0) \equiv 1$. The last two conditions do not restrict generality to any significant extent: the coefficients of H are typically rational functions of μ and, moreover, $\mu = 0$ corresponds to $\Delta t = 0$ and it is only natural to expect that $H(z,0)$ reduces to unity. Canonical full discretizations can be associated with certain semi-discretized schemes: set

$$h(z) = \left.\frac{\partial H(z,\mu)}{\partial \mu}\right|_{\mu=0}.$$

We say that h is **associated** with H.

Lemma 6.3 Let a full discretization with the characteristic function H be canonical and let h be associated with H. Then:

(a) Provided that $H(z,\mu) = \sum_{-r}^{s} \delta_j z^j / \sum_{-\bar{r}}^{\bar{s}} \gamma_j z^j$, the function h is of the form $\sum_{-\max\{r,\bar{r}\}}^{\max\{s,\bar{s}\}} \beta_j z^j / \sum_{-\bar{r}}^{\bar{s}} \alpha_j z^j$.

(b) If H is of order p then $h(z) = \log z + \mathcal{O}(|z-1|^{q+1})$ for some $q \geq p$ and the underlying semi-discretization is of order q.

(c) If both H and h are good then stability of the full discretization for $\mu \downarrow 0$ implies stability of the associated semi-discretization.

Proof. Let $H(z,\mu) = P(z,\mu)/Q(z,\mu)$, where P and Q are Laurent polynomials (in z). Since the method is canonical, we have $P(z,0) \equiv Q(z,0)$, thus

$$h(z) = \frac{\frac{\partial}{\partial \mu}\left(P(z,0) - Q(z,0)\right)}{Q(z,0)}.$$

This proves (a). Furthermore,

$$H(z,\mu) = z^{\mu} + c(\mu)(z-1)^{p+1} + \mathcal{O}(|z-1|^{p+2})$$

implies that

$$h(z) = \log z + c'(0)(z-1)^{p+1} + \mathcal{O}(|z-1|^{p+2})$$

and the order of h is at least p (it may increase if $c'(0) = 0$). Finally, let us assume that H is stable for all μ in an interval of the form $(0,\varepsilon)$. Thus, by Lemma 6.2,

$$|H(e^{i\theta},\mu)| \leq 1, \quad |\theta| \leq \pi, \ 0 < \mu < \varepsilon. \tag{6.15}$$

But

$$H(e^{i\theta},\mu) = 1 + \mu h(e^{i\theta}) + \mathcal{O}(\mu^2).$$

Consequently,

$$|H(e^{i\theta},\mu)| = 1 + \mu \mathrm{Re}\, h(e^{i\theta}) + \mathcal{O}(\mu^2)$$

and comparison with (6.15), in tandem with Lemma 6.1, furnish the proof of (c). □

Example 6.4 It is straightforward to verify that $\delta_j'(0) = \beta_j$, where the $\delta_j(\mu)$'s and the β_j's were given in (6.14) and (6.9) respectively. Thus, $h^{(r,s)}$ are associated with $H^{(r,s)}$.

An example of a noncanonical full discretization is the first-order Friedrichs scheme

$$u_\ell^{k+1} = \frac{1-\mu}{2} u_{\ell-1}^k + \frac{1+\mu}{2} u_{\ell+1}^k,$$

since $H(z,0) = \frac{1}{2}(z + z^{-1})$. In this case, actually, noncanonicity spells no harm: setting $\mu = 0$ in the μ-derivative of H produces the $h^{(1,1)}$ semi-discretization. Since H is stable for all $|\mu| < 1$ and $h^{(1,1)}$ is stable and second-order, the statement of Lemma 6.3 remains correct. Note that the order of the semi-discretization exceeds that of the Friedrichs method.

The truly harmful instances of noncanonicity occur when H is implicit or when differentiability with respect to μ breaks down at $\mu = 0$. We provide examples of neither – the first instance leads to quite complicated expressions and the second is highly contrived.

As a final illustration of the subject matter of Lemma 6.3, we note that it is necessary to require in (c) that both H and h are good. For example, the second-order box scheme

$$(1+\mu)u_\ell^{k+1} + (1-\mu)u_{\ell+1}^{k+1} = (1-\mu)u_\ell^k + (1+\mu)u_{\ell+1}^k$$

is good, as well as stable for $0 < \mu < 1$, but its associated scheme,

$$u_\ell' + u_{\ell+1}' = \frac{2}{\Delta x}(-u_\ell + u_{\ell+1}),$$

fails to be good. Consequently, this second order (part (b) of the lemma remains valid!) semi-discretization is unstable. ◇

6.2 The influence of boundaries on stability

Study of stability is considerably complicated by the presence of boundaries. Fourier analysis, which was central to the work of the last section, is too weak on its own to take care of this problem.

This is the moment to remind the reader that the partial differential equation (6.1) requires just one boundary condition (imposed at the right-hand side) on a finite interval. Unfortunately, this is not the case with finite-difference equations (unless $r = \bar{r} = 0$ and $s, \bar{s} \leq 1$). Thus, we

need to augment these equations with extra non-physical boundary conditions. These 'numerical' conditions are added solely for the sake of the well-posedness of the solution (otherwise we lack the necessary data!) and the main motivation in their choice is the retention of stability. Much effort has been expanded in the last two decades on stability analysis of initial-boundary-value problems (Gustafsson *et al.*, 1972; Osher, 1969; Trefethen, 1982).

Imposition of boundaries follows three steps of increasing analytic complexity: firstly, we assume that (6.1) is given on a singly-infinite interval, $[0, \infty)$, say, with zero boundary conditions at $x = 0$; secondly, we consider the initial-boundary-value problem in a finite interval, but still with zero boundary conditions; and, finally, we allow non-zero boundary conditions. Happily, we can disregard the last two steps in our exploration of stability barriers for the schemes (6.2) and (6.11). Stability conditions for a singly-infinite interval and a finite interval are the same, subject to zero boundary conditions (Strang, 1964b). Moreover, the imposition of non-zero boundary conditions means amending some of our discretized equations – after all, imposing artificial boundary conditions is the same as using different discretization near the boundary. Stability can no longer be analysed solely in terms of the functions h and H.

Let (6.1) be given in the interval $[0, \infty)$ with zero boundary conditions at $x = 0$. We reformulate the semi-discretization (6.2) in the language of Toeplitz matrices: essentially, this formulation is equivalent to the Fourier framework, but it allows an important generalization. Our exposition is based on the work of Strang (1964b) and its extension by Iserles and Strang (1983).

A matrix $A = (A_{i,j})$ (which may be finite, singly-infinite or bi-infinite) is said to be a **Toeplitz matrix** if there exist numbers a_i such that $A_{i,j} = a_{j-i}$ for all i and j in the range. In other words, the elements of A are constant along diagonals. The set of singly-infinite Toeplitz matrices will be denoted by $\mathbf{T}$ and we will follow the convention that the lower-case letters denote elements along diagonals, whereas upper-case letters are reserved for matrices: $B = (b_{j-i})$ etc. The formal Laurent series $\mathbf{a}(z) = \sum_{k=-\infty}^{\infty} a_k z^k$ is called the **symbol** of $A \in \mathbf{T}$ and will be forthwith denoted by a bold lower-case letter.

Proposition 6.4 Let $A, B \in \mathbf{T}$, where either A is upper triangular or B is lower triangular. Then $C := AB$ is in $\mathbf{T}$ and $\mathbf{c} = \mathbf{ab}$.

Proof. We prove the proposition for a lower triangular B, since the other case follows by an identical argument. Since $b_\ell = 0$ for $\ell \geq 1$, we have

$$C_{k,\ell} = \sum_{j=0}^{\infty} A_{k,j} B_{j,\ell}$$

$$= \sum_{j=0}^{\infty} a_{j-k} b_{\ell-j} = \sum_{j=0}^{\infty} a_{\ell-k-j} b_j$$

and $C_{k,\ell}$ depends only on $\ell - k$. Thus, $C \in \mathbf{T}$ and $\mathbf{c} = \mathbf{ab}$. □

Given $C \in \mathbf{T}$, we say that $C = AB$ is a **Wiener–Hopf factorization** of C if the matrices $A, B \in \mathbf{T}$ are upper and lower triangular respectively.[4]

Let $\mathbf{c}$, the symbol of C, be analytic in an open annulus $\mathcal{U}$ that envelopes the complex unit circle. Thus, its Laurent series converges on $|z| = 1$ and we can always factorize

$$\mathbf{c}(z) = \mathbf{a}(z)\mathbf{b}(z), \quad z \in \mathcal{U},$$

where a is analytic in $|z| \leq 1$ and b is co-analytic there (i.e. analytic in $|z| \geq 1$). $\mathbf{a}$ and $\mathbf{b}$ are sometimes called the inner and the outer function respectively.

Proposition 6.5 Let $\mathbf{c} = \mathbf{ab}$, where $\mathbf{a}$ and $\mathbf{b}$ are analytic and co-analytic respectively. Then $C = AB$, where A is upper triangular with symbol $\mathbf{a}$ and B is lower triangular with symbol $\mathbf{b}$.

Proof. The proof proceeds by straightforward comparison of the product of series $\mathbf{ab}$ with the elements of C from the proof of Proposition 6.4. □

We conclude that, subject to the analyticity of $\mathbf{c}$ on $\mathcal{U}$, a Wiener–Hopf factorization always exists. Moreover, it is easy to deduce that it is unique, up to a non-zero multiplicative constant.

Let $\mathbf{a}$ be analytic in $|z| \leq 1$. Thus, its Laurent and Taylor series coincide. Moreover, $\mathbf{a}^{-1}$ is analytic in $\mathcal{U}$ if and only if it is analytic in $|z| \leq 1$, and that, in turn, is equivalent to the absence of zeros of $\mathbf{a}$ there. In other words, all the zeros of $\mathbf{a}$ are banished to $|z| > 1$. Moreover, since $\mathbf{a}^{-1}$ is analytic in $|z| \leq 1$, A^{-1} must be upper triangular. Likewise, for $\mathbf{b}^{-1}$ to be analytic in $\mathcal{U}$, subject to the analyticity of $\mathbf{b}$ in $|z| \geq 1$, it is necessary and sufficient that all the zeros of $\mathbf{b}$ reside inside the unit disc. In that case B^{-1} is lower triangular.

The emphasis on the analyticity of the symbol is crucial: Let A be upper triangular. Choose $\eta \in [0,1)$ and $|\theta| \leq \pi$ and set $v_k := \eta^k e^{ik\theta}$, $k = 0,1,\ldots$. It is easy to verify that, for every $A \in \mathbf{T}$, $A\mathbf{v} = \mathbf{a}(\eta e^{i\theta})\mathbf{v}$, where $\mathbf{v} = [v_0, v_1, \ldots]^T$. Thus, and since the spectrum of an operator is a closed set,

$$\{\mathbf{a}(z) : |z| \leq 1\} \subseteq \sigma(A)$$

(actually, the two sets coincide (Krein, 1958) – it is easy to deduce this by showing that for $\zeta \in \{\mathbf{a}(z) : |z| > 1\}$ the matrix $A - \zeta I$ is invertible –

[4] Note that the order is opposite to the more familiar LU factorization of linear algebra.

but this is not necessary to our argument). However, the ℓ_2 norm of A is bounded below by the spectral radius, therefore

$$\|A\|_2 \geq \sup\{|\mathbf{a}(z)| : |z| \leq 1\}.$$

Thus, unless the power series $\mathbf{a}$ is analytic in the unit disc, $\|A\|_2$ is infinite. As a similar argument can be applied to lower triangular matrices (by using left eigenvectors: the adjoint of a lower triangular matrix is upper triangular and they share the same norm!), it is quite clear that analyticity is essential.

It should be clear by now that the ℓ_2-bounded portion of $\mathbf{T}$ and the algebra of functions analytic on an annulus are isomorphic, linked by the correspondence of a matrix in $\mathbf{T}$ to its symbol.

The next proposition follows quite easily from the preceding discussion.

Proposition 6.6 Let $C = AB$ be a Wiener–Hopf factorization. Then C^{-1} exists if and only if both A^{-1} and B^{-1} exist – and this, in turn, is equivalent to $\mathbf{a}$ being non-zero in $|z| \leq 1$ and $\mathbf{b}$ being non-zero in $|z| \geq 1$. □

The semi-discretization (6.2) can be now written as $C\mathbf{u}' = D\mathbf{u}$, where $C, D \in \mathbf{T}$ and $\mathbf{u}$ is a singly-infinite vector. Clearly, C should be invertible. Denote its Wiener–Hopf factorization by AB and set $\mathbf{w} := B\mathbf{u}$. Then

$$\mathbf{w}' = A^{-1}DB^{-1}\mathbf{w}. \tag{6.16}$$

We now invoke Propositions 6.4 and 6.5: since A is upper triangular, so is A^{-1}, hence $A^{-1}D$ is in $\mathbf{T}$. By the same token, B is lower triangular $\Rightarrow B^{-1}$ is lower triangular $\Rightarrow A^{-1}DB^{-1} \in \mathbf{T}$. Moreover, the symbol of this matrix is $\mathbf{d}/(\mathbf{ab}) = \mathbf{d}/\mathbf{c}$, and this coincides with the characteristic function of the scheme (6.2). Consequently, (6.16) can be solved stably if and only if the relevant conditions of Lemma 6.1 are maintained, i.e. the real part of the symbol is non-positive on the unit circle. Since

$$\|\mathbf{u}\| \leq \|B\| \times \|B^{-1}\| \times \|\mathbf{w}\|,$$

this is also the stability condition for the original semi-discretized system – all this, of course, if and only if C can be inverted.

The matrix C is not just Toeplitz but also banded: its symbol $\mathbf{c}$ is a Laurent polynomial. Specifically,

$$\begin{aligned} \mathbf{c}(z) &= z^{-\bar{r}} \sum_{j=0}^{\bar{r}+\bar{s}} \alpha_{j-\bar{r}} z^j \\ &= \alpha_{\bar{s}} z^{-\bar{r}} \prod_{k=1}^{\bar{r}+\bar{s}} (z - \xi_k), \end{aligned}$$

where, without loss of generality, $|\xi_1| \leq |\xi_2| \leq \ldots \leq |\xi_{\bar{r}+\bar{s}}|$. Thus, a Wiener–Hopf factorization is $C = AB$, where

$$\begin{aligned} \mathbf{a}(z) &= \alpha_{\bar{s}} \prod_{k=1}^{\bar{s}} (z - \xi_k), \\ \mathbf{b}(z) &= \prod_{k=\bar{s}+1}^{\bar{r}+\bar{s}} \left(1 - \frac{\xi_k}{z}\right). \end{aligned}$$

Lemma 6.7 The semi-discretization (6.2) is stable for the single-boundary problem if and only if $\mathrm{Re}\, h(e^{i\theta}) \leq 0$ for all $|\theta| \leq \pi$ and h has exactly $\bar{r}$ poles in $|z| < 1$ and $\bar{s}$ poles in $|z| > 1$.

Proof. The proof follows at once by applying Proposition 6.6 to the explicit form of **a** and **b**. □

Similar analysis can be extended to the fully discretized equations (6.11). Again, stability in the singly-infinite interval requires $\bar{r}$ poles of H inside and $\bar{s}$ outside the unit circle. Moreover, our results are valid when the equation (6.1) is given in a finite interval (Strang, 1964b). Moreover, if H is canonical then Lemma 6.3 is valid on a half-line, in the sense that if the pole property holds uniformly in a two-sided neighbourhood of $\mu = 0$ then it is inherited by the associated semi-discretization.

Note that the condition on the poles implies *a fortiori* that h (or H, as the case might be) is good. Indeed, 'goodness' is precisely the condition for bounded invertibility of the bi-infinite operator.

The condition on the poles of the characteristic function looks, at a first glance, rather strange. After all, we can multiply the numerator and the denominator by the same integer power of z, thereby shifting $\bar{r}$ and $\bar{s}$. However, the values of $\bar{r}$ and $\bar{s}$ determine the 'centre' of the method. Their meaning (which becomes apparent only in the presence of boundary conditions) is that we must specify exactly $\bar{r}$ boundary conditions on the left and $\bar{s}$ conditions on the right. If stability fails due to bad 'balancing' of zeros, it can be recovered by redistributing the boundary conditions.

6.3 A barrier for semi-discretizations

Let h be the characteristic function of the semi-discretized scheme (6.2). We assume that it is of order $p \geq 1$, thus $h(z) = \log z + \mathcal{O}(|z-1|^{p+1})$. Moreover, we assert that it is stable in the sense of Section 6.2: $\mathrm{Re}\, h(e^{i\theta}) \leq 0$ for all θ and h has precisely $\bar{r}$ poles inside and $\bar{s}$ poles outside the unit circle.

This section adopts the approach of Iserles and Williamson-Renaut (1984) to derive barriers on p, subject to stability and given $r, s, \bar{r}$ and

$\bar{s}$.[5]

Lemma 6.8 Subject to stability and $p \geq 1$, h has at most r zeros in $0 < |z| < 1$ and at most $s - 1$ zeros in $1 < |z| < \infty$.

Proof. As often in zero-counting proofs in the complex plane, we wish to employ the argument principle. Unfortunately, $p \geq 1$ implies that $h(1) = 0$ and we are unable to implement the argument principle along $|z| = 1$ for the function h. Instead, we choose $0 < \varepsilon \ll 1$ and define

$$h^*(z,\varepsilon) := h(z) - \varepsilon.$$

Note that the poles of h and h^* coincide, whereas their zeros are located near to each other.

Since $\mathrm{Re}\, h(e^{i\theta}) \leq 0$ for all θ, we have $\mathrm{Re}\, h^*(e^{i\theta},\varepsilon) < 0$. Thus, the variation of the argument of h^* when we revolve once along $|z| = 1$ in the positive direction is nil and, by the argument principle, this meromorphic function has the same number of zeros and poles inside the unit disc. By Lemma 6.7 there are exactly $\bar{r}$ poles in the disc. Moreover, there is a zero of multiplicity $\bar{r} - r$ at the origin,[6] implying that there must be exactly r zeros in the punctured disc $0 < |z| < 1$ for every $\varepsilon \downarrow 0$. Some of these zeros might migrate to the perimeter as ε vanishes, but as none can migrate from $|z| = 1$, the first statement of the lemma is true.

To count the zeros in $|z| > 1$ we map conformally $z \mapsto 1/z$ – by the analysis of the last paragraph, there are at most s bounded zeros outside the unit circle. However, we know that $h(1) = 0$ and this zero must be accounted by the migration to $|z| = 1$ as $\varepsilon \downarrow 0$. Herewith we prove that this migration occurs from the exterior. For small $\varepsilon > 0$ there exists a zero z_ε of h^* which is near 1. It must be real, otherwise it would have had a conjugate and, in the limit, we would have had a double zero at 1. But $p \geq 1$ implies that $h'(1) = 1$ and the zero at 1 is simple. Thus, we can expand $z_\varepsilon = 1 + c\varepsilon + \mathcal{O}(\varepsilon^2)$. Moreover, $h^*(z_\varepsilon,\varepsilon) = 0$ implies that $h(z_\varepsilon) = \varepsilon$. But $p \geq 1$, hence $h(z) = z - 1 + \mathcal{O}(|z-1|^2)$. Consequently $c = 1$, $z_\varepsilon = 1 + \varepsilon + \mathcal{O}(\varepsilon^2) > 1$ and the zero at 1 migrates from the exterior. This leaves out at most $s - 1$ bounded zeros in $|z| > 1$. □

Corollary: No method (6.2) with $s = 0$ can be stable. □

We set an **order star of the second kind** by specifying $f(z) = z$ and $R(z) = h(e^z)$. Hence $\tilde{\rho}(z) = h(e^z) - z$. Since R is periodic with period $2\pi i$, we restrict our attention to the strip

$$\mathcal{S} := \{z \in \mathcal{C} : |\mathrm{Im}\, z| \leq \pi\}.$$

[5] A barrier for $r = \bar{r} = \bar{s} = 0$ has been derived by Engquist and Osher (1981) by an entirely different technique.

[6] If $\bar{r} - r < 0$, we have at the origin a pole of multiplicity $r - \bar{r}$.

$\mathcal{S}_+$ and $\mathcal{S}_-$ denote the restrictions of $\mathcal{S}$ to the right and left open half-planes respectively.

Example 6.5 We examine in Figure 6.1 order stars of four different functions h, each time choosing coefficients so as to maximize the order. We have $p = r + s + \bar{r} + \bar{s}$.

(a) $r = s = 1$, $\bar{r} = \bar{s} = 0$:
This is the $H^{(1,1)}$ interpolatory scheme: $h(z) = \frac{1}{2}(z - z^{-1})$. As proved in Example 6.1, $\mathrm{Re}\, h(e^{i\theta}) \leq 0$ for all θ. Since $\bar{r} = \bar{s} = 0$, the pole condition is true by default and the scheme is stable.

(b) $r = 0$, $s = 2$, $\bar{r} = \bar{s} = 1$:

$$h(z) = \frac{-27 + 24z + 3z^2}{-1/z + 14 + 17z},$$

both poles are inside the unit disc and $\mathrm{Re}\, h(i) > 0$. Thus, stability fails on both counts.

(c) $r = s = \bar{s} = 1$, $\bar{r} = 3$:
The characteristic function

$$h(z) = \frac{-2430/z + 1440 + 990z}{11/z^3 - 104/z^2 + 1176/z + 2056 + 281z}$$

has three poles inside and one outside the unit circle, thus obeying the pole condition. None the less, $\mathrm{Re}\, h(-1) > 0$ and the scheme is unstable.

(d) $r = 1$, $s = \bar{s} = 0$, $\bar{r} = 3$:
Since

$$h(z) = \frac{-24/z + 24}{1/z^3 - 5/z^2 + 19/z + 9},$$

we have

$$\mathrm{Re}\, h(e^{i\theta}) = -\frac{12(1 - \cos\theta)^3}{65 + 11\cos\theta - 13\cos^2\theta + 9\cos^3\theta} \leq 0, \quad |\theta| \leq \pi.$$

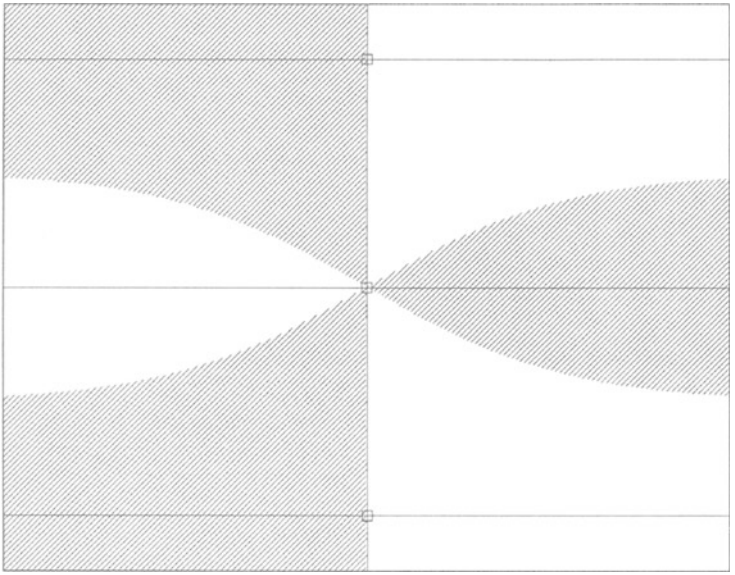
However, h has two poles in the unit disc, the pole condition is violated, and the scheme is unstable. $\diamond$

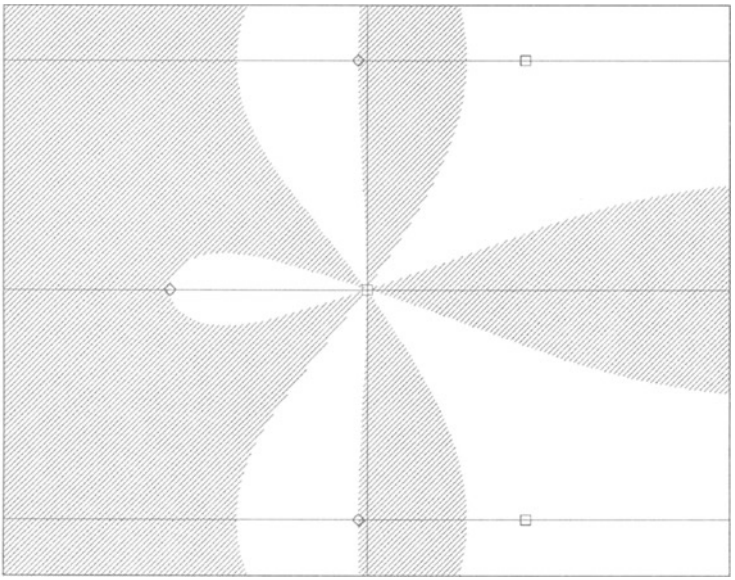
Seven properties of the underlying order stars are of interest. They are stated in the following lemma.

Lemma 6.9 The order star of h possesses the following properties:

(a) $\iota(0) = p + 1$ and the origin is a regular member of $\tilde{\mathcal{A}}_0$.

(b) If $s > \bar{s}$ then for $\mathrm{Re}\, z \gg 0$ the line segment $[\mathrm{Re}\, z - i\pi, \mathrm{Re}\, z + i\pi]$ is composed of $2(s - \bar{s}) + 1$ distinct intervals of $\tilde{\mathcal{A}}_-$ and $\tilde{\mathcal{A}}_+$, whereas if $s \leq \bar{s}$ then it belongs to $\tilde{\mathcal{A}}_-$.





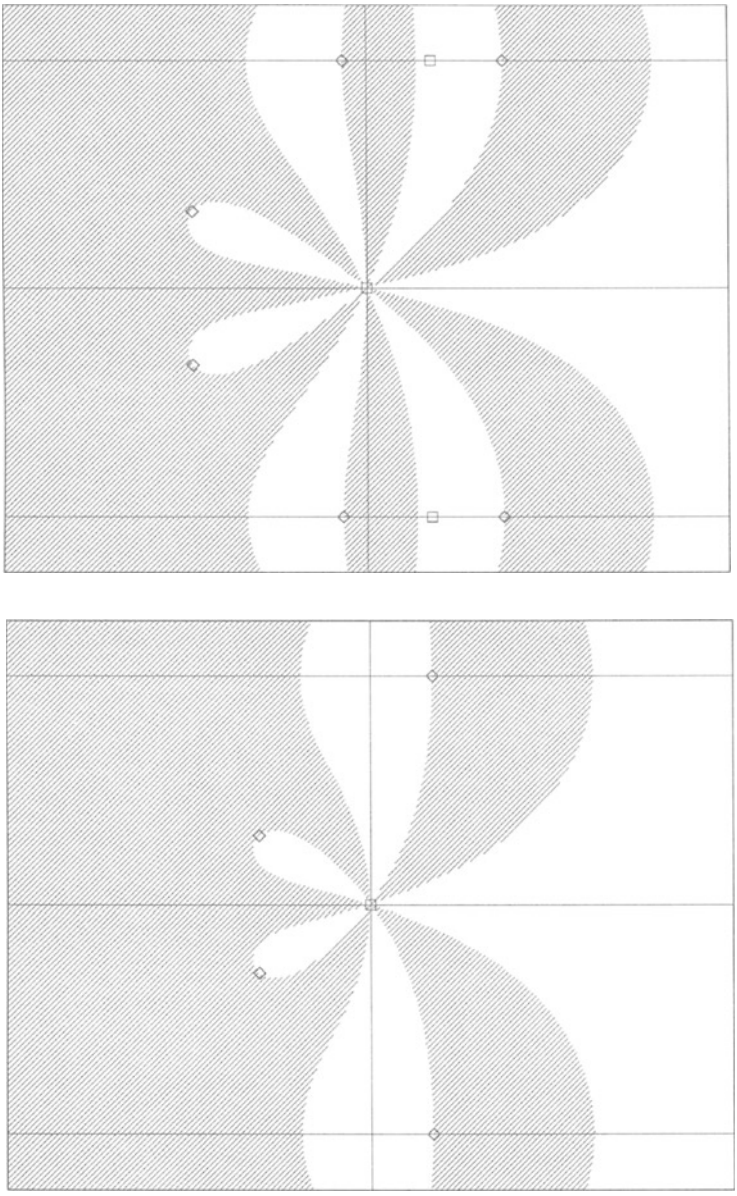


Figure 6.1 Order stars for the semi-discretizations from Example 6.5.

(c) If $r > \bar{r}$ then for $\mathrm{Re}\, z \ll 0$ the line segment $[\mathrm{Re}\, z - i\pi, \mathrm{Re}\, z + i\pi]$ is composed of $2(r - \bar{r}) + 1$ distinct intervals of $\tilde{\mathcal{A}}_-$ and $\tilde{\mathcal{A}}_+$, whereas if $r \leq \bar{r}$ then it belongs to $\tilde{\mathcal{A}}_+$.

(d) Between any two interpolation points along a loop there lies a point z such that e^z is a pole of h.

(e) Stability $\Rightarrow \tilde{\mathcal{A}}_+ \cap [-i\pi, i\pi] = \emptyset$.

(f) Stability $\Rightarrow h(e^z)$ has $\bar{r}$ bounded essential singularities in $\mathcal{S}_-$ and $\bar{s}$ bounded essential singularities in $\mathcal{S}_+$.

(g) Stability and $p \geq 1 \Rightarrow h(e^z)$ has at most r zeros in $\mathcal{S}_-$ and at most $s - 1$ zeros in $\mathcal{S}_+$.

Proof. Property (a) follows at once from Proposition 2.10. (b) and (c) are a simple consequence of Proposition 2.6 and periodicity of $h(e^z)$. Since the bounded essential singularities of $\tilde{\rho}$ are precisely the poles of h, mapped $z \mapsto e^z$, property (d) follows from Proposition 2.11. The remaining three properties can be derived at once from the definition of the order star. □

Clearly, to maximize the order of a stable scheme the multiplicities of loops need be as large as possible. Thus, according to property (d), poles of h act as a natural 'bottleneck' – in other words, the order is likely to be restricted by the availability of poles. However, by the periodicity of e^z, a negative pole is mapped to both lines $\{\mathrm{Im}\, z = \pi\}$ and $\{\mathrm{Im}\, z = -\pi\}$, hence it can be counted twice. Such poles will play a central role in our estimates. Specifically, we say that a bounded essential singularity (i.e. a pole of h, mapped $z \mapsto \log z$) is **efficient** if it lies on $\mathcal{R} \pm i\pi$, belongs to loops that approach the origin and there are no extra essential singularities along these loops.

Proposition 6.10 The number N of efficient essential singularities in an open line segment of the form $(x_1 \pm i\pi, x_2 \pm i\pi)$ is bounded by $Z + 1$, where Z is the number of zeros of $h(e^z)$ along the interval.

Proof. Let x be real. Then $\mathrm{Im}\, \tilde{\rho}(x \pm i\pi) = \mp\pi$. Since, by the proof of Proposition 2.11, the imaginary part of $\tilde{\rho}$ is strictly monotonically increasing along $\tilde{\mathcal{A}}_-$-loops, it follows that an efficient essential singularity lies between an $\tilde{\mathcal{A}}_-$-region to the left and an $\tilde{\mathcal{A}}_+$-region to the right. Moreover, $\mathrm{Re}\, \tilde{\rho}(x \pm i\pi) = h(-e^x) - x$, and it follows from the definition of the order star that $h(-e^x) > x$ in an $\tilde{\mathcal{A}}_+$-region and $h(-e^x) < x$ in an $\tilde{\mathcal{A}}_-$-region. Finally, $h(-e^x)$ becomes unbounded precisely at essential singularities and it is continuous elsewhere. It follows that any two efficient essential singularities along the line-segment are separated by (at least one) zero of $h(-e^x)$. The bound in the statement of the proposition follows. □

We assume that h represents a stable semi-discretization of order $p \geq 1$.

According to Lemma 6.9(e), the interval $[-i\pi, i\pi]$ separates $\tilde{\mathcal{A}}_+$-regions and it is meaningful to speak about the number of sectors of $\tilde{\mathcal{A}}_+$ that approach the origin from within $\mathcal{S}_-$ and from within $\mathcal{S}_+$. We denote these quantities by ω_- and ω_+ respectively. According to Lemma 6.9(a,e) it is true that

$$\omega_- + \omega_+ = p + 1, \quad \omega_- - 1 \le \omega_+ \le \omega_- + 1. \tag{6.17}$$

Further, we designate by $I_\pm$ and $N_\pm$ the number of sectors of $\tilde{\mathcal{A}}_+$ reaching infinity and the number of efficient essential singularities in $\mathcal{S}_\pm$ respectively.

Each efficient essential singularity 'contributes' (in the sense of Lemma 6.9(d)) to at most two bounded sectors of $\tilde{\mathcal{A}}_+$ that approach the origin. Each such sector contains at least one essential singularity along each of the loops of its boundary. Finally, if $J \ge I_-$ unbounded sectors of $\tilde{\mathcal{A}}_+$ approach the origin from within $\mathcal{S}_-$, say, then they envelop at least $J - I_-$ bounded $\tilde{\mathcal{A}}_-$-regions. Since the boundary of each such $\tilde{\mathcal{A}}_-$-region contains at least one essential singularity, and because, by Lemma 6.9(f), there are $\bar{r}$ bounded essential singularities, we have $J - I_- \le \bar{r} - N_-$. This implies that

$$\omega_- \le 2N_- + (\bar{r} - N_-) + I.$$

An identical argument is valid in $\mathcal{S}_+$, therefore

$$\omega_- \le I_- + N_- + \bar{r} \quad \text{and} \quad \omega_+ \le I_+ + N_+ + \bar{s}. \tag{6.18}$$

We first look at $\mathcal{S}_-$. Since the sign of $\operatorname{Re} h(e^z)$ changes from negative to positive when passing (with positive orientation) through an efficient essential singularity on the line $\operatorname{Im} z = \pi$, and since stability requires $h(-1) \le 0$, the bound of Proposition 6.10 can be further tightened and every efficient essential singularity must be 'accounted for' by at least one zero. Since the number of efficient essential singularities in $\mathcal{S}_-$ is bounded (according to Lemma 6.9(f)) by $\bar{r}$, we have $N_- \le \min\{r, \bar{r}\}$. Moreover, Lemma 6.9(c) implies that $I_- \le (r - \bar{r})_+ + 1$. Thus, (6.18) yields

$$\omega_- \le r + \bar{r} + 1. \tag{6.19}$$

Next we examine $\mathcal{S}_+$. If $s \le \bar{s}$ then Lemma 6.9(b) implies that $I_+ = 0$. Moreover, by Lemma 6.9(f,g) and Proposition 6.10 it is true that $N_+ \le \min\{s, \bar{s}\}$. Thus,

$$\omega_+ \le s + \bar{s}. \tag{6.20}$$

If $s > \bar{s}$ then there are two possibilities: the lines $\operatorname{Im} z = \pm\pi$ belong either to $\tilde{\mathcal{A}}_-$ or to $\tilde{\mathcal{A}}_+$ for $\operatorname{Re} z \gg 0$. In the first case $N_+ \le \min\{s, \bar{s} - 1\}$ and $I_+ = s - \bar{s}$. In the second case the unbounded $\tilde{\mathcal{A}}_+$-region that surrounds the border of the strip at infinity must account (since $\operatorname{Im} \tilde{\rho}$ is monotone along $\tilde{\mathcal{A}}_0$, cf. the proof of Proposition 2.11) for a bounded essential singularity

that, consequently, cannot be efficient. Therefore $N_+ \leq \min\{s, \bar{s} - 1\}$ and $I_+ = s - \bar{s} + 1$. In both cases (6.18) implies that the inequality (6.20) is true.

We now substitute (6.19) and (6.20) into (6.17). This results in the inequality

$$p \leq 2\min\{r + \bar{r} + 1, s + \bar{s}\}. \tag{6.21}$$

Our stability barrier is almost complete, except that we need a further result on the attainable order.

Lemma 6.11 If either $\{r \geq \bar{r},\ s \geq \bar{s}\}$ or $\{\bar{r} \geq r,\ \bar{s} \geq s\}$ then $p \leq r+s+\bar{r}+\bar{s}$, irrespective of stability. Moreover, if (6.2) is stable then $p \leq r + s + \bar{r} + \bar{s}$ holds unconditionally.

Proof. We again use the order star, bounding the number ω of sectors of $\tilde{\mathcal{A}}_-$ and $\tilde{\mathcal{A}}_+$ that approach the origin. By Lemma 6.9(d) and Proposition 6.10 the number N of efficient essential singularities is bounded by $\min\{r + s, \bar{r} + \bar{s}\}$ (since the zero $h(1)$ cannot contribute). Furthermore, Lemma 6.9(b,c) implies that exactly $2(s - \bar{s})_+ + 2(r - \bar{r})_+ + 2$ sectors of $\tilde{\mathcal{A}}_-$ and $\tilde{\mathcal{A}}_+$ approach $\pm\infty$. Thus, as every efficient essential singularity can contribute to at most four bounded sectors that reach the origin and the remaining essential singularities can contribute to just two,

$$\begin{aligned} \omega &\leq 4N + 2(\bar{r} + \bar{s} - N) + 2(s - \bar{s})_+ + 2(r - \bar{r})_+ + 2 \\ &= 2\,(\bar{r} + \bar{s} + N + (s - \bar{s})_+ + (r - \bar{r})_+ + 1)\,. \end{aligned}$$

Hence, if either $\{r \geq \bar{r},\ s \geq \bar{s}\}$ or $\{\bar{r} \geq r,\ \bar{s} \geq s\}$ then $\omega \leq 2(r+s+\bar{r}+\bar{s}+1)$ But, by Lemma 6.9(a), $\omega = p + 1$ and the proof is complete.

In the remaining cases we need to assume stability. Thus, it follows from our analysis that $N = N_- + N_+ \leq \min\{r, \bar{r}\} + \min\{s, \bar{s}\}$. Consequently,

$$\begin{aligned} p + 1 &= \frac{\omega}{2} \leq \bar{r} + \bar{s} + N + (r - \bar{r})_+ + (s - \bar{s})_+ + 1 \\ &\leq \bar{r} + \bar{s} + \min\{r, \bar{r}\} + \min\{s, \bar{s}\} \\ &\quad + (r - \bar{r})_+ + (s - \bar{s})_+ + 1 \\ &= r + s + \bar{r} + \bar{s} + 1. \end{aligned}$$

The lemma is true. □

It is interesting to note that the maximal order of approximation (disregarding stability) is the same as that of the $[(r+s)/(\bar{r}+\bar{s})]$ Padé approximant to $(1 - z)^{r-\bar{r}} \log(1 - z)$. The case of $\{r \geq \bar{r},\ s \geq \bar{s}\}$ is known, under this disguise, in the theory of Padé approximants (Baker, 1975).

We can now assemble our results into the main result of the present section.

Theorem 6.12 The order of a stable semi-discretization (6.2) is bounded by

$$\min\{r+s+\bar{r}+\bar{s}, 2(r+\bar{r}+1), 2(s+\bar{s})\}. \tag{6.22}$$

Moreover, stability requires $s \geq 1$. □

Note that the first term in the upper bound (6.22) is almost always submerged in the remaining terms: it comes into its own if and only if $s+\bar{s} = r+\bar{r}+1$.

The explicit interpolatory schemes were introduced in Example 6.1. It was proved there that the values $r \leq s \leq r+2$ lead to stability. A straightforward application of Theorem 6.12 resolves completely the stability of $h^{(r,s)}$ for all r and s. Its original proof by Iserles (1982) was the first instance of an application of order stars to stability barriers for partial differential equations. It can be also proved, quite easily, by order stars of the first kind (Iserles, 1986a).

Theorem 6.13 Explicit semi-discretizations (6.2) of maximal order $r+s$ are stable if and only if $r \leq s \leq r+2$. □

6.4 Stable full discretizations

Our intention is to do to full discretizations what we have done to semi-discretizations in Theorem 6.13: bound the order of stable schemes. However, first we need to debate stability of maximal-order methods for some choices of r, s, $\bar{r}$ and $\bar{s}$ and review some results of Iserles and Strang (1983).

Let $m := r+s$, $n := \bar{r}+\bar{s}$ and $\lambda := \mu + r - \bar{r}$. Because of Lemma 6.2, finding a method of order $p = r+s+\bar{r}+\bar{s}$ is equivalent to determining the coefficients of the $[m/n]$ Padé approximant $P_{m/n}(z,\lambda)/Q_{m/n}(z,\lambda)$ to the function $(1-z)^{\lambda}$ and setting

$$\begin{aligned} \sum_{j=-\bar{r}}^{\bar{s}} \gamma_j(\mu) z^j &= z^{-\bar{r}} Q_{m/n}(1-z,\lambda), \\ \sum_{j=-r}^{s} \delta_j(\mu) z^j &= z^{-r} P_{m/n}(1-z,\lambda). \end{aligned}$$

Our first step is to derive $P_{m/n}$ and $Q_{m/n}$ explicitly, essentially by following a technique similar to that in Example 3.2. We recall from Section 3.4 the definition of a **hypergeometric function**: $f(z) = \sum_{\ell=0}^{\infty} f_\ell/\ell! z^\ell$, where $f_{\ell+1}/f_\ell = r_f(\ell)$, r_f being a rational function. Throughout this chapter $r_f \in \pi_{2/1}$. In Example 6.3 we introduced the Pochhammer symbol, which can presently be employed to write the $\{2/1\}$ hypergeometric function in a

compact manner,

$$ {}_2F_1\left[\begin{matrix} a,b; \\ c; \end{matrix} \quad z\right] = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{k!(c)_k} z^k, $$

where c is neither zero nor a negative integer (Rainville, 1967).[7] Among the myriad of relations obeyed by hypergeometric functions we exploit the **Euler identity**

$$ {}_2F_1\left[\begin{matrix} c-a, c-b; \\ c; \end{matrix} \quad z\right] = (1-z)^{a+b-c} {}_2F_1\left[\begin{matrix} a,b; \\ c; \end{matrix} \quad z\right]. $$

We choose a small $\varepsilon > 0$ and set $a := -n$, $b := \lambda - m + \varepsilon$, $c := -n - m + \varepsilon$. The ${}_2F_1$ function on the right is an nth degree polynomial – and it remains so even when $\varepsilon \to 0$. Moreover, as $\varepsilon \to 0$, the terms corresponding to z^k for $k = m+1, m+2, \ldots, n+m$ on the left vanish, whereas the other terms tend to bounded values. Thus, we obtain there an mth degree polynomial and a perturbation of order $\mathcal{O}(z^{n+m+1})$. It is now an easy exercise to calculate

$$ \begin{aligned} P_{m/n}(z,\lambda) &= \sum_{k=0}^{m} \binom{m}{k} \frac{(n+m-k)!(-n-\lambda)_k}{(n+m)!} z^k; \qquad (6.23) \\ Q_{m/n}(z,\lambda) &= \sum_{k=0}^{n} \binom{n}{k} \frac{(n+m-k)!(-m+\lambda)_k}{(n+m)!} z^k \end{aligned} $$

and to verify that

$$ c_{m/n}(\lambda) := (-1)^{n+1} \frac{m!n!(-n-\lambda)_{n+m+1}}{(n+m)!(n+m+1)!} \qquad (6.24) $$

is the error constant:

$$ P_{m/n}(z,\lambda) - (1-z)^{\lambda} Q_{m/n}(z,\lambda) = c_{m/n}(\lambda) z^{n+m+1} + \mathcal{O}(z^{n+m+2}). $$

Note further that

$$ P_{m/n}(z,\lambda) = Q_{n/m}(z,-\lambda). \qquad (6.25) $$

Full discretizations whose coefficients are in conformity with (6.23) are called **Padé methods.**

[7] The vigilant reader will observe that every $\{p,q\}$ hypergeometric function can be written in the ${}_pF_q$ notation, which is, in fact, the standard in the theory of special functions. This is beyond our requirements in this chapter. Having said this, we certainly wish to encourage more familarity with special functions in general and hypergeometric functions in particular. In this brave day and age of functional-analytic and topological techniques, special functions are sometimes deemed to be old-fashioned, quaint and plain boring. Nothing can be further from the truth!

The function h should obey two stability requirements: its modulus must be bounded by unity along $|z| = 1$ and the number of poles must be distributed correctly between $|z| < 1$ and $|z| > 1$. We wish to express this in convenient terms for methods originating in the Padé approximants (6.23). First we evaluate $\tilde{q}_{m/n}(\theta, \lambda) := |Q_{m/n}(e^{i\theta}, \lambda)|^2$. To this end we exploit a product formula for ${}_2F_1$ functions due to Burchnall and Chaundy (Erdélyi *et al.*, 1953):

$$
\begin{aligned}
&{}_2F_1\left[\begin{matrix} a, b; \\ c; \end{matrix}\; z_1\right] {}_2F_1\left[\begin{matrix} a, b; \\ c; \end{matrix}\; z_2\right] \\
&\quad = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k (c-a)_k (c-b)_k}{k!(c)_k (c)_{2k}} (z_1 z_2)^k \\
&\qquad \times {}_2F_1\left[\begin{matrix} a+k, b+k; \\ c+2k; \end{matrix}\; z_1 + z_2 - z_1 z_2\right]
\end{aligned}
$$

As before, we set $a = -n$, $b = \lambda - m + \varepsilon$ and $c = -n - m + \varepsilon$. Moreover, since $z_1 = 1 - e^{i\theta}$, $z_2 = \bar{z}_1 = 1 - e^{-i\theta}$, it is true that

$$
Z := z_1 z_2 = z_1 + z_2 = 2(1 - \cos\theta).
$$

In particular, $z_1 + z_2 - z_1 z_2 = 0$ and all the hypergeometric functions on the right reduce to unity. Obviously, this greatly simplifies the expression:

$$
\tilde{q}_{m/n}(\theta, \lambda) = \lim_{\varepsilon \to 0} \sum_{k=0}^{n} \frac{(-n)_k (\lambda - m + \varepsilon)_k (-m + \varepsilon)_k (-n - \lambda)_k}{k!(-n - m + \varepsilon)_k (-n - m + \varepsilon)_{2k}} Z^k.
$$

The limiting value is easy to evaluate when $m \geq n$, since the denominator is bounded away from zero. Let

$$
v_k := \frac{m!n!}{((n+m)!)^2} (-n - \lambda)_k (-m + \lambda)_k, \quad k = 0, \ldots, n.
$$

The equation for $\tilde{q}_{m/n}$ reads

$$
\tilde{q}_{m/n}(\theta, \lambda) = \sum_{k=0}^{n} (-1)^k \frac{(n+m-k)!(n+m-2k)!}{k!(n-k)!(m-k)!} v_k Z^k. \tag{6.26}
$$

If $m < n$ then $(-m + \varepsilon)_k$ approaches zero in the numerator for $k = m + 1, \ldots, n$, as does the term $(-n - m + \varepsilon)_{2k}$ in the denominator for $k = [(n+m)/2] + 1, \ldots, n$. Their ratio is bounded:

$$
\lim_{\varepsilon \to 0} \frac{(-m + \varepsilon)_k}{(-n - m + \varepsilon)_{2k}} = (-1)^n \frac{m!(k-m-1)!}{(n+m)!(2k-n-m+1)!}.
$$

We obtain

$$\tilde{q}_{m/n}(\theta,\lambda) = \sum_{k=0}^{m}(-1)^k \frac{(n+m-k)!(n+m-2k)!}{k!(n-k)!(m-k)!} v_k Z^k \tag{6.27}$$

$$+ (-1)^n \sum_{k=[\frac{n+m}{2}]+1}^{n} \frac{(n+m-k)!(k-m-1)!}{k!(2k-n-m+1)!(n-k)!} v_k Z^k.$$

Bearing in mind the complexity of the above expressions, it is fortunate that, by virtue of (6.25), the formulae for $\tilde{p}_{m/n}(\theta,\lambda) := |P_{m/n}(e^{i\theta},\lambda)|^2$ can be obtained at once from (6.26) and (6.27) and

$$\tilde{p}_{m/n}(\theta,\lambda) \equiv \tilde{q}_{n/m}(\theta,-\lambda).$$

The end result is formulated as a lemma.

Lemma 6.14 The function

$$T_{m/n}(\theta,\lambda) := \frac{((n+m)!)^2}{m!n!}(\tilde{q}_{m/n}(\theta,\lambda) - \tilde{p}_{m/n}(\theta,\lambda))$$

equals

$$(-1)^n \sum_{k=[\frac{n+m}{2}]+1}^{n} \frac{(n+m-k)!(k-m-1)!(-n-\lambda)_k(-m+\lambda)_k}{k!(2k-n-m-1)!(n-k)!} Z^k$$

for $m < n$,

$$(-1)^{m+1} \sum_{k=[\frac{n+m}{2}]+1}^{m} \frac{(n+m-k)!(k-n-1)!(-n-\lambda)_k(-m+\lambda)_k}{k!(2k-n-m-1)!(m-k)!} Z^k$$

for $m > n$ and it vanishes identically for $m = n$. □

Example 6.6 We return to the schemes $H^{(r,s)}$ from Example 6.3. Since the method is explicit, the pole condition is satisfied by default. We have

$$T_{(r+s)/0}(\theta, r+\mu) = (-1)^{r+s+1} \sum_{k=[\frac{r+s}{2}]+1}^{r+s} \frac{(-r-\mu)_k(-s+\mu)_k}{k(2k-r-s-1)!} Z^k.$$

Note that $Z = 4\sin^2(\theta/2) \geq 0$.

Letting $s = r$, we have a linear combination with nonnegative coefficients of the terms $-(-r-\mu)_k(-r+\mu)_k = -(1+\mu)_r(1-\mu)_r(\mu)_{k-r}(-\mu)_{k-r} \geq 0$ for all $k = r+1,\ldots,2r$ and $|\mu| \leq 1$. Thus stability.

Stability for $s = r+1$ and $s = r+2$ follows similarly, but μ must now be restricted to $(0,1)$ and $(0,2)$ respectively. Thus, the 'positive' announcement

from Example 6.3 has been justified. The 'negative' statement therein, namely that $r \le s \le r+2$ exhausts the range of stable values, will be made good in the next section. $\diamond$

To locate the poles we recognize $Q_{m/n}$ as a Möbius transform of a generalized **Jacobi polynomial**. We recall that $P_n^{(\alpha,\beta)}$, $\alpha,\beta > -1$, is a Jacobi polynomial if it is orthogonal with respect to the weight function $(1-x)^\alpha(1+x)^\beta$ in the interval $(-1,1)$, appropriately normalized. A hypergeometric representation of a Jacobi polynomial takes the form

$$P_n^{(\alpha,\beta)}(x) = \frac{(1+\alpha+\beta)_{2n}}{n!(1+\alpha+\beta)_n}\left(\frac{x+1}{2}\right)^n {}_2F_1\left[\begin{matrix} -n, -\beta-n; \\ -\alpha-\beta-2n; \end{matrix} \frac{2}{x+1}\right] \tag{6.28}$$

(Rainville, 1967). Formula (6.28) remains valid when

$$\min\{\alpha,\beta\} < -1$$

(although, of course, orthogonality is lost), except when either α or β is a negative integer.

Proposition 6.15 The pole condition is satisfied if and only if the Jacobi polynomial $P_{\bar r+\bar s}^{(r-\bar r+\mu, s-\bar s-\mu)}$ has exactly $\bar r$ zeros in $\mathcal{C}^+$ and $\bar s$ zeros in $\mathcal{C}^-$.

Proof. Comparing (6.28) with (6.23) verifies that

$$Q_{m/n}(z,\lambda) = \frac{m!n!}{(m+n)!}(1-z)^n P_n^{(\lambda, m-n-\lambda)}\left(\frac{1+z}{1-z}\right).$$

The proposition follows by the virtue of the identity

$$z^{\bar r}\sum_{j=-\bar r}^{\bar s}\delta_j(\mu)z^j = Q_{(r+s(/(\bar r+\bar s)}(1-z, r-\bar r+\mu).$$

□

We say that a Padé method is **diagonal** if $m = n$, i.e. $r+s = \bar r+\bar s$. Stable diagonal schemes play central role in the next section, in the study of stability barriers for full discretizations. Hence the motivation to focus at present on such methods.

According to Lemma 6.14 it is true that $T_{m/m} \equiv 0$. Thus, a diagonal scheme is stable if and only if the pole condition holds. In other words, for every given n there are only certain values of $r+s = \bar r+\bar s = n$ that balance the poles correctly with respect to the unit circle.

Proposition 6.16 The equation

$$\psi(x) = P(x) - Q(x)x^\lambda = 0,$$

where P and $Q \not\equiv 0$ are polynomials of degree m and n respectively and $\lambda < m+1$ is noninteger, has at most $n+m+1$ real roots, counted with their multiplicity.

Proof. The proposition resembles the maximal interpolation theorem from Chapter 3. However, unlike in the proof of Theorem 3.7, we resort to elementary means. Repeatedly differentiating ψ yields

$$\frac{\partial^k}{\partial x^k}\psi(x) = P_k(x) - x^{\lambda-k}Q_k(x),$$

where $\deg P_k = m-k$ and $\deg Q_k = n$. Letting $k = m+1$ leads to

$$\frac{\partial^{m+1}}{\partial x^{m+1}}\psi(x) = -x^{\lambda-m-1}Q_{m+1}(x).$$

Suppose first that $Q_{m+1} \equiv 0$. Since $Q_{m+1}(x) = \lambda Q_m(x) + xQ'_m(x)$, this implies $Q_m(x) = Cx^{-\lambda}$ for some constant C. Since λ is not an integer, this can be true only if $C = 0$, thus $Q_m \equiv 0$. We proceed by induction on decreasing k, finally reaching $Q_0 = Q \equiv 0$, a contradiction. Consequently, our assumption that Q_{m+1} vanishes identically was false.

We now exploit the fact that the zeros of $\partial^{m+1}\psi/\partial x^{m+1}$ and Q_{m+1} coincide (since $\lambda < m+1$), hence the number of real zeros is bounded by n. Finally, $m+1$ consecutive applications of the Rolle theorem 'strip' the derivatives and produce the required bound on the number of real zeros of ψ. □

Corollary: Suppose that $T_{m/n}(\theta,\lambda) \geq 0$ for all $|\theta| \leq \pi$ and $\lambda \in (L, L+1)$, where $L \leq m$ is an integer. Then $\tilde{q}_{m/n}$ does not vanish in that range. Thus, stability for $\lambda \downarrow L$ implies stability for all $\lambda \in (L, L+1)$.

Proof. Suppose that $\tilde{q}_{m/n}(\theta^*,\lambda^*) = 0$ for some $\theta^* \in [-\pi,\pi]$, $\lambda^* \in (L, L+1)$. Thus

$$0 \leq T_{m/n}(\theta^*,\lambda^*) = -\tilde{p}_{m/n}(\theta^*,\lambda^*) \leq 0$$

implies that also $\tilde{p}_{m/n}$ vanishes there. It follows that $P_{m/n}$ and $Q_{m/n}$ share a nontrivial common factor. Dividing by this factor yields polynomials P^* and Q^* such that $\deg P^* \leq m-1$, $\deg Q^* \leq n-1$ and $P^*(1-x) = x^{\lambda^*}Q^*(1-x) + \mathcal{O}((1-x)^{m+n+1})$. Since this contradicts Proposition 6.16, the existence of such θ^* and λ^* is ruled out and the proof is complete. □

Returning to the case $m = n$, we note that the conditions of the last corollary hold, since $T_{n/n} \equiv 0$. Thus, by Proposition 6.15, we need only count the number of zeros of the Jacobi polynomial $P_n^{(r-\bar{r}+\mu,\bar{r}-r+\mu)}$ as $\mu \downarrow 0$. This process is complicated by the fact that, unless $r = \bar{r}$, one of the parameters is in $(-\infty,-1]$ and orthogonality is lost. In particular, the Jacobi polynomial might have zeros away from $(-1,1)$.

Fortunately, zeros of Jacobi polynomials, even with parameters in the 'forbidden' range, can be counted when these parameters are integers. To this end we exploit the identity

$$P_n^{(-K,\beta)}(z) = \frac{(-n-\beta)_K}{(-n)_K}\left(\frac{z-1}{2}\right)^K P_{n-K}^{(K,\beta)}(z), \quad K = 0,\ldots,n$$

(Szegő, 1939). Since $P_n^{(\alpha,\beta)}(z) = (-1)^n P_n^{(\beta,\alpha)}(-z)$, we also have

$$P_n^{(\alpha,-K)}(z) = \frac{(-n-\alpha)_K}{(-n)_K}\left(\frac{z+1}{2}\right)^K P_{n-K}^{(\alpha,K)}(z), \quad K = 0,\ldots,n.$$

Finally, using both formulae, we derive

$$P_n^{(-K,-L)}(z) = \frac{(-n+L)_K(-n)_L}{(-n)_K(-n+K)_L}\left(\frac{z-1}{2}\right)^K\left(\frac{z+1}{2}\right)^L P_{n-K-L}^{(K,L)}(z)$$

for all $K + L \leq n$. Thus,

$$P_n^{(r-\bar{r},\bar{r}-r)}(z) = C(z+1)^{(r-\bar{r})_+}(z-1)^{(\bar{r}-r)_+} P_{n-|r-\bar{r}|}^{(|r-\bar{r}|,|r-\bar{r}|)}(z), \tag{6.29}$$

where

$$C = \frac{(r+\bar{s})!(s+\bar{r})!}{2^{|r-\bar{r}|}((r+s)!)^2} > 0.$$

The zeros of $P_n^{(\alpha,\alpha)}$ (the **ultraspherical polynomial**, sometimes known, under different normalization, as the **Gegenbauer polynomial** (Rainville, 1967)) for $\alpha > -1$ are situated in $(-1,1)$ and symmetric with respect to the origin. This is quite obvious, since the polynomial is orthogonal in $(-1,1)$ with respect to the even weight function $(1-x^2)^\alpha$. Denote by $\xi_+(\mu)$, $\xi_-(\mu)$ and $\xi_0(\mu)$ the number of zeros of the Jacobi polynomial (with varying μ) in $\mathcal{C}^+$, $\mathcal{C}^-$ and $i\mathcal{R}$ respectively and let $\Xi(\mu) := (\xi_-(\mu), \xi_0(\mu), \xi_+(\mu))$. It follows from (6.29) that the zeros at $\mu = 0$ are in $[-1,1]$ and

$$\begin{aligned}
\xi_-(0) &= (r-\bar{r})_+ + \left[\tfrac{1}{2}(\bar{r}+\bar{s}-|r-\bar{r}|)\right] = \left[\tfrac{1}{2}(r+\bar{s})\right];\\
\xi_+(0) &= (\bar{r}-r)_+ + \left[\tfrac{1}{2}(\bar{r}+\bar{s}-|r-\bar{r}|)\right] = \left[\tfrac{1}{2}(s+\bar{r})\right];\\
\xi_0(0) &= \bar{r}+\bar{s}-\xi_-(0)-\xi_+(0).
\end{aligned}$$

Moreover, since zeros of orthogonal polynomials are simple, necessarily $\xi_0(0) \leq 1$, implying that $\bar{s}-1 \leq \xi_-(0) \leq \bar{s}$ and $\bar{r}-1 \leq \xi_+(0) \leq \bar{r}$. This singles out exactly three choices of r and s:

$$\begin{array}{llll}
r = \bar{s}-1, & s = \bar{r}+1 & \Rightarrow & \Xi(0) = (\bar{s}-1, 1, \bar{r});\\
r = \bar{s}, & s = \bar{r} & \Rightarrow & \Xi(0) = (\bar{s}, 0, \bar{r});\\
r = \bar{s}+1, & s = \bar{r}-1 & \Rightarrow & \Xi(0) = (\bar{s}, 1, \bar{r}-1).
\end{array} \tag{6.30}$$

The second choice remains, by the corollary to Proposition 6.16, invariant for all $\mu \in (0,1)$. In the remaining two cases we have a zero at the origin, which could have migrated there from either side. To pinpoint it for $\mu \in (0,1)$ we repeat the counting argument at $\mu = 1$. Thus, we check the zeros of $P_n^{(r-\bar{r}+1,\bar{r}-r-1)}$ and, similarly to (6.30), obtain $\xi_-(1) = \left[\frac{1}{2}(r+\bar{s}+1)\right]$ and $\xi_+(1) = \left[\frac{1}{2}(s+\bar{r}-1)\right]$. The only values of r and s that obey $\bar{s}-1 \leq \xi_-(1) \leq \bar{s}$, $\bar{r}-1 \leq \xi_+(1) \leq \bar{r}$ are

$$\begin{array}{lll} r=\bar{s}-2, & s=\bar{r}+2 & \Rightarrow \quad \Xi(1)=(\bar{s}-1,1,\bar{r}); \\ r=\bar{s}-1, & s=\bar{r}+1 & \Rightarrow \quad \Xi(1)=(\bar{s},0,\bar{r}); \\ r=\bar{s}, & s=\bar{r} & \Rightarrow \quad \Xi(1)=(\bar{s},1,\bar{r}-1). \end{array} \tag{6.31}$$

Theorem 6.17 The diagonal Padé methods are stable for $\mu \in (0,1)$ if and only if either $\{r=\bar{s},\ s=\bar{r}\}$ or $\{r=\bar{s}-1,\ s=\bar{r}+1\}$.

Proof. These are the only sets of values of r and s that appear in both (6.30) and (6.31). Moreover, as in each case $\xi_0(0)\xi_0(1)=0$, the zero at the origin migrates to $\mathcal{C}^+$ for $r=\bar{s}$ and to $\mathcal{C}^-$ for $r=\bar{s}-1$, precisely as required by Proposition 6.15. □

Many further examples of stable Padé methods are provided in (Iserles and Strang, 1983), including a complete characterization of stable Padé methods for $r+s=\bar{r}+\bar{s}\pm 1$ and for $\bar{r}=\bar{s}$, $r+s \geq 2\bar{r}$. Stable semi-discretizations can be derived by means of Lemma 6.3 and by utilizing the remark following Lemma 6.7, differentiating stable H with respect to μ and setting $\mu=0$. Care must be exercised to prevent migration of poles to the unit circle. For example, (6.30) verifies that the choice $\{r=\bar{s},\ s=\bar{r}\}$ is safe but $\{r=\bar{s}-1,\ s=\bar{r}+1\}$ is not: we have already encountered this in Example 6.4, with regard to the box scheme. It is a consequence of Lemma 6.3 that the 'association' technique produces only semi-discretizations with $r \geq \bar{r}$, $s \geq \bar{s}$. Further examples of stable semi-discretizations are presented by Iserles and Williamson-Renaut (1984) and least error constants have been analysed by Jeltsch and Strack (1985).

6.5 A barrier for full discretizations

In this section we follow the elegant technique of Jeltsch and Smit (1987): instead of considering H as an approximant to z^μ, we regard it as an approximant to a stable diagonal Padé scheme.[8] The bane of multivalued functions is eliminated at a stroke!

[8] A similar idea has been already encountered in Chapter 4, in analysing rational approximants to $\exp z$ *via* Hairer's representation, and in Chapter 5, in the derivation of the least error constant in multistep approximants to $\exp z$.

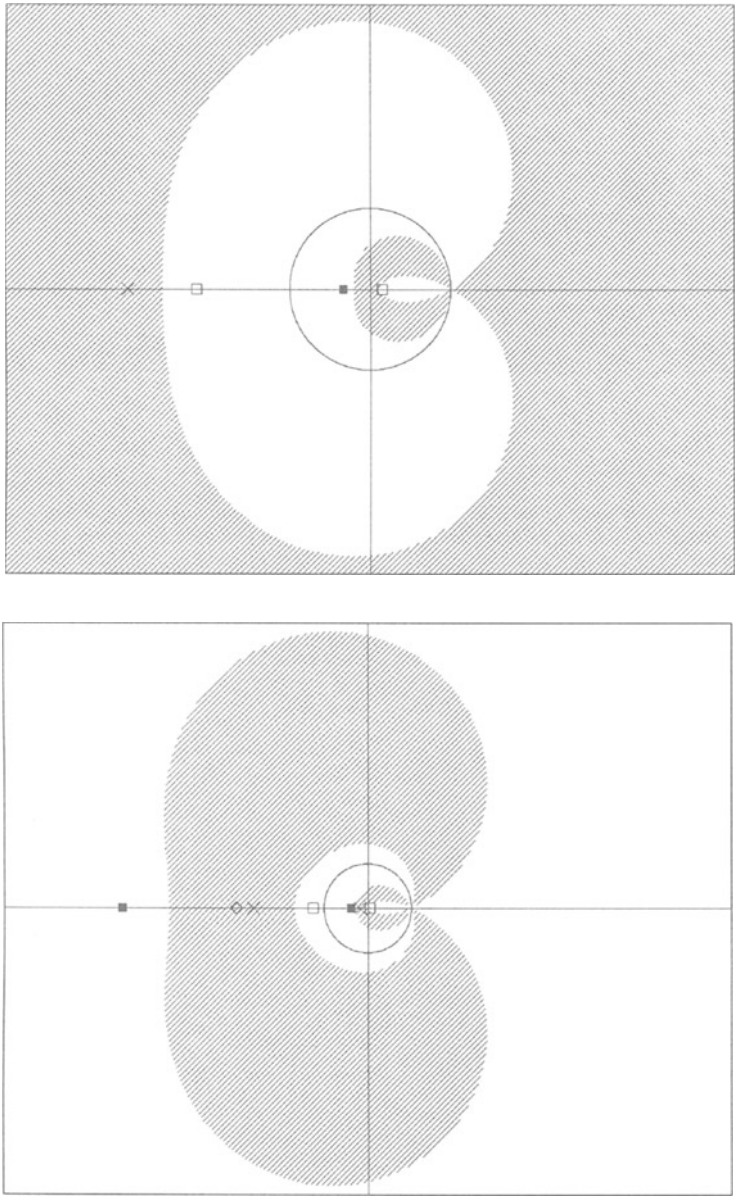


Figure 6.2 Order stars for the 'upwind' part of full discretizations for $\mu = \frac{1}{2}$ and $\mu = \frac{1}{4}$ respectively.

Fixing $\mu \in (0,1)$, we let H be a characteristic function of a stable (in the sense of Section 6.2) full discretization (6.11) of order p. We denote by $\tilde{H}$ the characteristic function of the Padé method with s and $\bar{s}$ unamended, r replaced with $\bar{s}$ and $\bar{r}$ replaced with s. As a consequence of Theorem 6.17, $\tilde{H}$ corresponds to a stable diagonal Padé method. Herewith we assume that $H \not\equiv \tilde{H}$.

We define an order star of the first kind by specifying $R = H$, $f = \tilde{H}$, hence $\rho = H/\tilde{H}$.

Example 6.7 Figure 6.2 displays two order stars, both for stable Padé methods. In the first, the explicit method

$$r = s = 1,\ \bar{r} = \bar{s} = 0, \quad H(z,\mu) = -\frac{1}{2}\mu(1-\mu)\frac{1}{z} + (1-\mu^2) + \frac{1}{2}\mu(1+\mu)z$$

is considered as an approximant of

$$r = \bar{s} = 0,\ \bar{r} = s = 1, \quad \tilde{H}(z,\mu) = \frac{(2-\mu)+\mu z}{\mu z^{-1} + (2-\mu)}.$$

In the second plot, we 'compare'

$$H(z,\mu) = \frac{-\frac{1}{2}\mu(1-\mu)z^{-1} + (1-\mu)(2+\mu) + \frac{1}{2}(1+\mu)(2+\mu)z}{(2+\mu)+(1-\mu)z}$$

(i.e. $r = s = \bar{s} = 1$, $\bar{r} = 0$) with

$$\tilde{H}(z,\mu) = \frac{(1-\mu)(2-\mu)z^{-1} + 2(4-\mu^2) + (1+\mu)(2+\mu)z}{(1+\mu)(2+\mu)z^{-1} + 2(4-\mu^2) + (1-\mu)(2-\mu)z}$$

(that is, $r = s = \bar{r} = \bar{s} = 1$) Zeros and poles of H and $\tilde{H}$ are denoted consistently with Table 2.1. ◇

The 'critical' portion of the complex plane is the exterior of the unit disc, which we henceforth denote by $\mathcal{D}_+$. We note the following points for future reference:

(a) $\iota(1) = \min\{p, 2(s+\bar{s})\} + 1$;

(b) H has $\bar{s}$ poles in $\mathcal{D}_+$;

(c) $\tilde{H}$ has s zeros in $\mathcal{D}_+$ (the proof is identical to the technique that led to Theorem 6.17: we identify the numerator with a Jacobi polynomial);

(d) ρ has no singularities in $\bar{\mathcal{D}}_+$, except at the poles of H and zeros of $\tilde{H}$;

(e) The error constant of $\tilde{H}$ obeys, according to (6.24), the inequality

$$(-1)^{s+\bar{s}} c_{(s+\bar{s})/(s+\bar{s})}(\bar{s} - s + \mu) > 0.$$

Proposition 6.18 Stability implies that $p \leq 2(s + \bar{s})$.

Proof. Let us assume that $p > 2(s+\bar{s})$. Since $|\tilde{H}(e^{i\theta}, \lambda)| \equiv 1$, the unit circle is a natural 'barrier' for $\mathcal{A}_+$-regions: according to (a) and Propositions 2.1 and 2.2, there are at least $s + \bar{s}$ sectors of $\mathcal{A}_+$ that approach $z = 1$ in $\mathcal{D}_+$. Taking further into account the observations (b), (c) and (d), in tandem with Proposition 2.3, we note that there are exactly $s + \bar{s}$ such sectors.

Since p exceeds the order of the diagonal Padé method, we have

$$\rho(z) = 1 + c^*(z-1)^{2(s+\bar{s})+1} + \mathcal{O}(|z-1|^{2(s+\bar{s}+1)}),$$

where the sign of $c^* \equiv c_{(s+\bar{s})/(s+\bar{s})}(\bar{s} - s + \mu)$ is, according to (e), $(-1)^{s+\bar{s}}$ for all $\mu \in (0,1)$.

We distinguish between two cases. If $s + \bar{s}$ is even then symmetry of $\mathcal{A}_+$ with respect to the real axis implies that for $0 < \varepsilon \ll 1$ the interval $(1, 1+\varepsilon)$ belongs to $\mathcal{A}_-$. This, together with

$$|\rho(1+x)| = 1 + c^* x^{2(s+\bar{s})+1} + \cdots, \quad 0 < x \ll 1$$

implies $c^* < 0$, a contradiction. The case of odd $s + \bar{s}$ responds to a similar argument. Since there are an odd number of sectors of $\mathcal{A}_+$ that approach $z = 1$ in $\mathcal{D}_+$, the interval $(1, 1+\varepsilon)$ lies in $\mathcal{A}_+$. This, in turn, implies that $c^* > 0$, an impossibility.

Either way we reach a contradiction. Consequently, $p \leq 2(s + \bar{s})$ and the proof is complete. □

Corollary: Given a stable scheme H with the error constant c and order $p = 2(s + \bar{s})$, it is true that $|c| \geq |c^*|$ and equality is attained if and only if $H \equiv \tilde{H}$.

Proof. Since the orders of H and $\tilde{H}$ coincide, we have

$$\rho(z) = 1 + (c^* - c)(z-1)^{2(s+\bar{s})+1} + \mathcal{O}((1-z)^{2(s+\bar{s}+1)}), \quad z \to 1.$$

It follows from the method of proof of Proposition 6.18 that $(-1)^{s+\bar{s}}(c^* - c) \geq 0$. The proof is complete unless $c = c^*$. Moreover, equality is impossible, otherwise $\rho(z) = 1 + \mathcal{O}((1-z)^{2(s+\bar{s}+1)})$ and, in line with Propositions 2.1 and 2.2, there must be at least $s+\bar{s}+1$ sectors of $\mathcal{A}_+$ approaching $z = 1$ in $\mathcal{D}_+$. This is prevented by (d) and Proposition 2.3. □

To bound the order by the 'downwind' part of the scheme we replace $\tilde{H}$ by the characteristic function of the diagonal Padé method with r and $\bar{r}$ unchanged, $s = \bar{r} + 1$ and $\bar{s} = r + 1$. Its order is $2(r + \bar{r} + 1)$ and, according to Theorem 6.17, it is stable. Defining the order star as before in terms of $\rho = H/\tilde{H}$, we amend (a)–(e):

(a′) $\iota(1) = \min\{p, 2(r + \bar{r} + 1)\} + 1$;

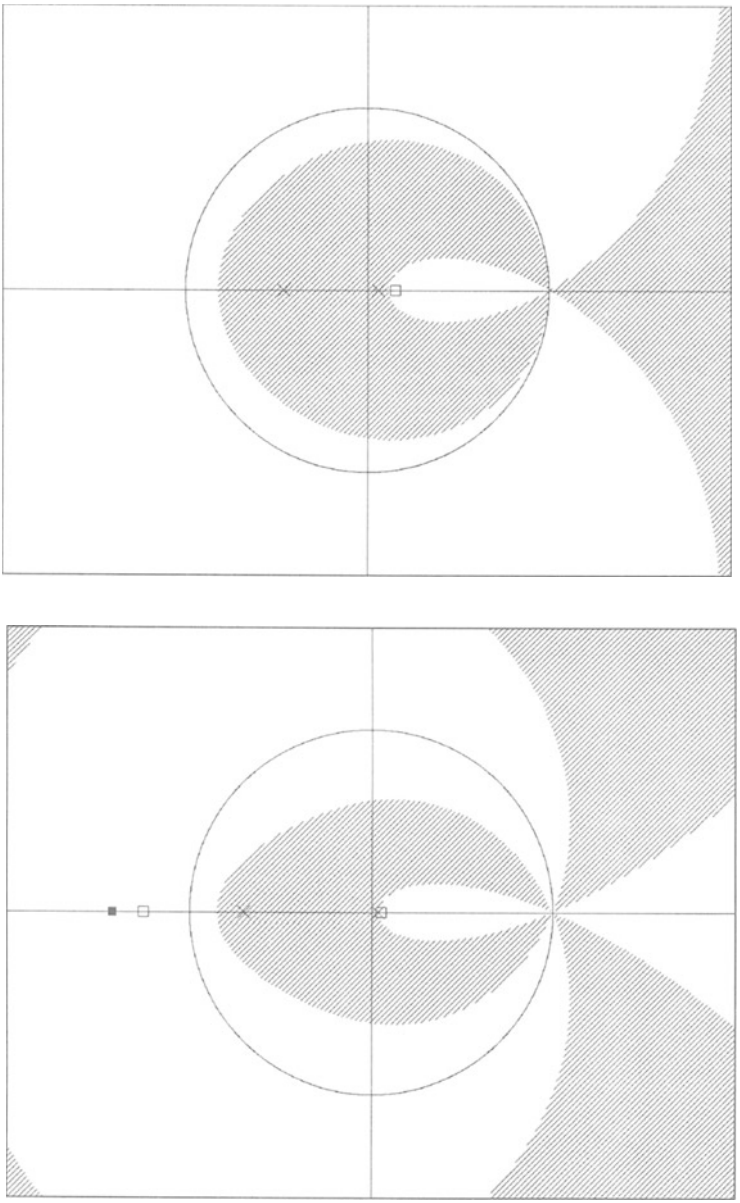


Figure 6.3 Order stars for the 'downwind' part of full discretizations.

(b′) H has $\bar{r}$ poles in the open unit disc $\mathcal{D}_-$;

(c′) $\tilde{H}$ has r zeros in $\mathcal{D}_-$;

(d′) ρ has no singularities in $\bar{\mathcal{D}}_-$, except at the poles of H and zeros of $\tilde{H}$;

(e′) The error constant of $\tilde{H}$ obeys the inequality

$$(-1)^{r+\bar{r}+1} c_{(r+\bar{r}+1)/(r+\bar{r}+1)}(r-\bar{r}+\mu) > 0.$$

Figure 6.3 illustrates the above observations. It displays order stars for the two approximants from Example 6.7. Note that both approximants give raise to the same $\tilde{H}$, namely

$$\tilde{H}(z,\mu) = \frac{-\mu(1-\mu)z^{-1} + 2(1-\mu)(3+\mu) + (2+\mu)(3+\mu)z}{(2+\mu)(3+\mu) + 2(1-\mu)(3+\mu)z - \mu(1-\mu)z^2},$$

with $r = s = 1$, $\bar{r} = 0$, $\bar{s} = 2$.

Proposition 6.19 Stability $\Rightarrow p \leq 2(r+\bar{r}+1)$.

Proof. We replicate the proof of Proposition 6.18, replacing $s+\bar{s}$ with $r+\bar{r}+1$, $\mathcal{D}_+$ with $\mathcal{D}_-$ and (a)–(e) with (a′)–(e′). □

Corollary: The least error constant for all stable methods (6.11) that attain the order barrier $2(r+\bar{r}+1)$ is provided by the unique choice of the diagonal Padé method which is defined by $\tilde{H}$. □

The last two propositions can be summed up in the statement that the order of a stable full discretization is bounded for all $\mu \in (0,1)$ by $\min\{2(r+\bar{r}+1), 2(s+\bar{s}), r+s+\bar{r}+\bar{s}\}$. In other words, the barrier of Theorem 6.13 persists. A more powerful result, valid for all non-integers μ, can be obtained by adopting a formalism due to Jeltsch and Smit (1987). We define $r^*(\mu)$ and $s^*(\mu)$ as the number of downwind and upwind points, respectively. The separating line in the (x,t)-plane is the characteristic curve $x+t =$ constant, drawn through the 'centre' at the old time level. It is easy to verify that

$$r^*(\mu) = \begin{cases} \bar{r} & : \mu < -r, \\ [r+\mu+1]+\bar{r} & : -r < \mu < s, \\ r+s+\bar{r}+1 & : s < \mu \end{cases};$$

$$s^*(\mu) = \begin{cases} r+s+\bar{s}+1 & : \mu < -r, \\ [s-\mu+1]+\bar{s} & : -r < \mu < s, \\ \bar{s} & : s < \mu \end{cases}.$$

Note that $r^*(\mu)+s^*(\mu) \equiv r+s+\bar{r}+\bar{s}+1$ for all non-integers μ. Consequently, $p \leq r^*(\mu)+s^*(\mu)-1$.

Recall that μ has been defined as $\Delta t/\Delta x$. Thus, although it is clear why we might be interested in stability for relatively large μ, the interest in negative μ requires further motivation. It is provided by replacing the differential equation (6.1) with the d-dimensional system

$$\frac{\partial}{\partial t}\mathbf{u} = A\frac{\partial}{\partial x}\mathbf{u},$$

where all the eigenvalues $\{\lambda_1, \ldots, \lambda_d\}$, say, of the matrix A are real (this means that the system is hyperbolic). Let (μ_-, μ_+) be the range of μ's such that the method (6.11) is stable. Then, amending the method to cater for the hyperbolic system, we need for stability that $\lambda_k \Delta t/\Delta x \in (\mu_-, \mu_+)$ for all $k = 1, \ldots, d$ – and this may include negative λ_k's!

The results of Propositions 6.18 and 6.19 can be recast in the language of r^* and s^*. In principle, it means considering non-integers μ outside $(0, 1)$. In practice, we replace s with s^* and r with r^* in the definitions of the 'upwind' and 'downwind' $\tilde{H}$ respectively.

Theorem 6.20 The order of a stable full discretization (6.11) may not exceed $\min\{r^*(\mu) + s^*(\mu) - 1, 2r^*(\mu), 2s^*(\mu)\}$ for all non-integer μ.

Proof. The proof of Propositions 6.18 and 6.19 carries through, with obvious changes of notation. □

The formulation of the stability barrier in terms of r^* and s^* has the virtue of having the 'slant' built into the inequality implicitly for all non-integers μ. Moreover, there are indications (which are debated in Chapter 10) that this is the correct formulation for multistep methods for the advection equation.

7

The diffusion equation

Space and Time! now I see it is true, what
I guess'd at,
What I guess'd when I loaf'd on the grass,
What I guess'd while I lay alone in my bed,
And again as I walk'd the beach under the
paling stars of the morning.

From *Song of Myself* by Walt Whitman (1819–1892).

7.1 Methods of optimal order

The **Crank–Nicolson** method is one of the oldest and most trustworthy workhorses of numerical mathematics, the method of choice in the eyes of many numerical analysts. The secret of its success is actually very simple: it is a combination of two techniques that are distinguished by their simplicity, computational robustness and stability. In a nutshell, we first replace the space derivatives in a partial differential equation of evolution by central finite differences.[1] This yields a system of ordinary differential equations, which is subsequently solved by the trapezoidal rule.

The Crank–Nicolson method was originally formulated for the diffusion equation

$$\frac{\partial}{\partial t}u = c\frac{\partial^2}{\partial x^2}u, \quad c > 0, \tag{7.1}$$

given with initial conditions along a finite interval $a \leq x \leq b$ and Dirichlet boundary conditions at $x = a$ and $x = b$ for all $t \geq 0$. We will assume henceforth that these boundary conditions are $u(a,t) = u(b,t) \equiv 0$, $t \geq 0$: this leads to considerable simplification in notation and the general case can be recovered quite painlessly.

Denoting by u_ℓ^k the numerical solution at $t = k\Delta t$, $x = \ell\Delta x$, we have

$$-\frac{\mu}{2}u_{\ell-1}^{k+1} + (1+\mu)u_\ell^{k+1} - \frac{\mu}{2}u_{\ell+1}^{k+1} = \frac{\mu}{2}u_{\ell-1}^k + (1-\mu)u_\ell^k + \frac{\mu}{2}u_{\ell+1}^k, \tag{7.2}$$

[1] We might also use a finite-element Galerkin scheme to get rid of the space derivatives. This is sometimes called the Crank–Nicolson–Galerkin method.

where $\mu = c\Delta t/(\Delta x)^2$ is the **Courant number**. Note that μ is likely to be quite large: choosing Δt and Δx of the same order of magnitude inevitably leads to $\mu = \mathcal{O}(1/\Delta x)$. This is a pointer to the important role stability is likely to play in the analysis of (7.2) and of other methods for the diffusion equation.

Our purpose in this chapter is to present stable methods that achieve a better order of accuracy than Crank–Nicolson, but first we need to set the stage for our analysis. Specifically, we review formally the twin concepts of stability and order. Much of our exposition parallels the work of Section 6.1 on fully discretized schemes for the advection equation.

At the onset of our discussion we assume that (7.1) is a Cauchy problem, i.e. that it is given along the whole real line and without any boundary conditions. We consider a general full discretization of the form

$$\sum_{j=-r}^{r} \gamma_j(\mu) u_{\ell+j}^{k+1} = \sum_{j=-r}^{r} \delta_j(\mu) u_{\ell+j}^{k}. \tag{7.3}$$

Note that we tacitly assume that the formula is centred: unlike in the case of the hyperbolic equation (6.1), no space direction is privileged and there is little point in upwinding. Moreover, our analysis remains correct if some of the coefficients are set identically to zero – this, in fact, introduces through the back door the possibility of studying unbalanced schemes.[2]

Recalling the definition of the Fourier transform in (6.3), we first act on the differential equation (7.1):

$$\frac{\partial}{\partial t}\hat{u}(\theta,t) = -\theta^2 \hat{u}(\theta,t), \quad |\theta| \le \pi.$$

Integration in t from $k\Delta t$ to $(k+1)\Delta t$ and the definition of the Courant number give

$$\hat{u}(\theta,(k+1)\Delta t) = e^{\mu(\theta\Delta x)^2}\hat{u}(\theta,k\Delta t).$$

Next, we Fourier-transform the finite-difference formula (6.3). This yields

$$\hat{u}^{k+1}(\theta) = H(e^{-i\theta\Delta x},\mu)\hat{u}^k(\theta), \tag{7.4}$$

where H is, as in Chapter 6, the characteristic function

$$H(z,\mu) := \frac{\sum_{j=-r}^{r} \delta_j(\mu) z^j}{\sum_{j=-r}^{r} \gamma_j(\mu) z^j}.$$

Consequently, denoting by $\hat{e}^k$ the Fourier transform of the error vector $(u_\ell^k - u(\ell\Delta x, k\Delta t))_{\ell=-\infty}^{\infty}$, we obtain

$$\hat{e}^{k+1}(\theta) = \left(H(e^{i\theta\Delta x},\mu) - e^{-\mu(\theta\Delta x)^2}\right)\hat{e}^k(\theta), \quad |\theta| \le \pi.$$

[2]More importantly, the last remark can be applied to allow for different values of r in the 'implicit' and the 'explicit' part, inclusive of explicit methods.

In line with the analysis of Chapter 6 and bearing in mind the isometry property of the Fourier transform, we deduce that the local error in the method (7.3) is $\mathcal{O}(\Delta x)^{p+1}$, where p is the order of approximation of $\exp(\mu(\log z)^2)$ by the rational function H at $z = 1$:

$$H(z,\mu) = e^{\mu(\log z)^2} + \mathcal{O}((z-1)^{p+1}), \quad z \to 1.$$

We say that (6.3) is of **order** p. Note that we have expressed the local error solely in terms of Δx. Of course, Δt has not gone away: it is hidden in the Courant number μ.

The isometry property of the Fourier transform is exploited again, to express **stability** as an attribute of the function H. We deduce by induction from (7.4) that

$$\hat{u}^k(\theta) = H(e^{-i\theta\Delta x}, \mu)^k \hat{u}^0(\theta),$$

where $\hat{u}^0$ is the Fourier-transformed initial condition. Therefore, provided that $|H(e^{i\theta}, \mu)| \leq 1$ for all $|\theta| \leq \pi$, $\hat{u}^k$ stays uniformly bounded for all $k \geq 0$ and $\Delta x \downarrow 0$ – again, as long as the Courant number is kept constant.[3] Moreover, if the modulus exceeds unity for some θ^* then we can always pick an initial condition that causes instability. The simplest such choice is $e^{i\theta^* x}$, but, unfortunately, it is not an L_2 function and its Fourier transform is the delta function – a generalized function. The correct course of action is to approximate $e^{i\theta^* x}$ by L_2 functions. The conclusion is that stability is equivalent to $|H(z,\mu)| \leq 1$ along the unit circle. Again, the two essential properties of a numerical method, order and stability, can be expressed in the 'language' of the characteristic function!

The aforementioned stability analysis is valid for a Cauchy problem and we would like to extend it to (7.1), supported by a finite interval. Fortunately, parabolic equations require no directional 'slant' in the finite difference method. All methods of interest, when written in vector form

$$C\mathbf{u}^{k+1} = D\mathbf{u}^k,$$

yield symmetric Toeplitz matrices C. The uniform boundedness of **symmetric** matrices follows at once from the analysis of Section 6.2 (as long as H is good – but this is required by stability anyway for the Cauchy problem).

Theorem 7.1 Given $\mu > 0$, the fully discretized method (7.3) is of order p (in Δx) if and only if $H(z,\mu) = \exp(\mu(\log z)^2) + \mathcal{O}((1-z)^{p+1})$, $z \to 1$, and it is stable if and only if $|H(z,\mu)| \leq 1$ for all $|z| = 1$. □

[3]The insistence on the constancy of the Courant number is in line with the Lax stability theory (Richtmyer and Morton, 1967).

Many methods yield stability for some Courant numbers, e.g. the forward Euler scheme

$$u_\ell^{k+1} = \mu u_{\ell-1}^k + (1-2\mu)u_\ell^k + \mu u_{\ell+1}^k$$

(for $\mu \leq \frac{1}{2}$). However, restricting stability to small Courant numbers drastically limits Δt, severely circumventing the practicability of the underlying method. Our interest is focused on **unconditionally stable** methods, that maintain stability for all $\mu > 0$. The Crank–Nicolson method (7.2) is an example of such a method: it gives

$$H(z,\mu) = \frac{z + \frac{1}{2}\mu(z-1)^2}{z - \frac{1}{2}\mu(z-1)^2}$$

and unconditional stability is easy to verify. Moreover, it follows, by virtue of Theorem 7.1, that (7.2) is a third-order method.

Can we do better than Crank–Nicolson with the same computational cost? The answer (which might come as a surprise to some numerical mathematicians) is a most emphatic *yes!* Consider the scheme

$$\begin{aligned} &\left(\frac{1}{12} - \frac{\mu}{2}\right) u_{\ell-1}^{k+1} + \left(\frac{5}{6} + \mu\right) u_\ell^{k+1} + \left(\frac{1}{12} - \frac{\mu}{2}\right) u_{\ell+1}^{k+1} \\ &= \left(\frac{1}{12} + \frac{\mu}{2}\right) u_{\ell-1}^k + \left(\frac{5}{6} - \mu\right) u_\ell^k + \left(\frac{1}{12} + \frac{\mu}{2}\right) u_{\ell+1}^k, \end{aligned} \tag{7.5}$$

originally introduced by Crandall (1955).[4] The first item of interest is that it uses exactly the same amount of information as Crank–Nicolson and leads to linear algebraic systems with exactly the same sparsity structure. Thus, as far as the computational expense is concerned, there is not much to choose between the two methods. However, the characteristic function of Crandall's scheme being

$$H(z,\mu) = \frac{z + \left(\frac{1}{12} + \frac{\mu}{2}\right)(z-1)^2}{z + \left(\frac{1}{12} - \frac{\mu}{2}\right)(z-1)^2},$$

it is easy to deduce from Theorem 7.1 that the method is both unconditionally stable and of order 5!

Of course, a method to solve numerically a constant-coefficient diffusion equation in a single space dimension is of little interest *per se*. Fortunately, Crandall's scheme (7.5) can be generalized to cater for more realistic equations. Thus, we replace (7.1) with

$$\frac{\partial}{\partial t} u = c(x,t) \frac{\partial^2}{\partial x^2} u,$$

[4] It has been derived by a multitude of techniques, cf. (Douglas, 1961; Samarskĭ, 1964). It appears as the sixth entry in an exhaustive table of methods for (7.1) in (Richtmyer & Morton, 1967, pp. 189–191).

where c is a positive function of sufficient smoothness. The Crank–Nicolson method can be extended readily, by embracing the principle of using central differences on space derivatives and subsequently applying the trapezoidal rule to the time-like ODE system.[5] The extension of Crandall's scheme takes slightly more effort, but the outcome,

$$\left(\frac{5}{6}+\mu c_\ell^{k+\frac{1}{2}}\right)u_\ell^{k+1}+c_\ell^{k+\frac{1}{2}}\left(\frac{1}{12c_{\ell+1}^{k+\frac{1}{2}}}-\frac{\mu}{2}\right)(u_{\ell-1}^{k+1}+u_{\ell+1}^{k+1})$$
$$=\left(\frac{5}{6}-\mu c_\ell^{k+\frac{1}{2}}\right)u_\ell^{k}+c_\ell^{k+\frac{1}{2}}\left(\frac{1}{12c_{\ell+1}^{k+\frac{1}{2}}}+\frac{\mu}{2}\right)(u_{\ell-1}^{k}+u_{\ell+1}^{k}),$$

where $\mu := \Delta t/(\Delta x)^2$ and $c_j^\alpha := c(j\Delta x, \alpha\Delta t)$, is fifth-order in Δx (Mitchell and Griffiths, 1980). The last formula can be further generalized to several space variables (Gourlay and Mitchell, 1968), while retaining the same order.

The Crandall method and its generalizations, here and throughout this chapter, underpins the remark from Chapter 6 that there is more to full discretization than a conjunction of a semi-discretization and a solution of ordinary differential equations, since they cannot be represented in this fashion.

The high order of the Crandall method is perhaps counter-intuitive: Given $r = 1$, we have six coefficients at our disposal. One can be normalized, e.g. by setting $\gamma_{-1}+\gamma_0+\gamma_1 \equiv 1$, leaving out five. But, to obtain order 5, we need to annihilate six powers of Δx (starting from zero). Hence, the order of accuracy exceeds the number of degrees of freedom, and it makes sense to investigate whether this situation persists for other values of r. We will demonstrate in this chapter that this is, indeed, the case and prove that the resulting methods are unconditionally stable.

The function $\exp(\mu(\log z)^2)$ from Theorem 7.1 is nothing but innocent and its Taylor expansion is very intricate. Its first terms, listed in Table 7.1, were generated by a symbolic manipulator.

The derivation of explicit formulae that yield highest-order methods for $r \geq 2$ and all $\mu > 0$ is not a task for the faint-hearted. It is seemingly enough to find the numerator P and the denominator Q of the $[(2r)/(2r)]$ Padé approximant to $\exp(\mu(\log(1+z)^2))$: the characteristic function of the optimal scheme is of the form

$$H(z,\mu) := \frac{z^{-r}P(z-1)}{z^{-r}Q(z-1)}.$$

[5] We are saying here – and elsewhere in this book – absolutely nothing about the computational expense of solving sparse algebraic systems that occur in this procedure. It is perhaps in order to emphasize that the conceptual generalization to several dimensions might be easy but the cost of the solution becomes much more substantial as the dimensionality grows.

k	Taylor coefficients α_k:
0	1
1	0
2	μ
3	$-\mu$
4	$\frac{11}{12}\mu + \frac{1}{2}\mu^2$
5	$-\frac{5}{6}\mu - \mu^2$
6	$\frac{137}{180}\mu + \frac{17}{12}\mu^2 + \frac{1}{6}\mu^3$
7	$-\frac{7}{10}\mu - \frac{7}{4}\mu^2 - \frac{1}{2}\mu^3$
8	$\frac{363}{560}\mu + \frac{967}{480}\mu^2 + \frac{23}{24}\mu^3 + \frac{1}{24}\mu^4$
9	$-\frac{761}{1260}\mu - \frac{89}{40}\mu^2 - \frac{3}{2}\mu^3 - \frac{1}{6}\mu^4$
10	$\frac{7129}{12600}\mu + \frac{4523}{1890}\mu^2 + \frac{3013}{1440}\mu^3 + \frac{29}{72}\mu^4 + \frac{1}{120}\mu^5$
11	$-\frac{671}{1260}\mu - \frac{7645}{3024}\mu^2 - \frac{781}{288}\mu^3 - \frac{55}{72}\mu^4 - \frac{1}{24}\mu^5$
12	$\frac{83711}{166320}\mu + \frac{341747}{129600}\mu^2 + \frac{242537}{72576}\mu^3 + \frac{10831}{8640}\mu^4 + \frac{35}{288}\mu^5 + \frac{1}{720}\mu^6$
13	$-\frac{6617}{13860}\mu - \frac{412009}{151200}\mu^2 - \frac{48035}{12096}\mu^3 - \frac{299}{160}\mu^4 - \frac{13}{48}\mu^5 - \frac{1}{120}\mu^6$

Table 7.1 Coefficients in the Taylor expansion of the function $\sum_{k=0}^{\infty} \alpha_k (z-1)^k = \exp(\mu(\log z)^2)$.

Unfortunately, it is evident from Table 7.1 that complexity grows fast with r. Thus, $r = 2$, the simplest case beyond that corresponding to the Crandall method, yields

$$\begin{aligned} P(z,\mu) &= \left(\tfrac{1}{57600} - \tfrac{\mu^2}{1440} + \tfrac{\mu^4}{144}\right) + \left(\tfrac{1}{28800} - \tfrac{\mu^2}{720} + \tfrac{\mu^4}{72}\right) z \\ &\quad + \left(\tfrac{19}{907200} + \tfrac{\mu}{115200} - \tfrac{299\mu^2}{362880} - \tfrac{\mu^3}{2880} + \tfrac{7\mu^4}{864} + \tfrac{\mu^5}{288}\right) z^2 \\ &\quad + \left(\tfrac{13}{3628800} + \tfrac{\mu}{115200} - \tfrac{47\mu^2}{362880} - \tfrac{\mu^3}{2880} + \tfrac{\mu^4}{864} + \tfrac{\mu^5}{288}\right) z^3 \\ &\quad + \left(\tfrac{23}{217728000} + \tfrac{31\mu}{29030400} - \tfrac{71\mu^2}{43545600} - \tfrac{13\mu^3}{362880} - \tfrac{\mu^4}{25920} \right. \\ &\qquad \left. + \tfrac{\mu^5}{3456} + \tfrac{\mu^6}{1728}\right) z^4; \\ Q(z,\mu) &= P(z,-\mu). \end{aligned}$$

The expression for $r = 3$ is, unsurprisingly, even more complex. It is a tribute to the power of the order star technique that we are able in the remainder of this chapter to derive the highest order and determine stability of the underlying methods for all $r \geq 1$. We call the methods that attain

maximal order for a given choice of r the **optimal** methods.

The considerable intricacy of optimal methods is not, in principle, a deterrent to their applicability: all we need in a specific situation is the value of the coefficients for a fixed $\mu > 0$, and this can be evaluated more easily from the Taylor expansion than from any general formula. However, generalizations that take care of variable coefficients and higher dimensions are not available at present. Moreover, it is fair to say that, as far as finite difference methods are concerned, $r \geq 3$ is only of academic interest. Its sole promise lies in its possible use in spectral and pseudospectral methods.

In Chapter 6 we discussed a link between full discretizations and their associated semi-discretizations. This correspondence persists in the present framework – we leave the proof to the reader. Thus, differentiating the pth-order characteristic function $H(z,\mu)$ with respect to μ and setting $\mu = 0$ yields an approximant $h(z)$ to $(\log z)^2$ (at $z = 1$) of order $q \geq p$ and of the same 'size' r. Moreover, stability[6] is inherited by h.

As was the case with full discretizations, we can derive semi-discretizations of maximal order from $[(2r)/(2r)]$ Padé approximants, this time to $(\log(1+z))^2$. Again, we term them **optimal**. Denoting characteristic functions by $h = P/Q$, $r = 1$ yields

$$\begin{aligned} P(z) &= z^2; \\ Q(z) &= 1 + z + \frac{1}{12}z^2 \end{aligned}$$

and order 5, whereas $r = 2$ produces a 9th order approximant

$$\begin{aligned} P(z) &= z^2 + z^3 + \frac{31}{252}z^4; \\ Q(z) &= 1 + 2z + \frac{76}{63}z^2 + \frac{13}{63}z^3 + \frac{23}{3780}z^4. \end{aligned}$$

Again, the order exceeds the number of degrees of freedom.

Semi-discretizations are, needless to say, interesting in their own right. More important from our point of view is that their study facilitates the understanding of full discretizations.

7.2 Padé approximants to $f((\log z)^2)$

In the present and the next sections we anticipate a central technique of Chapter 8 in a much simplified framework: we bound the order by studying the structure of the Padé tableau. Our ultimate goal is to investigate how well we can approximate $\exp(\mu(\log z)^2)$ or $(\log z)^2$ with $2r$-by-$2r$ rational approximants. The relevance to the design of optimal methods for

[6] Stability of a semi-discretized scheme is consistent with all the conditions of Lemma 6.7, with $\bar{r} \equiv r$.

the diffusion equation (7.1) rests upon Theorem 7.1 and its semi-discrete 'equivalent'. Our analysis throughout the sequel of this chapter follows in the footsteps of Iserles (1985c).

Let f be an entire function, $f \not\equiv 0$, and set

$$g(z) := f((\log z)^2). \tag{7.6}$$

Thus, g is a slit function and is analytic at $z = 1$. Moreover, it is easy to verify that it obeys the functional equation

$$g(z) = g\left(\frac{1}{z}\right), \quad |z-1| < 1. \tag{7.7}$$

As a matter of interest, we point out that for any g that obeys (7.7), is analytic away from the slit $(-\infty,0)$ and entire in the covering Riemann surface, there exists an entire f so that (7.6) holds. This can be seen at once by setting $\tilde{f}(z) := g(e^z)$. The function $\tilde{f}$ is entire and it follows from (7.7) that it is even. Hence its Taylor expansion is of the form

$$\sum_{k=0}^{\infty} f_k z^{2k}.$$

We now set $f(z) := \sum_{k=0}^{\infty} f_k z^k$. Both (7.6) and analyticity follow at once.

Proposition 7.2 Let P and Q be two polynomials in $\pi_n[z]$ such that $Q(1) = 1$ and

$$R(z) := \frac{P(z)}{Q(z)} = g(z) + C(z-1)^{p+1} + \mathcal{O}(|z-1|^{p+2}), \quad c \neq 0 \tag{7.8}$$

for some $p \geq 2n$. Then necessarily p is odd.

Proof. Set $P^*(z) := z^n P(z^{-1})$ and $Q^*(z) := z^n Q(z^{-1})$. It is a consequence of (7.7) that

$$R\left(\frac{1}{z}\right) = \frac{P^*(z)}{Q^*(z)} = g(z) + C(1-z)^{p+1} + \mathcal{O}(|z-1|^{p+1}). \tag{7.9}$$

Since $Q(1) = Q^*(1) = 1$, cross-multiplying (7.8) and (7.9) yields

$$P(z)Q^*(z) - P^*(z)Q(z) = \mathcal{O}(|z-1|^{p+1}). \tag{7.10}$$

The inequality $p \geq 2n$, in unison with the identity (7.10), implies $PQ^* - P^*Q \equiv 0$. We know that $Q \not\equiv 0$ and $f \not\equiv 0$ implies that $P \not\equiv 0$. Consequently, $P/Q \equiv P^*/Q^*$ and we have

$$R(z) = R\left(\frac{1}{z}\right). \tag{7.11}$$

Comparing (7.8) with (7.9) we notice that all the terms there coincide, except that the error constant in (7.9) is multiplied by $(-1)^{p+1}$. For this to be true, p must be odd. □

In Chapter 8 we introduce the Padé tableau and investigate in detail its structure by using order stars. Presently we require a single result that follows easily from Theorem 8.1:

Proposition 7.3 Let the order $p(n)$ of every $[n/n]$ Padé approximant be bounded by $2n+1$, $n \geq 0$. Moreover, assume that if $p(n) \geq 2n$ then it is necessarily odd. Then the following alternative holds: either $p(2n) = p(2n+1) = 4n+1$ for all $n \geq 0$ or $p(2n-1) = p(2n) = 4n-1$ for all $n \geq 1$. □

7.3 Order of optimal methods

Equipped with Propositions 7.2 and 7.3, we focus our discussion on the functions $f(z) = z^2$ and $f(z) = \exp\left(\mu^{\frac{1}{2}} z\right)$. In this section we derive an upper bound that allows us to deduce that the order of accuracy of the optimal scheme (7.3) is $4r+1$, possibly except for a finite subset of μ's. Along the way we prove a similar result for semi-discretizations.

Lemma 7.4 No $\pi_{n/n}$ function may approximate $(\log z)^2$ at $z = 1$ to order exceeding $2n+1$.

Proof. We define an order star of the second kind by letting $\tilde{\rho}(z) := h(e^z) - z^2$, where the rational function $h \in \pi_{n/n}$ approximates $(\log z)^2$ up to $\mathcal{O}(|z-1|^{p+1})$ at $z = 1$. Note that transforming $z \mapsto \log z$ maps the approximation point to the origin. Moreover, finite essential singularities (i.e. the mapped poles of $R(z)$) are periodic with period $2\pi i$.

As usual in order star proofs, we first examine thoroughly a few plots of order stars. These are presented in Figure 7.1 for

$$h(z) = \frac{(z-1)^2}{1 + (z-1) + \frac{1}{12}(z-1)^2}$$

and

$$h(z) = \frac{(z-1)^2 + (z-1)^3 + \frac{31}{252}(z-1)^4}{1 + 2(z-1) + \frac{76}{63}(z-1)^2 + \frac{13}{63}(z-1)^3 + \frac{23}{3780}(z-1)^4},$$

of orders 5 and 9 respectively. Next we assemble (and prove as necessary!) all the observations that are required to bound the number of sectors that approach the origin:

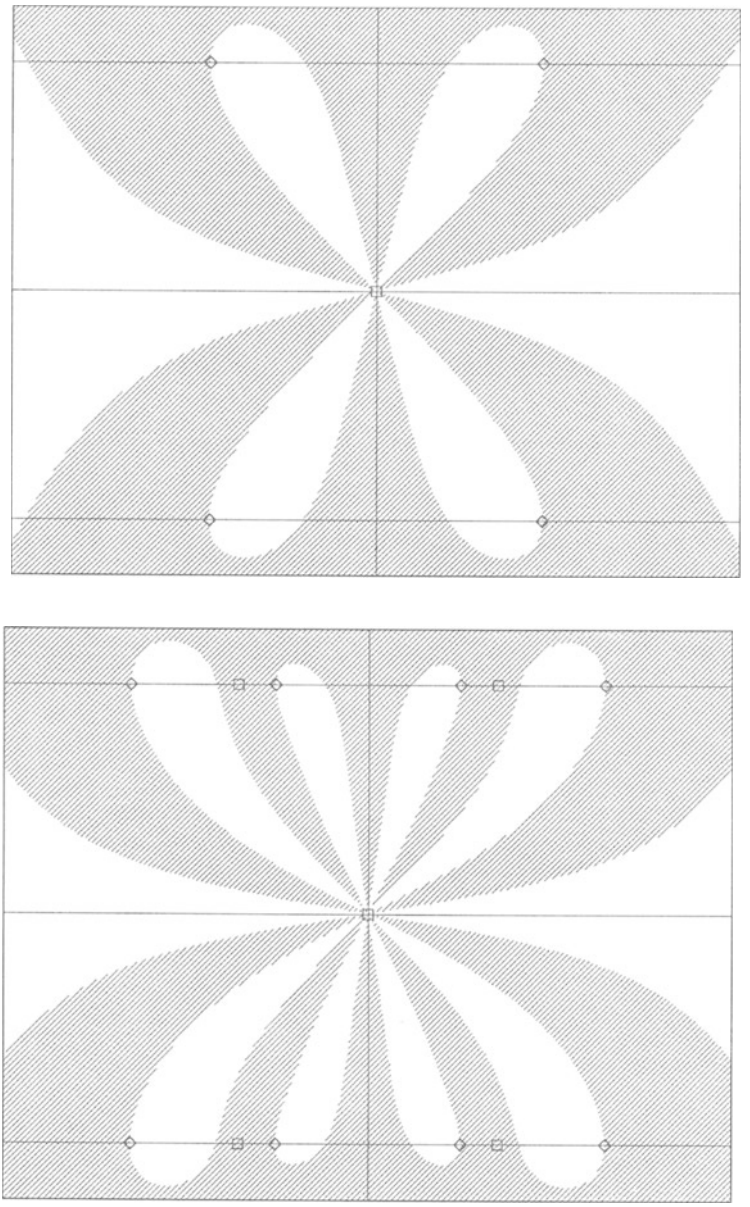


Figure 7.1 Order stars of the second kind for optimal semi-discretizations, $r = 1,2$. The top and bottom horizontal lines denote the boundary of $\mathcal{S}$.

(a) If $h(z)$ is bounded away from 0 and ∞ as $|z| \to 0, \infty$, then $\iota(\infty) = 2$, the point is regular and the real axis is asymptotically covered by an $\tilde{\mathcal{A}}_-$-region. Otherwise the index might decrease to 1 but it may never exceed 2.

(b) Suppose that $z \in \tilde{\mathcal{A}}_+$ and $\operatorname{Im} z \geq -\pi$. Then

$$|\tilde{\rho}(z + 2\pi i)| = |\tilde{\rho}(z)| + 4\pi(\pi + \operatorname{Im} z),$$

consequently $z + 2\pi i \in \tilde{\mathcal{A}}_+$.

(c) Symmetry with respect to the real axis and (b) imply that

$$z \in \tilde{\mathcal{A}}_+, \ \operatorname{Im} z < \pi \quad \Rightarrow \quad z - 2\pi i \in \tilde{\mathcal{A}}_+.$$

(d) Let $x + (\pi - \nu)i \in \tilde{\mathcal{A}}_+$, where $0 < \nu \leq \pi$. By (c) it follows that $x - (\pi + \nu)i \in \tilde{\mathcal{A}}_+$. Thus, by symmetry with respect to the real axis, $x + (\pi + \nu)i \in \tilde{\mathcal{A}}_+$. This provides the theoretical underpinning to the impression, evident from Figure 7.1, that $\tilde{\mathcal{A}}_+$ 'grows' as we increase the imaginary part away from the strip $\mathcal{S} = \{z \in \mathcal{C} : |\operatorname{Im} z| \leq \pi\}$.

(e) Let Γ be a bounded $\tilde{\mathcal{A}}_-$-loop that adjoins the origin. According to Proposition 2.11 there must be an essential singularity along Γ. Suppose that it lies outside $\mathcal{S}$. By symmetry, and without loss of generality, we may assume that it is of the form $x^* + (\pi + \nu^*)i$, where $\nu^* > 0$. Thus, by (d), the essential singularity

$$\hat{z} := x^* + \left(\pi\left(\left[\frac{\nu^*}{\pi}\right] + 1\right) - \nu^*\right)$$

also lies on Γ. But $\hat{z}$ lies in $\mathcal{S}$ and we have demonstrated that, although essential singularities are repeated infinitely, only those in $\mathcal{S}$ play a role in counting bounded $\tilde{\mathcal{A}}_-$-loops that adjoin the origin.

(f) Suppose that there exists an $\tilde{\mathcal{A}}_+$-region, $\mathcal{U}$, say, that adjoins the origin and extends beyond $\mathcal{S}$. By symmetry, we may confine our attention to the upper half-plane. It follows that there exists a continuous curve $\{\Gamma(\tau) : 0 \leq \tau \leq 1\}$, such that $\Gamma(0) = 0$ and $\Gamma(\tau) \in \mathcal{U}$ for all $\tau \in (0, 1]$. We use (d) to argue that the reflection of this curve about $\operatorname{Im} z = \pi$ also belongs to $\mathcal{U}$. Denote by Γ^* the curve together with its reflection: as a point-set we have $\Gamma^* = \Gamma \cup \{\bar{\Gamma} + 2\pi i\}$. In particular, Γ^* extends continuously up to the point $2\pi i$, which belongs to $\tilde{\mathcal{A}}_+$ by the definition of the order star. Next we use (b) inductively to argue that replicas of Γ^* mapped by $2\pi K i$ for $K \geq 1$ also belong to $\mathcal{U}$. It follows that $\mathcal{U}$ cannot be bounded. In other words, no bounded $\tilde{\mathcal{A}}_+$-region may extend beyond $\mathcal{S}$.

The remainder of the proof is straightforward: we count $\tilde{\mathcal{A}}_-$-loops that adjoin the origin. Suppose that the origin is adjoined by L bounded $\tilde{\mathcal{A}}_+$-

loops. Thus, according to (a), at most $L + 2$ $\tilde{A}_-$-loops there may be unbounded. Moreover, let us suppose that K bounded $\tilde{A}_-$-loops touch 0. According to (e) and (f), all the bounded loops at the origin must be 'supported' by essential singularities in $\mathcal{S}$. As in Section 6.3, we may double the number of available poles, since, by periodicity, negative poles are mapped to both $\mathrm{Im}\, z = \pi i$ and $\mathrm{Im}\, z = -\pi i$. Thus, we have $K + L \leq 2n$, hence

$$p \leq \iota(0) - 1 \leq 2n + 1, \tag{7.12}$$

and the proof follows. □

Corollary The order of the $[2r/2r]$ Padé approximant to $(\log z)^2$ (about $z = 1$) is precisely $4r + 1$ for all $r \geq 0$.

Proof. Putting together the statements of Propositions 7.2 and 7.3 and of Lemma 7.4, we deduce the following alternative: either the present corollary holds or the $[(2r + 1)/(2r + 1)]$ Padé approximants are of order $4r + 3$ for all $r \geq 0$. The second option is ruled out by examining the simplest case, $r = 0$. □

Another simple consequence of the lemma is of no immediate concern to us but is, none the less, interesting within the framework of the Padé theory:

Corollary All the poles of the $[2r/2r]$ Padé approximant to $(\log z)^2$ lie along the branch cut $(-\infty, 0)$ and are distinct.

Proof. Inequality (7.12) becomes an equality to allow order $2n + 1$. Therefore, it follows from the proof of Lemma 7.4 that all the poles of h must be 'efficient'. In other words, bearing in mind the definition of the order star, they are negative and distinct. □

The connection between a fully discretized scheme (7.3) and its associated semi-discretization is similar to that in Lemma 6.3, with obvious corrections. In particular, the order of the associated semi-discretized scheme is at least that of the fully discretized one. Of course, the order of a fully discretized scheme depends on μ – the last statement merely implies that, in the case of a maximal-order method, it is at most $4r + 1$ for sufficiently small $\mu > 0$. In other words, there exists a μ^* such that for all $\mu \in (0, \mu^*)$

$$H(z, \mu) = e^{\mu(\log z)^2} + C(\mu)(z - 1)^{4r+2} + \mathcal{O}(|z - 1|^{4r+3}, \quad C(\mu) \neq 0.$$

It can be easily proved that the Taylor coefficients of the approximated function are polynomials in μ – in the next section we will take a close interest in these coefficients, but for the time being we will simply refer to Table 7.1. Thus, it follows from the representation of Padé approximants as a quotient of two determinants (Baker, 1975) that $C(\mu)$ is rational in

μ. In particular, it can be extended analytically to all $\mu > 0$ (in fact, to all complex μ), except for a finite set of points. Likewise, the coefficients of Padé approximants are rational in μ and can be extended analytically, except that, normalizing $Q(0) = 1$, they may become unbounded at a finite number of points. This extends (with a *caveat*) the statement of Lemma 7.4 to $\exp(\mu(\log z)^2)$.

Theorem 7.5 The maximal order that is attainable by the fully discretized method (7.3) is $4r + 1$, except for a finite subset of $\mu \in \mathcal{C}$.

Proof. We proceed exactly as in the corollary to Lemma 7.4. □

Example 7.1 Given $r = 1$, the error constant is

$$C(\mu) = -\frac{\mu}{12}\left(\mu^2 - \frac{1}{20}\right),$$

hence the order exceeds 5 for $\pm\sqrt{5}/10$ – actually, it goes up to 7 (cf. Figure 7.2). The investigation of $r = 2$ is more complicated: the error constant is

$$C(\mu) = \frac{\mu}{103680}\left(\mu^2 - \frac{1}{20}\right)\left(\mu^2 - \frac{1214}{44025}\right)\left(\mu^4 - \frac{3930}{14427}\mu^2 + \frac{367}{16950}\right)$$

and it vanishes at eight points – two positive, two negative and four complex – except $\mu = 0$ which, of course, does not count. However, unlike the case of $r = 1$ (when order $p \geq 5$ persists for all $\mu \neq 0$), the coefficients of the [4/4] approximant fail when $\mu = \pm\sqrt{5}/10$. At this point the approximant 'collapses' to a seventh-order $\pi_{2/2}$ function. ◇

The procedure of considering an order star of the associated scheme to investigate the order of a full discretization might seem rather indirect and convoluted. Nevertheless, it makes perfect sense. Firstly, it allows two results – for semi-discretizations, as well as for full discretizations – for the price of one. Secondly (and more importantly), proving Lemma 7.4 for $\exp(\mu(\log z)^2)$ is a much more formidable task. Consider the μ-dependent order star of the first kind with

$$\rho(z) := e^{-\mu z^2} H(e^z, \mu),$$

where H is a rational approximant. Figure 7.2 displays the case $n = 2$ with four distinct values of $\mu > 0$: $\mu = \frac{1}{10}$ produces a picture similar to that in Figure 7.1 – not very surprising, considering the asymptotic relationship between the two order stars. At $\mu = \frac{1}{6}$ the [2/2] approximant reduces to a [2/0] approximant. In other words, the Crandall method becomes explicit. The value $\mu = \frac{\sqrt{5}}{10}$ yields order 7. Finally, at $\mu = \frac{1}{4}$ the order star has already settled to its 'large μ asymptotics': not much will change qualitatively as we increase μ further. Figure 7.2, with its special values of μ, demonstrates

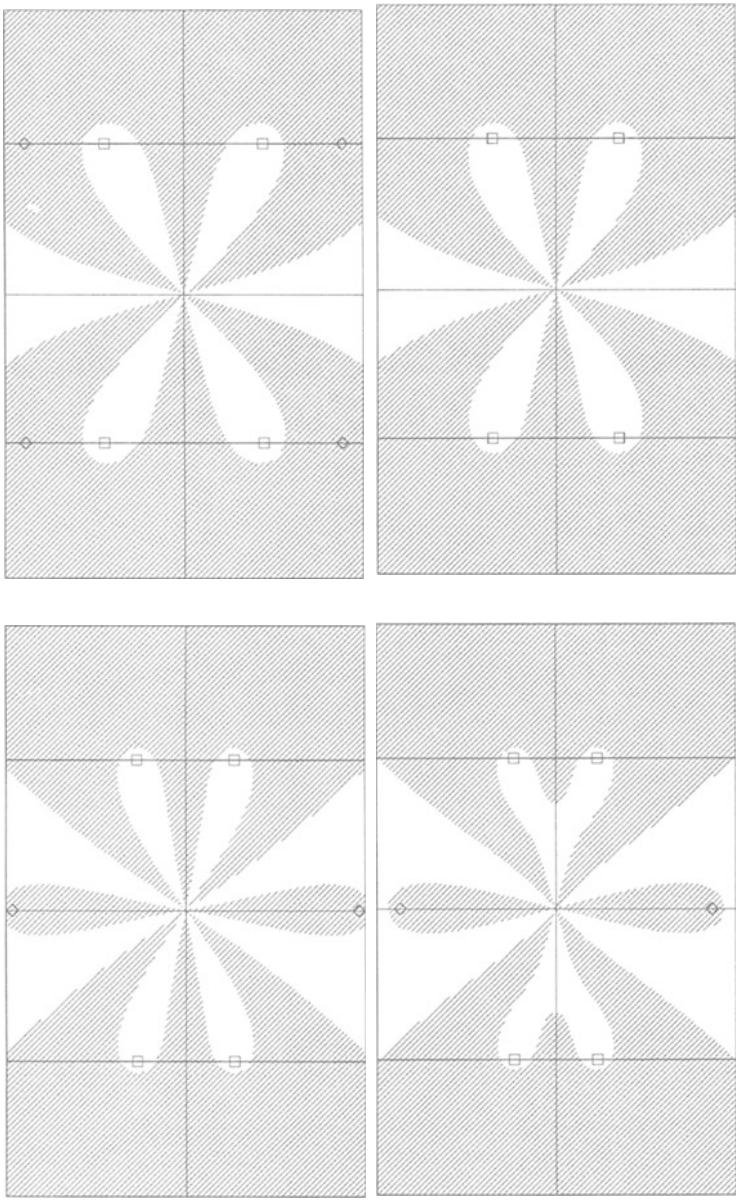


Figure 7.2 Order stars of the first kind for the Crandall method for different values of μ.

that it is considerably more difficult to work with the underlying order star and vindicates the approach adopted in Lemma 7.4.

There is one exception to the statement that the μ-dependent order star is more complicated, namely when μ is pure imaginary. This case is of intrinsic interest, since (7.1) with imaginary c is a linear version of the Schrödinger equation. Perhaps more importantly, its consideration leads to a result that is necessary to the stability analysis of Section 7.4.

Lemma 7.6 The maximal order that is attainable by the fully discretized method (7.3) with $\mathrm{Re}\,\mu = 0$ is exactly $4r + 1$.

Proof. We investigate the order star of the first kind with respect to $\rho(z) := e^{-i\nu z^2} H(e^z, i\nu)$, where $H = P/Q$ is a $\pi_{(2n)/(2n)}$ function and

$$H(z, i\nu) = e^{i\nu(\log z)^2} + \mathcal{O}(|z-1|^{p+1}), \quad , p \geq 4n$$

(cf. Figure 7.3, last plot). Letting $z = e^{i\theta}$ in the last expression gives

$$|P(e^{i\theta}, i\nu)|^2 - |Q(e^{i\theta}, i\nu)|^2 = \mathcal{O}(\theta^{p+1}) = \mathcal{O}((1-\cos\theta)^{[p/2]+1}).$$

But both $|P|^2$ and $|Q|^2$ are $(2n-1)$-degree polynomials in $(1-\cos\theta)$ and $2n - 1 \leq [p/2] + 1$, consequently

$$|P(e^{i\theta}, i\nu)| \equiv |Q(e^{i\theta}, i\nu)|, \quad \theta \in \mathcal{R}.$$

The order condition implies that

$$|P(e^x, i\nu)|^2 - |Q(e^x, i\nu)|^2 = \mathcal{O}(x^{p+1}) = \mathcal{O}((e^x - 1)^{p+1})$$

for real x and similar argument implies that

$$|P(e^x, i\nu)|^2 \equiv |Q(e^x, i\nu)|, \quad x \in \mathcal{R}.$$

Therefore

$$\mathcal{R}, i\mathcal{R} \in \mathcal{A}_0. \tag{7.13}$$

Incidentally, this reaffirms the result of Proposition 7.2 in the present context, namely that p is odd, by counting sectors at the origin.

We first note that

$$|\rho(x + 2\pi i)| = e^{4\pi x}|\rho(x)|, \quad x \in \mathcal{R}.$$

Thus, (7.13) implies

$$x + 2\pi i \in \begin{cases} \mathcal{A}_+ & : x > 0, \\ \mathcal{A}_- & : x < 0. \end{cases} \tag{7.14}$$

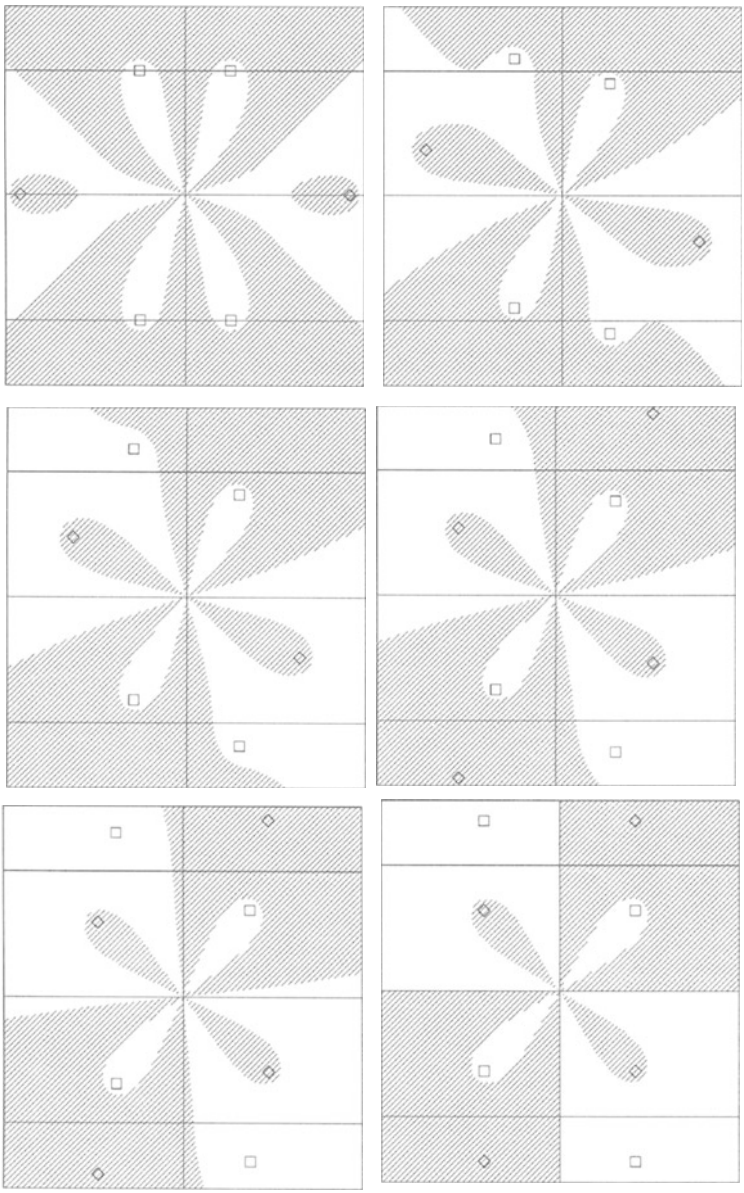


Figure 7.3 Order stars of the first kind for optimal methods with $r = 1$ and $\mu = \frac{1}{5}\exp\left(\frac{i\pi(k-1)}{10}\right)$, $k = 1,\ldots,6$. The last figure displays $\mathrm{Re}\,\mu = 0$.

Next we observe that, by Proposition 2.6 and since ρ has an essential singularity of exponential type 2 at ∞, $\iota(\infty) = 2$ and ∞ is regular. Furthermore, assuming, without loss of generality, that $\nu > 0$, we have

$$|\rho(z)| = e^{2\nu(\mathrm{Re}\, z)(\mathrm{Im}\, z)}(1 + o(1)), \quad |z| \gg 0,$$

hence $z \in \mathcal{C}$ belongs, far away from the origin, to $\mathcal{A}_+$ in the $(+,+)$ and $(-,-)$ quadrants and to $\mathcal{A}_-$ in the $(+,-)$ and $(-,+)$ quadrants.[7] In particular, (7.13) implies that all the $\mathcal{A}_-$-regions in the $(\pm,\pm)$ quadrants and all $\mathcal{A}_+$-regions in the $(\pm,\mp)$ quadrants are bounded. We focus our attention on sectors that adjoin the origin and belong to bounded regions. According to (7.14), these cannot 'utilize' zeros or poles that are repeated, by periodicity, outside $\mathcal{S}$. Since all bounded regions are analytic, we can have at most $2n$ sectors of $\mathcal{A}_-$ approaching the origin in the $(\pm,\pm)$ quadrants and at most $2n+2$ in the $(\pm,\mp)$ quadrants (since they envelop at most $2n$ sectors of $\mathcal{A}_+$). The inequality $p \leq 4n+1$ follows at once from Proposition 2.1.

We complete the proof by proceeding exactly as in the proof of the first corollary to Lemma 7.4. □

7.4 Stability of optimal methods

The differential equation (7.1) with complex constant c is ill posed for $\mathrm{Re}\, c < 0$, conservative for $\mathrm{Re}\, c = 0$ (the linearized Schrödinger equation) and dissipative for $\mathrm{Re}\, c > 0$ (in particular for $c > 0$, the diffusion equation).

Stability of optimal methods is easy when c, hence μ, is imaginary. It follows from the proof of Lemma 7.6 that the characteristic function is itself conservative – it equals 1 in modulus along the unit circle. Moreover, since all the poles are required to 'support' bounded sectors of $\mathcal{A}_+$ that adjoin the origin, symmetry implies that they split equally between the interior and the exterior of $|z| = 1$ and stay clear of the perimeter. Stability follows by Theorem 7.1.

Time and again we have encountered the situation whereby conservative schemes approximating conservative equations are relatively easy to analyse: diagonal Padé approximants to $\exp z$, conservative methods for the advection equation, etc. Dissipativity presents more of a challenge, *inter alia* since Proposition 2.2 is more difficult to interpret and apply.

To prove stability of the optimal methods for all $\mathrm{Re}\, \mu \geq 0$ we proceed in a manner which is at odds with the usual approach to stability analysis. It is standard to fix μ and prove that the magnitude of the characteristic function H is bounded by unity along the perimeter of the unit disc. In this section we fix a point $e^{i\theta}$ and prove that $|H(e^{i\theta}, \mu)|$ is bounded uniformly in $\mathrm{Re}\, \mu \geq 0$.

[7] It makes sound sense to follow various features of the order star, as they unravel in our exposition, in the bottom right order star displayed in Figure 7.3.

Let $n = 2r$ and $H(z,\mu)$ approximate $\exp(\mu(\log z)^2)$ to order $2n+1$ at $z = 1$. Thus, we have

$$H(e^{i\theta},\mu) = e^{-i\theta^2} + \mathcal{O}(\theta^{2n+2})$$

and

$$\frac{1}{H(e^{i\theta},-\mu)} = e^{-i\theta^2} + \mathcal{O}(\theta^{2n+2}).$$

Letting $H = P/Q$, we obtain

$$\begin{aligned} P(e^{i\theta},\mu)P(e^{i\theta},-\mu) - Q(e^{i\theta},\mu)Q(e^{i\theta},-\mu) &= \mathcal{O}(\theta^{2n+2}) \\ &= \mathcal{O}((1-e^{i\theta})^{2n+2}). \end{aligned}$$

Thus, $P(e^{i\theta},\mu)P(e^{i\theta},-\mu) - Q(e^{i\theta},\mu)Q(e^{i\theta},-\mu)$ being in $\pi_{2n}[1-e^{i\theta}]$, degree considerations imply the identity

$$P(e^{i\theta},\mu) = Q(e^{i\theta},-\mu). \tag{7.15}$$

According to Theorem 7.5, order $2n+1$ is assured for all $\mu \in \mathcal{C}$, except, possibly, for an exceptional set. We can readily extend (7.15) to that set (except $\mu = 0$) by noting that the coefficients of P and Q are rational functions of μ.

To stress the dependence of H upon μ we write

$$\Lambda_\theta(\mu) := H(e^{i\theta},\mu), \quad \mu \in \mathcal{C},\ |\theta| \le \pi.$$

The identity (7.15) implies that

$$\Lambda_\theta(\mu) = \frac{\lambda_\theta(\mu)}{\lambda_\theta(-\mu)},$$

where λ_θ is a polynomial in $e^{i\theta}$.

We need to investigate the structure of λ_θ as a function of μ. This necessitates taking a closer look at the Taylor coefficients of the function $g(z) = \exp(\mu(\log z)^2)$. 'Taking a closer look' is a knowingly vague statement: we mean here both an examination of Table 7.1, to provide intuition, and formal proofs to anchor it in mathematical certainty.

Proposition 7.7 Let $g(z) = \sum_{k=0}^{\infty} \alpha_k(\mu)(z-1)^k$. Then each α_k is a polynomial of degree $[k/2]$ in μ.

Proof. Since

$$g'(z) = \frac{2\mu \log z}{z} g(z), \tag{7.16}$$

it follows by repeated differentiation that

$$\frac{d^k}{dx^k} g(z) = \frac{1}{z^k} \left(\sum_{\ell=0}^{k} \beta_\ell^{(k)} (\log z)^\ell \right) g(z), \tag{7.17}$$

where the $\beta_\ell^{(k)}$'s obey for all $k = 0, 1, \ldots$ the recurrence relation

$$\begin{aligned} \beta_0^{(k+1)} &= -k\beta_0^{(k)} + \beta_1^{(k)}; \\ \beta_\ell^{(k+1)} &= -2\mu\beta_{\ell-1}^{(k)} - k\beta_\ell^{(k)} + (\ell+1)\beta_{\ell+1}^{(k)}, \quad 1 \le \ell \le k; \\ \beta_{k+1}^{(k+1)} &= 2\mu\beta_k^{(k)}. \end{aligned} \tag{7.18}$$

Moreover, $\beta_0^{(0)} \equiv 1$, and it follows at once from (7.16) that $\beta_\ell^{(k)} \in \pi_{[(k+\ell)/2]}[\mu]$ for all $k, \ell \ge 0$. In particular, letting $z = 1$ in (7.17) affirms that $\alpha_k = k!\beta_0^{(k)}$ is of the stipulated degree. □

Corollary The degree of each α_k is exactly $[k/2]$.

Proof. We need to demonstrate that the coefficient of $\mu^{[k/2]}$ does not vanish. Recall that each $\beta_\ell^{(k)}$ is a polynomial of degree $[(k+\ell)/2]$ and denote the coefficient of $\mu^{[(k+\ell)/2]}$ therein by $\chi_\ell^{(k)}$. It is a consequence of (7.18) that

$$\chi_\ell^{(k+1)} = \begin{cases} 2\chi_{\ell-1}^{(k)} - k\chi_\ell^{(k)} + (\ell+1)\chi_{\ell+1}^{(k)} & : k+\ell \text{ even}; \\ 2\chi_{\ell-1}^{(k)} + (\ell+1)\chi_{\ell+1}^{(k)} & : k+\ell \text{ odd}. \end{cases}$$

We now prove that $(-1)^{k+\ell}\chi_\ell^{(k)} > 0$: proceeding by induction on $k+\ell$, we have

$$\begin{aligned} \chi_{k-2j}^{(k+1)} &= -\Big(2|\chi_{k-1-2j}^{(k)}| + k|\chi_{k-2j}^{(k)}| \\ &\qquad + (k-2j+1)|\chi_{k+1-2j}^{(k)}|\Big) < 0 \end{aligned}$$

for all $j = 0, \ldots, [k/2]$ and

$$\chi_{k+1-2j}^{(k+1)} = 2|\chi_{k-2j}^{(k)}| + (k-2j+2)|\chi_{k-2j+2}^{(k)}| > 0$$

for all $j = 0, \ldots, [(k+1)/2]$. The corollary follows, since the coefficient of $\mu^{[k/2]}$ in α_k is $k!\chi_0^{(k)}$. □

It is known from the theory of Padé approximants (Baker, 1975) that $\lambda_\theta(\mu) = Q(e^{i\theta}, -\mu)$ equals

$$\det \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_n & \alpha_{n+1} \\ \alpha_2 & \alpha_3 & \cdots & \alpha_{n+1} & \alpha_{n+2} \\ \vdots & \vdots & & \vdots & \vdots \\ \alpha_n & \alpha_{n+1} & \cdots & \alpha_{2n-1} & \alpha_{2n} \\ e^{-in\theta} & e^{-i(n-1)\theta} & \ldots & e^{-i\theta} & 1 \end{bmatrix}.$$

Let $\lambda_\theta(\mu) = \sum_{k=0}^n q_k(\mu)e^{-ik\theta}$. The coefficients q_k can be derived by expanding the determinant in the bottom row. It follows that each q_k is a

polynomial. The range of powers of μ present in q_k (i.e. the value of the highest and the lowest power there) is of importance to our analysis. Recalling that $n = 2r$, we have

$$\begin{aligned} q_0 &= \det \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_n \\ \alpha_2 & \alpha_3 & \cdots & \alpha_{n+1} \\ \vdots & \vdots & & \vdots \\ \alpha_n & \alpha_{n+1} & \cdots & \alpha_{2n-1} \end{bmatrix} \\ &= C_1 \mu^{2r^2} + \text{lower-order terms.} \\ &= C_2 \mu^n + \text{higher-order terms..} \end{aligned}$$

This follows since, by Proposition 7.7, $\alpha_k \in \pi_{[k/2]}$ and it is easy to prove that the highest-order coefficient is obtained by expanding along the anti-diagonal: the highest power of the (k, ℓ) element of the determinant is $a_{k,\ell} = [(k + \ell - 1)/2]$. Thus, the highest-order coefficient in q_0 is

$$a^\star := \max \sum_{k=1}^{n} \left[\frac{k + j_k - 1}{2} \right],$$

where the n-tuple $(j_1, j_2, \ldots, j_n)$ ranges over all the $n!$ permutations of $(1, 2, \ldots, n)$. Clearly,

$$\begin{aligned} a^\star &\leq \max \left\{ \frac{1}{2} \sum_{k=1}^{n} (k + j_k + 1) : j_1 + j_2 + \cdots + j_n = \frac{1}{2} n(n+1) \right\} \\ &= \frac{1}{2} n^2 = 2r^2. \end{aligned}$$

The power $2r^2$ is attainable in $a^\star$ by choosing $j_k = n + 1 - k$, $k = 1, \ldots, n$. Moreover, it is easy to see that any other permutation yields a smaller sum, because it leads to chopping due to taking an integer value – the terms in the numerator are no longer even. Since the highest power is obtained from a unique permutation, it follows from the corollary to Proposition 7.7 that $C_1 \neq 0$. Finally, each row in the determinant is a multiple of μ, hence the lowest-order term is $\mathcal{O}(\mu^n)$.

Similar reasoning extends to q_k for all $k = 0, \ldots, n$. Specifically,

$$\begin{aligned} q_k &= \chi_0^{(1)} \chi_0^{(3)} \cdots \chi_0^{(2(n-k)-1)} \chi_0^{(2(n-k+1))} \cdots \chi_0^{(2(n-1))} \chi_0^{(2n)} \\ &\quad \times \mu^{r(2r-1)+k} + \text{lower-order terms.} \end{aligned}$$

Thus, q_k is of exact degree $r(2r-1) + k$ for all $k = 1, \ldots, n$. Likewise, it is easy to see that the least power of μ present in the expression is n.

The factor of μ^n repeats itself in both the numerator and the denominator and can be factored out. Easy reasoning leads to our next result:

Proposition 7.8 The function Λ_θ belongs to $\pi_{m/m}[\mu]$, where $m = r(2r-1) = \frac{1}{2}(n-1)n$. Moreover, the coefficients of μ^m in the numerator and the denominator do not vanish. □

In Chapters 3 and 4 we devoted much attention to Padé approximants to the exponential. In particular, as a consequence of Theorem 4.1, all the poles of the $[m/m]$ diagonal approximant $R_{K/K}$ reside in the right half-plane and all the zeros lie to the left of $i\mathcal{R}$. Combining

$$\Lambda_\theta(\mu) = e^{-\mu\theta^2} + \mathcal{O}(\theta^{2n+2})$$

and

$$R_{m/m}(-\mu\theta^2) = e^{-\mu\theta^2} + \mathcal{O}(\theta^{2m+2})$$

for small $|\theta|$, we obtain

$$\Lambda_\theta(\mu) = R_{m/m}(-\mu\theta^2) + \mathcal{O}(\theta^{2\min\{n,m\}+1}). \tag{7.19}$$

Suppose that ξ° is a zero of $R_{m/m}$ and set $\mu^\circ = -\xi^\circ/\theta^2$. Consequent on (7.19), we have $\lambda_\theta(\mu^\circ) = o(\theta)$: as $|\theta| \downarrow 0$, the zeros of $\lambda_\theta(\mu)$ and of $R_{m/m}(-\mu\theta^2)$ become arbitrarily close. The choice of m and Proposition 7.8 imply that the degrees of λ_θ (as a polynomial in μ) and of the numerator of $R_{m/m}$ coincide. Thus, the aforementioned statement is valid for all the zeros of λ_θ.

Proposition 7.9 The function Λ_θ is analytic for all $|\theta| \leq \pi$ and $\operatorname{Re}\mu \geq 0$.

Proof. We have $\operatorname{Re}\xi^\circ > 0$, therefore $\operatorname{Re}\mu^\circ < 0$ for small $|\theta| > 0$. Consequently, at least for $0 < |\theta| \ll 1$ all the poles of the rational function Λ_θ are in the left half-plane and it is analytic for $\operatorname{Re}\mu > 0$. It remains to extend the proof to all θ within the range.

Suppose that there exists $\theta_0 \in (0,\pi)$, say, such that Λ_{θ_0} has a pole in the right half-plane. Since the degree of g_θ is constant and independent of θ, no pole may approach infinity (unless $\theta \to 0$, which is not the case!) and the poles are continuous in the parameter θ. In other words, there exists $\theta_1 \in (0,\theta_0)$ such that Λ_{θ_1} has an imaginary pole. We now invoke Lemma 7.6. Since $\Lambda_\theta(i\nu) \equiv H(e^{i\theta}, i\nu)$, the proof of the lemma implies that $|\Lambda_\theta(i\nu)| \equiv 1$ for all $\nu \in \mathcal{R}$. In particular, Λ_θ may not become unbounded along the pure imaginary axis, and this rules out a pole there – unless it coalesces with a zero. But in that case $\Lambda_\theta \in \pi_{s/s}[\mu]$ for some $s \leq m-1$. Moreover, in addition the degree of Λ_θ as a rational function of $e^{i\theta}$ is lowered, as can be seen readily from the argument that led to Proposition 7.8. This contradicts the statement of Lemma 7.6. Therefore no such θ_1 may exist and the proposition is true. □

The scene has now been set for the final result of this chapter, the stability proof for optimal methods.

Theorem 7.10 The optimal methods are stable for all $r \geq 1$ and $\mathrm{Re}\,\mu \geq 0$.

Proof. According to Proposition 7.9 the function Λ_θ is analytic for all 'unexceptional' μ's in the closed right half-plane. Being a rational function, it remains analytic also at the (finite) number of 'exceptional' μ's there. Moreover, $|\Lambda_\theta(i\nu)| \equiv 1$ for all $\nu \in \mathcal{R}$. We now apply the maximal modulus principle (in the μ-plane) to argue that

$$|H(e^{i\theta}, \mu)| \equiv |\Lambda_\theta(\mu)| < 1, \quad \mathrm{Re}\,\mu > 0. \tag{7.20}$$

Since this is true for every $|\theta| \leq \pi$, Theorem 7.1 affirms stability. □

An interesting consequence of inequality (7.20) is that optimal methods are dissipative for $\mathrm{Re}\,\mu > 0$: the energy is not just uniformly bounded (as is required by stability) but it is, actually, decreased as the time evolves. This is, of course, entirely in line with the behaviour of the exact solution of equation (7.1). Moreover, it is easy to prove by any of several arguments that these methods are accretive for $\mathrm{Re}\,\mu < 0$, when equation (7.1) is ill posed – they increase the energy unboundedly. Therefore, optimal methods mimic correctly the growth properties of the underlying differential equations.

8

Padé approximants

Continuous as the stars that shine
And twinkle on the Milky Way,
They stretch'd in a never-ending line
Along the margin of the bay:
Ten thousand saw I at a glance,
Tossing their heads in sprightly dance.

From *Daffodils* by William Wordsworth
(1770–1850).

8.1 Block numbers of Padé tableaux

In the present chapter we abandon numerical analysis of differential equations, the focus of our exposition so far, for approximation theory. Specifically, we use order stars to derive information on the structure of Padé tableaux of entire functions.

We have already defined Padé approximants (in Chapter 1) and their tableaux (in Chapter 3). Nevertheless, it is valuable to repeat these definitions. Let f be a function which is analytic in the neighbourhood of the origin (or, for that matter, any other complex point, but restricting our attention to the origin entails no loss of generality). We say that $R_{m/n} \in \pi_{m/n}[z]$ is the $[m/n]$ **Padé approximant** if

$$R_{m/n}(z) = f(z) + \mathcal{O}(z^{p+1})$$

and there is no function $R \in \pi_{m/n}$ such that

$$R(z) = f(z) + \mathcal{O}(z^{q+1})$$

with $q > p$. It is easy to see that if $p \geq m + n$ then $R_{m/n}$ is unique. We say that $p \equiv p_{m/n}$ is the **order** of the $[m/n]$ approximant. We denote the numerator and the denominator of $R_{m/n}$ by $P_{m/n}$ and $Q_{m/n}$ respectively and stipulate that $Q_{m/n}(0) = 1$.

We arrange Padé approximants in an infinite two-dimensional array

$$\mathcal{T}(f) = \begin{bmatrix} R_{0/0} & R_{1/0} & R_{2/0} & \cdots \\ R_{0/1} & R_{1/1} & R_{2/1} & \cdots \\ R_{0/2} & R_{1/2} & R_{2/2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

called the **Padé tableau**. Note that, strictly speaking, if $p_{m/n} < m+n$ for some m and n then $\mathcal{T}(f)$ is not uniquely defined. This will be 'fixed' soon.

Example 8.1 We follow Gragg (1972) in examining the Padé tableau of

$$f(z) = \frac{1 - z + z^3}{1 - 2z + z^2}.$$

Since f itself is $\pi_{3/2}$, necessarily $R_{m/n} \equiv f$ when both $m \geq 3$ and $n \geq 2$. Otherwise, we have

$$\begin{aligned}
R_{0/0}(z) &\equiv 1, \\
R_{1/0}(z) &= 1 + z, \\
R_{2/0}(z) &= 1 + z + z^2, \\
R_{3/0}(z) &= 1 + z + z^2 + 2z^3, \\
R_{4/0}(z) &= 1 + z + z^2 + 2z^3 + 3z^4, \\
R_{5/0}(z) &= 1 + z + z^2 + 2z^3 + 3z^4 + 4z^5, \ \ldots \\
R_{0/1}(z) &= R_{1/1}(z) = R_{0/2}(z) = R_{1/2}(z) = \frac{1}{1 - z}, \\
R_{2/1}(z) &= \frac{1 - z - z^2}{1 - 2z}, \\
R_{3/1}(z) &= \frac{1 - \frac{1}{2}z - \frac{1}{2}z^2 + \frac{1}{3}z^3}{1 - \frac{3}{2}z}, \\
R_{4/1}(z) &= \frac{1 - \frac{1}{3}z - \frac{1}{3}z^2 + \frac{2}{3}z^3 + \frac{1}{3}z^4}{1 - \frac{4}{3}z}, \\
R_{5/1}(z) &= \frac{1 - \frac{1}{4}z - \frac{1}{4}z^2 + \frac{3}{4}z^3 + \frac{1}{2}z^4 + \frac{1}{4}z^5}{1 - \frac{5}{4}z}, \ \ldots \\
R_{2/2}(z) &= \frac{1 - z^2}{1 - z - z^2}, \\
R_{0/3}(z) &= R_{1/3}(z) = R_{2/3}(z) = R_{0/4}(z) = R_{1/4}(z) = R_{2/4}(z) \\
&= R_{0/5}(z) = R_{1/5}(z) = R_{2/5}(z) = \frac{1}{1 - z - z^3}, \\
R_{0/6}(z) &= \frac{1}{1 - z - z^3 + z^6},
\end{aligned}$$

$$R_{1/6}(z) = R_{2/6}(z) = R_{1/7}(z)$$
$$= R_{2/7}(z) = \frac{1-z}{1-2z+z^2-z^3+z^4+z^6}, \cdots$$

When $p_{m/n} < m+n$ and uniqueness cannot be taken for granted, we take $R_{m/n}$ which is identical to an $[m'/n]$ approximant for some $m' > m$. This is always possible (Baker, 1975) and will be assumed henceforth.

Let us arrange the $p_{m/n}$'s in a two-dimensional array, paralleling the Padé tableau:

0	1	2	3	4	5	⋯
2	2	3	4	5	6	⋯
2	2	4	∞	∞	∞	⋯
5	5	5	∞	∞	∞	⋯
5	5	5	∞	∞	∞	⋯
5	5	5	∞	∞	∞	⋯
6	8	8	∞	∞	∞	⋯
7	8	8	∞	∞	∞	⋯
⋮	⋮	⋮	⋮	⋮	⋮	

A curious observation is that identical numbers appear in square blocks. Thus, we have two 2×2 blocks and a 3×3 block, as well as the infinite-sized block that corresponds to $R_{m/n} \equiv f$ and plenty of 1×1 blocks. This, in fact, illustrates a general phenomenon, the theme of Theorem 8.1. ◇

Example 8.2 We have already encountered Padé approximants to $\exp z$ (in Chapters 1, 3 and 4) and to $(1-z)^\alpha$ (in Chapter 6). In all these cases $p_{m/n} = m+n$ for all $m, n \geq 0$. ◇

Nontrivial blocks in Example 8.1 occurred for two reasons. Firstly (and quite obviously this is a general rule), if f is itself rational then, for sufficiently large m and n, the Padé approximant is nothing other than the function itself. More importantly, nontrivial blocks are 'associated' with values of m and n such that $p_{m/n} \neq m+n$. Specifically, the following theorem, due to Padé (Baker, 1975; Brezinski, 1980), regulates the relationship between the values of $p_{m/n}$'s and the block-structure of the tableau.

Theorem 8.1 The Padé tableau can be decomposed into $d \times d$ blocks. Moreover, given any finite block with vertices at $((m-d+1)/n)$, (m/n), $((m-d+1)/(n+d-1))$ and $(m/(n+d-1))$, such that $p_{m'/n'} \equiv p_{m/n}$ for all $m-d+1 \leq m' \leq m$, $n \leq n' \leq n+d-1$, and which is not embedded in a larger block of this form, then $R_{m'/n'} \equiv R_{m/n}$ for all m' and n' within the range, the degree of $P_{m/n}$ is exactly m and the degree of $Q_{m/n}$ is exactly n. □

There is, thus, a 'compensation': what is won on the swings – the upper-left part of the square – is lost on the roundabouts – the bottom-right triangle, while nothing is lost or gained across the anti-diagonal.

We say that a function f is **normal** if all the blocks in $\mathcal{T}(f)$ are 1×1 (Baker, 1975). There are important advantages to normal functions. In particular, most methods for rapid and robust calculation of Padé approximants rely on the distinctiveness of these in a portion of $\mathcal{T}(f)$ (Brezinski, 1980). Note that both $\exp z$ and $(1-z)^{\alpha}$ are normal, unlike the function from Example 8.1. A more general example of normal functions are the **Pólya frequency series** (Baker, 1975), which include $\exp z$ as a (very) special case and which occur in several branches of analysis (Karlin, 1968).

The natural number

$$\beta(f) := \sup\left\{p_{m/n} - m - n + 1 \,:\, m,n \geq 0\right\} \tag{8.1}$$

provides the maximal size of a block in $\mathcal{T}(f)$. Accordingly, we call it the **block number** of f. In particular, f is normal if and only if $\beta_f = 1$.

Example 8.3 The block number need not be bounded, even if f is not a rational function. For example, it is easy to see that the function

$$f(z) = \sum_{\ell=0}^{\infty} z^{\ell^2},$$

which is analytic in $|z| < 1$, admits blocks of increasing size along the top edge of $\mathcal{T}(f)$. Actually, the structure of f is quite complicated – the following is a table of the $p_{m/n}$'s:

0	3	3	3	8	8	8	8	⋯
1	3	3	3	8	8	8	8	⋯
2	3	3	3	8	8	8	8	⋯
4	4	5	6	8	8	8	8	⋯
4	4	7	7	8	8	8	8	⋯
5	6	7	7	9	10	11	12	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Quite clearly, nontrivial blocks occur all over the place! ◇

Let f be an entire function of bounded order $\lambda(f)$. Then, according to the Hadamard factorization theorem (Hille, 1962), it can be written in the form

$$f(z) = z^M e^{p(z)} \prod_{\ell=1}^{N} E\left(\frac{z}{\kappa_\ell}, q\right),$$

where E is the Weierstrass prime factor,

$$\begin{aligned} E(z,0) &:= 1-z, \\ E(z,q) &:= (1-z)\exp\left(\sum_{k=1}^{q}\frac{1}{k}z^k\right), \quad q=1,2,\ldots. \end{aligned}$$

Here M is a nonnegative integer, p is a polynomial, N is either a nonnegative integer or ∞, and $q, \deg p \leq \lambda(f)$ (actually, if $\lambda(f)$ is an integer then $\max\{q, \deg p\} = \lambda(f)$). Thus, the behaviour of any entire function of finite order can be fully explained by the nature and location of its zeros and its behaviour at infinity. The same holds for **essentially analytic** functions (cf. Section 2.1), except that now we have isolated zeros, poles and essential singularities to reckon with.

Whatever determines f also determines its block number $\beta(f)$. In other words, zeros, poles and essential singularities of an essentially analytic function exert their pull from afar and influence its 'approximability' by rational functions at the origin. Order stars are the natural tool to study this action-at-a-distance, since they link behaviour at salient points – zeros, poles, essential singularities and interpolation points – in a single geometric construct. This has been done by Iserles (1985b). Our exposition follows closely that work, but also presents some new material.

8.2 Nonvanishing functions

Let f be an essentially analytic function, analytic at the origin, which is is not a rational function. We assume that it has $L \geq 0$ essential singularities in $\bar{\mathcal{C}}$. In order to avoid spurious blocks in the first few columns of $\mathcal{T}(f)$, we stipulate that $f(0) \neq 0$. We denote

$$\begin{aligned} &\kappa_1, \kappa_2, \ldots : \text{ zeros of } f; \\ &\zeta_1, \zeta_2, \ldots : \text{ poles of } f; \\ &\vartheta_1, \ldots, \vartheta_L : \text{ essential singularities of } f \end{aligned}$$

and set

$$I(f) := \sum_{\ell=1}^{L} \iota(\vartheta_\ell),$$

the **essential singularity index** of f.

The following theorem serves as a relatively gentle introduction to the application of order stars to the derivation of bounds on the block number.

Theorem 8.2 Let $\vartheta_1 = \infty$ and suppose that the function f is analytic and nonzero in $\mathrm{cl}\,\mathcal{C} \setminus \{\vartheta_1, \vartheta_2, \ldots, \vartheta_L\}$, where $\vartheta_2, \ldots, \vartheta_L \in \mathcal{C} \setminus \{0\}$ are distinct. Then

$$\beta(f) \leq I(f) + L - 1. \tag{8.2}$$

Proof. We consider the order star (of the first kind) of $\{f, R\}$, where $R \in \pi_{m/n}$ approximates f at the origin to order p. Since $f(0) \neq 0$, Proposition 2.1 implies that $0 \in \mathcal{A}_0$ and $\iota(0) = p + 1$.

Let us first assume that $m = 0$. The function f does not vanish, hence the poles of $\rho := R/f$ are precisely the n poles of R. Consequently, according to Proposition 2.3, at most n of the $p + 1$ sectors of $\mathcal{A}_+$ at the origin may belong to analytic $\mathcal{A}_+$-regions. To account for remaining sectors, we need help from essential singularities. Specifically, the remaining $p - n + 1$ sectors of $\mathcal{A}_+$ at the origin must belong to $\mathcal{A}_+$-regions that have essential singularities on their boundaries. Let us count these sectors, bearing in mind that $m = 0$ implies an absence of zeros (and hence, by Proposition 2.3, analytic $\mathcal{A}_-$-regions): At most $\iota(\infty)$ can be accounted for by $\vartheta_1 = \infty$. Bounded essential singularities are slightly more complicated, because a single $\mathcal{A}_+$-region may approach the origin twice (cf. Figure 8.1). Thus, there are at most $\iota(\vartheta_\ell) + 1$ sectors of $\mathcal{A}_+$ at the origin that can be 'explained' by ϑ_ℓ, $\ell = 2, \ldots, L$. Altogether, the number of sectors of $\mathcal{A}_+$ at the origin may not exceed $I(f) + L + n - 1$.

Our argument remains valid when $m \geq 1$, except that now there might be analytic $\mathcal{A}_-$-regions. These can approach the origin from within nonanalytic $\mathcal{A}_+$-regions, adding to our count of sectors of $\mathcal{A}_+$ there. Because of Proposition 2.3, there can be at most m sectors of $\mathcal{A}_-$ at the origin. Hence the bound $I(f) + L + n + m - 1$ on the number of sectors of $\mathcal{A}_+$ and, bearing in mind that $\iota(0) = p + 1$, we obtain the inequality

$$p \leq I(f) + L + n + m - 2$$

which is valid for all $R \in \pi_{m/n}$. In particular,

$$p_{m/n} - m - n + 1 \leq I(f) + L - 1.$$

The theorem now follows at once from definition (8.1). □

Functions that obey the conditions of Theorem 8.2 can be characterized quite easily. Since f is nonzero away from the isolated essential singularities, $\log f(z)$ is well defined for all $z \in \mathcal{C} \setminus \{\vartheta_1, \ldots, \vartheta_L\}$. It follows that $f(z) = \exp S(z)$, where the function S is analytic together with f. Moreover, the requirement that $\lambda(f, \vartheta_\ell) < \infty$ for all $\ell = 1, \ldots, L$ can be easily seen to imply that S must be a rational function in $\pi_{\tilde{m}/\tilde{n}}$, say, with $\tilde{m} \geq \tilde{n} + 1$ (because $\vartheta_1 = \infty$) and with poles precisely at $\vartheta_2, \ldots, \vartheta_L$. We assume without loss of generality that S is irreducible – all this means is that its poles are distinct from its zeros.

Let us suppose that ϑ_ℓ is a pole (of S) of multiplicity μ_ℓ. It follows at once that $\iota(\vartheta) = \mu_\ell$. Hence, $\sum_{\ell=2}^{L} \iota(\vartheta_\ell) = \tilde{n}$. Moreover, $\iota(\infty) = \tilde{m} - \tilde{n}$, consequently $I(f) = \tilde{m}$ and (8.2) can be rephrased as

$$\beta(f) \leq \tilde{m} + L - 1. \tag{8.3}$$

It is easy to generalize this result. For example, we can abandon the requirement that $\vartheta_1 = \infty$. Revisiting the proof of Theorem 8.2, we can see at once that this increases the bound in (8.2) by one. In terms of S, we now have $\tilde{m} \leq \tilde{n}$ and $I(f) = \tilde{n}$. Therefore,

$$\beta(f) \leq \tilde{n} + L. \tag{8.4}$$

Comparison of (8.3) with (8.4) proves at once a reformulation and generalization of Theorem 8.2.

Theorem 8.3 Suppose that the irreducible function $S \in \pi_{\tilde{m}/\tilde{n}}$ is analytic at the origin and set $f(z) := \exp S(z)$. Then

$$\beta(f) \leq \max\{\tilde{m} - 1, \tilde{n}\} + L, \tag{8.5}$$

where L is the number of distinct poles of S. □

Example 8.4 The bound (8.5) is always attainable by $S(z) = z^{\tilde{m}}$. We then have $L = 1$, $\vartheta_1 = \infty$, $\tilde{m} = 0$, hence $\beta(f) \leq \tilde{m}$. But, obviously, the Padé approximant to $\exp z^{\tilde{m}}$ is nothing other than an appropriate Padé approximant to $\exp z$, evaluated at $z^{\tilde{m}}$. This, clearly, leads to $\tilde{m} \times \tilde{m}$ blocks. ◇

Example 8.5 Let $S(z) = \gamma z/(1 - z)$, where $\gamma \in \mathcal{C} \setminus \{0\}$. Thus,

$$f(z) = \exp\left(\frac{\gamma z}{1 - z}\right)$$

and (8.5) yields $\beta(f) \leq 2$.

To examine Padé approximants to f, it is convenient first to derive its Taylor expansion. This can be done in several equally tedious ways: expanding the argument and manipulating the ensuing series, deriving a differential equation that is obeyed by f and solving it by comparing Taylor coefficients, etc. The simplest approach relies on a well-known generating function for Laguerre polynomials,

$$\sum_{\ell=0}^{\infty} L_\ell^{(\alpha)}(x) z^\ell = \frac{1}{(1 - z)^{1+\alpha}} \exp\left(\frac{-xz}{1 - z}\right)$$

(Rainville, 1967). Integration and the substitutions $\alpha = 1$, $x = -\gamma$ produce at once

$$f(z) = 1 + \gamma \sum_{\ell=1}^{\infty} \frac{1}{\ell} L_{\ell-1}^{(1)}(-\gamma) z^\ell.$$

We can now derive, for example, the [2/1] Padé approximant:

$$R_{2/1}(z) = \frac{1 - \frac{1 - \frac{1}{3}\gamma^2}{1 + \frac{1}{2}\gamma} z + \frac{1}{2} \frac{\gamma^3}{1 + \frac{1}{2}\gamma} z^2}{1 - \frac{1 + \gamma + \frac{1}{6}\gamma^2}{1 + \frac{1}{2}\gamma} z},$$

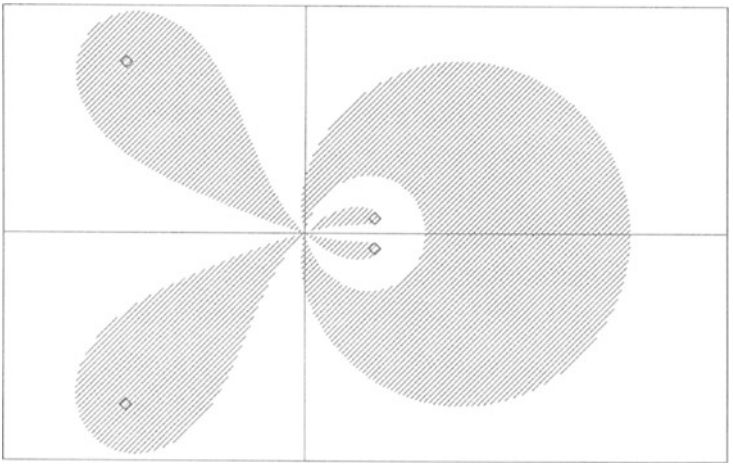


Figure 8.1 Order star of the [0/4] Padé approximant to $\exp(\gamma z/(1-z))$, $\gamma \approx 2.57163500764628$, of order 5.

except when $\gamma = -2$, when

$$R_{2/1}(z) = 1 - 2z.$$

Moreover,

$$p_{2/1} = \begin{cases} 2 & : \gamma = -2, \\ 3 & : \gamma \notin \{-2, -3 \pm i\sqrt{3}\}, \\ 4 & : \gamma = -3 \pm i\sqrt{3} \end{cases} .$$

It now follows from (8.1) that $\beta(f) = 2$ for $\gamma \in \{-2, -3 \pm \sqrt{3}\}$.

Other values of γ can also lead to $\beta(f) = 2$. In particular, if $-\gamma$ is a zero of $L_m^{(1)}$ (and, a Laguerre polynomial being orthogonal in $(0, \infty)$, there are m distinct, negative γ's of that kind) then $p_{m/0} = m + 1$ and $\beta(f) = 2$. Moreover, if γ, rather than $-\gamma$, is a zero of $L_m^{(1)}$ then $p_{0/m} = m + 1$. This is so because $1/f(z)$ is of the same form as $f(z)$, except that $-\gamma$ replaces γ. Figure 8.1 displays the order star for $R_{0/4}$ when $\gamma \approx 2.57163500764628$, with $p_{0/4} = 5$. Note how all the 'action' is confined to a few blobs in the relative vicinity of the origin, while ∞ is just another point of analyticity on the Riemann sphere, of no special significance. $\diamond$

Theorem 8.3 can be, in turn, generalized further, by allowing f a finite number of zeros and poles. It is seen at once that

$$f(z) = T(z)\exp S(z),$$

where T is itself a rational function. Suppose that $T \in \pi_{\tilde{m}/\tilde{n}}$. Since zeros of T (hence, of f) play the same role in the order star as poles of R, while poles of T can be counted together with zeros of R, we obtain the bound

$$\beta(f) \leq \max\{\tilde{m}-1, \tilde{n}\} + L + \hat{m} + \hat{n}. \tag{8.6}$$

Example 8.6 Let K be a natural number and set

$$f(z) = e^z \sum_{\ell=0}^{K} \frac{(-1)^\ell}{\ell!} z^\ell.$$

Thus, f is entire, $L = 1$, $\vartheta_1 = \infty$ with a unit index (hence $\tilde{m} = 1$, $\tilde{n} = 0$), and f has exactly K zeros in $\mathcal{C}$. Consequently, (8.6) produces the bound

$$\beta(f) \leq K + 1.$$

Clearly, the bound is attainable already by $R_{0/0} \equiv 1$, since $f(z) = 1 + \mathcal{O}(z^{K+1})$. $\diamond$

8.3 Functions with an infinite number of zeros

Throughout this section we assume that the function f is real analytic (that is, $f(\bar{z}) = \overline{f(z)}$ throughout the domain of analyticity) and nonzero in $\mathcal{C}$, except for the following points:

Essential singularities at the distinct points $\vartheta_1 = \infty, \vartheta_2, \ldots, \vartheta_L$;

Real zeros at the points $\kappa_1, \kappa_2, \ldots$; and

Poles at the distinct points $\zeta_1, \ldots, \zeta_N$, of multiplicities $\mu_1, \ldots, \mu_N$ respectively.

Of course, the origin is a point of analyticity (otherwise there is no point in approximating there!) and we assume that $f(0) \neq 0$. Note that the zeros (but not the poles!) of f are real and that there might be an infinity of them.

Given such a function, we set the quantity $\Upsilon(f) \in \{0,1,2\}$ in the following manner. Initially, we set $\Upsilon(f) = 0$. If there exists a real sequence $\{x_k\}_{k=1}^{\infty}$ such that

$$\lim_{k\to\infty} x_k = +\infty, \qquad \lim_{k\to\infty} |f(x_k)| < 1,$$

we increase $\Upsilon(f)$ by one. Likewise, we increase it by one if there exists a real sequence $\{y_k\}_{k=0}^{\infty}$ such that

$$\lim_{k\to\infty} y_k = -\infty, \qquad \lim_{k\to\infty} |f(y_k)| < 1.$$

Finally, we replace $\Upsilon(f)$ by $\min\{\Upsilon(f), \iota(\infty)\}$.

Theorem 8.4 The block number of f obeys the bound

$$\beta(f) \le I(f) + \sum_{k=1}^{N} \mu_k + L - \Upsilon(f) + 1. \tag{8.7}$$

Proof. Let $R \in \pi_{m/n}$ be given. Similarly to the proof of Theorem 8.3, we consider the order star of $\{f, R\}$ and count the number of sectors of $\mathcal{A}_+$ that approach the origin. Note that the order star is symmetric with respect to the real axis, since f is a real analytic function.

We assume that the origin is adjoined by:

(a) ω_- sectors of $\mathcal{A}_+$ that contain points in $(-\infty, 0)$;

(b) ω_+ sectors of $\mathcal{A}_+$ that contain points in $(0, \infty)$;

(c) ω_0 sectors of $\mathcal{A}_+$ that are wholly in $\mathcal{C} \setminus \mathcal{R}$ and belong to analytic $\mathcal{A}_+$-regions; and

(d) ω_* sectors of $\mathcal{A}_+$ that are wholly in $\mathcal{C} \setminus \mathcal{R}$ and belong to nonanalytic $\mathcal{A}_+$-regions.

Note that

$$\omega_0 \le n. \tag{8.8}$$

Moreover, proceeding as in the proof of Theorem 8.3, we can prove that

$$\omega_* \le I(f) + L - \Upsilon(f) - 1 \tag{8.9}$$

(since $\Upsilon(f)$ sectors of $\mathcal{A}_+$ at infinity are not available to 'support' $\mathcal{A}_+$-regions away from $\mathcal{R}$).

The ω_+ sectors of $\mathcal{A}_+$ approaching the origin and containing points on $(0, \infty)$ enclose $\omega_+ - 1$ sectors of $\mathcal{A}_-$. Likewise, $\omega_- - 1$ sectors of $\mathcal{A}_-$ are enveloped by the ω_- sectors of $\mathcal{A}_+$ to the right. We may assume without loss of generality that all these sectors belong to analytic $\mathcal{A}_-$-regions, since otherwise the upper bound in (8.9) need be decreased accordingly. Thus, there are at most $m + \sum_{\ell=1}^{N} \mu_\ell$ such sectors, providing the estimate

$$\omega_- + \omega_+ \le m + \sum_{\ell=1}^{N} \mu_\ell + 2. \tag{8.10}$$

A little extra care is required when $\omega_+ = 0$, say, i.e. there are no sectors of $\mathcal{A}_+$ adjoining the origin and having points in $(0, \infty)$. In that case the bound (8.10) is actually lower.

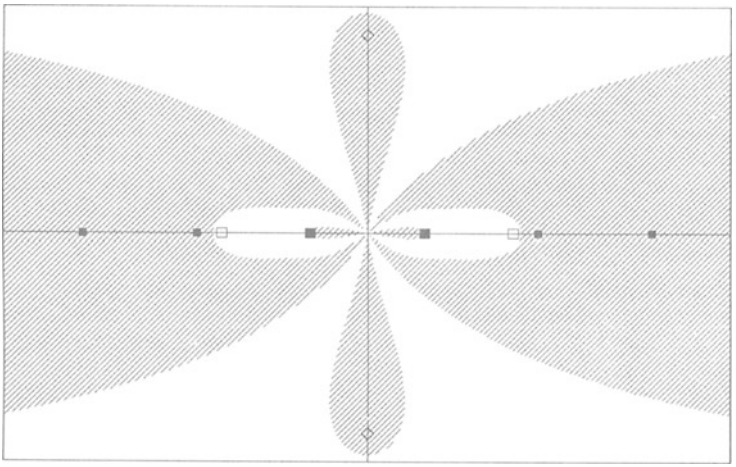


Figure 8.2 Order star of the [4/2] Padé approximant to $\cos z$, of order 7.

Putting (8.8)–(8.10) together, we have from Proposition 2.1

$$p+1 = \iota(0) = \omega_- + \omega_+ + \omega_\circ + \omega_\star \le m + n + I(f) + L + \sum_{\ell=1}^{N} \mu_\ell - \Upsilon(f) + 1$$

and the bound (8.7) follows at once from (8.1). □

Example 8.7 Let $f(z) = \cos z$, an entire function (thus $L = 1$ and $N = 0$), nonzero in $\mathcal{C} \setminus \mathcal{R}$, symmetric with respect to the real axis. Moreover, $\iota(\infty) = 2$. Clearly, $\lim_{r\to\infty} |f(re^{i\theta})| = \infty$ for all $0 < |\theta| < \pi$, while $|f(x)| \le 1$ along the real axis, with zeros for arbitrarily large $|x|$. Of course, $\Upsilon(f) = 2$ (we can choose, for example, $x_k := \left(k + \frac{1}{2}\right)\pi$, $y_k = -x_k$, $k = 1, 2, \ldots$). Thus, (8.7) gives $\beta(f) \le 2$. Actually, since f is an even function, so are its Padé approximants, hence $\beta(f)$ must itself be an even number. It follows that $\beta(f) = 2$ and $\mathcal{T}(f)$ is composed of 2×2 blocks.

Figure 8.2 displays the underlying order star for $R_{4/2}$. ◇

Example 8.8 We choose $f(z) = (z/2)^{-\nu} J_\nu(z)$, where J_ν is the Bessel function. Here $\nu \in \mathcal{R} \setminus \{-1, -2, -3, \ldots\}$. Since

$$J_\nu(z) = \left(\frac{z}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(k+1+\nu)} \left(-\frac{z^2}{4}\right)^k$$

(Rainville, 1967), it follows that f is an entire even function, $f(0) = 1$ and f is symmetric about $\mathcal{R}$. By using the standard asymptotic formula for J_ν with a large argument (Erdélyi *et al.*, 1953) we obtain the estimate

$$f(z) = \pi^{\frac{1}{2}} \left(\frac{2}{z}\right)^{\nu-\frac{1}{2}} \cos\left(z - \left(\frac{\nu}{2} + \frac{1}{4}\right)\pi\right)(1 + o(1)),$$

which is valid for $|z| \gg 0$ in the wedge $|\arg z| < \pi$. This, in tandem with $f(z) = f(-z)$, implies that $\iota(\infty) = 2$. Moreover, J_ν has an infinity of zeros in $\mathcal{R}$, with accumulation points at $\pm\infty$ (hence $\Upsilon(f) = 2$), no zeros in $\mathcal{C} \setminus \mathcal{R}$ and (8.7) implies that $\beta(f) \leq 2$. As in the previous example, we can deduce at once that $\beta(f) = 2$ and that every block in $\mathcal{T}(f)$ is 2×2, since f is even.

The structure of the order star is remarkably similar to that of $\cos z$ from Example 8.7. This is the whole point: the cosine and the (normalized) Bessel function might be totally different creatures, but they share several important features – evenness, real zeros, behaviour at ∞ – that determine the geometry of the order star.

Figure 8.3 presents the order star corresponding to $R_{2/4}$. ◇

Example 8.9 Recall the Mittag-Leffler function

$$E_\alpha(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + 1)}, \quad \alpha > 0,$$

from Example 2.3. It is entire, symmetric with respect to $\mathcal{R}$ and $\lambda(E_\alpha) = \alpha^{-1}$. Moreover, if $\alpha > 2$ then all the zeros of E_α are negative (Erdélyi *et al.*, 1953). Let $f = E_\alpha$. It is easy to see that $\iota(\infty) = 0$ and $\Upsilon(f) = 1$, therefore, by (8.7), $\beta(f) = 1$ and E_α is normal. ◇

Example 8.10 Let

$$f(z) = \prod_{k=1}^{\infty}(1 - q^k z), \quad q \in (0, 1).$$

We have already seen in Example 2.1 that f is entire and $\iota(\infty) = 0$ and have had an opportunity to examine its order star in Figure 2.1. Clearly, it is symmetric with respect to the real axis. Moreover, $\Upsilon(f) = 1$ and, as a consequence of Theorem 8.4, it transpires that f is normal.

It is a matter of minor curiosity that in the present case we can write the Padé approximants explicitly (Iserles, 1985b): $R_{m/n} = P_{m/n}/Q_{m/n}$, where

$$\begin{aligned} P_{m/n}(z) &= \frac{(q;q)_m}{(q;q)_{m+n}} \sum_{\ell=0}^{m} (-1)^\ell \frac{(q;q)_{m+n-\ell}}{(q;q)_\ell (q;q)_{m-\ell}} q^{\frac{1}{2}\ell(\ell+1)} z^\ell, \\ Q_{m/n}(z) &= \frac{(q;q)_n}{(q;q)_{m+n}} \sum_{\ell=0}^{n} \frac{(q;q)_{m+n-\ell}}{(q;q)_\ell (q;q)_{m-\ell}} z^\ell. \end{aligned}$$

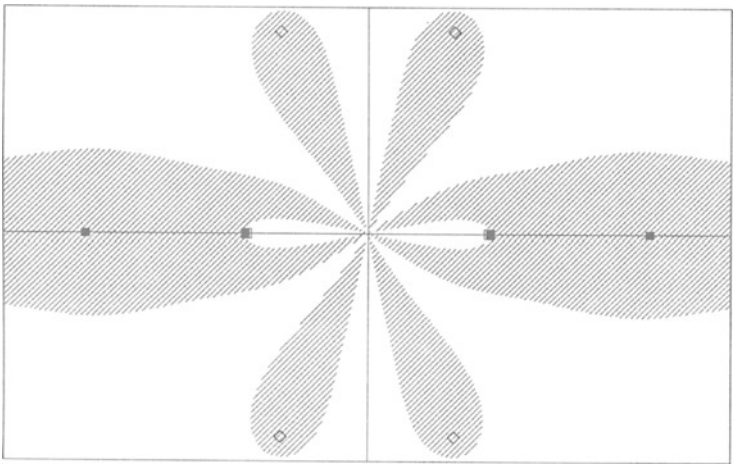


Figure 8.3 Order star of the [2/4] Padé approximant to the Bessel function J_0.

The q-factorial coefficient $(q;q)_k = (1-q)(1-q^2)\cdots(1-q^k)$ was defined in Chapter 2. Note some similarity with the explicit form of Padé approximants to $\exp(-z)$. This is not surprising to workers in the theory of special functions, where the function f is sometimes dubbed the **q-exponential** (Gasper and Rahman, 1990). $\diamond$

Example 8.11 In Examples 8.7–8.10 an application of (8.7) produced $\beta(f)$ exactly, not just an upper bound. This is not always the case. For example, let

$$f(z) = \frac{1}{\Gamma(z+1)},$$

where Γ is the familiar **Gamma function** (Rainville, 1967). It is entire and symmetric about the real axis. Moreover, according to the **Stirling formula** it is true that

$$f(z) = \exp\left(z - \left(z + \frac{1}{2}\right)\log z\right)(1 + o(1))$$

for $|z| \to \infty$, as long as z is confined to a wedge of the form $|z - \pi| > \delta$ for some $\delta > 0$. It is possible to deduce from the latter (or by other means) that $\iota(\infty) = 1$ and that the underlying order star contains a single nonanalytic $\mathcal{A}_+$ region which lies (asymptotically) to the right of $i\mathcal{R}$ and

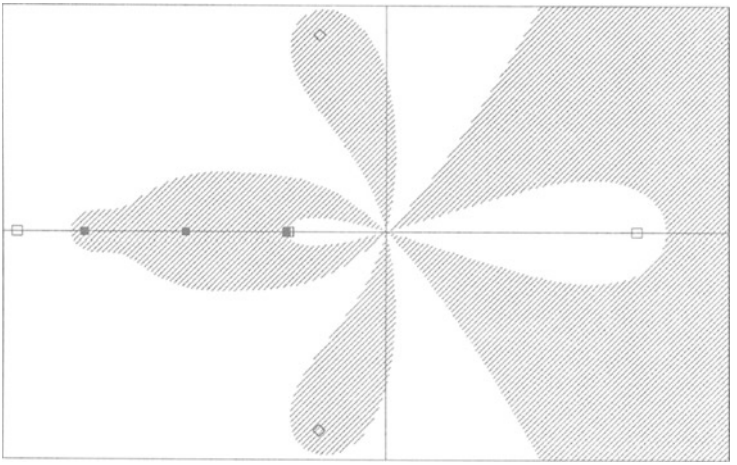


Figure 8.4 Order star of the [3/2] Padé approximant to $1/\Gamma(z+1)$.

a single nonanalytic $\mathcal{A}_-$-region that approaches ∞ to the left of the pure imaginary axis. Moreover, $\Upsilon(f) = 1$: although we can easily find both $\{x_k\}_{k=0}^{\infty}$ and $\{y_k\}_{k=0}^{\infty}$, the number $\Upsilon(f)$ is reduced since $\iota(\infty) = 1$. Hence, (8.7) produces the bound $\beta(f) \leq 2$.

The precise block number of f is unknown, but computer experiments suggest that the function is normal. In other words, not all available zeros and/or poles 'contribute' towards increasing the order and there is a measure of redundancy in the order star. This can be seen in Figure 8.4, where $R = R_{3/2}$: one of the zeros of $R_{3/2}$ is negative and does not 'support' a sector of $\mathcal{A}_-$ (or, at least, it is not required to support it, since it is 'wasted' inside the nonanalytic $\mathcal{A}_-$-region). Had we had, instead of that zero and the positive zero, a complex conjugate pair enveloped by the nonanalytic $\mathcal{A}_+$-region, there would have been enough scope to 'support' an extra sector of $\mathcal{A}_-$ at the origin, thereby boosting the order by one. $\diamond$

Theorem 8.4 can be easily generalized, allowing, for example, zeros in $\mathcal{C} \setminus \mathcal{R}$. Moreover, it is possible to consider functions with an infinity of zeros on several rays emerging from the origin, not just two: it requires an appropriate redefinition of $\Upsilon(f)$ and some amendments to the statement and proof of the theorem. A case in point is the function $f(z) = z^{-2} \sin z^2$, with zeros along both $\mathcal{R}$ and $i\mathcal{R}$. Figure 2.3 displays the order star of its

[4/0] Padé approximant. Note that $p_{4/0} = 7$, thus $\beta(f) \geq 4$. We leave it as an exercise to the reader for prove that $\beta(f) = 4$.

More interesting is the special case where f is entire, $\iota(\infty) = 1$ and all the zeros $\kappa_1, \kappa_2, \ldots$ are negative. The bound (8.7) is in that case $\beta(f) \leq 3 - \Upsilon(f)$. This can be significantly reduced.

Theorem 8.5 Suppose that f is entire, $f(0) \neq 0$, $\overline{f(z)} = f(\bar{z})$ for all $z \in \mathcal{C}$, $\iota(\infty) = 1$, all the zeros of f are negative and there exists a positive number r such that $(r, \infty) \in \mathcal{A}_-$ in the order star of $\{f, R\}$ for any rational function R. Then $\beta(f) = 1$ and f is a normal function.

Proof. Suppose that $R \in \pi_{m/n}$ and denote the order by p. Since $\iota(\infty) = 1$, exactly one unbounded (hence nonanalytic) $\mathcal{A}_+$-region exists in the order star. We denote it by $\mathcal{A}_+^\infty$ and note that it contains arbitrarily small points in $(-\infty, 0)$ but that $(r, \infty) \cap \mathcal{A}_+^\infty = \emptyset$.

If no sector of $\mathcal{A}_+$ at the origin belongs to $\mathcal{A}_+^\infty$ then at most n sectors of $\mathcal{A}_+$ there can be 'supported' (in the sense of Proposition 2.3) by poles of R. All remaining sectors of $\mathcal{A}_+$ must be 'supported' by zeros of f, hence belong to $\mathcal{A}_+$-regions that intersect with $(-\infty, 0)$. The order star is symmetric with respect to the real axis. Thus, if there are $\omega_0 \geq 1$ such sectors, they must necessarily envelop $\omega_0 - 1$ sectors of $\mathcal{A}_-$ that belong to bounded – hence analytic – $\mathcal{A}_-$-regions. Because of Proposition 2.3, the number of such sectors of $\mathcal{A}_-$ cannot exceed the number of zeros of R/f, therefore $\omega_0 - 1 \leq m$. We have

$$p + 1 = \iota(0) \leq n + \omega_0 \leq n + m + 1,$$

hence the order is bounded by $m + n$.

The remaining possibility is that $\omega_\infty \geq 1$ sectors of $\mathcal{A}_+$ at the origin lie in $\mathcal{A}_+^\infty$. Moreover, the origin can be adjoined by up to n sectors of $\mathcal{A}_+$ that are 'supported' by poles of R and by $\omega_0 \geq 0$ sectors of $\mathcal{A}_+$ that belong to analytic $\mathcal{A}_+$-regions and are accounted for by zeros of f.

Since $\kappa_1, \kappa_2, \ldots < 0$, all the $\mathcal{A}_+$-regions that contain these points are either $\mathcal{A}_+^\infty$ itself (hence nonanalytic) or are enveloped by $\mathcal{A}_+^\infty$. Thus, sectors of $\mathcal{A}_+$ at the origin that belong either to $\mathcal{A}_+^\infty$ or to analytic $\mathcal{A}_+$-regions that intersect with $(-\infty, 0)$ envelop $\omega_0 + \omega_\infty - 1$ sectors of $\mathcal{A}_-$ that all belong to analytic $\mathcal{A}_-$-regions. Consequently, by Proposition 2.3, $\omega_0 + \omega_\infty \leq m + 1$. Counting sectors of $\mathcal{A}_+$, we obtain the bound

$$p + 1 = \iota(0) \leq n + \omega_0 + \omega_\infty \leq n + m + 1$$

and, again, $p \leq m + n$. This proves $\beta(f) = 1$ and completes the proof. □

Note that $p \leq m + n$ for all $R \in \pi_{m/n}$, $m, n \geq 0$, implies that $p_{m/n} = m + n$, hence that the bound is attainable. This paradoxical fact – an upper

bound implying a lower bound – follows from Theorem 8.1. Recall our remark about swings and roundabouts.

It is an interesting consequence of the proof of Theorem 8.5 that if $R = R_{m/n}$, hence $p = m + n$, all the inequalities hold as equalities. Thus, exactly n sectors of $\mathcal{A}_+$ at the origin are 'supported' by poles of $R_{m/n}$. Moreover, the underlying $\mathcal{A}_+$-regions cannot be enveloped by either $\mathcal{A}_+^\infty$ or by any $\mathcal{A}_+$-regions that contain zeros of f, otherwise there will not be enough zeros to account for all the sectors of $\mathcal{A}_-$ at the origin. It follows that $R_{m/n}$ may not have any negative poles.

Theorem 8.6 Every function of the form

$$f(z) = \prod_{k=1}^{\infty} \left(1 - \frac{z}{\kappa_k}\right),$$

where

$$\kappa_k < 0, \quad k = 1, 2, \ldots, \quad \sum_{k=1}^{\infty} |\kappa_k|^{-1} < \infty, \quad \sum_{k=1}^{\infty} |\kappa_k|^{-\alpha} = \infty \; \forall \, \alpha < 1,$$

is normal.

Proof. According to Proposition 2.8 and a subsequent remark, f is entire and $\iota(\infty) = 1$. Moreover, it follows at once from its proof that, given any rational function R, there is precisely one nonanalytic $\mathcal{A}_+$-region and one nonanalytic $\mathcal{A}_-$-region in the order star of $\{f, R\}$, the $\mathcal{A}_+$ region lying to the left. This implies that all the conditions of Theorem 8.5 hold, hence the normalcy of f. □

Example 8.12 An example of a function that obeys all the conditions of Theorem 8.5 has been already presented in Example 2.4, namely

$$f(z) = \prod_{k=2}^{\infty} \left(1 + \frac{z}{k(\log k)^b}\right),$$

for an arbitrary $b > 1$. ◇

Contractive approximation

See! that huge circle, like a necklace, stares
With thousands of bold eyes to heaven, and dares
The golden stars

From *Impression de Nuit* by Lord Alfred Douglas (1870–1945).

9.1 Zeros and contractions

Time and again throughout this book we have encountered the following paradigm. A function f is analytic in the open domain $\mathcal{V}$, $|f| \equiv 1$ along $\partial\mathcal{V}$, and it is approximated in $\mathrm{cl}\,\mathcal{V}$ by another analytic function g. The opportunity to express several problems in numerical mathematics in this form is the secret behind the applicability of order stars.

In the present chapter we develop further the logic of the aforementioned paradigm. Throughout the present chapter $\mathcal{V} \subset \mathrm{cl}\,\mathcal{C}$ is an open, connected (but not necessarily simply connected) domain with a Jordan boundary, which is neither the empty set nor all of $\mathrm{cl}\,\mathcal{C}$. The nonconstant function f is analytic in $\mathcal{V}$ (but not necessarily along its boundary) and obeys $|f(z)| < 1$ inside the domain, $|f(z)| \equiv 1$ along the boundary.[1] We denote by $\mathcal{V}_A$ the set

$$\{z \in \mathrm{cl}\,\mathcal{V} : f \text{ is analytic at } z\}.$$

The task in hand is to interpolate f at some points in $\mathcal{V}_A$ by a $\mathcal{V}$-contractive function g. It turns out, with a little help from order stars, that there is an intimate relationship between specific properties of f and the extent to which it can be contractively interpolated.

Let us denote by M the number of zeros of f in $\mathcal{V}$, counted with their multiplicities. Note that, as long as f is analytic along $\partial\mathcal{V}$, necessarily $M \geq 1$. For suppose that this is not the case. Then $1/f$ is analytic in the domain and, by the maximum principle, its modulus is maximized on $\partial\mathcal{V}$.

[1] There is no contradiction between loss of analyticity on the boundary and $|f| \equiv 1$ there, although, obviously, poles are ruled out. We will return to this point later.

Consequently, f being nonconstant, $|1/f(z)| < 1$ inside $\mathcal{V}$ and we are faced with a contradiction.

Given a point $z \in \mathcal{V}$, we set $\Delta(z) := 1$. The function Δ is extended to $z \in \operatorname{cl} \mathcal{V}$ in the following fashion: we say that the **multiplicity** $\mu(z)$ is the number of branches of $\partial\mathcal{V}$ that intersect at z. Note that usually (always if $\mathcal{V}$ is simply connected) $\mu(z) = 1$, but Figure 9.1 displays a domain with a point such that $\mu(z) = 2$. Since the boundary is Jordan, $\mu(z)$ left and right tangents to $\partial\mathcal{V}$ are well defined at z, and we let $\Delta(z)$ be the sum of inner angles spanned by these tangents at z, divided by 2π. It follows at once that $0 \leq \Delta(z) \leq 1$ – if $\partial\mathcal{V}$ is smooth at z then $\Delta(z) = \frac{1}{2}$. Actually, much more can easily be deduced: since $z \in \mathcal{V}_A$, we can expand

$$f(\xi) = f(z) + \frac{1}{j!} f^{(j)}(z)(\xi - z)^j + \mathcal{O}(|\xi - z|^{j+1}), \quad f^{(j)}(z) \neq 0,$$

for some $j \geq 1$. It is easy to observe by employing the method of proof of Proposition 2.1 that $\Delta(z) = \mu(z)/j \in (0, 1)$. In particular, the extreme cases $\Delta(z) = 0$ and $\Delta(z) = 1$ may not occur at analytic point of $\partial\mathcal{V}$.

Example 9.1 Figure 9.1 displays sets $\mathcal{V}$ for three functions f. The first is $f(z) = (1 + z)e^{-z}$. Clearly, $\mu(z) \equiv 1$ for all boundary points, since $\mathcal{V}$ is simply connected. Moreover, $\partial\mathcal{V}$ is smooth (hence $\Delta(z) = \frac{1}{2}$ there), except at the origin, where $f(z) = 1 + \mathcal{O}(|z|^2)$ and $\Delta(0) = \frac{1}{4}$. Finally, f has a single zero in $\mathcal{V}$, thus $M = 1$.

The second figure was generated by

$$f(z) = \frac{z(1 + z^2)}{1 - 2z + 3z^2},$$

a function with simple zeros at 0 and $\pm i$. It is easy to verify that $f(z) = 1 + \mathcal{O}(|1 - z|^3)$. Moreover, the point $z = 1$ is an example of nontrivial multiplicity – specifically, $\mu(1) = 2$ and it follows that $\Delta(1) = \frac{2}{3}$.

Another phenomenon of interest is displayed in the bottom figure, where $f(z) = z^2 \sin z^{-1}$. The set $\mathcal{V}$ has two 'holes' and it approaches the origin along two cusps. To realize why, map $z \mapsto z^{-1}$: the origin travels to infinity and we obtain the function $\tilde{f}(z) = z^{-2} \sin z$. As long as we stay in a strip along the real axis, it is easy to see that $\tilde{f}(z)$ tends to zero as $|z| \to \infty$. However, as soon as we let z tend to infinity along a nonhorizontal ray, the exponential growth of $\sin z$ takes over and $|\tilde{f}(z)| \to \infty$. Mapping back, we obtain the aforementioned cusps. Of course, the origin is an essential singularity of f and $\mathcal{V}_A = \operatorname{cl} \mathcal{V} \setminus \{0\}$. $\diamond$

The sets $\mathcal{V}$ in Figure 9.1 look remarkably like order stars, for the simple reason that they *are* order stars – or portions theoreof! Recall that both $\mathcal{A}_-$, say, and $\mathcal{V}$ are nothing other than level sets of essentially analytic functions...

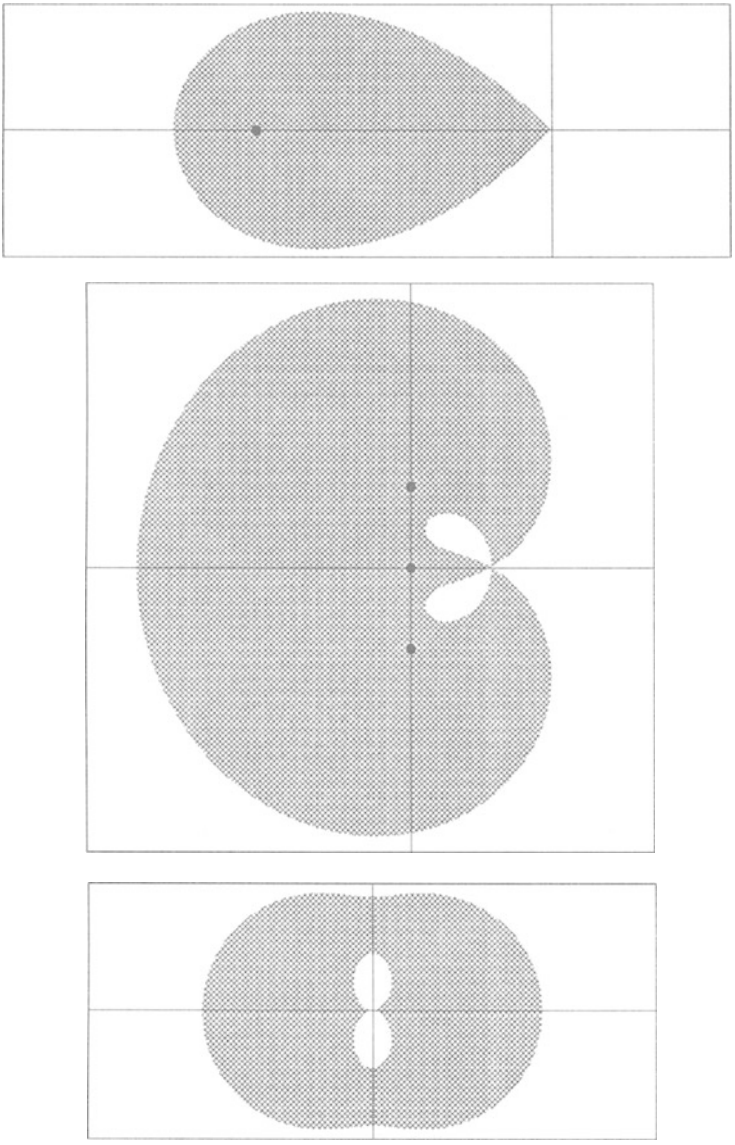


Figure 9.1 The sets $\mathcal{V}$ for three functions from Example 9.1. The sets are cross-hatched and zeros are denoted by dots, except for the bottom figure (where zeros accumulate at the origin).

Theorem 9.1 (Iserles, 1984; 1985b) If

$$g(z) = f(z) + C(z - z^*)^p + \mathcal{O}(|z - z^*|^{p+1}), \quad C \neq 0,\ z^* \in \mathcal{V}_A, \tag{9.1}$$

for some $p \geq 1$, is a $\mathcal{V}$-contraction then necessarily

$$p \leq \frac{M + \frac{1}{2}}{\Delta(z^*)}. \tag{9.2}$$

Proof. We consider the order star of the first kind for

$$\rho(z) := \frac{g(z)}{f(z)}.$$

According to Proposition 2.2, $\mathcal{V}$-contractivity of g is tantamount to the function being analytic in $\mathcal{V}$ and $\mathcal{A}_+ \cap \partial\mathcal{V} = \emptyset$. In other words, ∂V separates $\mathcal{A}_+$-regions.

Assume first that $f(z^*) \neq 0$. Therefore, by (9.1), $\iota(z^*) = p$ and, by Proposition 2.1, z^* is approached by p sectors of $\mathcal{A}_+$. If $z^* \in \mathcal{V}$ then the sum of multiplicities of all $\mathcal{A}_+$-regions inside the domain is at least p. Since all these regions are analytic and $\Delta(z^*) = 1$, inequality (9.2) follows at once from Proposition 2.3. If $z^* \in \partial\mathcal{V}$ then some sectors of $\mathcal{A}_+$ may approach from without $\mathcal{V}$. Specifically, if j sectors of $\mathcal{A}_+$ approach the point from within the domain, then the definition of Δ, regularity and Proposition 2.1 imply that

$$\frac{2j-1}{2p} \leq \Delta(z^*) \leq \frac{2j+1}{2p},$$

hence $p \leq \left(j + \frac{1}{2}\right)/\Delta(z^*)$. The lemma follows again from Proposition 2.3,

Finally, consider the case of $f(z^*) = 0$ and denote by $L \geq 1$ the multiplicity. Note that necessarily $z^* \in \mathcal{V}$, because $|f| \equiv 1$ along the boundary. By Proposition 2.1, $\iota(z^*) = p - L$. However, we now have only $M - L \geq 0$ zeros of f in $\mathcal{V} \setminus \{z^*\}$ and (9.2) follows, as before, from Propositions 2.2 and 2.3. □

Inequality (9.2) can be tightened somewhat for $z \in \partial V$ such that $\mu(z) \geq 2$, since not every configuration of sectors of $\mathcal{A}_+$ can be arranged to 'fit' the available angles.

It is important to emphasize that Theorem 9.1 is, in one sense, quite trivial: as long as we interpolate *inside* $\mathcal{V}$, its statement follows at once and quite easily from the Rouche theorem (Ahlfors, 1966). However, the various complex-analytic methods based on the Cauchy theorem, the argument principle, the Rouche theorem etc., all fail when the interpolation point is at the boundary. Order stars (which, after all, also heavily exploit Cauchy theorem-type formalism, cf. Chapter 2) provide a framework to extend the discussion right to the boundary.

Example 9.2 We return to the function $f(z) = (1+z)e^z$, whose level set $\mathcal{V}$ has been displayed in the top of Figure 9.1. Since $\Delta(0) = \frac{1}{4}$, (9.2) predicts $p \leq 6$ for $z^* = 0$. This is attainable by the [0/5] Padé approximant

$$g(z) = \frac{1}{1 + \frac{1}{2}z^2 - \frac{1}{3}z^3 + \frac{3}{8}z^4 - \frac{11}{30}z^5}.$$

The underlying order star is displayed in the top of Figure 9.2.

At the remaining points of the boundary we have $\Delta(z) \equiv \frac{1}{2}$, hence $p \leq 3$. The middle of Figure 9.2 depicts the [2/0] Padé approximant (that is, a quadratic truncation of the Taylor series) about $z^* = -1.2784645427611\ldots$, the intersection of ∂V with $(-\infty, 0)$. Note that the approximant is not $\mathcal{V}$-contractive. The lesson of this example ought to be obvious, to the point of self-evidence: the inequality (9.2) is always necessary, but by no means sufficient for $\mathcal{V}$-contractivity.

Finally, at the bottom of Figure 9.2 we display the order star with respect to $g(z) \equiv \frac{1}{2}e^{\frac{1}{2}}$, that interpolates f at $z^* = -0.5$ with $p = 1$. This rather simple function is, of course, $\mathcal{V}$-contractive (as will be any constant interpolant within $\mathcal{V}$, for any analytic f – why?). $\diamond$

There are two possible ways to the extend the framework of Theorem 9.1: by allowing essential singularities along the boundary and by considering a more intricate interpolation pattern.

Theorem 9.2 (Iserles, 1985b) Suppose that $K \geq 1$ sectors of $\mathcal{A}_+$ adjoin $\partial\mathcal{V}$ from within $\mathcal{V}$ at essential singularities of f. If g is a $\mathcal{V}$-contraction with m zeros in $\mathcal{V}$ that obeys (9.1), it is true that

$$p \leq \frac{K + M + m + \frac{1}{2}}{\Delta(z^*)}. \tag{9.3}$$

Proof. We again derive a bound on the sum of the multiplicities of $\mathcal{A}_+$-regions in $\mathcal{V}$. It is already known from the proof of Theorem 9.1 that zeros of f contribute M to our count. The only other contribution follows from the presence of essential singularities on the boundary, which allows some the $\mathcal{A}_+$-regions to be nonanalytic. Since $|f(z)| < 1$ in $\mathcal{V}$, an essential singularity can be approached by at most a single $\mathcal{A}_+$-region from within the domain. It follows that there are at most K nonanalytic $\mathcal{A}_+$-regions in $\mathcal{V}$.

Suppose that $z^* \in \mathcal{V}$ is approached by $J \geq 2$ sectors of $\mathcal{A}_+$ that belong to the same nonanalytic $\mathcal{A}_+$-region $\mathcal{F}$, say. Then $\mathcal{F}$ must surround at least $J - 1$ $\mathcal{A}_-$-regions. All these $\mathcal{A}_-$-regions are, by necessity, analytic, hence they 'cost' $J - 1$ zeros of g. Since we have m zeros of g and at most K nonanalytic $\mathcal{A}_+$-regions, it follows that $p \leq K + M + m$.

The proof for $z \in \partial\mathcal{V} \cap \mathcal{V}_A$ proceeds similarly. $\square$

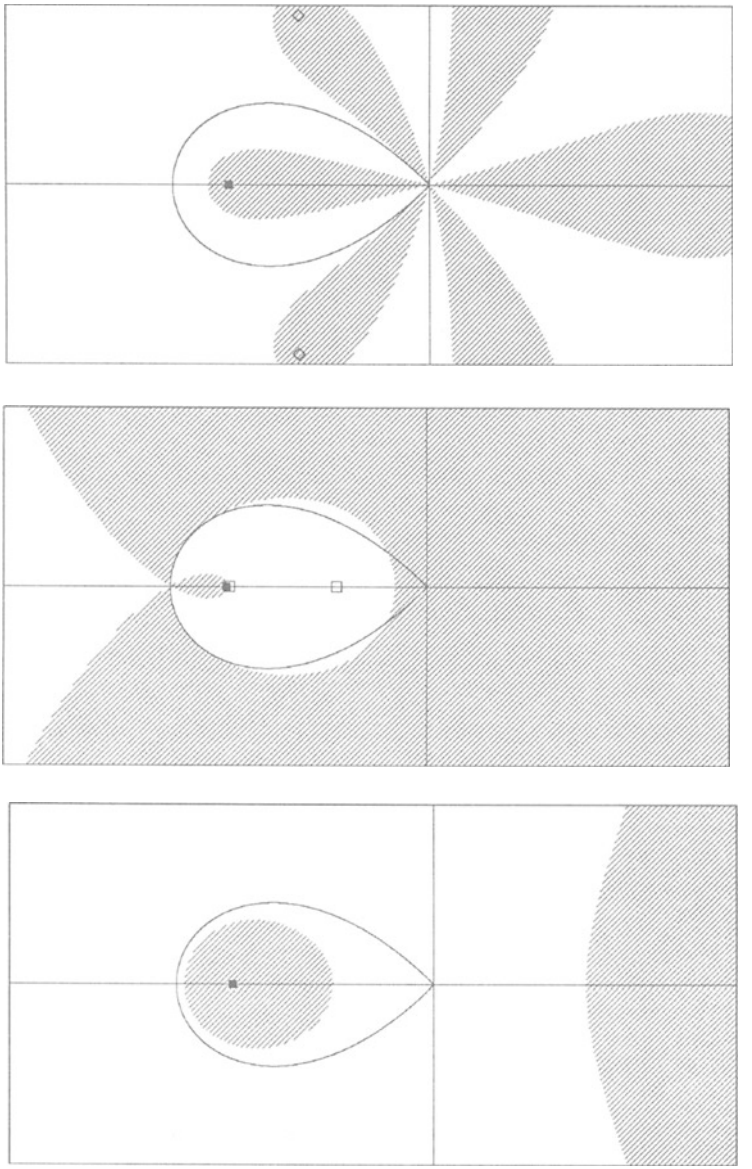


Figure 9.2 Order stars for the function $f(z) = (1 + z)e^{-z}$ from Example 9.2.

Example 9.3 We began this book, in Chapter 1, with the proof of the first Ehle conjecture: Padé approximants to the exponential can be A-acceptable only if $m \leq n \leq m+2$.

We have $f(z) = \exp z$, $\mathcal{V} = \{z \in C : \mathrm{Re}\, z < 0\}$ and $g(z) = R_{m/n}(z)$. Hence $M = 0$, $K = 1$ (essential singularity at infinity!), $p = m + n + 1$, $\Delta(0) = \frac{1}{2}$ and substitution into (9.3) gives $n \leq m+2$. Obviously, $m \leq n$ (we do not need order stars for that...) and the proof of the first Ehle conjecture is complete.

The aforementioned conjecture was for a decade one of the more celebrated open problems in numerical mathematics and its solution was the seed that sprouted the whole theory of order stars. It is a tribute to the generality of (9.3) that it can be proved by the simple expedient of 'painting by numbers'. ◇

Example 9.4 Following Iserles (1985b), we consider the function $f(z) = \exp(2z/(1+z))$. The set $\mathcal{V}$ is the unit disc, f has an essential singularity at $z = -1$ and is otherwise analytic and nonzero in the closed disc. Letting $z^* = 0$ we have $p \leq m + 1$. Figure 9.3 displays order stars corresponding to three rational approximants:

$$\begin{aligned} g_1(z) &= R_{0/2}(z) = \frac{1}{1 - 2z + 4z^2}, \quad (p = 3), \\ g_2(z) &= R_{3/0}(z) = 1 + \tfrac{2}{3}z^3, \quad (p = 4), \\ g_3(z) &= \frac{1 + \frac{5}{2}z}{\left(1 + \frac{1}{4}z\right)^2}, \quad (p = 2). \end{aligned}$$

Clearly, g_1 breaches the barrier $p \leq m + 1$, since $m = 0$. An examination of Figure 9.3 reveals the reason for the failure in V-contractivity: g_1 is not analytic within the disc.

The function g_2 has three complex zeros, hence, at first glance, there is hope for $\mathcal{V}$-contractivity. However, it is easy to verify that a single zero resides in each of the intervals $(-\infty, -1)$, $(-1, 0)$ and $(1, \infty)$ – only one inside the disc – and $p \leq m + 1$ is breached. As Figure 9.3 affirms, $\mathcal{V}$-contractivity fails since $|g_2|$ exceeds unity in portions of the unit circle. Of course, g_2 being analytic, this is the only possible avenue to non-$\mathcal{V}$-contractivity.

Finally, g_3 has a single zero in the disc and the inequality $p \leq m + 1$ is satisfied. This, of course, is just a necessary condition, but an examination of Figure 9.3 – or an easy calculation – is sufficient to verify that g_3 is, indeed, $\mathcal{V}$-contractive. ◇

Our final generalization of (9.2) allows interpolation at several points in $\mathcal{V}_A$. Thus, we replace (9.1) with

$$g(z) = f(z) + c_i(z - z_i^*)^{p_i} + \mathcal{O}(|z - z_i^*|^{p_i+1}), \quad c_i \neq 0, \quad i = 1, 2, \ldots, I, \quad (9.4)$$

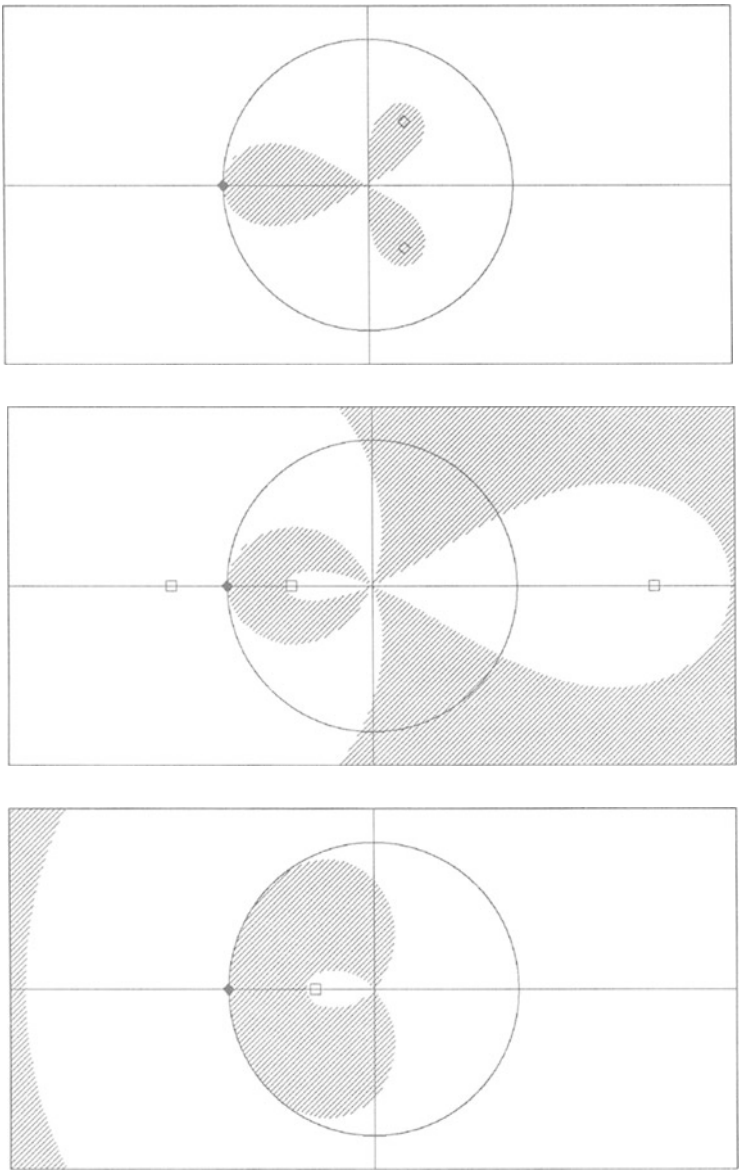


Figure 9.3 Order stars for the function $f(z) = \exp(2z/(1+z))$ from Example 9.3.

where $z_1^*, \ldots, z_I^*$ are distinct points in $\mathcal{V}_A$. Again, we count the number of sectors of $\mathcal{A}_+$ that approach $z_1^*, \ldots, z_I^*$. The number of such sectors at z_i^* depends both on p_i and on $\Delta(z_i^*)$ and we can deduce quite easily by the method of proof of Theorem 9.1 that it is at least

$$\left[\Delta(z_i^*)\left(p_i - \frac{1}{2}\right)\right],$$

except when $\Delta(z_i^*) = 1$ (i.e. the point is inside $\mathcal{V}$), when there are p_i such sectors. We denote by M and K^* the number of zeros of f in $\mathcal{V}$ and the number of interpolation points from $\{z_1^*, \ldots, z_I^*\}$ that can be adjoined by nonanalytic $\mathcal{A}_+$-regions.[2] It follows identically to the proof of Theorem 9.2 that at most $K^* + M + m$ sectors of $\mathcal{A}_+$ can adjoin interpolation points within $\mathcal{V}$.

Theorem 9.3 Let the $\mathcal{V}$-contraction g interpolate f to degree p_i at z_i^*, $i = 1, 2, \ldots, I$. Then necessarily

$$\sum_{i=1}^{I} \Upsilon_i \leq K^* + M + m. \tag{9.5}$$

where

$$\Upsilon_i = \begin{cases} p_i & : z_i^* \in \mathcal{V}, \\ \left[\Delta(z_i^*)\left(p_i - \frac{1}{2}\right)\right] & : z_i^* \in \mathcal{V}_A \setminus \mathcal{V}. \end{cases}$$

□

Example 9.5 We return to the framework of Example 9.3 (and much of this book), rational approximation of $\exp z$. Again $f(z) = \exp z$, $\mathcal{V}$ is the left half-plane and g is a rational m/n function. We stipulate that g obeys (9.4) with $\sum_{i=1}^{I} p_i = n + m + 1$, $z_1^* = 0$ and real $z_i^* < 0$, $i = 2, 3, \ldots, I$. Since $K^* = 1$ (by symmetry of the order star), $M = 0$ and $\sum_{i=2}^{I} p_i = n + m + 1 - p_1$, (9.5) yields

$$p_1 \geq \left[\frac{1}{2}\left(p_1 - \frac{1}{2}\right)\right] + n,$$

which, upon further examination, gives $p_1 \geq 2n - 1$. This implies that $m \leq n \leq m + 2$ and we recover Theorem 4.5. ◇

Having, in Examples 9.3 and 9.5, derived two major results by straightforward substitution in inequalities (9.3) and (9.5), there is a natural temptation to deduce other material in this volume in a similar way. This, unfortunately, is not as straightforward as it may seem – typical order-star results hinge on more than just contractivity.

[2] In general $K \leq K^* \leq \infty$, since a nonanalytic $\mathcal{A}_+$-region can adjoin any – finite or even infinite – number of interpolation points. Fortunately, it is sometimes possible to obtain less generous estimates of K^*. Thus, $I = 1$ implies that $K^* = 1$. Another instance of restricting K^* is provided in Example 9.5.

9.2 The Pick–Nevanlinna interpolation problem

The **Pick–Nevanlinna theory** is a powerful approach to complex approximation, with important applications in such areas as control theory (Glover, 1984) and electronics (Ball and Helton, 1979), as well as in rational approximation theory (Trefethen, 1981). Although it has been known since 1916, when Pick published the original version of the theorem bearing his name,[3] this knowledge has been confined to a relatively narrow group of specialists. In this section we quite shamelessly exploit a minor application of order stars to introduce the reader to some beautiful results. To delve into deep function-analytic results that have called in the past upon the skills of the likes of Akhiezer, Kakeya, Krein, Takagi and Walsh, is beyond the scope of this book. The present section is just an appetizer and the reader is encouraged to consult Sarason's (1967) survey, the remarkable dissertation of Scales (1982) or, best of all, the classical monograph of Walsh (1956) for the full story.

Let $z_1^*, z_2^*, \ldots, z_I$ be distinct points in the open complex unit disc, and $\omega_i^{(j)}$, $j = 0, 1, \ldots, p_i - 1$, $i = 1, 2, \ldots, I$, arbitrary complex numbers. We denote by $\mathcal{F}$ the set of all functions f that are analytic in the open disc and obey the interpolation conditions

$$\frac{\mathrm{d}^j}{\mathrm{d}z^j} f(z_i^*) = \omega_i^{(j)}, \quad j = 0, 1, \ldots, p_i - 1, \quad i = 1, 2, \ldots, I. \tag{9.6}$$

The simplest version of the Pick interpolation problem consists of determining the function $\tilde{f} \in \mathcal{F}$ that minimizes the L_∞ norm:

$$\sup_{|z| \leq 1} |\tilde{f}(z)| = \inf_{f \in \mathcal{F}} \sup_{|z| \leq 1} |f(z)|.$$

Of course, since we are dealing here with analytic functions, the maximum is always reached on the boundary and $\sup_{|z| \leq 1}$ can be replaced with $\sup_{|z|=1}$.

An interesting fact about the Pick problem is that its solution can be derived constructively (Walsh, 1956). A complex function

$$B(z; \alpha_1, \ldots \alpha_q) := \prod_{k=1}^{q} \frac{z - \alpha_k}{1 - \bar{\alpha}_k z},$$

where $\alpha_1, \ldots, \alpha_q \in C$, is said to be a **Blaschke product** of degree q.[4] Two basic facts need be highlighted. Firstly, $B(\cdot; \alpha_1, \ldots, \alpha_q)$ is analytic in the closed unit disc, provided that $|\alpha_k| < 1$ for all $k = 1, 2, \ldots, q$ (of course, we may assume, without loss of generality, that $|\alpha_1|, \ldots, |\alpha_q| \neq 1$, otherwise the rational function B is reducible). Secondly, $|B(e^{i\theta}; \alpha_1, \ldots, \alpha_q)| \equiv 1$ for

[3] A weaker version, due to Carathéodory and Fejér, was known even earlier.

[4] It is perfectly possible to analyse – with care – infinite Blaschke products (i.e. $q = \infty$).

all $0 \leq \theta \leq 2\pi$. Remarkably, the solution of the Pick interpolation problem is a scaled Blaschke product of degree $q \leq p-1$.

The proof comes in two parts. Firstly, it is demonstrated that the underlying data can always be interpolated by a scaled Blaschke product. In particular, a scaling factor can be derived explicitly, by evaluating the largest root of the equation $\det A(\xi) = 0$, where $A(\xi) = (a_{i,j})_{i,j=1}^{p}(\xi)$ and

$$a_{i,j}(\xi) := \left(\frac{\xi^2 - \omega_i \bar{\omega}_j}{1 - z_i^* \bar{z}_j^*}\right)^p, \quad i,j = 1,2,\ldots,p.$$

A being self-adjoint, we note that such a positive root ξ^* exists. It is now possible to demonstrate that there exist $\alpha_1, \alpha_2, \ldots, \alpha_q$, in the unit disc so that the minimal function in $\mathcal{F}$ is $\xi^* B(\,\cdot\,; \alpha_1, \ldots, \alpha_q)$. Here $q \leq p^* - 1$, where $p^* := \sum_{i=1}^{I} p_i$ the the sum of multiplicities of interpolation points.

The next step consists of proving that the scaled Blaschke product is optimal – no other function in $\mathcal{F}$ can have ∞-norm ξ^* or less. Although this can be done by more straightforward means, we will use order stars to that end (cf. Figure 9.4). First, we normalize data points by replacing $\omega_i^{(j)}$ with $\omega_i^{(j)}/\xi^*$. Thus, the Pick procedure produces an *unscaled* Blaschke product f, say, of unit modulus on the unit circle. We are within the framework of Section 9.1, $\mathcal{V}$ being the unit disc. Suppose that a function $g \in \mathcal{F} \setminus \{f\}$ is a $\mathcal{V}$-contraction. Since it also obeys (9.6), it follows from Theorem 9.1 that f must have at least p^* zeros in the disc, and this contradicts $q \leq p^* - 1$.

An important generalization of the Pick–Nevanlinna interpolation problem retains the conditions (9.6), but insists on a different interpolatory function space. The paradigm originates in the theory of electrical circuits (Ball and Helton, 1979): zeros of a meromorphic function correspond to sources, poles to sinks and the L_∞ norm on the unit circle[5] is the noise. The task in hand is to minimize noise, subject to a specific number of sources and sinks and provided that (9.6) is obeyed – this simply means that the system is performing its alloted task. It has been known for a long time that a Blaschke product fails, in general, to solve this problem. In an important work, Scales (1982) identified the general solution, which turns out to be a function of constant modulus along the boundary – but only at points of analyticity!

Without going in any significant detail into Scales' work, we sketch the heuristic argument that leads to the optimal function. This has the interesting side-effect of illuminating the nature of essential singularities and the difference between the sets $\mathrm{cl}\,\mathcal{V}$ and $\mathcal{V}_A$ from the previous section. Any function f which is meromorphic in the unit disc with exactly m zeros and n poles therein can be written formally as

$$f(z) = \xi^* B(z; \alpha_1, \alpha_2, \ldots, \alpha_{m+n}) F(z),$$

[5]For meromorphic functions this differs from the L_∞ norm in a disc.

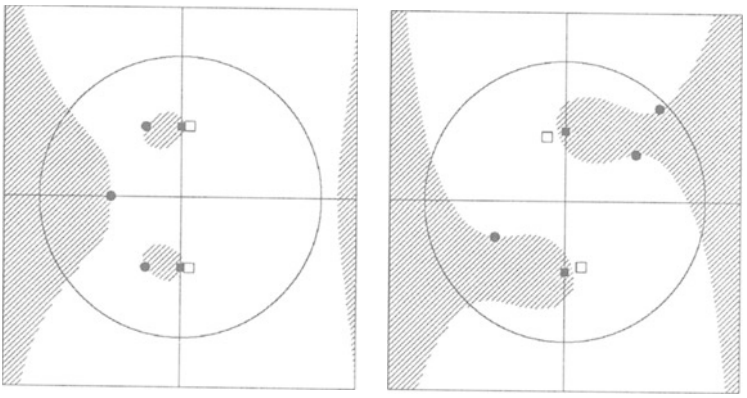


Figure 9.4 Order stars for two candidate solutions of the Pick–Nevanlinna problem: the optimal Blaschke product and a polynomial interpolant.

where $|\alpha_1|,\ldots,|\alpha_m| < 1$, $|\alpha_{m+1}|,\ldots,|\alpha_{m+n}| > 1$ and F is nonzero and analytic in $\mathcal{V}$, such that $\|F\|_\infty = 1$. It is easy to observe that there exists a function $\tilde{F}$, analytic in $\mathcal{V}$, such that $F = \exp \tilde{F}$ and

$$\operatorname{Re} \tilde{F}(z) \leq 0, \quad |z| < 1.$$

Analytic functions of nonpositive real part in the unit disc can be represented explicitly by the **Riesz–Herglotz formula** (Henrici, 1977): there exists a real Borel measure μ, supported by the interval $[-\pi,\pi]$, such that

$$\tilde{F}(z) = \int_{-\pi}^{\pi} \frac{z + e^{i\phi}}{z - e^{i\phi}} \mathrm{d}\mu(\phi).$$

Assembling our results, we obtain the representation

$$f(z) = \xi^* B(z;\alpha_1,\alpha_2,\ldots,\alpha_{m+n}) \exp\left\{ \int_{-\pi}^{\pi} \frac{z + e^{i\phi}}{z - e^{i\phi}} \mathrm{d}\mu(\phi) \right\}, \tag{9.7}$$

that is obeyed by all candidates for the optimal solution. The simplest Borel measure is an **atomic measure**, supported by a finite number of mass points,

$$\int_{-\pi}^{\pi} h(\phi) \mathrm{d}\mu(\phi) = \sum_{k=1}^{s} \mu_k h(\phi_k),$$

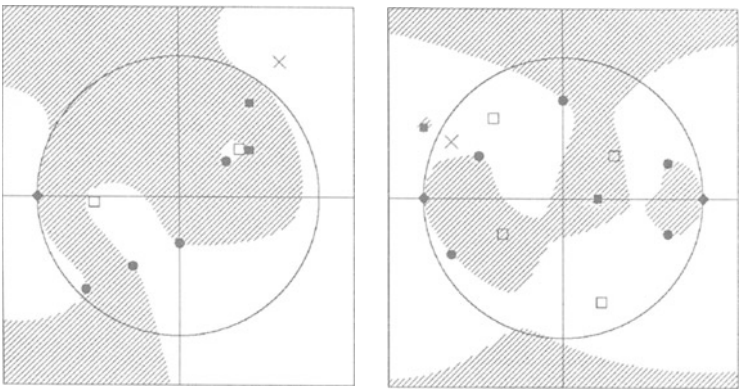


Figure 9.5 Order stars for solutions – optimal and polynomial – of the generalized Pick–Nevanlinna problem.

where $-\pi \leq \phi_1 < \cdots < \phi_s < \pi$ and $\mu_k > 0$, $k = 1, \ldots, s$. Substitution into (9.7) yields

$$f(z) = \xi^* B(z; \alpha_1, \alpha_2, \ldots, \alpha_{m+n}) \prod_{k=1}^{s} \exp\left\{\mu_k \frac{z + e^{i\phi_k}}{z - e^{i\phi_k}}\right\}. \tag{9.8}$$

The function f has an essential singularity of exponential type at the s points $e^{i\phi_1}, e^{i\phi_2}, \ldots, e^{i\phi_s}$ along the perimeter of the unit disc, and otherwise is of unit modulus there. Moreover, it follows from our argument that every function with the aforementioned features – meromorphic in the open disc, essentially analytic on the boundary and of unit modulus at points of analyticity there – can be represented in the form (9.8) with $|\xi^*| = 1$.

It is but a short step to a representation of all functions that fit the framework of Section 9.1, provided that the domain $\mathcal{V}$ is simply connected. In that case there exists a conformal mapping of $\mathcal{V}$ onto the open unit disc, i.e. an analytic, univalent function ψ such that $\psi(\mathcal{V}) = \{z \in \mathcal{C} : |z| < 1\}$. Any function that is meromorphic in $\mathcal{V}$, essentially analytic on $\partial\mathcal{V}$ and of unit modulus in $\mathcal{V}_A \setminus \mathcal{V}$ can be written as a superposition $f \circ \psi$, where f obeys (9.8). This provides an alternative avenue that might lead to singularities on the boundary, since ψ may lose analyticity on nonsmooth points of $\partial\mathcal{V}$. It is important no to confuse these with the exponential-type essential singularities from the aforementioned analysis: Extending ψ beyond $\operatorname{cl}\mathcal{V}$, it can be affirmed that, typically, its singularities on the boundary are branch points (Gaier, 1964).

We conclude this chapter on a downbeat note: the bound (9.3) is too weak and cannot be used to provide a short proof of Scales' extension to the Pick–Nevanlinna interpolation problem. Figure 9.5, which displays order stars of optimal solutions *vis-à-vis* polynomial interpolants, falls short of providing sufficient insight. However, hope unfulfilled is not hope abandoned: further research on the behaviour of order stars should lead to more profound contributions to the theory of contractive complex approximation. One highly promising avenue of approach, which we further highlight in the next chapter, emphasizes the one aspect of order stars that made them so attractive in the context of Section 9.1, namely that they allow interpolation points on the boundary.

10

Open problems

No unregarded star
Contracts its light
Into so small a character,
Removed far from our human sight,
But if we steadfast look
We shall discern
In it, as in some holy book,
How man may heavenly knowledge learn.

From *Nox Nocti Indical Scientiam* by William Habington (1605–1654).

Tales of fiction, especially those of the Hollywood variety, are likely to conclude with all loose ends tied, i's dotted, t's crossed and the brave hero and heroine living happily ever-after. This, needless to say, is far less likely in accounts of mathematical research. A new theory is spurred by old problems and well-matured conjectures, but it always leads not just to new insights, but also to new problems and new questions. The opportunistic spirit that pervades mathematics is one of its great assets and, indeed, charms: theories and concepts that might have been designed to solve one kind of problem frequently throw light on problems of an entirely different and unexpected kind.

Thus, the right way to conclude an exposition of a technique that was originally developed to answer questions in the numerical analysis of ordinary differential equations, but has been ultimately applied to a wide spectrum of targets of opportunity, is not by listing its achievements. It is much more fitting – and interesting – to speculate freely and to discuss several open problems which can be cast into the basic order-star paradigm.

This chapter is devoted to open problems and speculative ideas. Some of these are already subject to intense scrutiny, while others may well originate in a recklessly unreined imagination. Although we strive to quantify the 'speculation index' of each open problem, this is, of course, a matter for inexact and subjective judgement. *Caveat emptor!*

Problem 1: $A(\alpha)$-acceptability of Padé approximants to $\exp z$

We know well by now that the $R_{m/n}$ Padé approximant to $\exp z$ is A-acceptable if and only if $n-2 \leq m \leq n$. However, not all is lost underneath the second subdiagonal of the Padé tableau, because approximants therein are $A(\alpha)$-acceptable for some $\alpha \in (0, \pi/2)$ (cf. Section 5.3). It follows at once from the maximal modulus theorem that $A(\alpha)$-acceptability implies $A(\beta)$-acceptability for all $0 \leq \beta < \alpha$. Thus, for every $m \leq n-2$ there exists a maximal $\alpha_{m/n} < \frac{1}{2}\pi$ such that $R_{m/n}$ is $A(\alpha_{m/n})$-acceptable. What is the value of $\alpha_{m/n}$?

In a sequence of important papers, Saff and Varga (1975, 1977, 1978) analyzed the location of zeros and poles of Padé approximants to $\exp z$. They proved that these are confined to specific portions of the complex plane as m and n increase in a prescribed fashion and that there exist in $\mathcal{C}$ a parabolically-shaped zero-free and pole-free domains. For example, all the zeros of $R_{m/n}$ reside in

$$\left\{ z \in \mathcal{C} \, : \, |\arg z| > \cos^{-1}\left(\frac{m-n-2}{m+n}\right) \right\}.$$

Moreover, let

$$\mathcal{G}_\kappa := \left\{ z \in \mathcal{C} \, : \, |\arg z| > \cos^{-1}\left(\frac{1-\kappa}{1+\kappa}\right) \right\}, \quad \kappa \geq 0,$$

and $\{m_\ell\}_{\ell=0}^{\infty}$, $\{n_\ell\}_{\ell=0}^{\infty}$ be two monotonically increasing sequences of natural numbers such that

$$\lim_{\ell \to \infty} n_\ell = \infty, \quad \lim_{\ell \to \infty} \frac{m_\ell}{n_\ell} = \kappa.$$

Then, for every $\varepsilon \in (0, \kappa)$, the sequence $\{R_{m_\ell/n_\ell}\}_{\ell=0}^{\infty}$ has only a finite number of zeros in $\mathcal{C} \setminus \mathcal{G}_\kappa$. Moreover, $\mathcal{G}_\kappa$ is the smallest wedge of the form $\{z \in \mathcal{C} : |\arg z| > c\}$ with that property. Of course, results on zeros are readily 'translatable' to statements on poles, reflecting all sets in $i\mathcal{R}$.

Asymptotic location of zeros and poles is important to the study of convergence along rays in the Padé tableau. Intuitively speaking, since $\exp z$ is an entire function that never vanishes in the complex plane, an accumulation point of zeros or poles of $\{R_{m_\ell/n_\ell}\}_{\ell=0}^{\infty}$ cannot be a point of convergence. Another significance of the Saff–Varga results is to the search for $\alpha_{m/n}$. As it stands entirely to reason that, in a strictly asymptotic sense, poles are the main impediment to acceptability, the aforementioned results suggest the value

$$\lim_{\substack{\ell \to \infty \\ \frac{m_\ell}{n_\ell} \to \kappa}} \alpha_{m_\ell/n_\ell} = \cos^{-1}\left(\frac{1-\kappa}{1+\kappa}\right)$$

for all $\kappa \in [0,1]$. We credit Edward Saff[1] with the original formulation of this conjecture, which is sustained by extensive computer experimentation.

Two main difficulties in using order stars to affirm – or otherwise – the Saff conjecture are firstly, its asymptotic nature and, secondly, the difficulty in attributing a significance to $\partial(\mathcal{C} - \mathcal{G}_\kappa)$, as far as the behaviour of the order star is concerned. The first difficulty is, probably, less severe, since the conjecture might be replaced with 'finite' statements, e.g. that $\mathcal{C} - \mathcal{G}_\kappa$ belongs to the linear stability domain of $R_{m/n}$ for some $c \geq 0$. It is the difficulty to interpret the boundary of the reflected wedge that poses the more formidable problem. ▷

Problem 2: Restricted Padé approximants to $\exp z$

Restricted Padé approximants were introduced in Example 3.3 and their A-acceptability was discussed at some length in Section 4.4. The underlying assumption was that the rational function is single p-restricted – it possesses a single real pole of sufficiently high multiplicity.

Insisting on confluence of poles makes sense from purely computational point of view, since the corresponding Runge–Kutta methods are easier to solve. However, there are three possible reasons why the discussion of non-confluent poles is valuable. Firstly, as we have already seen in Section 4.4, very few singly p-restricted approximants are A-acceptable (cf. Table 4.1). Secondly, although it follows from Theorem 3.6 that the order of an m/n p-restricted approximant is at most $m+1$ (and this is true regardless of the confluence of zeros), it might well be that the local error constant can be minimized by allowing distinct poles. The third reason is different in kind and has to do with Runge–Kutta methods, not with rational approximants. As mentioned in Example 3.3, the original impetus for the study of p-restricted approximants came from the study of **Singly-Diagonally Implicit Runge–Kutta** (SDIRK) methods (Nørsett, 1974a), of the form

$$\begin{aligned}
\mathbf{k}_1 &= \mathbf{f}(a + \alpha h, \mathbf{y}_0 + c_1 h \mathbf{k}_1), \\
\mathbf{k}_i &= \mathbf{f}\left(a + c_i, \mathbf{y}_0 + h \sum_{i=1}^{i-1} a_{i,j} \mathbf{k}_j + c_1 h \mathbf{k}_i\right), \quad i = 2, 3, \ldots, n, \\
\mathbf{y}_1 &= \mathbf{y}_0 + h \sum_{i=1}^{n} b_i \mathbf{k}_i.
\end{aligned}$$

The main advantage of such methods stems from their 'decomposability': when evaluated in succession, each stage depends implicitly only upon itself. Since singly p-restricted approximants (with a pole at α^{-1}) are the underlying linear stability functions, it follows that the order of SDIRK methods

[1] in a personal communication to one of the authors

cannot exceed $n + 1$. Nørsett (1974a) proves that for $n \leq 4$ optimal-order methods exist, but this is not the case for $5 \leq n \leq 30$.[2] A possible means of producing n-stage order-$(n + 1)$ methods is by considering **Diagonally Implicit Runge–Kutta** (DIRK) methods

$$\mathbf{k}_i = \mathbf{f}\left(a + c_i h, \mathbf{y}_0 + h\sum_{j=1}^{i} a_{i,j}\mathbf{k}_j\right), \quad i = 1, 2, \ldots, n,$$

$$\mathbf{y}_1 = \mathbf{y}_0 + h\sum_{i=1}^{n} b_i \mathbf{k}_i.$$

Here we no longer insist that the $a_{i,i}$'s coincide, hence the linear stability function is, in general, multiply p-restricted. Can we do better with DIRK than with SDIRK? The question is open and outside the scope of this book but, at the very least, it motivates the attention that we pay to multiply p-restricted approximants.

Initial inroads into the problem were made in (Nørsett and Wanner, 1979) and further insight has been gained by the work of Orel (1990, 1991). By using the theory of biorthogonal polynomials (Iserles and Nørsett, 1988) he refutes our second 'motivation', proving that the least local error is obtained when all the poles coalesce and the approximant is singly p-restricted. Applying order stars, he derives bounds on the number of zeros to the left and to the right of the **external boundary** as the poles vary along a parametrized line and A-acceptability inequalities similar to (4.12). Unfortunately, these are necessary but, by no means, sufficient. An alternative approach, using optimization to study boundedness at ∞, has been promoted by Keeling (1989), who has characterized all A-acceptable n/n p-restricted Padé approximants – these coincide with the four A-acceptable choices for singly p-restricted approximants from Theorem 4.13 and nothing can be gained by allowing nonconfluence.

Extensive computer experimentation led Orel to conjecture that for every $m \geq 1$ there exists $n_0 \geq m$ such that for all $n \geq n_0$ there exists an m/n A-acceptable p-restricted approximant of order $m + 1$.[3] In other words, descending sufficiently 'deep' along a 'restricted Padé tableau', all approximants become A-acceptable. Furthermore, he also conjectures that the opposite is true if the tableau is travelled along diagonals, namely that for every $k \geq 0$ there exists n_1 such that the $(n - k)/n$ p-restricted Padé approximant fails to be A-acceptable (for all possible choices of poles) for $n \geq n_1$.

[2] A conjecture therein predicts that no n-stage SDIRK method of order $n+1$ may exist for $n \geq 5$.

[3] The conjecture is trivially true for $m = 0$ and every $n \geq 1$ by taking $R(z) = \prod_{j=1}^{n}(1 - \gamma_j z)^{-1}$ with $\gamma_1, \ldots, \gamma_n > 0$ and $\gamma_1 + \cdots + \gamma_n = 1$.

It seems highly unlikely that order stars can make a significant impact on the Orel conjectures, unless our understanding transcends the contents of Section 4.4. An alternative – and probably more promising – approach is by expanding on the work of Keeling (1989). ▷

Problem 3: Polynomial approximants to $\exp z$

It is characteristic of semi-discretized parabolic partial differential equations that they are stiff. Hence, their numerical solution leads in a natural way to rational (or algebraic, implicit) approximants to $\exp z$. This is typically not the case with semi-discretized hyperbolics – instead of large (in magnitude) eigenvalues deep inside the complex left half-plane, they possess moderate eigenvalues near the imaginary axis. Thus, it makes sense to approximate them with explicit methods that have a relatively generous stability interval along $i\mathcal{R}$,[4] and this leads in a natural way to the following problem: Given two integers $n, p \geq 1$, find the nth-degree polynomial $P_{n,p}$ that approximates $\exp z$ to order p and, among all such polynomials, has the longest interval of acceptability along $i\mathcal{R}$. The problem has been posed by van der Houwen (1972) and optimal explicit formulae for $p \leq 4$ have been provided by Kinnmark and Gray (1984a,b) in terms of Chebyshev polynomials. Thus, for $n \geq 2$

$$P_{n,1}(z) = (-1)^n \left\{ iT_{n-1}\left(\frac{iz}{n-1}\right) - \left(1 + \frac{z^2}{(n-1)^2}\right) U_{n-2}\left(\frac{iz}{n-1}\right) \right\},$$

where

$$T_m(\cos\theta) = \cos m\theta, \qquad U_m(\cos\theta) = \frac{\sin(m+1)\theta}{\sin\theta}$$

are Chebyshev polynomials of the first and the second kind respectively.[5] The interval of acceptability is $[-i(n-1), i(n-1)]$ and no nth degree polynomial $p(z) = 1 + z + \mathcal{O}(|z|^2)$ can improve upon this. Even more interestingly, if $n \geq 3$ is odd then $P_{n,1} \equiv P_{n,2}$ – the most stable polynomial is, actually, a second-order approximant!

The explicit expressions for $p = 3$ are considerably more complicated and the only matter of interest is that, similarly to $P_{n,1}$, they are conveniently written in terms of Chebyshev polynomials. The optimal interval of acceptability is bounded by $\pm i\sqrt{(n-1)^2 - 1}$. And here the plot thickens, since for every even $n \geq 4$ we have, again, an extra 'unit' of order and $P_{n,3} \equiv P_{n,4}$.

The aforementioned results are suggestive of the belief that the optimal polynomial is quite structured, that its interval of acceptability is bounded

[4] It is typical to use such methods only with the onset of asymptotic behaviour.

[5] It is easy to verify that, i's notwithstanding, $P_{n,1}$ is a real polynomial.

by 'nice' numbers and that occasionally optimality boosts the order. There are two obvious ways to investigate this problem, either by trying to find an explicit formula (expressible in Chebyshev polynomials?) or by proving general barrier theorems. Clearly, order stars are of little help in the search for explicit formulae. Their most significant promise is in proving optimality or establishing barriers. The latter task does not require an explicit knowledge of $P_{n,p}$, just an educated guess on the size of the largest acceptability interval. ▷

Problem 4: Multipoint Padé approximants to $\exp z$

One-step, multistage (that is, Runge–Kutta) and one-step, multiderivative (e.g. the Obrechkoff schemes of Section 3.3) methods for ordinary differential equations lead to rational approximants to $\exp z$. By the same reasoning, multistep methods yield algebraic approximants to the same function. It is only natural to try and combine both frameworks into 'multistep, multistage, multiderivative' methods or **general linear methods** (Butcher, 1987). We have already seen in Chapter 5 (in the instance of the multistep multiderivative method (5.11)) that this leads to the **algebraic approximant**

$$M(z,w) := Q_N(z)w^N + Q_{N-1}(z)w^{N-1} + \cdots + Q_1(z)w + Q_0(z),$$

where each Q_ℓ is a polynomial of degree n_ℓ, say. Here $n_\ell \geq -1$, where we adopt the convention that $n_\ell = -1$ corresponds to $Q_\ell \equiv 0$. It is necessary (and occasionally, e.g. for multiderivative methods, sufficient) for order of accuracy p that M approximates $\exp z$ to order p,

$$M(z,e^z) = cz^{p+1} + \mathcal{O}(|z|^{p+2}), \quad c \neq 0.$$

Moreover, the approximant is A-acceptable (and the underlying method is A-stable) if for all $z \in \mathcal{C}$ such that $\mathrm{Re}\, z < 0$ it is true that all the zeros of $M(z,w) = 0$ lie in the open unit disc.

The function M possesses $\sum_{\ell=0}^{N} n_\ell + N$ free parameters, hence we might hope to attain order of approximation $p_{\max} := \sum_{\ell=0}^{N} n_\ell + N - 1$. This expectation is fully justified and explicit formulae for the Q_ℓ's which are consistent with order $p_{\max}$ have been presented by Butcher and Chipman (1989):[6] Let

$$D := \frac{\mathrm{d}}{\mathrm{d}z}$$

[6] Different formulae for the case $N = 2$, involving biorthogonal polynomials, are given in (Iserles and Nørsett, 1987).

denote the differential operator. Then

$$Q_\ell(z) = \frac{(-1)^{n_N+1} n_N!}{(N-\ell)^{n_N-n_\ell} n_\ell!} \left(\prod_{\substack{j=0 \\ j\neq\ell}}^{N-1} \left(\frac{N-\ell}{\ell-j}\right)^{n_j+1} \left(1+\frac{D}{\ell-j}\right)^{-n_j-1} \right) \times \left(1-\frac{D}{N-\ell}\right)^{-n_N-1} z^{n_\ell}, \quad \ell = 0,1,\ldots,N-1,$$

$$Q_N(z) = \prod_{j=0}^{N-1} \left(1+\frac{D}{N-j}\right)^{-n_j-1} z^{n_N}.$$

The action of D should be understood in the formal sense. Thus, for all $m \geq 0$, $a \in R$ and a C^∞ function f,

$$\begin{aligned}(1+\alpha D)^{-m} f(z) &= \left(1+\sum_{\ell=1}^{\infty} \binom{m+\ell-1}{\ell} \alpha^\ell D^\ell\right) f(z) \\ &= f(z) + \sum_{\ell=1}^{\infty} \binom{m+\ell-1}{\ell} \alpha^\ell f^{(\ell)}(z).\end{aligned}$$

Note that the above operator maps the set of nth degree polynomials into itself.

Although Butcher and Chipman refer to **generalized Padé approximants**, we opt here for the name of **multipoint Padé approximants**, which is consistent with its usage in (Baker, 1975).

Although naive reading of Theorem 5.14 provides the bound

$$p \leq 2\max\{n_0, n_1, \ldots, n_N\}$$

for the order of A-acceptable multipoint Padé approximants, it is easy to see that the stricter bound $p \leq 2n_N$ applies. The main reason is that the proof of Theorem 5.14 rests on a pole-counting argument. Thus the inequality

$$\sum_{\ell=0}^{N-1} (n_\ell + 1) \leq n_N + 1. \tag{10.1}$$

Letting $N = 1$ we recover the inequality $n_0 \leq n_1$, the trivial part of the exact range $n_1 - 2 \leq n_0 \leq n_1$ (cf. Theorem 4.5 and the corollary to Theorem 4.1). It stands to reason that (10.1) should be supplemented by further conditions to characterize all the multipoint Padé approximants. Butcher and Chapman have checked A-acceptability for a wide range of 'candidates', coming up with the conjecture that, in the case $N = 2$, it is equivalent to

$$1 \leq n_2 - n_0 - n_1 \leq 3. \tag{10.2}$$

Arranging multipoint Padé approximants for all $n_0, n_1, n_2 \geq 0$ in a three-dimensional table, similarly to the Padé tableau, we obtain in (10.2) a layer of 'thickness' 3. This parallels the characterization of A-acceptable Padé approximants.

It is tempting to generalize (10.2) to general $N \geq 1$, into

$$N - 1 \leq n_N - \sum_{\ell=0}^{N-1} n_\ell \leq N + 1.$$

The first inequality is a consequence of the theory of Section 5.3. As Butcher and Chipman (1989) have demonstrated by extensive computer experimentation, the second inequality is, in all likelihood, necessary – but, by no means, sufficient – for A-acceptability.

A likely attempt to apply order stars to derive (10.2) should use Riemann surfaces. The main difficulty is that only the zeros of Q_0 and Q_2 are easily expressible in the geometry of the order star, whereas a successful proof should somehow take account of the degree of Q_1. This might follow from firmer knowledge of the influence of branch points on the order star. A naive approach (in the case $N = 2$) is to consider each branch point as a 'half-zero' (because of its contribution to the increase of the argument), hence overall branch points contribute $\max\{n_1, [(n_0 + n_2)/2]\}$ 'zero-equivalents'. This leads, after a standard order-star type of argument, to (10.2), except that this method of 'proof' is unsafe: Because there are two sheets in the Riemann surface, it is possible, for all we know, for the boundary to wind twice round a branch point, invalidating this line of reasoning. ▷

Problem 5: Multipoint restricted Padé approximants to $\exp z$

Multipoint Padé approximants are of limited utility, since the best ratio of order *versus* the total degree $\max\{n_0, \ldots, n_N\}$ for A-acceptable approximants is already attained by one-step methods. However, multipoint approximation comes into its own when the poles of Q_N are restricted. We have already seen in Chapter 3 that singly p-restricted n/n Padé approximants confer important advantages on the underlying numerical schemes (specifically, Runge–Kutta). Unfortunately, they suffer from two important shortcomings. Firstly, the order is restricted to $n + 1$, even without the imposition of A-acceptability. Secondly, as soon as we require A-acceptability, Theorem 4.13 narrows the field down to just four approximants. The positive features of both forms of approximation can be recovered by considering multistep, multiderivative methods (hence multipoint approximants) that, while attaining the Daniel–Moore A-stability barrier, fit into a restricted form.

Letting $Q_N(z) = (1 - \alpha z)^{n_N}$ and choosing the degrees of freedom in

$Q_0, \ldots, Q_{N-1}$ to maximize order yields **multipoint restricted Padé approximants**, that were debated by Butcher and Chipman (1989). The combination of α and $n_0, \ldots, n_N$ that yields A-acceptability is an open question that might be tractable by order-star techniques.

Another type of multipoint approximants is produced when, instead of maximizing order, we let $n_0 = n_1 = \cdots = n_N = n$ and impose order $2n$ – the largest choice which is consistent with A-acceptability. The remaining degrees of freedom are used to ensure that the approximant is, indeed, A-acceptable. The case $N = 2$ has been analyzed in detail in (Iserles, 1981d), where explicit expressions are presented. There are two degrees of freedom (including α) and it is possible to show that for $n \leq 3$ they have A-acceptable choices. However, in the case $n = 4$ a computer search failed to produce any parameters that lead to A-acceptability. Two possible remedies are either to let $N \geq 3$ or to compromise on lower, suboptimal order. Both ideas might be amenable to order-star treatment, subject to our earlier remark about better understanding of order stars on Riemann surfaces. ▷

Problem 6: The root condition for BDF

Backward differentiation formulae (BDF) are ideally suited for the integration of stiff ordinary differential equations, since they share good order of accuracy with damping of stiff components. They can be written conveniently in terms of backward differences,

$$\sum_{\ell=1}^{N} \frac{1}{\ell} \nabla_-^{\ell} \mathbf{y}_i = h\mathbf{f}_i \tag{10.3}$$

and are of order N. It has been known for a long time, and proved first by Cryer (1971), that they are zero-stable (in other words, that the polynomial r (cf. Section 5.1) obeys the root condition) if and only if $N \leq 6$. A short and beautiful proof has been provided by Hairer and Wanner (1983), using a complex-theoretical argument. Since

$$r(w) = \sum_{\ell=1}^{N} \frac{(-1)^{\ell}}{\ell} w^{N-\ell}(1-w)^{\ell},$$

transforming $w = (1-\omega)^{-1}$ results in

$$\hat{r}(\omega) := (1-\omega)^N r\left((1-\omega)^{-1}\right) = \sum_{\ell=1}^{N} \frac{\omega^{\ell}}{\ell}.$$

It is easy to see that the root condition is transformed to the requirement that all the zeros of $\hat{r}$ reside outside the disc $\mathcal{S}_1 := \{\omega \in \mathcal{C} : |z-1| \leq 1\}$,

with simple zeros along the perimeter. Hairer and Wanner express $\hat{r}$ in an integral form,

$$\hat{r}(re^{i\theta}) = e^{i\theta} \int_0^r \left(1 - e^{iN\theta}\tau^N\right) \frac{d\tau}{1 - e^{i\theta}\tau}$$

and show that the argument of $\hat{r}$ varies along the boundary of a wedge-like sector

$$\left\{ \omega = re^{i\theta} \,:\, r_- \le r \le r_+, \quad \frac{2\pi\left(m - \frac{1}{2}\right)}{N} \le \theta \le \frac{2\pi\left(m + \frac{1}{2}\right)}{N} \right\},$$

where $0 < r_- \ll 1 \ll r_+$ and $m \in \{0, 1, \ldots, N-1\}$, so as to allow exactly one zero there. Moreover, they show that $\hat{r}$ has no zeros outside a curve that asymptotically hugs the perimeter of the unit disc and that, for $N \ge 12$, forces zeros inside $\mathcal{S}_1$. This leaves out the five cases $N = 7, 8, \ldots, 11$, that can be checked easily, and determines completely the convergence properties of (10.3).

The BDF methods can be written in the form

$$\sum_{\ell=0}^{N} \alpha_\ell \mathbf{y}_{i+\ell-N} = h\beta_N \mathbf{f}_i$$

and there are valid reasons to consider a more general framework, namely

$$\sum_{\ell=0}^{N} \alpha_\ell \mathbf{y}_{i+\ell-N} = h \sum_{\ell=M}^{N} \beta_\ell \mathbf{f}_{i+\ell-N}. \tag{10.4}$$

Here $\beta_M, \ldots, \beta_{N-1}$ are fixed, whereas $\alpha_0, \alpha_1, \ldots, \alpha_N, \beta_N$ are chosen so as to maximize the order. An interesting case has been discussed in (Iserles and Stuart, 1990). It is possible to show that, when implemented with constant step size, multistep methods (5.1) may display oscillation on a grid scale (which is always a numerical artifact, hence is most unwelcome), unless $s(-1) = 0$ (cf. Section 5.1 for the definition of the polynomial s). This can be achieved in (10.4) by letting $M = N - 1$ and $\beta_{N-1} = \beta_N$. The proof of Hairer and Wanner can be extended to this case. However, it is unlikely that it is robust enough to cater for the case of general (10.4).

There are sound theoretical reasons to expect (10.4) to fail the convergence test for sufficiently large N. The order conditions mean that $\hat{r}$ (whose definition is the same as for (10.3)) is a truncated Taylor expansion of $\log\omega/Q(\omega)$ about the origin. Here Q is a rational function, $Q(0) \neq 0$. According to the Jentzsch theorem (Walsh, 1956), each point on the perimeter of the disc of convergence of an analytic function is an accumulation point of zeros of truncated Taylor expansions. Therefore zeros congregate on $|\omega| = 1$ and, ultimately, they venture into $\mathcal{S}_1$ and the root condition is lost. However, this asymptotic argument tells us nothing about the size of the critical N.

Since the problem is neatly expressible in a complex approximation-theoretical framework, order stars come to mind as a natural technique. We have already seen in this book order stars of approximants to $\log z$, both in Chapter 2 and in Chapters 5–6.[7] Unfortunately, all efforts to apply order stars to (10.4) (and even to (10.3)) have been unsuccessful. ▷

Problem 7: The root condition for general algebraic functions

In Section 5.1 order stars have been employed to prove the first Dahlquist barrier for multistep methods. This states that the root condition restricts the order p of (5.1) to

$$p \leq \begin{cases} N & : \text{explicit method}, \\ N+1+\frac{1}{2}\left(1-(-1)^{N+1}\right) & : \text{general method}. \end{cases} \tag{10.5}$$

How better can we do with a multistep, multiderivative method (5.11), or, for that matter, with any multistep, multiderivative, multistage[8] method with the stability function

$$M(z;w) = \sum_{\ell=0}^{N} \sum_{j=0}^{n} q_{\ell,j} z^j w^\ell ?$$

Dahlquist(1959) himself considered the case of $n = 2$, producing the bound

$$p \leq \begin{cases} 2N & : \text{explicit method}, \\ 2N+2 & : \text{general method}. \end{cases} \tag{10.6}$$

The inequalities (10.5) and (10.6) hint that there might exist a 'master inequality' valid for all N and n. Unfortunately, the original method of proof in (Dahlquist, 1956, 1959) did not lend itself easily to $n \geq 3$. Indeed, the general inequality has been established by Reimer (1968) by a very different approach. It reads

$$p \leq \begin{cases} nN & : \text{explicit method}, \\ n(N+1)+\frac{1}{2}(1-(-1)n(N+1)) & : \text{general method}. \end{cases} \tag{10.7}$$

Moreover, for every choice of N and n the barrier (10.7) is attainable.[9]

An interesting new proof of the **generalized first Dahlquist barrier** (10.7) appears in the work of Jeltsch and Nevanlinna (1983, 1984), who

[7] Figures 5.1d, 5.4 and 5.5 display three different order stars for the BDF method (10.3) with $N = 2$.

[8] As we have already mentioned, a multistep multistage method is a hybrid between multistep and Runge–Kutta formulae.

[9] Although, if $n(N+1)$ is even, the highest-order method obeys the root condition only marginally: all the zeros reside on the unit disc.

marry together Dahlquist's and Reimer's techniques. They also pose – and answer – further questions about the highest order (subject to the satisfaction of the root condition) of methods that obey additional 'side conditions', e.g. damping at ∞.

In the context of this book, the challenge is to produce a proof of (10.7) by order stars, possibly building upon the work in (Iserles and Nørsett, 1984) and Section 5.1. The aim is not simply to demonstrate that order stars can be used to prove each and every important stability barrier for numerical ordinary differential equations but, mainly, the hope that such a proof can provide us with much additional insight. The obstacle, that looms large, is the familiar bugbear of Riemann surfaces. ▷

Problem 8: Jeltsch–Nevanlinna comparison theorems

The theme of Section 5.4 was a brief exposition of a remarkable theorem of Jeltsch and Nevanlinna[10] (1981). Many other powerful and beautiful results *á la* Theorem 5.18 abound and it is possible to bring implicit methods into the discussion. Here we pose just two interesting open problems. The first has to do with property C. Although Jeltsch and Nevanlinna (1981) demonstrated that it is satisfied by a wide range of methods, it should be nonetheless interesting to ascertain whether we can dispose of this condition altogether, possibly by acquiring better knowledge on order stars defined on Riemann surfaces. Alternatively, it might well be the case that the condition is genuine and that the Jeltsch–Nevanlinna results can be breached by designing some strange-looking methods that contravene property C. The second problem adds an extra ingredient, the approximation order of the algebraic function w (as an approximant to $\exp z$ on the principal branch), seeking comparison results on methods of equal (or similar) orders. Evidently, order stars are the natural tool in this work. However, one should not be overly optimistic that the above questions are amenable to an easy treatment, since they have been already subjected to a very skillful analysis. ▷

Problem 9: Order reduction for Runge–Kutta methods

The order of Runge–Kutta methods provides a reliable yardstick to their local performance, as long as the underlying ordinary differential system is not excessively stiff. However, it has been observed by Frank, Schneid and Ueberhuber (1985b) that, for acutely stiff systems, the 'real' order of a method falls short of the formal order of accuracy. The intuitive reason is

[10]Different Nevanlinna to that 'from' the Pick–Nevanlinna interpolation theory and Section 9.2, although a member of the same illustrious mathematical family.

easy to grasp: each quantity

$$\mathbf{v}_i := \mathbf{y}_0 + h\sum_{j=1}^{n} a_{i,j}\mathbf{k}_j, \quad i = 1,2,\ldots,n,$$

of the n-stage method (3.4) can be thought as an approximation of the solution value at $a + c_i h$. This is obvious either from the collocation formulation of Section 3.2 or by considering a Runge–Kutta method as a generalization of a quadrature scheme, applied to the integral equation

$$\mathbf{y}(t) = \mathbf{y}(a) + \int_a^t \mathbf{f}(\tau, \mathbf{y}(\tau))\mathrm{d}\tau$$

(which, of course, is (3.1) under disguise) with abscissae at $c_1, c_2, \ldots, c_n$. Thus, it makes sense to define the **stage order** p_i *via*

$$\mathbf{v}_i = \mathbf{y}(a + c_i h) + C_i h^{p_i+1} + \mathcal{O}(h^{p_i+2}), \quad i = 1,2,\ldots,n,$$

and let

$$\bar{p} := \max_{i=1,2,\ldots,n} p_i \leq p,$$

where p is the 'classical' order of accuracy. It is proved in (Frank *et al.*, 1985b) that for very stiff systems the effective order of the method is just $\bar{p}$.

It is of interest to analyze the connection between effective order and A-stability.[11] To this end we apply (3.4) to the linear test equation $y' = \lambda y$, $y(0) = 1$. Each stage value $\mathbf{v}_i$ becomes a rational function of $z = h\lambda$, which we denote by $R_i(z)$. The linear stability function R can be obtained at once from the R_i's,

$$R(z) = 1 + \lambda\sum_{i=1}^{n} b_i R_i(z).$$

Consequently, all the R_i's share the same denominator as R (allowing for possible common factors in the numerator and the denominator). In other words,

$$R(z) = \frac{P(z)}{Q(z)} \qquad \text{and} \qquad R_i(z) = \frac{P_i(z)}{Q(z)}, \quad i = 1,2,\ldots,n.$$

For example, the fourth-order two-stage Gauss–Legendre method

$$\mathbf{k}_1 = \mathbf{f}\left(a - \frac{\sqrt{3}}{6}h, \mathbf{y}_0 + h\left(\frac{1}{4}\mathbf{k}_1 + \left(\frac{1}{4} - \frac{\sqrt{3}}{6}\right)\mathbf{k}_2\right)\right),$$

[11] It is imperative to exercise some care here: According to Frank *et al.*, order reduction occurs for nonlinear problems only. Thus, paradoxically, linear analysis is of utility, but only for nonlinear problems!

$$\begin{aligned} \mathbf{k}_2 &= \mathbf{f}\left(a + \frac{\sqrt{3}}{6}h, \mathbf{y}_0 + h\left(\left(\frac{1}{4} + \frac{\sqrt{3}}{6}\right)\mathbf{k}_1 + \frac{1}{4}\mathbf{k}_2\right)\right), \\ \mathbf{y}_1 &= \mathbf{y}_0 + \frac{h}{2}(\mathbf{k}_1 + \mathbf{k}_2) \end{aligned}$$

yields

$$R_1(z) = \frac{1 - \frac{\sqrt{3}}{z}}{1 - \frac{1}{2}z + \frac{1}{12}z^2}, \quad R_2(z) = \frac{1 + \frac{\sqrt{3}}{z}}{1 - \frac{1}{2}z + \frac{1}{12}z^2}, \quad R(z) = \frac{1 + \frac{1}{2}z + \frac{1}{12}z^2}{1 - \frac{1}{2}z + \frac{1}{12}z^2}.$$

Note that

$$\begin{aligned} R_1(z) - \exp\left(\left(\frac{1}{2} - \frac{\sqrt{3}}{6}\right)z\right) &= -\frac{\sqrt{3}}{216}z^3 + \mathcal{O}(|z|^4), \\ R_2(z) - \exp\left(\left(\frac{1}{2} + \frac{\sqrt{3}}{6}\right)z\right) &= \frac{\sqrt{3}}{216}z^3 + \mathcal{O}(|z|^4), \end{aligned}$$

hence $\bar{p} = 2$ and the effective order is just half the formal order.

The $(n+1)$-tuple $[R_1, R_2, \ldots, R_n, R]$ can be considered as a rational approximant to the vector function $[\exp c_1 z, \exp c_2 z, \ldots, \exp c_n z, \exp z]$ in which all the components share the same denominator. Similar approximants have been a focus of much attention in rational approximation theory, where they are known as **simultaneous approximants**. In general, they involve approximation to a vector function $[g_1, g_2, \ldots, g_s]$, say, by an s-tuple of rational functions with the same denominator. The highest-order approximants of this form are called **simultaneous Padé approximants** (Baker, 1975)[12] and they have been analyzed in the present context (i.e. $s = n+1$, $g_i(z) = \exp c_i z$, $i = 1, \ldots, n$, $g_{n+1}(z) = \exp z$) in (Iserles, 1981e). It is demonstrated there that approximation order n is possible for every nth degree denominator, but that there exists a unique choice of denominator that leads to approximation order $n+1$. It corresponds to collocation at **Gauss–Lobatto points** (Butcher, 1987) which, consequently, is superior (in the present narrow context) to the more familiar Gauss-Legendre scheme.

The phenomenon of order reduction is just one result in a remarkable sequence of papers by Frank, Schneid and Ueberhuber (1981, 1985a, 1985b). Equally significant are their results on the impact of individual stages on nonlinear stability features of the underlying Runge–Kutta scheme and the introduction of the concept of **BSI-stability**. These impose additional conditions on $R_1, \ldots, R_n$.

[12]There are two distinct entities that have been dubbed 'simultaneous Padé', the other being the **Hermite–Padé approximants**: $s+1$ rational functions, $\tilde{Q}, \tilde{P}_1, \tilde{P}_2, \ldots, \tilde{P}_s$, say, such that $Q(z) + \sum_i \tilde{P}_i(z) g_i(z) = \mathcal{O}(|z|^{\max})$.

Analysis of simultaneous Padé approximants by order stars requires a considerable extension of their definition. Specifically, we need to examine s-tuples of order stars, linked together by having identical poles. This, rather than the 'translation' of the theory of Frank *et al.* to the language of rational approximation, is the most formidable difficulty and, to the best of our knowledge, this approach has not been to date subjected to serious scrutiny. ▷

Problem 10: Multistep methods for the advection equation

The theme of Chapter 6 being the discussion of numerical schemes for the advection equation (6.1), much of the attention therein has been devoted to fully-discretized schemes of the form

$$\sum_{j=-\bar{r}}^{\bar{s}} \gamma_j(\mu) u_{\ell+j}^{k+1} = \sum_{j=-r}^{s} \delta_j(\mu) u_{\ell+j}^{k}, \quad \mu = \frac{\Delta t}{\Delta x}. \tag{10.8}$$

Prior to embarking on further debate of the problem, it behoves us well to set first its goals. Thus, we wish to derive methods that are **stable** for a large choice of μ, of high **order of accuracy** and preferably **explicit**. This becomes of the utmost importance when, instead of the equation $\partial u/\partial t = \partial u/\partial x$, we solve the hyperbolic system

$$\frac{\partial \mathbf{u}}{\partial t} = A \frac{\partial \mathbf{u}}{\partial x},$$

where $\sigma(A)$ is real, or variations thereof.[13] A particular focus of attention is the interplay between order and the length of the interval of 'stable' values of μ. Quite obviously, we can trade one off for the other: it makes sense in a transient stage of solution (e.g. in a boundary layer) to stake all on increasing the order of accuracy, even if this spells a severe restriction to the Courant number. On the other hand, asymptotic stages of evolution call for large solution steps with modest accuracy.[14]

It is a good strategy – in mathematics, not just in life – to optimize final outcome by increasing initial choice. Thus, instead of (10.8), we consider the **multistep fully-discretized scheme**

$$\sum_{i=0}^{N} \sum_{j=-r_i}^{s_i} a_{i,j}(\mu) u_{\ell+j}^{k+i} = 0. \tag{10.9}$$

[13] An important application is the solution of nonlinear hyperbolic systems, that can be reduced to this form by some numerical methods.

[14] Needless to say, no matter how we reconcile order with the size of the Courant number, the underlying scheme must be stable in the sense of Lax, otherwise convergence is lost.

We stipulate that $a_{i,-r_i}, a_{i,s_i} \neq 0$ for $i = 0, 1, \ldots, N$. To analyze (10.9), we require the combined wisdom of Chapters 5 and 6. Let us assume that the function

$$M(z, \mu; w) := \sum_{i=0}^{N} \sum_{j=-r_i}^{s_i} a_{i,j}(\mu) z^j w^i = \sum_{i=0}^{N} a_i(z, \mu) w^i$$

is irreducible. Then (10.9) is of order p if

$$M(z, \mu; z^\mu) = c(z-1)^{p+1} + \mathcal{O}(|z-1|^{p+2}), \quad c \neq 0.$$

Stability requires the simultaneous satisfaction of both the **Fourier condition** and the **pole condition**. The first means that, for all $\theta \in [0, \pi]$, the polynomial (in w) $M(e^{i\theta}, \mu; w)$ obeys the **root condition**, whereas the second reduces to the statement that the polynomial (in z) $z^{r_N} a_N(z, \mu)$ has exactly r_N zeros inside and s_N zeros outside the complex unit circle.

It is only reasonable to stipulate that $0 < \mu \ll 1$ is stable. In that case, there exist numbers $\mu_- \leq 0 \leq \mu_+$ such that (10.9) is stable for all $\mu \in (\mu_-, \mu_+)$. Possible goals in choosing coefficients in (10.9) might be to maximize p, given $\mu_\pm$, or to maximize $\mu_+ - \mu_-$, given $p \geq 1$.

A classical example of a two-step method is the **leapfrog** scheme

$$u_\ell^{k+2} - \mu(u_{\ell+1}^{k+1} - u_{\ell-1}^{k+1}) - u_\ell^k = 0,$$

of order 2 and stable for all $\mu \in (-1, 1)$.[15] Trading off the length of $\mu_+ - \mu_-$ for order, while confining the attention to explicit, two-step schemes with $\max r_i = \max s_i = 1$ yields the fourth order method

$$u_\ell^{k+2} + \sum_{j=-1}^{1} a_{1,j} u_{\ell+j}^{k+1} + \sum_{j=-1}^{1} a_{0,j} u_{\ell+j}^{k} = 0,$$

with

$$a_{1,-1} = \frac{1}{12} \frac{(1-4\mu^2)(2-5\mu)}{1-\mu}, \qquad a_{1,0} = -\frac{1}{6} \frac{(1-4\mu^2)(8-5\mu^2)}{1-\mu^2},$$

$$a_{1,1} = \frac{1}{12} \frac{(1-4\mu^2)(2+5\mu)}{1+\mu}, \qquad a_{0,-1} = -\frac{1}{12} \frac{(1-2\mu)^2(2+\mu)}{1+\mu},$$

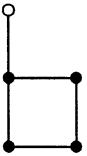
$$a_{0,0} = \frac{1}{6} \frac{(2+\mu^2)(1-4\mu^2)}{1-\mu^2}, \qquad a_{0,1} = -\frac{1}{12} \frac{(1+2\mu)^2(2-\mu)}{1-\mu},$$

which is stable for all $\mu \in \left(0, \frac{1}{2}\right)$.

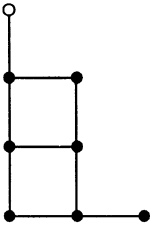
The relationship between order, stability and the shape of the 'computational stencil' has been investigated extensively by Jeltsch and his collaborators (Jeltsch, 1988; Jeltsch and Kiani, 1989; Jeltsch and Raczek, 1986;

[15] Generalizations of leapfrog have been investigated in (Iserles, 1986b).

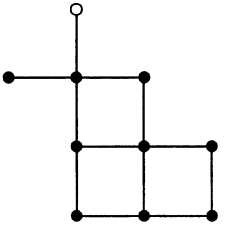
Jeltsch and Smit, 1987; Kiani, 1987). The obvious goal is to use all the available degrees of freedom to attain order conditions, in line with the logic behind the Padé schemes from Chapter 6. Thus, the order is two less than the number of points in a 'stencil'. As a sample, the following 'stencils' have been culled from a table in (Jeltsch and Raczek, 1986).



order 5,
$\mu_+ = \frac{1}{2}$;



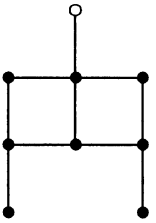
order 6,
$\mu_+ = \frac{1}{3}$;



order 8,
$\mu_+ = \frac{1}{3}$.

The left endpoint of the stability interval is $\mu_- = 0$ throughout and the hollow circle denotes the 'centre' u_ℓ^{k+N}.

Needless to say, not every highest-order method is stable (for any non-trivial choice of μ_- and μ_+. An alternative approach is to decide on a 'stencil' and optimize the order, subject to stability. It has been pursued by Kiani (1987), who provided the example of



which is of order 5 and stable in $\left(0, \frac{1}{3}\right)$. A more systematic attempt to prove stability for large families of methods has been presented in (Jeltsch and Kiani, 1989).

One-step methods are subject to the stability barrier of Theorem 6.20 and it is only fair to expect similar barriers in the multistep model. To remind the reader of the aforementioned result, we set $r^*(\mu)$ and $s^*(\mu)$ as the number of downwind and upwind points, respectively. In other words, in the 'stencil geometry', we draw a straight line $x + t = \text{const}$ through the point u_ℓ^{k+N-1} – $r^*(\mu)$ and $s^*(\mu)$ are the number of points to the left and to

the right of that line. Note that the definition is valid for multistep methods. Theorem 6.20 states that the order of a stable one-step full discretization is bounded by

$$\min\{r^*(\mu) + s^*(\mu) - 1, 2r^*(\mu), 2s^*(\mu)\}. \tag{10.10}$$

All the stable multistep schemes above are consistent with this barrier, as are all the other schemes investigated in (Jeltsch and Kiani, 1989; Jeltsch and Raczek, 1986) and elsewhere. This has prompted the **Jeltsch conjecture** (Jeltsch, 1988), that (10.10) is valid for multistep schemes.

Clearly, order stars offer the most promising line of attack, because of their success with one-step methods. However, the major difficulty is the same as with multistep schemes for ordinary differential systems, namely that not enough is known about order stars defined on Riemann surfaces. ▷

Problem 11: Advection with two space variables

The design of fully discretized schemes for

$$\frac{\partial u}{\partial t} = a\frac{\partial u}{\partial x} + b\frac{\partial u}{\partial y}, \tag{10.11}$$

where $a, b \in \mathcal{R}$ and the initial condition is given for all $(x, y) \in \mathcal{R}^2$, requires first to decide on the shape of the 'stencil'. Thus, a general explicit method of this kind can be written as

$$u_{k,\ell}^{n+1} = \sum_{(i,j)\in\mathcal{I}} \delta_{i,j} u_{k+i,\ell+j}^{n}, \tag{10.12}$$

where $u_{k,\ell}^n \approx u(k\Delta x, \ell\Delta y, n\Delta t)$ and we need first to choose the set $\mathcal{I}$. The reason why this point needs emphasizing is important: The purpose of linear analysis is, *inter alia*, to gain insight into nonlinear equations. The solution of nonlinear hyperbolic conservation laws involves spontaneously-arising discontinuities, that act as impenetrable barriers to the flow of information. It is important to mimic these barriers (as far as possible) in a numerical scheme, by locally choosing $\mathcal{I}$ that contains no points from 'the wrong end of the tracks'.[16]

It is elementary to extend the concepts of order and stability to (10.12) and the reader can easily verify that the simplest scheme,

$$u_{k,\ell}^{n+1} = (1 - 2\mu)u_{k,\ell}^n + a\mu u_{k+1,\ell}^n + b\mu u_{k,\ell+1}^n,$$

[16]It should be transparent that analyzing (10.11) for the sake of a nonlinear equation and letting $\mathcal{I}$ in (10.12) to be independent of k and ℓ is just the first step in a formidable programme of research. However, to paraphrase a well-known Chinese proverb into numerical terminology, every stable algorithm for asymptotic solution starts with a first step...

where $\mu := \Delta t/\Delta x$,[17] is of order 1 and, provided that $ab \geq 0$, it is stable as long as $0 \leq a\mu, b\mu \leq 1$. Stability analysis is surprisingly intricate for more complicated sets $\mathcal{I}$. For example, proving that the second-order scheme with $\mathcal{I} = \{(0,0),(1,0),(0,1),(1,1),(-1,0),(0,-1)\}$ and

$$\begin{aligned}
\delta_{0,0} &= 1 - \mu^2(a^2 - ab + b^2), \\
\delta_{1,0} &= \frac{1}{2}a\mu(1 + (a - 2b)\mu), \\
\delta_{0,1} &= \frac{1}{2}b\mu(1 + (-2a + b)\mu), \\
\delta_{1,1} &= ab\mu^2, \\
\delta_{-1,0} &= -\frac{1}{2}a\mu(1 - a\mu), \\
\delta_{0,-1} &= -\frac{1}{2}b\mu(1 - b\mu)
\end{aligned}$$

is stable for the same range of μ is very time consuming and results for more 'exotic' sets $\mathcal{I}$ are a matter for either computer experimentation or pure guesswork.

Chapter 6 makes the case for order stars as a very efficient tool in the analysis of the advection equation in one space variable, hence it is tempting to use them in a multivariate framework. This, however, necessitates the extension of the theory of Chapter 2 to two complex variables, a formidable exercise of great theoretical interest *per sé*. ▷

Problem 12: Padé approximants to slit functions

The class of **Stieltjes functions** has been already mentioned in Section 2.3. These are of the form

$$g(z) = \int_{-\infty}^{0} \frac{\mathrm{d}\tilde{\mu}(\tau)}{\tau - z},$$

where $\tilde{\mu}$ is a distribution, and their $(n-1)/n$ Padé approximants possess many wondrous properties. The denominator of $R_{(n-1)/n}$ can be written down explicitly as $z^n p_n(-z^{-1})$, where p_n is the nth degree orthogonal polynomial with respect to the distribution $\tilde{\mu}$. Therefore, all the poles of the approximant lie along the branch cut. It is possible to identify the numerator of $R_{(n-1)/n}$ with another orthogonal polynomial (with respect to a different distribution) and to prove that its zeros lie, as well, on the branch cut and interlace there with the poles. The first – and most trivial – conclusion is that the approximant belongs, like the function g itself, to the class $\mathbf{N}$ of functions that map analytically the open upper half-plane into its closure. More importantly, the interlace of zeros and poles implies that the residua

[17] For the sake of simplicity, we assume in the sequel that $\Delta y = \Delta x$.

at the poles are positive and this feature can be used to argue that the set $\{R_{(n-1)/n}(z)\}_{n=1}^{\infty}$ can be uniformly bounded on any compact domain $\mathcal{X}$ that does not include points on the branch cut. Suppose now that the branch cut does not extend all the way to the origin and that g is analytic there. Then there exists a neighbourhood $\mathcal{Y}$ about the origin where, by virtue of the Taylor theorem and the definition of Padé approximants,

$$\lim_{n\to\infty} R_{(n-1)/n} = g(z), \quad z \in \mathcal{Y} \subset \mathcal{X}.$$

The stage is now set to apply the Stieltjes–Vitali–Montel theorem, namely that if a sequence of analytic functions $\{h_n\}_{n=1}^{\infty}$, say, is uniformly bounded in a compact domain and converges to an analytic function in a neighbourhood then it uniformly converges to that function throughout the whole domain (Baker, 1975). It follows that the $R_{(n-1)/n}$'s converge uniformly to g in any compact domain, away from the 'bad interval', the branch cut.[18]

Lest it be assumed that there is something special about the first superdiagonal of the Padé tableau, we hasten to state that similar results are valid for all m/n approximants, for all $m \geq n-1 \geq 0$, although the proofs are slightly more intricate. The reader may consult (Baker, 1975) for an extensive exposition of the theory of Stieltjes functions and Padé approximants thereof. Moreover, it is perfectly allowed to abandon Padé approximation in favour of interpolation or L_∞ approximation without interfering with the interlace property and, by implication, with the convergence result (Barnsley, 1973; Blatt *et al.*, 1987).

Extension of the aforementioned analysis to the whole of class N and to more general slit functions (cf. Section 2.3) is an exciting prospect. The link between approximation at the origin and the configuration of zeros and poles lies comfortably within the order stars' paradigm. The main obstacle lies in the presence of branch cuts, where jumps in the argument may interfere with the all-important Proposition 2.3. Careful use of the Stieltjes–Perron formula for jumps across branch cuts (Akhiezer, 1965) might be helpful in that respect. ▷

Problem 13: Complex approximation and interpolation

The link between interpolation and the position of zeros and poles, so central to order stars, features prominently in L_p approximation and interpolation by polynomials or rational functions. Nonetheless, any possible application of order stars to complex approximation is highly speculative, unless the

[18]The theory of Padé approximants to Stieltjes functions has been described first in an – admittedly long – paper that also introduced the Stieltjes integral, proved the conditions for determinacy of the Stieltjes moment problem, described the Stieltjes–Wigert polynomials and the Stieltjes–Perron formula... Little credit for guessing the author of this marvelous paper, with unique impact on mathematical analysis (Stieltjes, 1894).

correct questions are asked. Order stars are an exclusively qualitative tool and they cannot be expected to answer questions on the approximation error or Lebesgue constants, say. They cannot prove existence either, and their main promise lies in analyzing uniqueness, location of zeros and poles, contractivity and other 'structural' features of approximating functions.

Contractive interpolation is a prime candidate for an order star treatment, going well beyond the material in Chapter 9. The first goal might be to prove the optimality of the Scales (1982) solution of the generalized Pick–Nevanlinna interpolation problem by using order stars. Our present understanding of the order star geometry falls short of that goal (cf. Figure 9.5). More tractable subject-matter might be the solution of the classical (that is, with analyticity assumed in the domain) Pick–Nevanlinna problem, but in a general complex domain $\mathcal{V}$ with a Jordan boundary, instead of the unit disc. As long as $\mathcal{V}$ is simply-connected and its boundary is analytic, a conformal mapping of $\mathcal{V}$ onto the unit disc can be continued analytically to the boundary. Consequently, results from the disc can be 'lifted' to $\mathcal{V}$, without any need for order stars. This is not the case, however, if $\mathcal{V}$ is multiply-connected (since it cannot be conformally mapped onto a disc) or if its boundary is nonanalytic (because the conformal mapping will have – typically logarithmic – singularities at the images of points of nonanalyticity). Only the future can tell if order stars can be applied to this set of problems.

Particular promise lies in investigating interpolation at boundary points, since the more conventional techniques, based on the Rouche theorem, fail there. Section 9.1 demonstrates that order stars can accommodate interpolation along the boundary and that, unsurprisingly, it leads to different formulae. Interpolation at the boundary obviously sets a lower bound on the L_∞ norm there, hence it does not sit easily with the classical Pick–Nevanlinna theory. However, the ability to follow interpolation right to the boundary opens up a whole new range of problems to mathematical scrutiny. As a sample of these, consider interpolation of total multiplicity p in the *closed* unit disc, under the extra condition that all the function values that are set on the unit circle share the same modulus. It is possible to apply the theory of Section 9.1 to prove that if the data are interpolated by a Blaschke product of sufficiently low degree then this is the optimal (in the L_∞ sense) interpolant. Of course, this says nothing about the existence of such a Blaschke product, that must be sought by more ortodox methods. ▷

Bibliography

Abramowitz, M. and Stegun, I.A. (1965), *Handbook of Mathematical Functions*, Dover, New York.

Ahlfors, L.V. (1966), *Complex Analysis*, McGraw–Hill, New York.

Akhiezer, N.I. (1965), *The Classical Moment Problem*, Oliver and Boyd, Edinburgh and London.

Arnold, V.I. (1978), *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York.

Baker, G.A. (1975), *Essentials of Padé Approximants*, Academic Press, New York.

Ball, J. and Helton, W. (1979), Interpolation with outer functions and gain equalization in amplifiers, in *Proceedings of the International Conference on Mathematical Theory of Circuits and Systems*, Delft.

Barnsley, M. (1973), The bounding properties of the multipoint Padé approximant to a series of Stieltjes on the real line, *J. Math. Phys.* **14**, 299–313.

Birkhoff, G. and Varga, R.S. (1965), Discretization errors for well-set Cauchy problems, *Int. J. Math. Phys.* **44**, 1–23.

Blatt, H.-P, Iserles, A. and Saff, E.B. (1987), Remarks on the behaviour of zeros of best approximating polynomials and rational functions, in *Algorithms for Approximation, Shrivenham 1985* (J.C. Mason and M.G. Cox, eds), Oxford University Press, Oxford, 437–445.

Brezinski, C. (1990), *Padé-Type Approximation and General Orthogonal Polynomials*, Birkhäuser-Verlag, Basel.

Burrage, K. (1988), Order properties of implicit multivalue methods for ordinary differential equations, *IMA J. Num. Anal.* **8**, 43–69.

Butcher, J.C. (1987), *The Numerical Analysis of Ordinary Differential Equations, Runge-Kutta Methods and General Linear Methods*, Wiley, Chichester.

Butcher, J.C. (1988), Towards efficient implementation of singly-implicit methods, *ACM Trans. on Math. Software* **14**, 68–75.

Butcher, J.C. and Chipman, F.H. (1989), Generalized Padé approximations to the exponential function, University of Acadia Technical Report.

Cartwright, M.L. (1935), Some inequalities in the theory of functions, *Math. Ann.* **111**, 98–118.

Cheney, E.W. (1966), *Introduction to Approximation Theory*, McGraw–Hill, New York.

Chihara, T.S. (1978), *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York.

Cody, W.J., Meinardus, G. and Varga, R.S. (1969), Chebyshev rational approximations to e^{-x} in $[0, +\infty)$ and applications to heat-conduction problems, *J. Approx. Th.* **2**, 50–65.

Crandall, S.H. (1955), An optimum implicit recurrence formula for the heat conduction equation, *J. Assoc. Comput. Mach.* **2**, 42–49.

Crouzeix, M. and Ruamps, F. (1977), On rational approximations to the exponential, *RAIRO Analyse Numérique* **11**, 241–243.

Cryer, C.W. (1971), A proof of the instability of backward-difference multistep methods for the numerical integration of ordinary differential equations, University of Wisconsin, Madison, report 117.

Dahlquist, G. (1956), Convergence and stability in the numerical integration of ordinary differential equations, *Math. Scand.* **4**, 33–53.

Dahlquist, G. (1959), Stability and error bounds in the numerical integration of ordinary differential equations, KTH Stockholm report 130.

Dahlquist, G. (1963), A special stability problem for linear multistep methods, *BIT* **3**, 27–43.

Daniel, J.W. and Moore, R.E. (1970), *Computation and Theory in Ordinary Differential Equations*, Freeman, San Francisco.

Douglas, J. (1961), A survey of numerical methods for parabolic differential equations, *Advances in Computers* **2**, 1–55.

Ehle, B.L. (1968), High order A-stable methods for the numerical solution of systems of D.E.'s, *BIT* **8**, 276–278.

Ehle, B.L. (1969), On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems, Ph.D. thesis, University of Waterloo.

Ehle, B.L. (1973), A-stable methods and Padé approximations to the exponential, *SIAM J. Math. Anal.* **4**, 671–680.

Ehle, B.L. (1976), On certain order constrained Chebyshev rational approximations, *J. Approx. Th.* **17**, 297–306.

Ehle, B.L. and Picel, Z. (1975), Two-parameter, arbitrary order, exponential approximations for stiff initial-value problems, *Math. Comput.* **19**, 501–511.

Engquist, B. and Osher, S. (1981), One-sided difference approximations for non-linear conservation laws, *Math. Comput.* **36**, 321–352.

Erdélyi, A., Magnus, W., Oberhettinger, F. and Tricomi, F.G. (1953), *Higher Transcendental Functions*, McGraw-Hill, New York.

Frank, R., Schneid, J. and Ueberhuber, C.W. (1981), The concept of B-convergence, *SIAM J. Num. Anal.* **18**, 753–780.

Frank, R., Schneid, J. and Ueberhuber, C.W. (1985a), Stability properties of implicit Runge–Kutta methods, *SIAM J. Num. Anal.* **22**, 497–514.

Frank, R., Schneid, J. and Ueberhuber, C.W. (1985b), Order results for implicit Runge–Kutta methods applied to stiff systems, *SIAM J. Num. Anal.* **22**, 515–534.

Gaier, D. (1964), *Konstruktive Methoden der konformen Abbildung*, Springer-Verlag, Berlin.

Gasper, G. and Rahman, M. (1990), *Basic Hypergeometric Series*, Cambridge University Press, Cambridge.

Genin, Y. (1974), An algebraic approach to A-stable linear multistep-multiderivative integration formulas, *BIT* **14**, 382–406.

Glover, K. (1984), All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds, *Int. J. Control* **39**, 1115–1193.

Goluzin, G.M. (1969), *Geometric Theory of Functions of Complex Variables*, AMS Trans. Math. Monographs 26.

Gonzales, C. (1987), On the A-acceptability of Padé-type approximants to the exponential with a single pole, *J. Comput. Appl. Math.* **19**, 133-140.

Gourlay, A.R. and Mitchell, A.R. (1968), High accuracy A.D.I. method for parabolic equations with variable coefficients, *Numer. Math.* **12**, 180–185.

Gragg, W.B. (1972), The Padé table and its relation to certain algorithms of numerical analysis, *SIAM Review* **14**, 1–62.

Guckenheimer, J. and Holmes, P. (1983), *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York.

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

Hairer, E. (1979), Unconditionally stable methods for second-order differential equations, *Numer. Math.* **32**, 373–379.

Hairer, E. (1982), Constructive characterization of A-stable approximations to $\exp(z)$ and its connection with algebraically stable Runge–Kutta methods, *Numer. Math.* **39**, 247–258.

Hairer, E., Iserles, A. and Nørsett, S.P. (1985), Rational approximations to the exponential function with two complex conjugate interpolation points, *SIAM J. Math. Anal.* **16**, 814–821.

Hairer, E., Nørsett, S.P. and Wanner, G. (1987), *Solving Ordinary Differential Equations I: Nonstiff Problems*, Springer-Verlag, Berlin.

Hairer, E. and Wanner, G. (1983), On the instability of the BDF formulas, *SIAM J. Num. Anal.* **20**, 1206–1209.

Henrici, P. (1962), *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, New York.

Henrici, P. (1977), *Applied and Computational Complex Analysis*, Vol. 2, Wiley, New York.

Hille, E. (1962), *Analytic Function Theory*, Blaisdell, Waltham.

van der Houwen, P.J. (1972), Explicit Runge–Kutta formulas with increased stability boundaries, *Numer. Math.* **20**, 149–164.

Iserles, A. (1979), On the generalized Padé approximations to the exponential function, *SIAM J. Num. Anal.* **16**, 631–636.

Iserles, A. (1981a), Rational interpolation to $\exp(-x)$ with application to certain stiff systems, *SIAM J. Num. Anal.* **18**, 1–14.

Iserles, A. (1981b), Generalized order star theory, in *Rational Approximation, Theory and Application* (H. van Rossum and M.G. de Bruin, eds.), Springer-Verlag LNiM 888, Berlin, 228–238.

Iserles, A. (1981c), Padé and rational approximations to the exponential and their applications in numerical analysis, in *Padé Approximation and Convergence Acceleration Techniques* (J. Gilewicz, ed.), Centre de Physique Théorique, Marseille, 52–59.

Iserles, A. (1981d), Two-step numerical methods for parabolic differential equations, *BIT* **21**, 80–96.

Iserles, A. (1981e), On multivalued exponential approximations, *SIAM J. Num. Anal.* **18**, 480–499.

Iserles, A. (1982), Order stars and a saturation theorem for first order hyperbolics, *IMA J. Num. Anal.* **2**, 49–61.

Iserles, A. (1983), Order stars and the structure of Padé tableaux, in *Padé Approximation, Bad Honnef 1983* (H. Werner, ed.), Springer-Verlag LNiM 1071, Berlin, 166–175.

Iserles, A. (1984), Order stars, contractivity and a Pick-type theorem, in *Rational Approximation and Interpolation* (P.R. Graves-Morris, E.B. Saff and R.S. Varga, eds), Springer-Verlag LNiM 1105, Berlin, 117–124.

Iserles, A. (1985a), Order stars, approximations and finite differences I. The general theory of order stars, *SIAM J. Math. Anal.* **16**, 559–576.

Iserles, A. (1985b), Order stars, approximations and finite differences II. Theorems in approximation theory, *SIAM J. Math. Anal.* **16**, 785–802.

Iserles, A. (1985c), Order stars, approximations and finite differences III. Finite differences for $u_t = \omega u_{xx}$, *SIAM J. Math. Anal.* **16**, 1020–1033.

Iserles, A. (1986a), Order stars and stability barriers, in *Numerical Analysis, Dundee 1985* (D.F. Griffiths and G.A. Watson, eds), Longman, Harlow, 98–111.

Iserles, A. (1986b), Generalized leapfrog methods, *IMA J. Num. Anal.* **6**, 381–392.

Iserles, A. (1990), Efficient Runge–Kutta methods for Hamiltonian equations, University of Cambridge report DAMTP 1990/NA10.

Iserles, A. and Nørsett, S.P. (1982), Rational approximations to the exponential function with two complex conjugate interpolation points, Report 7/82, Norwegian Institute of Technology, Trondheim.

Iserles, A. and Nørsett, S.P. (1983), Frequency fitting of rational approximations to the exponential function, *Math. Comput.* **40**, 547–559.

Iserles, A. and Nørsett, S.P. (1984), A proof of the first Dahlquist barrier by order stars, *BIT* **24**, 529–537.

Iserles, A. and Nørsett, S.P. (1985), A-acceptability of derivatives of rational approximations to $\exp(z)$, *J. Approx. Th.* **43**, 327–337.

Iserles, A. and Nørsett, S.P. (1987), Two-step methods and bi-orthogonality, *Math. Comput.* **49**, 543–552.

Iserles, A. and Nørsett, S.P. (1988), On the theory of bi-orthogonal polynomials, *Trans. Amer. Math. Soc.* **306**, 455–474.

Iserles, A. and Nørsett, S.P. (1989), Order stars and rational approximants to $\exp(z)$, *Appl. Num. Math.* **5**, 63–70.

Iserles, A., Peplow, A.T. and Stuart, A.M. (1990), A unified approach to spurious solutions introduced by time discretisation. Part I: Basic theory, University of Cambridge report DAMTP 1990/NA4.

Iserles, A. and Powell, M.J.D. (1981), On the A-acceptability of rational approximations that interpolate the exponential function, *IMA J. Num. Anal.* **1**, 241–251.

Iserles, A. and Strang, G. (1983), The optimal accuracy of difference schemes, *Trans. Amer. Math. Soc.* **277**, 779–803.

Iserles, A. and Stuart, A.M. (1990), A unified approach to spurious solutions introduced by time discretisation. Part II: BDF-like methods, University of Cambridge report DAMTP 1990/NA6.

Iserles, A. and Williamson-Renaut, R.A. (1984), Stability and accuracy of semi-discretized finite difference methods, *IMA J. Num. Anal.* **4**, 289–307.

Jeltsch, R. (1976), Note on A-stability of multistep-multiderivative methods, *BIT* **16**, 74–78.

Jeltsch, R. (1985), Stability and accuracy of difference schemes for hyperbolic problems, *J. Comput. Appl. Math.* **12–13**, 91-108.

Jeltsch, R. (1988), Order barriers for difference schemes for linear and nonlinear hyperbolic problems, in *Numerical Analysis, Dundee 1987* (D.F. Griffiths and G.A. Watson, eds), Longman, Harlow, 157–175.

Jeltsch, R. and Kiani, P. (1989), Stability of a family of multi-time-level difference schemes for the advection equations, RWTH Aachen report 59.

Jeltsch, R. and Nevanlinna, O. (1981), Stability of explicit time discretizations for solving initial value problems, *Numer. Math.* **37**, 61–91.

Jeltsch, R. and Nevanlinna, O. (1982), Stability of time discretizations for initial value problems, *Numer. Math.* **40**, 245–296.

Jeltsch, R. and Nevanlinna, O. (1983), Accuracy of multistage multistep formulas, RWTH Aachen report 23.

Jeltsch, R. and Nevanlinna, O. (1984), Dahlquist's first barrier for multistage multistep formulas, *BIT* **24**, 538–555.

Jeltsch, R. and Raczek, K. (1986), Counter examples to an order barrier for stable multistep discretizations of linear hyperbolic equations, RWTH Aachen report 39.

Jeltsch, R. and Smit, J.H. (1985), Bounds for the accuracy of difference methods for hyperbolic differential equations, RWTH Aachen report 31.

Jeltsch, R. and Smit, J.H. (1987), Accuracy bounds of two time level difference schemes for hyperbolic equations, *SIAM J. Num. Anal.* **24**, 1–11.

Jeltsch, R. and Strack, K.-G. (1985), Accuracy bounds for semidiscretizations of hyperbolic problems, *Math. Comput.* **45**, 365-376.

Karlin, S. (1968), *Total Positivity*, Vol. I, Stanford University Press, Stanford, Calif.

Keeling, S.L. (1989), On implicit Runge–Kutta methods with a stability function having distinct real poles, *BIT* **29**, 91–109.

Kiani, P. (1987), A high order stable and differentiable 3-step scheme for the linear constant coefficient advection equation, RWTH Aachen report 49.

Kinnmark, I.P.E. and Gray, W.G. (1984a), One step integration methods with maximum stability regions, *Math. & Comput. in Simulation* **26**, 87–92.

Kinnmark, I.P.E. and Gray, W.C. (1984b), One step integration methods of third–fourth order accuracy with large hyperbolic stability limits, *Math. & Comput. in Simulation* **26**, 181–188.

Krall, H.L. and Frink, O. (1949), A new class of orthogonal polynomials: the Bessel polynomials, *Trans. Amer. Math. Soc.* **65**, 100-115.

Krein, M.G. (1958), Integral equations on a half-line with kernel depending upon the difference of the arguments, *Uspehi Mat. Nauk* **13**, 3–120.

Lambert, J.D. (1973), *Computational Methods in Ordinary Differential Equations*, Wiley, London.

Lasagni, F. (1988), Canonical Runge–Kutta methods, *ZAMP* **39**, 952–953.

Lax, P.D. (1973), *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, SIAM, Philadelphia.

Liniger, W. and Willoughby, R.A. (1967), Efficient numerical integration of stiff systems of ordinary differential equations, IBM Research Report RL-1970.

Mitchell, A.R. and Griffiths, D.F. (1980), *The Finite Difference Method in Partial Differential Equations*, Wiley, Chichester.

Nørsett, S.P. (1974a), Multiple Padé approximations to the exponential function, Ph.D. thesis, University of Dundee.

Nørsett, S.P. (1974b), One-step methods of Hermite type for numerical integration of stiff systems, *BIT* **14**, 63–77.

Nørsett, S.P. (1975), C-polynomials for rational approximation to the exponential function, *Numer. Math.* **25**, 39–56.

Nørsett, S.P. (1978), Restricted Padé approximations to the exponential function, *SIAM J. Numer. Anal.* **15**, 1008–1029.

Nørsett, S.P. and Trickett, S.R. (1984), Exponential fitting of restricted rational approximations to the exponential function, in *Rational Approximation and Interpolation* (P.R. Graves-Morris, E.B. Saff and R.S. Varga, eds) Springer-Verlag LNiM 1105, Berlin, 466–476.

Nørsett, S.P. and Trickett, S.R. (1986), Order-constrained uniform approximations to the exponential based on restricted rationals, *Constr. Approx.* **2**, 189–195.

Nørsett, S.P. and Wanner, G. (1979), The real-pole sandwich for rational approximations and oscillation equations, *BIT* **19**, 89–94.

Nørsett, S.P. and Wolfbrandt, A. (1977), Attainable order of rational approximations to the exponential function with only real poles, *BIT* **17**, 200–208.

Obrechkoff, N. (1942), Sur les quadratures mécaniques, *Spisanie Bulgar. Akad. Nauk* **65**, 191–289 (in Bulgarian, French summary).

Orel, B. (1990), Real pole approximations to the exponential function, Norwegian Institute of Technology report 1/90.

Orel, B. (1991), Runge–Kutta methods with real eigenvalues, Ph.D. thesis, University of Ljubljana.

Osher, S. (1969), Systems of difference equations with general homogeneous boundary conditions, *Trans. Amer. Math. Soc.* **137**, 177–201.

Pólya, G. and Szegő, G. (1979), *Problems and Theorems in Analysis*, Springer-Verlag, Berlin.

Powell (1981), M.J.D. *Approximation Theory and Methods*, Cambridge University Press, Cambridge.

Rainville, E.D. (1967), *Special Functions*, Macmillan, New York.

Reimer, M. (1968), Finite difference forms containing derivatives of higher order, *SIAM J. Numer. Anal.* **5**, 725–738.

Richtmyer, R.D. and Morton, K.W. (1967), *Difference Methods for Initial Value Problems*, Wiley, New York.

Saff, E.B. and Varga, R.S. (1975), On the zeros and poles of Padé approximants to e^z, *Numer. Math.* **25**, 1–14.

Saff, E.B. and Varga, R.S. (1977), On the zeros and poles of Padé approximants to e^z. II, in *Padé and Rational Approximations: Theory and Applications* (E.B. Saff and E.S. Varga, eds), Academic Press, New York, 195–213.

Saff, E.B. and Varga, R.S. (1978), On the zeros and poles of Padé approximants to e^z. III, *Numer. Math.* **30**, 241–266.

Samarskĭ, A.A. (1964), Economical difference schemes for systems of equations of parabolic type, *Zh. Vychisl. Mat. i Fiz.* **4**, 927–930 (in Russian).

Sanz-Serna, J.M. (1988), Runge–Kutta schemes for Hamiltonian systems, *BIT* **28**, 877–883.

Sarason, D. (1967), Generalized interpolation in H^∞, *Trans. Amer. Math. Soc.* **127**, 179–203.

Scales, W.A. (1982), Interpolation with meromorphic functions of minimal norm, Ph.D. thesis, University of California, San Diego.

Slater, L.J. (1966), *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge.

Stetter, H.J. (1973), *Analysis of Discretization Methods for Ordinary Differential Equations*, Springer-Verlag, Berlin.

Stieltjes, T.J. (1894), Recherches sur les fractions continues, *Ann. Fac. Sci. Univ. Toulouse* **8J**, 1–122 and **9A**, 1–47.

Strang, G. (1962), Trigonometric polynomials and difference methods of maximum accuracy, *J. Math. Phys.* **41**, 147–154.

Strang, G. (1964a), Accurate partial difference methods. II. Non-linear problems, *Numer. Math.* **6**, 37–46.

Strang, G. (1964b), Wiener–Hopf difference equations, *J. Math. Mech.* **13**, 85–96.

Strang, G. (1966), Implicit difference methods for initial-boundary value problems, *J. Math. Anal. Appl.* **16**, 188–198.

Strang, G. and Iserles, A. (1983), Barriers to stability, *SIAM J. Num. Anal.* **20**, 1251–1257.

Suris, Y.B. (1989), Canonical transformations generated by methods of Runge–Kutta type for the numerical integration of the system $x'' = -\partial U/\partial x$, *Zh. Vychisl. Mat. i Mat. Fiz.* **29**, 202–211 (in Russian).

Szegő, G. (1939), *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ. 23, Providence, R.I.

Trefethen, L.N. (1981), Rational Chebyshev approximation on the unit disk, *Numer. Math.* **37**, 297–320.

Trefethen, L.N. (1982), Group velocity in finite difference schemes, *SIAM Review* **24**, 113–136.

Varga, R.S. (1962), *Matrix Iterative Analysis*, Prentice Hall, Englewood Cliffs, NJ.

Walsh, J.L. (1956), *Interpolation and Approximation by Rational Functions in the Complex Domain*, Amer. Math. Soc. Colloq. Publ. 20, Providence, RI.

Wanner, G. (1987), Order stars and stability, in *State of the Art in Numerical Analysis* (A. Iserles and M.J.D. Powell, eds), Oxford University Press, Oxford, 451–471.

Wanner, G., Hairer, E. and Nørsett, S.P. (1978), Order stars and stability theorems, *BIT* **18**, 475–489.

Widlund, O.B. (1967), A note on unconditionally stable linear multistep methods, *BIT* **7**, 65–70.

Wolfbrandt, A. (1977), A study of Rosenbrock processes with respect to order and stiff stability, Ph.D. thesis, Chalmers University of Technology, Gothenburg, Sweden.

Name index

Subject index