Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

Explicit Runge-Kutta methods for parabolic partial differential equations

J.G. Verwer

Department of Numerical Mathematics

NM-R9602 January 31, 1996

Report NM-R9602
ISSN 0169-0388

CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands

Explicit Runge-Kutta Methods for Parabolic Partial Differential Equations *

J.G. Verwer

CWI

P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract

Numerical methods for parabolic PDEs have been studied for many years. A great deal of the research focuses on the stability problem in the time integration of the systems of ODEs which result from the spatial discretization. These systems often are stiff and highly expensive to solve due to a huge number of components, in particular for multi-space dimensional problems. The combination of stiffness and problem size has led to an interesting variety of special purpose time integration methods. In this paper we review such a class of methods, viz. explicit Runge-Kutta methods possessing extended real stability intervals.

AMS Subject Classification (1991): 65M20, 65M12

CR Subject Classification (1991): G.1.8

Keywords & Phrases: Parabolic partial differential equations, Runge-Kutta methods.

1. Introduction

In this paper we review a class of special purpose methods for the time integration of parabolic PDEs, viz. explicit Runge-Kutta (RK) methods possessing extended real stability intervals. Following the method of lines approach, we assume that the parabolic problem has been converted to a system of ODEs, in $\Re^m$ say, by means of space discretization. We write this system as

$$\frac{dU(t)}{dt} = F(t, U(t)), \quad 0 < t \leq T, \quad U(0) = U_0, \tag{1.1}$$

where U_0 is a given initial vector and boundary conditions are supposed to be contained in (1.1). For the derivation and stability analysis of the stabilized methods there is no need to specify a particular class of PDEs or the space discretization. The only restrictions we assume are (i) the eigenvalues of the Jacobian matrix $F'(t,U) = \partial F(t,U)/\partial U$ lie in a long narrow strip along the negative axis of the complex plane, and (ii) $F'(t,U)$ is close to normal. These two properties trivially hold if $F'(t,U)$ is symmetric and non-positive definite, properties frequently encountered when discretizing elliptic operators. With regard to (i) it should be noted that the methods remain applicable if eigenvalues exist possessing a positive real part. As usual, these eigenvalues are not taken into account in the derivation and stability analysis.

Stabilized RK methods can be easily applied to large problem classes with only a low memory demand. Their advantage, when compared with fully implicit or partly implicit methods, is that they do not require the solution of large and complicated systems of nonlinear algebraic or transcendental equations when solving multi-space dimensional problems. Of course, they do possess a finite, real interval for absolute stability. However, this interval is much larger than for standard explicit RK methods because the real boundary, denoted by $\beta(s)$, is always of the quadratic form

[*] Paper prepared on the occasion of the symposium *One Hundred Years of Runge-Kutta Methods*, held at CWI, December 8, 1995. Submitted for publication to the related special issue of *Applied Numerical Mathematics*, editor J.C. Butcher.

$$\beta(s) = c(s)s^2, \tag{1.2}$$

where s is the number of stages and $c(s)$ is a nearly constant function of s. The quadratic dependence emanates from the use of first kind Chebyshev polynomials. Because the scaled stability interval $\beta(s)/s$, which takes into account the work per time step, linearly increases with s, it is attractive to take s as large as possible so as to ameliorate the conditional stability. Of course, this is practical only up to a point where the CPU costs are still acceptable in comparison with the CPU costs of implicit or partly implicit methods.

We advocate stabilized methods for multi-space dimensional parabolic PDEs which give rise to moderate stiffness. Often this will occur for problems whose solutions are of a travelling wave front type, as these normally require small step sizes. Generally, reaction-diffusion systems

$$\frac{\partial u}{\partial t} = \nabla\,(K\,\nabla u) + f(u,x,t), \quad u = u(x,t), \quad x \in \Re^d, \tag{1.3}$$

where f is a modestly stiff reaction term and $\|K\| \ll 1$, can be efficiently solved with stabilized methods. For such problems these methods offer a very attractive alternative for standard explicit and unconditionally stable implicit ones. In cases where f gives rise to severe stiffness, they also can prove useful as part of an operator splitting scheme which treats at any of the grid points the reaction part with a standard stiff ODE solver. Likewise, in combination with operator splitting they can prove useful for systems of transport-chemistry problems of the advection-diffusion-reaction type

$$\frac{\partial u}{\partial t} + \nabla(a\,u) = \nabla\,(K\,\nabla u) + f(u,x,t), \quad u = u(x,t), \quad x \in \Re^d. \tag{1.4}$$

Problems of this type play an important role in the modeling of pollution of the atmosphere, ground water and surface water and are subject of much current research.

2. Preliminaries

Within the explicit RK family the design of stabilized methods is rather special. Customary is to construct methods with a maximal order of consistency p for a minimal number of stages s. For stabilized methods the greatest interest lies in a maximal stability interval and a low order. A low order is more appropriate since for most parabolic PDEs only a modest accuracy is needed. As a rule, the accuracy of the spatial discretization is also modest to avoid expensive, fine grids. Another reason favouring lower order methods is that the stability constant $c(s)$ of $\beta(s)$ decreases for increasing order. Hence, given a step size τ and spectral radius $\sigma(F'(t,U))$, for a higher order method a larger s is required to meet the absolute stability restriction $\tau\sigma(F'(t,U)) \leq \beta(s)$. In this paper we will therefore restrict ourselves to methods of order of consistency $p \leq 2$, denoted by

$$\begin{aligned} Y_0 &= U_n, \\ Y_j &= U_n + \tau\sum_{l=0}^{j-1} a_{jl}F_l, \quad F_l = F(t_n + c_l\tau, Y_l), \quad 1 \leq j \leq s, \\ U_{n+1} &= Y_s, \end{aligned} \tag{2.1}$$

where $c_j = a_{j0} + \ldots + a_{jj-1}$, $\tau = t_{n+1} - t_n$ denotes the stepsize and U_n is the approximation to the exact solution $U(t)$ at time $t = t_n$.

The stability analysis is carried out for linear systems

$$\frac{dU(t)}{dt} = MU + g(t), \quad 0 < t \leq T, \quad U(0) = U_0, \tag{2.2}$$

where M is a symmetric, constant coefficient matrix with non-positive eigenvalues. From practical experience we know that with regard to stability, conclusions and results based on this linear test model largely carry over to the nonlinear problem (1.1), as long as the Jacobian matrix F' satisfies the assumptions (i)-(ii) made in the introduction. While the standard *absolute stability* analysis focuses on the error propagation resulting from a perturbation of the initial value U_0, for stabilized RK methods with a large number of stages it has appeared to be also necessary to take into account error propagation over the stages within a single integration step. The involved analysis is called *internal (numerical) stability* analysis.

We will thus examine the stability using the perturbed scheme

$$\begin{aligned} \tilde{Y}_0 &= \tilde{U}_n, \\ \tilde{Y}_j &= \tilde{U}_n + \tau \sum_{l=0}^{j-1} a_{jl} \tilde{F}_l + \tilde{r}_j, \quad \tilde{F}_l = F(t_n + c_l\tau, \tilde{Y}_l), \quad 1 \le j \le s, \\ \tilde{U}_{n+1} &= \tilde{Y}_s, \end{aligned} \tag{2.3}$$

where $\tilde{U}_n$ denotes a perturbation of U_n and $\tilde{r}_j$ a perturbation introduced at stage j (e.g. round-off). Let $e_n = \tilde{U}_n - U_n$, $d_j = \tilde{Y}_j - Y_j$ denote the errors introduced by these perturbations. For the linear system any explicit RK scheme is then easily seen to possess an error scheme of the general form

$$d_j = P_j(\tau M)\, e_n + \sum_{k=1}^{j} Q_{jk}(\tau M)\, \tilde{r}_k, \quad 1 \le j \le s, \tag{2.4}$$

where $P_j(\tau M)$ is a matrix polynomial of degree j and $Q_{jk}(\tau M)$ one of degree $j-k$. In particular,

$$e_{n+1} = P_s(\tau M)\, e_n + \sum_{j=1}^{s} Q_{sj}(\tau M)\, \tilde{r}_j. \tag{2.5}$$

This error scheme gives a complete description of the linear stability. The polynomial P_s is the absolute *stability polynomial.* The polynomials Q_{sj} $(1 \le j \le s)$ are called *internal stability polynomials*, as they determine the propagation of the stage perturbations $\tilde{r}_j$.

In the remainder, $\|\cdot\|$ denotes the common (appropriately weighted) Euclidean norm in $\Re^m$ or the associated spectral norm. Since M is normal, we have

$$\begin{aligned} \|e_{n+1}\| &\le \|P_s(\tau M)\|\,\|e_n\| + \sum_{j=1}^{s} \|Q_{sj}(\tau M)\|\,\|\tilde{r}_j\| \\ &= \sigma(P_s(\tau M))\,\|e_n\| + \sum_{j=1}^{s} \sigma(Q_{sj}(\tau M))\,\|\tilde{r}_j\| \\ &= \max_{z=\tau\lambda} |P_s(z)|\,\|e_n\| + \sum_{j=1}^{s} \max_{z=\tau\lambda} |Q_{sj}(z)|\,\|\tilde{r}_j\|, \end{aligned} \tag{2.6}$$

where λ runs through the spectrum of M. Hence if we satisfy the absolute stability condition

$$\tau\sigma(M) \le \beta(s) \doteq \max\,[-z : z \le 0, |P_s(z)| \le 1], \tag{2.7}$$

where z is now taken continuous, then $\|P_s(\tau M)\| \le 1$. Evidently, $\sigma(M)$ is large for parabolic PDEs, so we need stability polynomials P_s with absolute value ≤ 1 for an interval which is as long as possible on the negative half line.

With regard to consistency we recall that $P_s(z)$ approximates e^z for $z \to 0$. In particular, $P_s(z) = e^z + O(z^p)$ for $p \leq 2$ implies p-th order consistency of the RK method for the nonlinear problem (1.1). Hence, for the consistency analysis it suffices to study P_s since we restrict ourselves to $p = 2$. The polynomials $P_j(z)$ play a similar role for $e^{c_j z}$ and Y_j.

3. Absolute stability polynomials

In this section we will review optimal or near-to-optimal stability polynomials P_s for order $p = 1$ and $p = 2$. In later sections particular RK methods will then be discussed which generate these stability functions. In the remainder $P_s(z)$ is denoted by

$$P_s(z) = c_{0,s} + c_{1,s}z + c_{2,s}z^2 + \ldots + c_{s,s}z^s, \tag{3.1}$$

where $c_{0,s} = c_{1,s} = 1$ for $p = 1$ and, in addition, $c_{2,s} = \frac{1}{2}$ for $p = 2$. The free coefficients $c_{j,s}$ are used to maximize the real stability boundary $\beta(s)$.

3.1 First order polynomials

For 1st-order methods the optimal polynomial P_s, that is, the polynomial which maximizes $\beta(s)$, is known. It is given by the shifted Chebyshev polynomial of the first kind,

$$P_s(z) = T_s(1 + \frac{z}{s^2}) \quad \text{with} \quad \beta(s) = 2s^2, \tag{3.2}$$

a result which goes back to a very early paper by Markoff published in 1892 [22]. As pointed out in [18], its use for parabolic PDEs has been first discussed around 1960 in [39, 5, 7]. For $T_s(x)$ various representations exist. Mostly used are

$$T_s(x) = \cos(s \arccos(x)), \quad -1 \leq x \leq 1, \tag{3.3}$$

or, equivalently,

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_j(x) = 2xT_{j-1}(x) - T_{j-2}(x), \quad 2 \leq j \leq s. \tag{3.4}$$

This recursive definition holds for all complex-valued x. It trivially follows from (3.3) that $|P_s(z)| \leq 1$ for $z \in [-2s^2, 0]$ while on this interval P_s alternates $s+1$ times between $+1$ and -1. This latter property can be used to prove that the shifted Chebyshev polynomial is unique and optimal (see [12], Theorem 4.2.1). According to [1] (formulas 15.1.1 and 15.4.3), $T_s(x)$ can also be written as

$$T_s(x) = \sum_{i=0}^{s} \frac{(-s)_i\,(s)_i}{(\frac{1}{2})_i\, i!} \left(\frac{1-x}{2}\right)^i, \tag{3.5}$$

where, for $a \in \Re$, $(a)_i$ is defined as $(a)_0 = 1$ and $(a)_i = a(a+1)\ldots(a+i-1)$ for $i \geq 1$. A simple calculation then yields

$$c_{0,s} = c_{1,s} = 1, \quad c_{i,s} = \frac{1-(i-1)^2/s^2}{i(2i-1)}\, c_{i-1,s} \quad \text{for} \quad i = 2, \ldots, s. \tag{3.6}$$

By way of illustration we list

$$P_2(z) \quad = \quad 1 + z + \frac{1}{8}z^2, \tag{3.7}$$

$$\begin{aligned}
P_3(z) &= 1 + z + \frac{4}{27}z^2 + \frac{4}{729}z^3, \\
P_4(z) &= 1 + z + \frac{5}{32}z^2 + \frac{1}{128}z^3 + \frac{1}{8192}z^4, \\
P_5(z) &= 1 + z + \frac{4}{25}z^2 + \frac{28}{3125}z^3 + \frac{16}{78125}z^4 + \frac{16}{9765625}z^5,
\end{aligned}$$

and observe that for a given value of s the coefficients $c_{i,s}$ rapidly decrease with increasing i, which is a prerequisite for a large stability interval. For any fixed i, $c_{i,s}$ tends to the limit value $c_i = 2^i/(2i)!$ for $s \to \infty$. This follows trivially from induction on the limiting recurrence relation for c_i,

$$c_0 = 1, \quad c_i = \frac{1}{i(2i-1)}\, c_{i-1} \quad \text{for} \quad i \geq 1. \tag{3.8}$$

Consequently, we may deduce the following limiting relation [1] for $P_s(z)$ (see also [5]),

$$\lim_{s\to\infty} P_s(z) = \cos\left(\sqrt{-2z}\right) = \sum_{i=0}^{\infty} \frac{(2z)^i}{(2i)!}, \tag{3.9}$$

which shows that an RK scheme possessing P_s as stability function stands on its own in the sense that it cannot be interpreted as the result of a Chebyshev iterative method applied to a standard implicit ODE integration formula. From (3.9) it also follows that for s sufficiently large,

$$P_s(z) \approx e^z - \frac{1}{3}z^2 + O(z^3), \quad z \to 0.$$

3.2 Second order polynomials

Until now no explicit analytical solution for truly optimal polynomials seems to be known for orders of consistency $p \geq 2$. For $p = 2$ the author is aware of two approximate polynomials, given in analytic form, for arbitrary $s \geq 2$. These have been derived by Bakker [3] (see also [12]) and Van der Houwen and Sommeijer (see [14, 18]). The polynomial derived by Van der Houwen and Sommeijer is very close to the optimal one and defined by

$$P_s(z) = \frac{2}{2-z} - \frac{z}{2-z}\, T_s\left(\cos\left(\frac{\pi}{s}\right) + \frac{z}{2}\left(1 - \cos\left(\frac{\pi}{s}\right)\right)\right), \quad s \geq 2, \tag{3.10}$$

with

$$\beta(s) = \frac{2}{\left(\tan\left(\frac{\pi}{2s}\right)\right)^2} \approx 8\frac{s^2}{\pi^2} \approx 0.810\, s^2. \tag{3.11}$$

Numerical computations carried out by Sommeijer have led to the conjecture that, albeit very close, the polynomial is not the optimal one. The polynomial earlier derived by Bakker [3] is defined by

$$P_s(z) = \frac{2}{3} + \frac{1}{3s^2} + \left(\frac{1}{3} - \frac{1}{3s^2}\right) T_s\left(1 + \frac{3z}{s^2-1}\right), \quad \beta(s) \approx \frac{2}{3}(s^2 - 1), \quad s \geq 2, \tag{3.12}$$

and generates approximately 80% of the optimal interval. This is also very good, because by increasing s with only about 10% the same step size as would be allowed by the optimal polynomial will yield

[1] In a private communication, G. Wanner, Université de Genève, pointed out that this relation follows more directly by inserting into (3.2) and (3.3) the approximation $\arccos\left(1 - \frac{1}{2}x^2\right) \approx x$.

stability, due to the quadratic behaviour. Following [13, 26, 37], in the remainder we will proceed with the Bakker-Chebyshev polynomial for the 2nd-order methods, since this polynomial has been extensively tested and also seems to perform even somewhat better due to smaller error constants [18]. For an even degree, $\beta(s)$ equals $\frac{2}{3}(s^2-1)$ exactly, while for an odd degree $\beta(s)$ is slightly larger. This follows from the observation that $P_s(z)$ alternates between $\frac{1}{3}+\frac{2}{3s^2}$ and 1.0 for z between 0 and $-\frac{2}{3}(s^2-1)$, while for an odd degree the exact boundary is determined by the point z where $P_s(z)$ intersects the line -1. This point is slightly smaller than $-\frac{2}{3}(s^2-1)$.

Following the derivation in the 1st-order case, the coefficients $c_{i,s}$ of P_s are easily found to be

$$c_{0,s}=c_{1,s}=1, \quad c_{2,s}=1/2, \quad c_{i,s}=3\,\frac{1-(i-1)^2/s^2}{i(2i-1)(1-1/s^2)}\,c_{i-1,s} \quad \text{for} \quad i=3,\ldots,s. \tag{3.13}$$

In particular,

$$\begin{aligned} P_3(z) &= 1+z+\frac{1}{2}z^2+\frac{1}{16}z^3, \\ P_4(z) &= 1+z+\frac{1}{2}z^2+\frac{2}{25}z^3+\frac{1}{250}z^4, \\ P_5(z) &= 1+z+\frac{1}{2}z^2+\frac{7}{80}z^3+\frac{1}{160}z^4+\frac{1}{6400}z^5, \end{aligned} \tag{3.14}$$

and the limiting polynomial is given by

$$\lim_{s\to\infty} P_s(z)=\frac{2}{3}+\frac{1}{3}\cos\left(\sqrt{-6z}\right)=\frac{2}{3}+\frac{1}{3}\sum_{i=0}^{\infty}\frac{(6z)^i}{(2i)!}. \tag{3.15}$$

Notice that the coefficients are a factor 3^{i-1} larger than in the 1st-order case. For s sufficiently large,

$$P_s(z)\approx e^z-\frac{1}{15}z^3+O(z^4), \quad z\to 0,$$

which shows that the Bakker-Chebyshev polynomial possesses a small error constant.

Remark In 1972, Riha [24] has already proved existence of truly optimal polynomials for arbitrary orders of consistency $p\geq 1$. The polynomials are found in the form of a solution of a certain generalized Chebyshev differential equation. This solution does not seem to be computable however for $p>1$. In the same period, coefficients of approximate polynomials have been computed numerically, for various orders and number of stages. Specifically, for $p\leq 4$ and $s\leq 5$ in [23], for $p=2$ and $s\leq 10$ in [21] and for $p\leq 4$ and $s\leq 10+p$ in [10]. These approximate results are useful since they have resulted in accurate estimates of the optimal $\beta(s)$-values (see [12] for more details),

$$\beta(s)=c_p(s)s^2 \quad \text{as} \quad s\to\infty, \quad c_2(s)\approx 0.814,\ c_3(s)\approx 0.489,\ c_4(s)\approx 0.341. \tag{3.16}$$

Apparently, $c_p(s)$ decreases with the order. Hence for a higher order method a larger number of stages is required to meet the stability condition. This of course diminishes the advantage of the higher order. However, the greatest drawback of the approximate polynomials constructed in [23, 21, 10] is their low degree. A low degree limits practical use, as this prevents to fully exploit the quadratic behaviour of $\beta(s)$. Finally it is of interest to mention that [21] and [10] used least square techniques for the computation of coefficients. These techniques become sensitive for instabilities if the degree is increased. In [33], for a class of related stabilized three-step RK methods, coefficients of approximate polynomials were computed (up to a degree 12) using an optimization technique based on linear programming. Although analytically given polynomials are to be preferred, still interest exists in computing optimal or near-to-optimal polynomials numerically. [2]

3.3 Damping

Of practical importance is to slightly modify P_s so as to introduce a little damping for the higher harmonics. Adopting the choice made in [13], we will thus replace the 1st-order, shifted Chebyshev polynomial (3.2) by

$$P_s(z) = \frac{T_s(w_0 + w_1 z)}{T_s(w_0)}, \quad w_1 = \frac{T_s(w_0)}{T_s'(w_0)}, \tag{3.17}$$

when appropriate. The parameter $w_0 > 1$ is called the damping parameter and the parameter w_1 has been chosen such that for any w_0 we have $P_s'(0) = 1$, implying 1st-order consistency. With the exception of a small neighbourhood of $z = 0$, $P_s(z)$ now alternates between $T^{-1}(w_0)$ and $-T^{-1}(w_0)$ for $z \in [-\beta(s), 0]$, and the stability boundary $\beta(s)$ is now defined by the equality $w_0 + w_1 z = -1$. A convenient choice for the damping parameter w_0 is $w_0 = 1 + \epsilon/s^2$, where ϵ is a small positive number. Then, by using the derivative values $T_s'(1) = s^2, T_s''(1) = \frac{1}{3}s^2(s^2-1)$, a simple calculation yields,

$$\beta(s) = \frac{(w_0 + 1)T_s'(w_0)}{T_s(w_0)} \approx (2 - \frac{4}{3}\epsilon)s^2. \tag{3.18}$$

A suitable choice for ϵ is 0.05. Since $T_s^{-1} \approx 1 - \epsilon$, this yields about 5% damping with only a very little decrease of $\beta(s)$ to $\approx 1.93s^2$. Observe that for z close to zero, where $x = w_0 + w_1 z > 1$, T_s is defined by $T_s(x) = \cosh(s \operatorname{arccosh}(x))$.

In a similar manner, the 2nd-order Bakker-Chebyshev polynomial is damped,

$$P_s(z) = a_s + b_s T_s(w_0 + w_1 z), \quad w_1 = \frac{T_s'(w_0)}{T_s''(w_0)}, \quad a_s = 1 - b_s T_s(w_0), \quad b_s = \frac{T_s''(w_0)}{(T_s'(w_0))^2}, \tag{3.19}$$

with $\beta(s)$ now defined by

$$\beta(s) = \frac{(w_0 + 1)T_s''(w_0)}{T_s'(w_0)} \approx \frac{2}{3}(s^2 - 1)(1 - \frac{2}{15}\epsilon). \tag{3.20}$$

For this polynomial, $\epsilon = 2/13$ yields $\approx$ 5% damping ($a_s + b_s \approx 1 - \epsilon/3$) with a reduction in $\beta(s)$ of about 2%.

A positive effect of damping is that the stability region $S = [z : \text{Re}(z) \leq 0, |P_s(z)| \leq 1]$, contains a genuine narrow strip along the negative half line. Without damping the boundary of S touches the negative half line. Figures 1 and 2 illustrate S for $s = 5$ for, respectively, order $p = 1$ and $p = 2$ with the damping specified above. Figure 2 shows a real boundary slightly larger than 16.0, which is in contradiction with expression (3.20). This is due to the use of the odd degree 5 as already pointed out in the discussion of expression (3.12). To the best of our knowledge, the widening of S through damping has been suggested for the first time by Guillou and Lago [7]. In fact, for the 1st-order case they have also proposed formula (3.17).

[2] In a private communication, G. Wanner, Université de Genève, pointed out that the Remes' algorithm can be naturally used for computing coefficients of optimal polynomials. He mentioned results up to a degree 24. Numerical instability prevents a higher degree. In the second print of [8], which is currently prepared, this use of the Remes' algorithm will be illustrated. G. Wanner also mentioned an alternative algorithm recently proposed by V.I. Lebedev [20], which can be used for any degree.

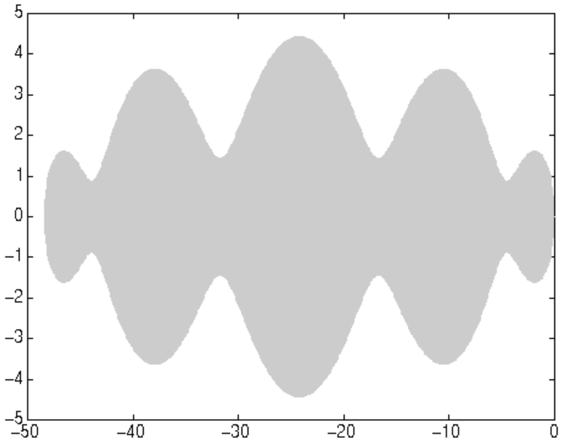

Figure 1: The stability region S of the 1st-order damped stability polynomial P_5.

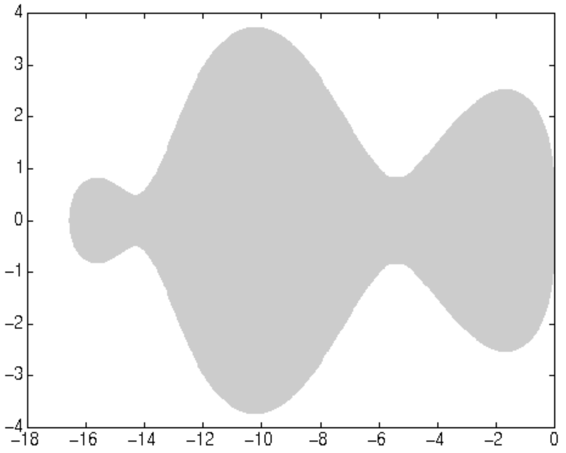

Figure 2: The stability region S of the 2nd-order damped stability polynomial P_5.

4. The diagonal and factorized method

Having selected stability polynomials, we are now ready to construct RK methods generating these polynomials. The explicit RK method which has all coefficients a_{jl} equal to zero except those on the first subdiagonal of the lower triangular Butcher matrix (a_{jl}), is called the *diagonal method*:

$$\begin{aligned} Y_0 &= U_n, \\ Y_j &= U_n + \tau a_{j,j-1} F(t_n + c_{j-1}\tau, Y_{j-1}), \quad 1 \leq j \leq s, \\ U_{n+1} &= Y_s. \end{aligned} \tag{4.1}$$

By defining

$$a_{j.j-1} = \frac{c_{s+1-j,s}}{c_{s-j,s}}, \tag{4.2}$$

the method generates the stability polynomial (3.1) for which any of the choices discussed can be made. This method is simple and requires a minimum amount of storage and has therefore been considered quite extensively [12]. However, the method is of limited practical use as it is severely internally unstable.

This follows immediately from the internal stability polynomial Q_{sj} which is easily found to be

$$Q_{sj}(z) = c_{s-j,s} z^{s-j}, \quad j = 1, \ldots, s. \tag{4.3}$$

The rapid growth of $|z|^{s-j}$ for $|z| \in [0, \beta(s)]$ renders the diagonal method useless for approximately $s \geq 12$. To illustrate this, we substitute the limiting value $c_{s-j} = 2^{s-j}/(2s-2j)!$ for the 1st-order case and $z = -\beta(s) = -2s^2$, which yields

$$|Q_{sj}(-\beta(s))| \approx \frac{(2s)^{2s-2j}}{(2s-2j)!}. \tag{4.4}$$

Hence, for j small, that is for the early stages, we get growth factors of $\approx \frac{(2s)^{2s}}{(2s)!}$, which are to be multiplied by the machine precision of the computer used. They increase so rapidly with s that in actual application only a limited number of stages can be used. For example, for $s = 12$ we have $\approx 10^9$ so that with a machine precision of 14 digits at most 5 digits remain. Although the above upper estimate is a conservative one, the internal accumulation of rounding errors observed in actual computations [12, 33] is in line with this estimate.

Of similar simplicity as the diagonal method is

$$\begin{aligned} Y_0 &= U_n, \\ Y_j &= Y_{j-1} + \tau a_{jj-1} F(t_n + c_{j-1}\tau, Y_{j-1}), \quad 1 \leq j \leq s, \\ U_{n+1} &= Y_s. \end{aligned} \tag{4.5}$$

This method is called the *factorized method*, since its stability function reads

$$P_s(z) = \prod_{j=1}^{s} (1 + a_{jj-1} z). \tag{4.6}$$

Identification with a suitable polynomial possessing real zeroes thus immediately defines the factorized method. For example, this is possible for the 1st-order polynomial (3.17) which is factorized as

$$P_s(z) = \prod_{j=1}^{s} \left(1 + \frac{\frac{w_1}{w_0}}{1 - \frac{1}{w_0}\cos(\frac{2j-1}{2s}\pi)} z \right). \tag{4.7}$$

However, the factorized method is not recommended for practical use either, since it also suffers from severe internal instability. This can be overcome by using special orderings of zeroes [6]. But since the ordering depends on s, we consider this approach too cumbersome. Notice that the 2nd-order polynomial (3.19) possesses complex conjugated zeroes, which means that the factorized method as it stands cannot be based on (3.19). For linear problems the factorized approach based on (4.7) has been proposed also in [7]. The internal stability problem, however, is not mentioned in this paper.

5. The method of Lebedev

Extensive research to stabilized, explicit methods has also been carried out by Lebedev and co-workers ([19] and references therein). The notation used in [19] differs from our general notation (2.1). Partly adopting their notation, their main scheme is given by

$$\begin{aligned} &Y_{k+\frac{1}{2}} = U_k + h_{k+1}\tau F(t_k, U_k), \quad t_{k+\frac{1}{2}} = t_k + h_{k+1}\tau, \\ &Y_{k+1} = Y_{k+\frac{1}{2}} + h_{k+1}\tau F(t_{k+\frac{1}{2}}, Y_{k+\frac{1}{2}}), \quad t_{k+1} = t_{k+\frac{1}{2}} + h_{k+1}\tau, \\ &U_{k+1} = Y_{k+1} - \gamma_{k+1}(Y_{k+1} - 2Y_{k+\frac{1}{2}} + U_k), \end{aligned} \tag{5.1}$$

where $k = 0, \ldots, N-1$. Hence the full integration step proceeds from (t_0, U_0) to (t_N, U_N) using N times two consecutive forward Euler steps with step size $h_{k+1}\tau$ followed by a correction step. The step size for the full step equals τ, imposing that $2h_1 + \ldots + 2h_N = 1$. The scheme can of course be written as an s-stage RK method with $s = 2N$.

The coefficients h_{k+1}, γ_{k+1} are again used to determine large real absolute stability boundaries. For this purpose the stability polynomial

$$P_{2N}(z) = \prod_{k=1}^{N} [(1 + h_k z)^2 - \gamma_k h_k^2 z^2], \tag{5.2}$$

is identified with a suitable polynomial to define 1st- and 2nd-order methods, in a similar way as for the factorized method. A difference is that we here have a factorization into quadratic terms which enables an identification with polynomials possessing complex conjugated zeroes. Like for the factorized method, an appropriate ordering of zeroes is needed for ensuring internal stability. This makes the construction of the schemes cumbersome since no best ordering exists for all stages.

The quadratic factorization exploits the fact that zeroes of even degree stability polynomials appear in pairs, in a certain way. We will illustrate this for the 1st-order method which is based on the first kind Chebyshev polynomial. According to [19], let us suppose that the factored form of the stability polynomial P_{2N} can be written as

$$P_{2N}(z) = \prod_{k=1}^{2N} \left(1 + \frac{h^0}{1 - ad_k} z\right) = \prod_{k=1}^{N} \left(1 + \frac{h^0}{1 - ad_k} z\right) \left(1 + \frac{h^0}{1 - ad_{k'}} z\right), \tag{5.3}$$

where $d_{k'} = -d_k$ and k' is determined by k. An elementary calculation then shows that (5.3) can be written in the form (5.2) if we define

$$h_k = \frac{h^0}{1 - a^2 d_k^2}, \quad \gamma_k = a^2 d_k^2. \tag{5.4}$$

Now consider the factorization (4.7) of the damped, first kind Chebyshev polynomial (3.17) with $s = 2N$. It immediately follows that this factorization fits in the form (5.3) and with the coefficients

$$h_k = \frac{\frac{w_1}{w_0}}{1 - \frac{1}{{w_0}^2}\cos^2(\frac{2k-1}{4N}\pi)}, \quad \gamma_k = \frac{1}{{w_0}^2}\cos^2(\frac{2k-1}{4N}\pi), \quad k = 1, \ldots, N, \tag{5.5}$$

the Lebedev method (5.1) possesses (3.17) as stability polynomial. For details on the specific ordering for k, as required for internal stability, we refer to [19] and the references therein.

The definition of h_k, γ_k for the 2nd-order methods is based on a so-called Zolotarev polynomial. Part of their derivation is done in a numerical way, which means that h_k, γ_k are not known in closed analytical form. This derivation requires a more lengthy discussion than in the 1st-order case and is omitted here. Observe that also the Bakker-Chebyshev polynomial (3.19) can be used to define h_k, γ_k. The methods have been implemented in a variable step size Fortran code, called DUMKA, the performance of which has been illustrated for a few test examples [19]. A comprehensive comparison with Sommeijer's code RKC discussed in Section 7 seems of interest.

6. The RKC method

To overcome the internal instability problem, Van der Houwen and Sommeijer [13] have developed a class of RK methods which is based on the three-term Chebyshev recursion (3.4). They derived 1st- and 2nd-order methods for which all coefficients are given as analytical expressions and which can be used for any (practical) value of s. These methods are called Runge-Kutta-Chebyshev (RKC) methods. In this section we discuss their derivation [13] and their internal stability and full convergence properties [37].

6.1 Derivation

The main idea of [13] is to construct the method in such a way that all polynomials P_j belonging to the intermediate stages (cf. the error scheme (2.4)) are defined by the three-term Chebyshev recursion in such a way that at the final s-th stage a selected stability polynomial P_s results. Thus the Ansatz is made that all P_j $(0 \le j \le s)$ are of the form

$$P_j(z) = a_j + b_j T_j(w_0 + w_1 z), \quad a_j = 1 - b_j T_j(w_0), \tag{6.1}$$

with P_s being one of the earlier derived stability polynomials. This means that w_0, w_1 and b_s are defined and that the parameters b_j are as yet undetermined for $0 \le j < s$. Imposing the three-term recursion (3.4), and using the property $P_j(0) = 1$, it then follows that the P_j satisfy

$$\begin{aligned} P_0(z) &= 1, \\ P_1(z) &= 1 + \tilde{\mu}_1 z, \\ P_j(z) &= (1 - \mu_j - \nu_j) + \mu_j P_{j-1}(z) + \nu_j P_{j-2}(z) + \tilde{\mu}_j P_{j-1}(z) z + \tilde{\gamma}_j z, \quad j = 2, \ldots, s, \end{aligned} \tag{6.2}$$

where

$$\tilde{\mu}_1 = b_1 w_1, \quad \mu_j = \frac{2 b_j w_0}{b_{j-1}}, \quad \nu_j = \frac{-b_j}{b_{j-2}}, \quad \tilde{\mu}_j = \frac{2 b_j w_1}{b_{j-1}}, \quad \tilde{\gamma}_j = -a_{j-1}\tilde{\mu}_j, \quad j = 2, \ldots, s. \tag{6.3}$$

From the relations for P_j we now deduce the RKC method for the general nonlinear problem (1.1) by associating P_j with the intermediate RK approximation Y_j and the occurrence of z with a function evaluation. This yields the scheme

$$Y_0 = U_n,$$

$$\begin{aligned} Y_1 &= Y_0 + \tilde{\mu}_1 \tau F_0, \\ Y_j &= (1-\mu_j-\nu_j)Y_0 + \mu_j Y_{j-1} + \nu_j Y_{j-2} + \tilde{\mu}_j \tau F_{j-1} + \tilde{\gamma}_j \tau F_0, \end{aligned} \tag{6.4}$$

$$U_{n+1} = Y_s, \quad j = 2, \ldots, s,$$

which is immediately recognized as an s-stage explicit RK method.

The free parameters b_j are chosen as follows. In the 1-st order case,

$$b_j = T_j^{-1}(w_0), \quad j = 0, \ldots, s. \tag{6.5}$$

This implies

$$P_j(z) = e^{c_j z} + O(z^2), \quad c_j = \frac{T_s(w_0)}{T_s'(w_0)} \frac{T_j'(w_0)}{T_j(w_0)} \approx \frac{j^2}{s^2} \quad (1 \le j \le s-1), \quad c_s = 1. \tag{6.6}$$

In the 2nd-order case the free parameters b_j can be chosen such that Y_j is of 2nd-order too for $2 \le j \le s$ [26]. Hence the expansion

$$P_j(z) = 1 + b_j w_1 T_j'(w_0)\, z + \tfrac{1}{2} b_j w_1^2 T_j''(w_0)\, z^2 + O(z^3), \quad j = 1, \ldots, s, \tag{6.7}$$

must be matched with $P_j(z) = 1 + c_j z + (c_j z)^2/2 + O(z^3)$. An elementary calculation yields

$$b_j = \frac{T_j''(w_0)}{(T_j'(w_0))^2}, \quad j = 2, \ldots, s. \tag{6.8}$$

Then

$$P_j(z) = e^{c_j z} + O(z^3), \quad c_j = \frac{T_s'(w_0)}{T_s''(w_0)} \frac{T_j''(w_0)}{T_j'(w_0)} \approx \frac{j^2-1}{s^2-1} \quad (2 \le j \le s-1), \quad c_s = 1. \tag{6.9}$$

For $j = 1$ only 1st-order is possible of course, which yields some freedom. As in [26] we put

$$b_0 = b_2, \quad b_1 = b_2, \tag{6.10}$$

so that

$$c_1 = \frac{c_2}{T_2'(w_0)} \approx \frac{c_2}{4}. \tag{6.11}$$

The 1st- and 2nd-order RKC methods we recommend are now defined. Note that the original 2nd-order method from [13] is different from the one chosen here. The main difference is that in [13] all intermediate approximations Y_j are of order one. The current choice is to be preferred, although with respect to accuracy and convergence no essential difference exists, as we will see later.

6.2 Internal stability

By applying the RKC method to the linear system (2.2), perturbing it in a similar way as we did with the RK method in Section 2, but now with perturbations $\hat{r}_j$, we find the error scheme [37]

$$e_{n+1} = P_s(\tau M)\, e_n + \sum_{j=1}^{s} Q_{sj}(\tau M)\, \hat{r}_j, \tag{6.12}$$

where

$$Q_{sj}(z) = \frac{b_s}{b_j} S_{s-j}(w_0 + w_1 z), \quad j = 1, \ldots, s, \tag{6.13}$$

$S_i(x)$ being the i-th degree Chebyshev polynomial of the second kind. For any of the choices for b_j made, one can prove that if $\tau\sigma(M) \leq \beta(s)$,

$$\|Q_{sj}(\tau M)\| \leq \frac{b_s}{b_j}(s - j + 1)\,(1 + C\epsilon), \quad j = 1, \ldots, s, \tag{6.14}$$

where ϵ determines the damping as defined in Section 3.3 and C is a constant of moderate size independent of M, τ, s. Consequently, if $\tau\sigma(M) \leq \beta(s)$,

$$\|e_{n+1}\| \leq \|e_n\| + \sum_{j=1}^{s} \frac{b_s}{b_j}(s - j + 1)\,(1 + C\epsilon)\|\hat{r}_j\|. \tag{6.15}$$

For both the 1st- and 2nd-order method, with and without damping, examination of the parameters b_j then leads to the error bound

$$\|e_{n+1}\| \leq \|e_n\| + \tilde{C} \sum_{j=1}^{s}(s - j + 1)\|\hat{r}_j\| \leq \|e_n\| + \tfrac{1}{2}s(s+1)\,\tilde{C}\,\max_j \|\hat{r}_j\|, \tag{6.16}$$

where $\tilde{C}$ is again a constant of moderate size independent of M, τ, s.

The following important conclusion can be made. Within one integration step the accumulation of internal perturbations, such as round-off errors, is *independent of the spectrum of M* as long as $\tau\sigma(M) \leq \beta(s)$. This obviously is a tremendous improvement over the diagonal and factorized methods. The estimate says that perturbations grow at most *quadratically* with the number of stages s. A numerical experiment in [37] shows that in actual computation this quadratric growth indeed takes place. For rounding errors this renders no problem at all. For example, if $s = 1000$, which for a serious application of course is a hypothetical value, the local perturbation is at most $\approx 10^6 \max \|\hat{r}_j\|$. If the machine precision of the computer is about 14 digits, a common value, this local perturbation still leaves 8 digits for accuracy which for PDEs is more than enough.

Interestingly, unbounded accumulation of round-off errors for the related three-term Chebyshev iterative method for elliptic problems does not occur. Woźniakowski [38] has proven that this accumulation is bounded and proportional to the condition number of the matrix under consideration. See the Appendix for more details on the Chebyshev iterative method, where also a self-contained proof of its stability for round-off errors is presented.

6.3 Full convergence

The property of internal stability has been shown to be of practical importance in connection with accumulation of round-off errors. In this section we will show, through results from a convergence analysis, that internal stability also plays a crucial role for the accuracy of the method. The analysis reveals that quadratic growth of local perturbations is natural, in the sense that it transfers local stage truncation errors to the final stage just in the right amount for obtaining the order of convergence one expects. The convergence analysis deals with the full error, that is, the difference between the solution of the original PDE problem and the RKC solution. The analysis is akin to the B-convergence analysis from the stiff ODE field [4] and establishes unconditional convergence under the assumption that $\tau\sigma(M) \leq \beta(s)$ [37]. The term unconditional refers to the usual asymptotic assumption that τ

and a grid size parameter h are allowed to tend to zero simultaneously and independently of each other. It is emphasized that this is unusual for an explicit method. For the RKC method we can obtain unconditional convergence since $\tau\sigma(M)$ is allowed to become arbitrarily large, stability being achieved by taking s sufficiently large. This, of course, is possible only if the internal stability of the method is in order.

Consider a semi-discrete, linear PDE problem of type (2.2),

$$\frac{du_h(t)}{dt} = M_h u_h(t) + g_h(t) + \alpha_h(t), \quad 0 \le t \le T, \tag{6.17}$$

where $u_h(t)$ denotes the exact PDE solution, restricted to a space grid, and $\alpha_h(t)$ the local space truncation error on this grid. Let $e_n = u_h(t_n) - U_n$ be the global error on this grid and r_j $(1 \le j \le s)$ the stage truncation errors obtained by substituting u_h into the RKC method (6.4). That is,

$$\begin{array}{rcl} u_h(t_n + c_1\tau) & = & u_h(t_n) + \tilde{\mu}_1 \tau F_0 + r_1, \\ u_h(t_n + c_j\tau) & = & (1 - \mu_j - \nu_j)\, u_h(t_n) + \mu_j u_h(t_n + c_{j-1}\tau) + \nu_j u_h(t_n + c_{j-2}\tau) + \\ & & \tilde{\mu}_j \tau F_{j-1} + \tilde{\gamma}_j \tau F_0 + r_j, \quad j = 2, \ldots, s, \end{array} \tag{6.18}$$

where $F_j = M_h u_h(t_n + c_j\tau) + g_h(t_n + c_j\tau)$. Because (6.18) can be seen as a perturbed RKC method, where the perturbations $\hat{r}_j$ have been replaced by the local errors r_j, bounds for the global error can be obtained by properly estimating r_j and using internal stability results.

Following [37], we will illustrate this for the undamped 1-st order scheme (6.4) defined by

$$\tilde{\mu}_1 = \frac{1}{s^2}, \; \mu_j = 2, \; \nu_j = -1, \; \tilde{\mu}_j = \frac{2}{s^2}, \; \tilde{\gamma}_j = 0, \quad 2 \le j \le s. \tag{6.19}$$

Let $u_h \in C^2[0,T]$. From (6.18) and the Taylor series expansion of $u_h, \dot{u}_h$ at the intermediate step points $t_n + c_{j-1}\tau$, it follows that

$$r_j = \tau^2 \rho_j + \tau \tilde{\mu}_j \alpha_h(t_n + c_{j-1}\tau), \quad 1 \le j \le s, \tag{6.20}$$

where the remainder terms ρ_j are given by

$$\rho_1 = \tfrac{1}{2} c_1^2 u_h^{(2)}(t_n), \quad \rho_j = \tfrac{1}{2}(c_j - c_{j-1})^2 u_h^{(2)}(t_*) + \tfrac{1}{2}(c_{j-2} - c_{j-1})^2 u_h^{(2)}(t_*), \quad 2 \le j \le s, \tag{6.21}$$

with t_* denoting some point $\in [t_n, t_{n+1}]$. Since $c_j = j^2/s^2$,

$$c_j - c_{j-1} = s^{-2}(2j - 1) \le 2s^{-1}, \quad 1 \le j \le s, \tag{6.22}$$

so that we easily find

$$\|r_j\| \le \frac{4\tau}{s^2} (\tau \max_{t_n \le t \le t_{n+1}} \|u_h^{(2)}(t)\| + \tfrac{1}{2} \max_{t_n \le t \le t_{n+1}} \|\alpha_h(t)\|), \quad 1 \le j \le s. \tag{6.23}$$

Inserting these bounds into (6.16) and adding up yields the aimed convergence result,

$$\|e_n\| \le C\, (\tau \max_{0 \le t \le T} \|u_h^{(2)}(t)\| + \max_{0 \le t \le T} \|\alpha_h(t)\|), \quad n = 1, 2, \ldots; n\tau \le T. \tag{6.24}$$

C is again a moderately sized constant independent of M, τ, s. Apparently, the s^2-growth of the local errors within a single step does not harm the convergence, since each local error r_j is proportional to s^{-2}.

For the undamped 2nd-order scheme (6.4) defined by (6.3), (6.8) and (6.10), the derivation of the global error bound is more complicated due to the fact that stage one is only 1st-order consistent. To save space we therefore only give the bound and refer to [37] for its derivation,

$$\|e_n\| \leq C\,(\frac{\tau}{s^3} \max_{0\leq t\leq T} \|u_h^{(2)}(t)\| + \tau^2 \max_{0\leq t\leq T} \|u_h^{(3)}(t)\| + \max_{0\leq t\leq T} \|\alpha_h(t)\|). \tag{6.25}$$

The 1-st order consistency of stage one introduces the $O(\tau)$-term. Hence for s small we actually have 1st-order convergence. However, in application s is sufficiently large to render this $O(\tau)$-term negligibly small, which means 2nd-order convergence in practice. In this connection it is of interest to also present a bound for the original 2nd-order scheme from [13], since here all intermediate stages are only 1st-order consistent. Hence this scheme is expected to possess a larger $O(\tau)$-term in the global error. For the undamped case ($w_0 = 1$), defined by (6.4) and

$$w_1 = \frac{3}{s^2-1},\ \tilde{\mu}_1 = \frac{1}{s^2},\ \mu_j = 2,\ \nu_j = -1,\ \tilde{\mu}_j = \frac{6}{s^2-1},\ \tilde{\gamma}_j = \frac{-4}{s^2}\,\frac{2s^2+1}{2s^2-2},\quad 2 \leq j \leq s, \tag{6.26}$$

we found

$$\|e_n\| \leq C\,(\frac{\tau}{s^2} \max_{0\leq t\leq T} \|u_h^{(2)}(t)\| + \tau^2 \max_{0\leq t\leq T} \|u_h^{(3)}(t)\| + \max_{0\leq t\leq T} \|\alpha_h(t)\|). \tag{6.27}$$

The $O(\tau)$-term is indeed a factor s larger. Consequently, the original 2nd-order scheme from [13] will normally be somewhat less accurate, but for s sufficiently large the difference in accuracy will not be noticeable. Finally we wish to emphasize that these convergence results are also valid in the presence of damping.

6.4 Related work on RKC methods

In Section 3 we have shown that the limiting RKC method for $s \to \infty$ differs from a standard, implicit integration method. This raises the question, why not adopting the iterative approach and applying existing three-term Chebyshev iteration to an appropriate implicit method. We address this question in the Appendix. Apparently, for parabolic problems the iterative Chebyshev approach should be used with care. The technique can be used however to construct explicit integration formulas with similar stability characteristics as the RKC method. Van der Houwen, Sommeijer and de Vries [15, 16, 17] have derived predictor-corrector methods with extended stability regions along the negative axis. By using a multistep formula and adjusting free parameters, it is possible to obtain a higher order of consistency and even larger real stability boundaries than for the one-step RKC method. This obviously renders the predictor-corrector methods in theory more promising. However, an actual accuracy-efficiency comparison with the one-step RKC methods has not been given [15, 16, 17].

The predictor-corrector methods can be interpreted as multistep, s-stage RK methods. Particular two-step and three-step, s-stage RK methods with extended stability regions along the negative axis were studied earlier by Verwer [31, 32, 33, 34, 35, 36]. The methods investigated in [31, 32, 33, 35] (see also [25]) are still of diagonal type for the RK part and suffer from the internal instability phenomenon discussed in Section 4. The number of stages is therefore limited to 12. The methods from [34, 36] were constructed along the same lines as the one-step RKC methods from [13] and share their favourable internal stability properties. As in [13], the order p is restricted to one and two. The corresponding stability boundaries $\beta_p(s)$ are notably larger than in the one-step case. The table below compares these values, for appropriately chosen damping parameters.

order	1-step	2-step	3-step
1	$1.94\ s^2$	$3.58\ s^2$	$5.17\ s^2$
2	$0.65\ (s^2-1)$	$1.19\ s^2$	$2.32\ s^2$

A numerical comparison between the methods from this table has been reported in [26]. Despite the smallest β-value, this comparison is in favour of the 2nd-order one-step formula owing to its smaller truncation errors. Of course, in test cases where the step size is really determined by stability, or when the spatial error would dominate accuracy, the 1st-order three-step formula would come out as the most efficient one. The report [26] presents also results for linearized schemes, adopting ideas from [27]. When using linearization the three-step schemes turn out to be superior.

7. Numerical illustration

Sommeijer [28] has developed a Fortran (research) code based on the 2nd-order formulas (6.3), (6.4), (6.8), (6.10). This code is also called RKC and is provided with a local error estimator and variable step size control, as is customary for ODEs. It adjusts the number of stages s automatically so as to satisfy in each integration step the absolute stability condition $\tau\sigma(F') \leq \beta(s)$. Hence τ is selected by the local error estimator and is not restricted by stability. RKC thus works as an unconditionally stable code. Owing to the explicitness, it uses only 6 arrays for storage, consists of only about 200 lines of Fortran and is very easy to use. The user has to provide an upper estimate of $\sigma(F')$, which normally renders no problem. The spectral radius estimation can also be carried out automatically using the power method. This leads to marginal overhead [35]. In this section we will apply RKC to a diffusion-reaction problem of the type mentioned in the introduction. The numerical performance of the RKC method has been illustrated before in [26, 27, 29, 30, 37]. More numerical experiments carried out with the code RKC can be found in [9].

The diffusion-reaction problem stems from combustion theory and has been borrowed from [2, 29]. The (hotspot) problem is a scalar 2D model for a reaction of a mixture of two chemicals, defined by

$$u_t = d\,\Delta u + f(u), \quad f(u) = \frac{R}{\alpha\delta}\,(1+\alpha-u)e^{\delta\,(1-1/u)}, \quad 0 < x, y < 1, \quad t > 0, \tag{7.1}$$

with the initial condition $u(x,y,0) = 1, 0 \leq x,y \leq 1$, and for $t > 0$ the zero Neumann condition for $x = 0$ and $y = 0$ and the Dirichlet condition $u(x,y,t) = 1$ for $x = 1$ and $y = 1$. The parameter values are $d = 1, \alpha = 1, \delta = 20, R = 5$. The solution u represents the temperature of the mixture. For small times u gradually increases in a circular region around the origin. Then ignition occurs, causing u to suddenly jump from near unity to $1+\alpha$, while simultaneously a reaction front is formed which circularly propagates towards the outer Dirichlet boundaries. When the front reaches the boundary, a steady state results. In [29] the diffusion coefficient $d = 0.1$ which yields a steeper front. Here we select $d = 1$ [2] to better illustrate the use of RKC. For $d = 1$ the ignition occurs at $t \approx 0.30$, while the front arrives at the Dirichlet boundaries at $t \approx 0.36$.

RKC is a natural candidate for the numerical integration. First, the travelling reaction front limits the step size of any integration scheme, be it implicit or explicit. Second, the problem becomes locally unstable in the course of time because $f'(u)$ varies between $\approx +1000$ (for $u \approx 1.6$) and -5500 (for $u \approx 2.0$). Consequently, irrespective the integrator used, rather small step sizes are required to maintain sufficient accuracy in the transient phase, especially during the ignition. Only during the start phase and in the near approach to steady state the step size can be increased to a level fully justifying an implicit method. The following experiment serves to illustrates this.

The problem is discretized on a uniform grid with grid size 10^{-2}, using 2nd-order central differences for the Laplacian and the Neumann boundary conditions. This results in 10^4 unknowns. The spectral radius $\sigma(F')$ is estimated as $9.0\,10^4$ for all t. Hence the required number of stages s and the step size τ relate as $s \approx \lceil 372\sqrt{\tau}\,\rceil$. Figure 3 shows the values of s for an integration over the interval $[0, 0.5]$, using $\tau = 10^{-4}$ as initial step size and $tol = 10^{-4}$ as local error tolerance (see [28] for details). We see that s becomes quite large in the start phase and in the near approach to steady state. At the onset of the front when the ignition occurs and during its propagation to the boundary, a notably smaller τ is required resulting in much lower values for s. The total number of integration steps amounts

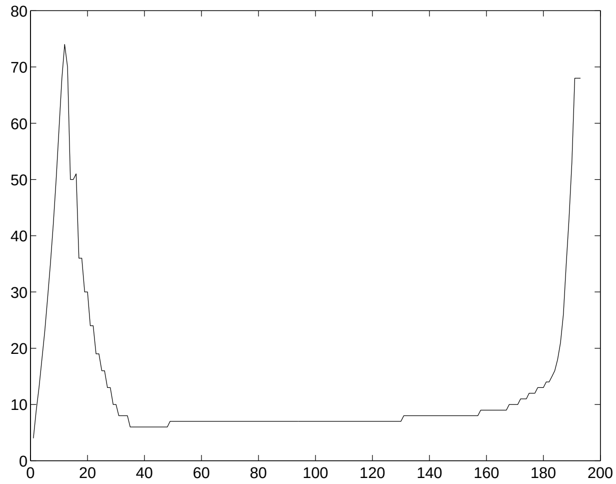

Figure 3: The number of stages s as a function of the step index.

to 203, including 10 rejected ones. The total number of derivative evaluations is 2803, yielding an average s of ≈ 14. Taking into account that there is no overhead whatsoever, these numbers clearly illustrate the effectiveness of the stabilized explicit approach (see also the table below). For example, any explicit, 2nd-order, 2-stage method allows a maximal step size of $\approx 2/9\,10^{-4}$ to ensure absolute stability. This would require at least 45000 derivative evaluations over the interval $[0, 0.5]$.

The table shows for $tol = 10^{-4}\,(10^{-1})\,10^{-7}$ with $\tau = 10^{-4}$ as initial step size, the maximum absolute time integration error at $t = 0.32$, as well as the numbers of integration steps, rejected steps, f-evaluations and average number of f-evaluations. The low accuracies in comparison with the values of tol are due to the local instability and the sudden ignition which at time $t = 0.32$ still dominate the temporal solution behaviour.

tol	$error$	steps	rejected	f-evals	average
10^{-4}	$6.8\,10^{-2}$	141	7	1790	12.7
10^{-5}	$1.6\,10^{-2}$	278	0	2373	8.5
10^{-6}	$3.2\,10^{-3}$	638	0	3731	5.8
10^{-7}	$5.7\,10^{-4}$	1480	1	6495	4.4

Acknowledgements I gratefully acknowledge Willem Hundsdorfer, Ben Sommeijer and Gerhard Wanner for their useful comments and remarks on a first version of the paper.

References

1. M. Abramowitz and I.A. Stegun (eds.), *Handbook of mathematical functions,* (Dover Publications Inc., New York, 1965).
2. S. Adjerid and J.E. Flaherty, A local refinement finite-element method for two-dimensional parabolic systems, Rept. 86 - 7, Dept. of Computer Sc., Rensselaer Polytechnic Institute, Troy, NY (1986).
3. M. Bakker, Analytical aspects of a minimax problem (Dutch), Technical Note TN 62, Mathematical Centre, Amsterdam (1971).

4. K. Dekker and J.G. Verwer, *Stability of Runge-Kutta methods for stiff nonlinear differential equations* (North-Holland, 1984).

5. J.N. Franklin, Numerical stability in digital and analogue computation for diffusion problems, *J. Math. Phys.* 37 (1959) 305 - 315.

6. W. Gentzsch and A. Schlüter, Über ein Einschrittverfahren mit zyklischer Schrittweitenänderung zur Lösung parabolischer Differentialgleichungen, *ZAMM* 58 (1978) T415 - T416.

7. A. Guillou and B. Lago, Domaine de stabilité associé aux formules d'intégration numérique d'équations différentielles, a pas séparés et a pas liés. Recherche de formules a grand rayon de stabilité, Ier Congr. Assoc. Fran. Calcul, AFCAL, Grenoble, Sept. 1960 (1961) 43 - 56.

8. E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems* (Springer-Verlag, 1991).

9. S. Hofmann, First and second order Runge-Kutta-Chebyshev methods for the numerical integration of parabolic differential equations and stiff differential equations (German), Master Thesis, University of Wuppertal, Germany (1992).

10. P.J. van der Houwen and J. Kok, Numerical solution of a minimax problem, Report TW 124/71, Mathematical Centre, Amsterdam (1971).

11. P.J. van der Houwen, Explicit Runge-Kutta methods with increased stability boundaries, *Numer. Math.* 20 (1972), 149 - 164.

12. P.J. van der Houwen, *Construction of integration formulas for initial value problems* (North-Holland, 1977).

13. P.J. van der Houwen and B.P. Sommeijer, On the internal stability of explicit m-stage Runge-Kutta methods for large values of m, *ZAMM* 60 (1980) 479 - 485.

14. P.J. van der Houwen, On the time integration of parabolic differential equations, *Numerical Analysis*, Proceedings of the 9th Biennial Dundee Conference, June 1981, Lecture Notes in Mathematics 912, ed. G.A. Watson (1982) 157 - 168.

15. P.J. van der Houwen and B.P. Sommeijer, A special class of multistep Runge-Kutta methods with extended real stability interval, *IMA J. Numer. Anal.* 2 (1982) 183 - 209.

16. P.J. van der Houwen and B.P. Sommeijer, Predictor-corrector methods with improved absolute stability region, *IMA J. Numer. Anal.* 3 (1983) 417 - 437.

17. P.J. van der Houwen, B.P. Sommeijer and H.B. de Vries, Generalized predictor-corrector methods of high order for the time integration of parabolic differential equations, *ZAMM* 66 (1986) 595 - 605.

18. P.J. van der Houwen, The development of Runge-Kutta methods for partial differential equations, CWI Report NM-R9420, CWI, Amsterdam (to appear in the proceedings of the 1994 IMACS World Congress, Atlanta, 1994) (1994).

19. V.I. Lebedev, How to solve stiff systems of differential equations by explicit methods, In *Numerical Methods and Applications*, ed. G.I. Marchuk, CRC Press Inc., Boca Raton-Ann Arbor-London-Tokyo (1994) 45 - 80.

20. V.I. Lebedev, Extremal polynomials with restrictions and optimal algorithms, Note (unpublished - 1995), Institute of Numerical Mathematics, Russian Academy of Sciences, Moscow.

21. H. Lomax, On the construction of highly stable, explicit numerical methods for integrating coupled ordinary differential equations with parasitic eigenvalues, NASA Technical Note NASATND/4547 (1968)

22. V.A. Markoff, Über Polynome, die in einem gegebenen Intervall möglichst wenig von Null abweichen, *Math.* Ann. 77 (1916) 213 - 258 (translation by J. Grossman of original Russian article in

Acad. Sc. St. Petersburg, 1892).

23. C.L. Metzger, Méthodes de Runge-Kutta de rang supérieur à l'ordre. Thèse (troisième cycle), Université de Grenoble (1967).

24. W. Riha, Optimal stability polynomials, Computing 9 (1972) 37 - 43.

25. K. Sepehrnoori and G.F. Carey, Numerical integration of semi-discrete evolution systems, *Computer Meth. in Appl. Mech. and Engn.* 27 (1981) 45 - 61.

26. B.P. Sommeijer and J.G. Verwer, A performance evaluation of a class of Runge-Kutta-Chebyshev methods for solving semi-discrete parabolic differential equations, Report NW91/80, Mathematisch Centrum, Amsterdam (1980).

27. B.P. Sommeijer and P.J. van der Houwen, On the economization of stabilized Runge-Kutta methods with applications to parabolic initial value problems, *ZAMM* 61 (1981) 105 - 114.

28. B.P. Sommeijer, RKC, a nearly-stiff ODE solver. Available from netlib@ornl.gov, send rkc.f from ode (1991).

29. R.A. Trompert and J.G. Verwer, A static-regridding method for two-dimensional parabolic partial differential equations, *Appl. Numer. Math.* 8 (1991) 65 - 90.

30. A.S. Vasudeva Murthy and J.G. Verwer, Solving parabolic integro-differential equations by an explicit integration method, *J. Comp. Appl. Math.* 39 (1992) 121 - 132.

31. J.G. Verwer, Multipoint multistep Runge-Kutta methods I: On a class of two-step methods for parabolic equations, Report NW30/76, Mathematisch Centrum, Amsterdam (1976).

32. J.G. Verwer, Multipoint multistep Runge-Kutta methods II: The construction of a class of stabilized three-step methods for parabolic equations, Report NW31/76, Mathematisch Centrum, Amsterdam (1976).

33. J.G. Verwer, A class of stabilized three-step Runge-Kutta methods for the numerical integration of parabolic equations, *J. Comp. Appl. Math.* 3 (1977) 155 - 166.

34. J.G. Verwer, On a class of explicit three-step Runge-Kutta methods with extended real stability intervals, Report NW77/79, Mathematisch Centrum, Amsterdam (1979).

35. J.G. Verwer, An implementation of a class of stabilized explicit methods for the time integration of parabolic equations, *ACM Trans. Math. Software* 6 (1980) 188 - 205.

36. J.G. Verwer, A note on a Runge-Kutta-Chebyshev method, *ZAMM* 62 (1982) 561 - 563.

37. J.G. Verwer, W.H. Hundsdorfer and B.P. Sommeijer, Convergence properties of the Runge-Kutta-Chebyshev method, *Numer. Math.* 57 (1990) 157 - 178.

38. H. Woźniakowski, Numerical stability of the Chebyshev method for the solution of large linear systems, *Numer. Math.* 28 (1977) 191 - 209.

39. C.-D. Yuan, Some difference schemes for the solution of the first boundary value problem for linear differential equations with partial derivatives (Russian), Thesis, Moscow State University (1958).

A. Appendix: Chebyshev iteration and absolute stability

The success of the RKC method emanates from the three-term Chebyshev recursion (3.4). The three-term recursion also plays a key role in the Chebyshev iterative solution method developed for elliptic problems (see e.g. [1,2]). This raises the question whether straightforward Chebyshev iteration, applied to a stable and accurate implicit integration method, will lead to an equally efficient explicit approach. We address this question in this appendix. In this connection it is of interest to emphasize once more that the limiting RKC method for $s \to \infty$ cannot be interpreted as the result of Chebyshev iteration applied to an implicit method (see Section 3).

For simplicity of presentation we consider the implicit Euler method

$$AU_{n+1} = B, \quad A = I - \tau M, \quad B = U_n, \tag{A.1}$$

applied to the linear problem (2.2) with $g(t) = 0$. Like the RKC method, Chebyshev iteration can also be applied to nonlinear problems, but analysis is feasible only for the linear model problem. The implicit Euler method is a natural choice as it of one-step type and unconditionally stable for (2.2). The conclusions we draw for implicit Euler carry over to other methods, like the trapezoidal rule.

For system (A.1), three-term Chebyshev iteration (also called semi-iterative Chebyshev, second-order Chebyshev or second-order Richardson iteration) is defined by

$$Y_1 = Y_0 - \tfrac{1}{2}\beta_0\,(AY_0 - B), \tag{A.2}$$

$$Y_j = \alpha_{j-1}Y_{j-1} + (1-\alpha_{j-1})Y_{j-2} - \beta_{j-1}\,(AY_{j-1} - B), \quad j \geq 2,$$

where Y_0 is a given initial guess for U_{n+1} and the parameters β_0 and α_j, β_j for $j \geq 1$ are defined by

$$\beta_0 = \frac{4}{a+b}, \quad \alpha_j = \frac{2\tilde{b}\,T_j(\tilde{b})}{T_{j+1}(\tilde{b})}, \quad \beta_j = \frac{4}{b-a}\,\frac{T_j(\tilde{b})}{T_{j+1}(\tilde{b})}, \tag{A.3}$$

$$\tilde{b} = \frac{b+a}{b-a}, \quad 0 < a \leq \lambda_1(A), \quad b \geq \sigma(A),$$

with $\lambda_1(A)$ denoting the smallest positive eigenvalue $\lambda(A)$ of A and a, b bounds for $\lambda(A)$ as indicated. Natural choices for Y_0, a and b are

$$Y_0 = U_n, \quad a = 1, \quad b = 1 + \tau\sigma, \quad \sigma = \sigma(M). \tag{A.4}$$

The iteration formula (A.2) shows the close connection with the RKC method. In fact, (A.2) fits in the RKC format (6.4) with $F(t,U) = MU$, $\tilde{\mu}_1 = 2/(2+\tau\sigma)$ and for $j \geq 2$,

$$\mu_j = \frac{2\,T_{j-1}(\tilde{b})}{T_j(\tilde{b})}, \quad \nu_j = -\,\frac{T_{j-2}(\tilde{b})}{T_j(\tilde{b})}, \quad \tilde{\mu}_j = \frac{4}{\tau\sigma}\,\frac{T_{j-1}(\tilde{b})}{T_j(\tilde{b})}, \quad \tilde{\gamma}_j = 0.$$

One of the obvious differences is that prior to the integration step, the RKC method fixes the number of stages whereas for (A.2) any number of iterations can occur, as determined by a stopping criterion. We will show that this pure iterative approach leads to conflicting requirements regarding consistency and stability.

Starting from any consistent integration formula, the iterate Y_j is consistent either at $t = t_{n+1}$ or at $t = t_n + \tau c_j$, where c_j is defined by the parameters of (A.2) and the choice for the initial iterate Y_0. That is, for all $j \geq 0$ a positive constant c_j exist such that

$$Y_j = U_n + \tau c_j \dot{U}_n + O(\tau^2), \quad \tau \to 0. \tag{A.5}$$

For a given c_0 these constants satisfy

$$c_1 = c_0 - \tfrac{1}{2}\beta_0(c_0 - 1), \quad c_j = \alpha_{j-1}c_{j-1} + (1-\alpha_{j-1})c_{j-2} - \beta_{j-1}(c_{j-1} - 1), \quad j \geq 2. \tag{A.6}$$

Hence if the initial estimate is already consistent at $t = t_{n+1}$, then $c_j = 1$ for all $j \geq 0$. If $c_0 = 0$, which is the case for the initial guess $Y_0 = U_n$, then c_j converges monotonically to 1, but $c_j < 1$ for any finite j. Consequently, the final iterate Y_j which is accepted as the new approximation then should be accepted as a result consistently defined at $t = t_n + c_j\tau$, rather than at the forward time level $t = t_{n+1}$ of the implicit Euler method. This suggests to use instead the explicit Euler result $Y_0 = U_n + \tau M U_n$ as initial guess for which $c_j = 1$ for all j. Unfortunately, this choice leads to poor absolute stability properties as we will show next.

Let $Y_0 = U_n + \delta\tau M U_n$ with either $\delta = 0$ or $\delta = 1$. Assuming exact arithmetic (no local perturbations such as round-off), the iteration error $U_{n+1} - Y_j$ satisfies (see [1,2])

$$U_{n+1} - Y_j = R_j(Z)(U_{n+1} - Y_0), \quad R_j(Z) = \frac{T_j(Z)}{T_j(\tilde{b})}, \quad Z = \frac{1}{b-a}[(b+a)I - 2A], \tag{A.7}$$

and eliminating Y_0 and U_{n+1} yields

$$Y_j = \left((I - \tau M)^{-1}(I - R_j(Z)) + R_j(Z)(I + \delta\tau M)\right) U_n. \tag{A.8}$$

Hence, to Y_j is associated the absolute stability polynomial

$$P_j(z) = \frac{1 + zR_j(\tilde{z})(\delta - 1 - \delta z)}{1 - z}, \quad \tilde{z} = \frac{\tau\sigma + 2z}{\tau\sigma}, \tag{A.9}$$

where, as before, $z = \tau\lambda(M)$ and $a = 1, b = 1 + \tau\sigma$, $-\tau\sigma \leq z \leq 0$. An elementary calculation reveals that $|P_j(z)| \leq 1$ for all j if $Y_0 = U_n$. This means that the unconditional absolute stability of the implicit Euler rule is maintained for any accepted Y_j. However, this attractive property is lost for the explicit Euler initial guess as shown by the following computation. For $\delta = 1$ and $-\tau\sigma \leq z < 0$,

$$|P_j(z)| \leq 1 \quad \Leftrightarrow \quad \frac{1}{z} \leq \frac{T_j(1 + \frac{2z}{\tau\sigma})}{T_j(1 + \frac{2}{\tau\sigma})} \leq \frac{2 - z}{z^2}. \tag{A.10}$$

The condition for absolute stability for Y_j can now be seen to be $T_j(1 + \frac{2}{\tau\sigma}) \geq \tau\sigma$, which follows by substitution of the most critical value $z = -\tau\sigma$. From computations and results found at page 181 of [1], we then can conclude that Y_j is absolutely stable if

$$j \geq j^*, \quad j^* \approx \tfrac{1}{2}\ln(2\tau\sigma)\sqrt{1 + \tau\sigma}, \tag{A.11}$$

supposing $\tau\sigma \gg 1$. Violation of this inequality can result in error growth in the time stepping process. Comparing (A.11) with $1 + \lceil(\tau\sigma/2)^{0.5}\rceil$, which equals the number of stages guaranteeing absolute stability of the undamped, 1st-order RKC method, convincingly shows that with regard to absolute stability the pure iterative approach will generally be much less efficient.

References

1. O. Axelsson (1994), *Iterative solution methods* (Cambridge University Press).

2. R.S. Varga (1962), *Matrix iterative analysis* (Prentice Hall).

B. Appendix: Numerical stability of the Chebyshev iterative solution method

We present a round-off error analysis for the three-term Chebyshev iterative solution method for symmetric, positive definite linear systems

$$AU = B. \tag{B.1}$$

An analysis of the type presented here has been given before in [4]. Our derivations are along the lines of the internal stability analysis of the Runge-Kutta-Chebyshev method for parabolic equations found in [3]. These are akin to those in [4], but shorter and self-contained.

For system (B.1), three-term Chebyshev iteration is defined by [1,2]

$$Y_1 = Y_0 - \tfrac{1}{2}\beta_0\,(AY_0 - B), \tag{B.2}$$

$$Y_j = \alpha_{j-1}Y_{j-1} + (1-\alpha_{j-1})Y_{j-2} - \beta_{j-1}\,(AY_{j-1} - B), \quad j \geq 2,$$

where Y_0 is a given initial guess for U and the parameters β_0 and α_j, β_j for $j \geq 1$ are defined by

$$\beta_0 = \frac{4}{a+b}, \quad \alpha_j = \frac{2\tilde{b}\,T_j(\tilde{b})}{T_{j+1}(\tilde{b})}, \quad \beta_j = \frac{4}{b-a}\,\frac{T_j(\tilde{b})}{T_{j+1}(\tilde{b})}, \tag{B.3}$$

$$\tilde{b} = \frac{b+a}{b-a}, \quad 0 < a \leq \lambda_1(A), \quad b \geq \sigma(A),$$

with $\lambda_1(A)$ denoting the smallest eigenvalue $\lambda(A)$ of A, $\sigma(A)$ the spectral radius, and a, b bounds for $\lambda(A)$ as indicated. The function T_j denotes the first kind Chebyshev polynomial.

To study the stability for round-off errors, we introduce the perturbed method

$$\tilde{Y}_1 = \tilde{Y}_0 - \tfrac{1}{2}\beta_0\,(A\tilde{Y}_0 - B) + r_1, \tag{B.4}$$

$$\tilde{Y}_j = \alpha_{j-1}\tilde{Y}_{j-1} + (1-\alpha_{j-1})\tilde{Y}_{j-2} - \beta_{j-1}\,(A\tilde{Y}_{j-1} - B) + r_j, \quad j \geq 2,$$

where $\tilde{Y}_0$ is the perturbed Y_0 and r_j a perturbation representing the round-off. Let

$$e_j = \tilde{Y}_j - U, \quad j \geq 0. \tag{B.5}$$

Since U satisfies (A.2), these errors are defined by the perturbed inhomogeneous recursion

$$e_1 = e_0 - \tfrac{1}{2}\beta_0\,Ae_0 + r_1, \tag{B.6}$$

$$e_j = \alpha_{j-1}e_{j-1} + (1-\alpha_{j-1})e_{j-2} - \beta_{j-1}\,Ae_{j-1} + r_j, \quad j \geq 2.$$

Elaborating this recursion shows that the error e_j can be written as

$$e_j = R_j(Z)\,e_0 + \sum_{k=1}^{j} Q_{jk}\,r_k, \quad j \geq 0, \tag{B.7}$$

where the matrices Z and $R_j(Z)$ are defined by

$$Z = \frac{1}{b-a}[(b+a)I - 2A], \quad R_j(Z) = \frac{T_j(Z)}{T_j(\tilde{b})}, \tag{B.8}$$

and $Q_{jk}(1 \le k \le j)$ are matrices determining the propagation of the perturbations r_j. The homogeneous part of (B.7) is well-known [1,2]. The following lemma specifies the inhomogeneous part.

Lemma 1 *The errors $e_j\,(j \ge 0)$ satisfy*

$$e_j = R_j(Z)\, e_0 + \sum_{k=1}^{j} \frac{\tilde{T}_k}{\tilde{T}_j}\, S_{j-k}(Z)\, r_k, \tag{B.9}$$

where $\tilde{T}_j = T_j(\tilde{b})$ and S_{j-k} is the second kind Chebyshev polynomial of degree $j-k$.

Proof. Substitute (B.7) into (B.6) and put $e_0 = 0$ for convenience. It follows that $Q_{11} = I$ and

$$\sum_{k=1}^{j} Q_{jk}\, r_k = (\alpha_{j-1} - \beta_{j-1}A) \sum_{k=1}^{j-1} Q_{j-1k}\, r_k + (1-\alpha_{j-1}) \sum_{k=1}^{j-2} Q_{j-2k}\, r_k + r_j, \quad j \ge 2. \tag{B.10}$$

Equating coefficients of r_k, in the order $k = j, \ldots, 1$, shows that

$$\begin{aligned} &Q_{jj} = I, \quad j \ge 1, \\ &Q_{jj-1} = \alpha_{j-1} - \beta_{j-1}A, \quad j \ge 2, \\ &Q_{jk} = (\alpha_{j-1} - \beta_{j-1}A)\, Q_{j-1k} + (1-\alpha_{j-1})Q_{j-2k}, \quad k = j-2, \ldots, 1, \quad j \ge 3, \end{aligned} \tag{B.11}$$

while, for a given iteration index j, the matrices Q_{lk} exist for all k, l satisfying $1 \le k \le l \le j$. Next we substitute the expressions for $\alpha_{j-1}, \beta_{j-1}$. This gives

$$\begin{aligned} &Q_{jj} = I, \quad j \ge 1, \\ &Q_{jj-1} = 2\tilde{b}\, \frac{\tilde{T}_{j-1}}{\tilde{T}_j}\, I - \frac{4}{b-a}\, \frac{\tilde{T}_{j-1}}{\tilde{T}_j}\, A, \quad j \ge 2, \\ &Q_{jk} = 2\, \frac{\tilde{T}_{j-1}}{\tilde{T}_j}\, Z Q_{j-1k} + (1 - 2\tilde{b}\, \frac{\tilde{T}_{j-1}}{\tilde{T}_j})Q_{j-2k}, \quad k = j-2, \ldots, 1, \quad j \ge 3. \end{aligned} \tag{B.12}$$

Using $\tilde{T}_j = 2\tilde{b}\tilde{T}_{j-1} - \tilde{T}_{j-2}$, we then can write

$$\begin{aligned} &Q_{jj} = I, \quad j \ge 1, \\ &Q_{jj-1} = 2\, \frac{\tilde{T}_{j-1}}{\tilde{T}_j}\, Z, \quad j \ge 2, \\ &Q_{jk} = 2\, \frac{\tilde{T}_{j-1}}{\tilde{T}_j}\, Z Q_{j-1k} - \frac{\tilde{T}_{j-2}}{\tilde{T}_j}\, Q_{j-2k}, \quad k = j-2, \ldots, 1, \quad j \ge 3, \end{aligned} \tag{B.13}$$

which is equivalent to

$$\begin{aligned}
&\tilde{Q}_{jj} = \tilde{T}_j\, I, \quad j \geq 1, \\
&\tilde{Q}_{jj-1} = 2\,\tilde{T}_{j-1}\, Z, \quad j \geq 2, \qquad \text{(B.14)} \\
&\tilde{Q}_{jk} = 2\, Z\, \tilde{Q}_{j-1k} - \tilde{Q}_{j-2k}, \quad k = j-2, \ldots, 1, \quad j \geq 3,
\end{aligned}$$

where $\tilde{Q}_{jk} = \tilde{T}_j\, Q_{jk}$. In the third line of this formula we recover the three-term Chebyshev recursion with j as index. Since $\tilde{Q}_{lk}$ exist for $1 \leq k \leq l \leq j$, this recursion starts from $\tilde{Q}_{kk} = \tilde{T}_k\, I$ and $\tilde{Q}_{k+1k} = 2\tilde{T}_k\, Z$. Due to the factor 2 present in the second starting value, instead of the first kind the second kind Chebyshev polynomial is involved. It thus follows that $\tilde{Q}_{jk} = \tilde{T}_k\, S_{j-k}(Z)$, so that

$$Q_{jk} = \frac{\tilde{T}_k}{\tilde{T}_j}\, S_{j-k}(Z). \quad \Box \qquad \text{(B.15)}$$

The following theorem reveals the stability of three-term Chebyshev iteration for round-off errors. It shows that accumulation of these errors is truly bounded in j and at most proportional to the condition number of the matrix A, here estimated by b/a.

Theorem 1 *Let $\|.\|$ denote the (appropriately weighted) Euclidean vector norm or the spectral matrix norm. Let r be an upper bound for $\|r_j\|$, $j \geq 1$. The errors e_j then can be bounded by*

$$\|e_j\| \leq \|R_j(Z)\|\, \|e_0\| + C\, \frac{b}{a}\, r, \qquad \text{(B.16)}$$

where C is a constant of moderate size independent of A, a, b and j.

Proof. Since all eigenvalues of Z lie in [-1,1], $\|S_{j-k}(Z)\| \leq j-k+1$. Hence we must bound

$$\sum_{k=1}^{j} \frac{\tilde{T}_k}{\tilde{T}_j}\, (j-k+1). \qquad \text{(B.17)}$$

From computations found at p. 181 of [1], it follows that

$$\tilde{T}_j = 2\, \frac{1+\eta^{2j}}{\eta^j}, \quad \eta = \frac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}} < 1. \qquad \text{(B.18)}$$

Hence

$$\frac{\tilde{T}_k}{\tilde{T}_j} = \eta^{j-k}\, \frac{1+\eta^{2k}}{1+\eta^{2j}} \leq 2\, \eta^{j-k}, \qquad \text{(B.19)}$$

since $k \leq j$ and $\eta < 1$. Thus, (B.17) is bounded by

$$2 \sum_{k=1}^{j} (j-k+1)\, \eta^{j-k} = 2\eta^{-1} \sum_{k=1}^{j} k\eta^k. \qquad \text{(B.20)}$$

Now consider the function $f(x) = x\eta^x$. For increasing x, starting in $x = 0$, f first monotonically increases and then monotonically decreases towards zero. Using $d(\eta^x)/dx = \eta^x \ln \eta$, a simple calculation then gives

$$2\eta^{-1} \sum_{k=1}^{j} k\eta^k \leq 4\eta^{-1} \int_0^\infty x\eta^x \, dx = 4\eta^{-1} \ln^{-2}\eta. \tag{B.21}$$

Finally, we use

$$\ln \eta = \ln \frac{1 - \sqrt{a/b}}{1 + \sqrt{a/b}} = -2\sqrt{a/b} + O(\sqrt{a/b})^3, \tag{B.22}$$

tacitly assuming that $b \gg a$. Consequently,

$$\ln^{-2} \eta = \frac{1}{4} \frac{b}{a} \left(1 + O(\frac{a}{b})\right), \tag{B.23}$$

which completes the proof. □

References

1. O. Axelsson (1994), *Iterative solution methods* (Cambridge University Press).

2. R.S. Varga (1962), *Matrix iterative analysis* (Prentice Hall).

3. J.G. Verwer, W.H. Hundsdorfer and B.P. Sommeijer (1990), Convergence properties of the Runge-Kutta-Chebyshev method, *Numer. Math.* 57, 157 - 178.

4. H. Woźniakowski (1977), Numerical stability of the Chebyshev method for the solution of large linear systems, *Numer. Math.* 28, 191 - 209.