TW

stichting
mathematisch
centrum

MC

AFDELING TOEGEPASTE WISKUNDE TW 119 APRIL

P.J. VAN DER HOUWEN

POLYNOMIAL METHODS: ONE STEP METHODS FOR LINEAR VALUE PROBLEM I

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

Contents

1. Introduction 2

2. The error of the difference scheme 5
   2.1 Construction of the difference scheme 5
   2.2 The discretization error 8
   2.3 The numerical error 9
   2.4 Convergence and stability 10
   2.5 Step-size control 14
   2.6 Numerical stability 15

3. Runge-Kutta methods 16
   31. Regions of stability 17

4. The use of Chebyshev polynomials 19
   4.1 Construction of the difference scheme 19
   4.2 Regions of stability 22

5. The case of purely imaginary eigenvalues 22
   5.1 A polynomial problem 23
   5.2 Regions of stability 24

6. Stabilization of higher order schemes 26
   6.1 Properties of the polynomial $B_q(x)$ 26
   6.2 Introduction of a single stability term 28
   6.3 The case of two stability terms 29
   6.4 Polynomials $B_q(x)$ of higher degree 32

References 36

1. Introduction

In this paper we have studied one-step difference methods for equations of the type

$$\frac{d\tilde{U}}{dt} = D\tilde{U} + F, \tag{1.1}$$

where $\tilde{U}$ and F are (vector) functions of the variable t, and D is a matrix with constant entries. We shall denote functions of the continuous variable t by capitals and the corresponding discretized functions, i.e. the functions which arise from restricting t to a discrete set of points, by the corresponding lower cases.
When F(t) and $\tilde{U}(0)$ are given the function $\tilde{U}(t)$ is uniquely defined by equation (1.1). In particular, we are interested in initial value problems of type (1.1) which arise from linear partial differential equations when the space variables are discretized. In such cases the matrix D usually is of very large order (100 or 1000). The study of difference methods for linear equations is also useful to attack non-linear differential equations of the type

$$\frac{d\tilde{U}}{dt} = H(t,\tilde{U}) \ , \tag{1.2}$$

since such equations locally have the form (1.1). This immediately follows from the Taylor expansion of $H(t,\tilde{U}(t))$ in a point $t = t_0$:

$$H(t_0 + \tau, \tilde{U}(t_0+\tau)) = H(t_0, \tilde{U}(t_0)) + \tau\, H_t(\bar{t}_0, \tilde{U}(t_0)) + D(t_0)\,(\tilde{U}(t_0+\tau) - \tilde{U}(t_0)) \ . \tag{1.3}$$

Here, $D(t_0)$ is the matrix (d_{ij}),

$$d_{ij} = \frac{\partial}{\partial \tilde{U}^{(j)}} H^{(i)}\,(t_0, \bar{U}(t_0)) \ ,$$

$\bar{t}_0 = t_0 + \theta\tau$, $\bar{U}(t_0) = \theta\tilde{U}(t_0+\tau) + (1-\theta)\,\tilde{U}(t_0)$, $0 \le \theta \le 1$, and $\tilde{U}^{(j)}$, $H^{(i)}$ are the j-th and i-th component of the vector functions U and H, respectively.

From (1.3) it follows that in the neighbourhood of $t = t_0$ equation (1.2) behaves like

$$(1.2') \qquad \frac{d\tilde{U}}{dt} = D(t_0)\tilde{U} + \left[\bar{H}(t_0, \tilde{U}(t_0)) + \tau\, \tilde{U}_t(\bar{t}_0, \tilde{U}(t_0)) - D(t_0)\, \tilde{U}(t_0)\right],$$

which is of type (1.1).

The one-step methods for solving numerically (1.1), which are most widely used, are of the Runge-Kutta type. However, when these methods are applied to systems arising from partial differential equations it turns out that the (time) step $\tau = \Delta t$ prescribed by accuracy considerations is considerably larger than the step prescribed by stability considerations.

The same situation is met when Runge-Kutta methods are applied to systems describing circuit simulations.

This may be explained as follows. The Runge-Kutta methods are based on the Taylor series approximation in $\tau = 0$ of the operator $\exp(\tau D)$. Let the approximating polynomial be $A_p(\tau D)$, thus

$$(1.4) \qquad A_p(\tau D) = 1 + \tau D + \frac{1}{2!}\tau^2 D^2 + \ldots + \frac{1}{p!}\tau^p D^p.$$

Then this approximation becomes better as the value of τ is smaller. In particular the effect of the operators $\exp(\tau D)$ and $A_p(\tau D)$ on eigenfunctions of D corresponding to eigenvalues with a large modulus is completely different, unless τ is relatively small. In the case of systems arising from partial differential equations or circuit simulations the eigenfunctions in the analytical solution corresponding to large modulus eigenvalues vanish rapidly. In actual computation, however, they are introduced at each step by round-off errors. This forces us to take small time steps in order to avoid instabilities.

In order to overcome these difficulties we have constructed polynomial approximations of $\exp(\tau D)$ which are more accurate for larger values of τ. We have distinguished the case where D has real eigenvalues and the case where D has imaginary eigenvalues. In the latter case the optimal polynomial approximations are hardly better than the Runge-Kutta methods. In the first case, however, a considerable improvement can be obtained by using Chebyshev polynomials. In principle, this fact is well-known and was first used by Franklin in 1958 (cf. reference [4]).

A disadvantage of the Chebyshev polynomial method is that it is only first order exact. Therefore, it is desirable to construct polynomial approximations which are accurate and stable as well. A method is given to construct such polynomials. In the second order case these polynomials are derived explicitly. When polynomials of sufficiently high degree are used the difference scheme is nearly six times cheaper than the second order Runge-Kutta method.

In references [6] and [7], which will appear in the near future, applications to stiff equations and numerical examples will be given.

Finally, the author wishes to acknowledge the work done by Mr. IJsselstein who programmed the plotting-program by which the figures 3.1, 4.1 and 5.1 were obtained.

2. The error of the difference scheme

2.1 Construction of the difference scheme

Suppose it is required to find the solution of the initial value problem

$$(2.1)\qquad \begin{cases} \dfrac{d\tilde{U}}{dt} = D\tilde{U} + F, \ t \geq 0, \\ \tilde{U} = \tilde{U}_0, \ t = 0, \end{cases}$$

where $\tilde{U}_0$ is a given initial function. Assuming that $\tilde{U}(t)$ has continuous derivatives of up to order $p + 1$ we may write

$$(2.2)\quad \tilde{U}(t+\tau) = \left[1+\tau \frac{d}{dt} + \frac{1}{2!}\tau^2 \frac{d^2}{dt^2} + \ldots + \frac{1}{p!}\tau^p \frac{d^p}{dt^p}\right] \tilde{U}(t) + \frac{1}{(p+1)!}\tau^{p+1}\tilde{U}^{(p+1)}(\bar{t}).$$

Here, $\bar{t}$ denotes a point in the interval $[t,t+\tau]$.
It is convenient to introduce the operator

$$(2.3)\qquad E_j = D^j + D^{j-1}\frac{d}{dt} + \ldots + D\frac{d^{j-1}}{dt^{j-1}} + \frac{d^j}{dt^j}, \ E_0 = 1 .$$

We may then write by virtue of the differential equation

$$(2.4)\qquad \frac{d^j}{dt^j}\tilde{U} = D^j\tilde{U} + E_{j-1}\mathbf{F}.$$

Equations (2.2)-(2.4) suggest the following difference scheme for an approximate difference solution u of the initial value problem (2.1):

$$(2.5)\qquad \begin{cases} u_0 = \tilde{U}_0, \\ u_{k+1} = u_k + \tau c_k^{(1)} + \frac{1}{2}\tau^2 c_k^{(2)} + \ldots + \frac{1}{p!}\tau^p c_k^{(p)}, \ k = 1, 2, \ldots, \\ c_k^{(j)} = D^j u_k + E_{j-1} f_k. \end{cases}$$

In these formulae u_k denotes the difference solution at $t = t_k = k\tau$ and $E_{j-1} f_k$ is defined as $E_{j-1} F(t)|_{t=t_k}$.

Scheme (2.5) defines u_{k+1} as a sum of corrections of increasing order of τ. Each correction term can be derived from the preceding one by a recurrence relation. To see this we observe that the operators E_j satisfy the recurrence relation

$$E_j = D\,E_{j-1} + \frac{d^j}{dt^j}. \tag{2.6}$$

Hence

$$\begin{aligned} c_k^{(j+1)} &= D^{j+1} u_k + E_j f_k = \\ &= D^{j+1} u_k + D\,E_{j-1} f_k + \frac{d^j}{dt^j} f_k = \\ &= D\,c_k^{(j)} + \frac{d^j}{dt^j} f_k, \end{aligned} \tag{2.7}$$

where $\frac{d^j}{dt^j} f_k$ denotes $\frac{d^j}{dt^j} F(t)|_{t=t_k}$. From this relation it can be deduced that (2.5) is equivalent to

$$\begin{cases} u_0 = \tilde{U}_0, \\ u_{k+1} = u_k + \tau c_k^{(1)} + \frac{1}{2!}\tau^2 c_k^{(2)} + \ldots + \frac{1}{p!}\tau^p c_k^{(p)}, \quad k = 0,1,2,\ldots, \\ c_k^{(1)} = Du_k + f_k, \\ c_k^{(2)} = Dc_k^{(1)} + \frac{d}{dt} f_k, \\ \cdots\cdots\cdots\cdots \\ c_k^{(p)} = Dc_k^{(p-1)} + \frac{d^{p-1}}{dt^{p-1}} f_k. \end{cases} \tag{2.8}$$

In this form the difference scheme is more appropriate in actual computation.

For theoretical considerations we shall employ another form of the difference scheme. It is easily verified that (2.5) can be written as

$$
(2.9)\qquad \begin{cases} u_0 = \tilde{U}_0, \\ u_{k+1} = A_p(\tau D)\, u_k + \tau g_k^{(p)}, \quad k = 0, 1, 2, \ldots, \\ A_p(\tau D) = 1 + \tau D + \frac{1}{2!}\,\tau^2 D^2 + \ldots + \frac{1}{p!}\,\tau^p D^p, \\ g_k^{(p)} = \left[E_0 + \frac{1}{2!}\,\tau E_1 + \ldots + \frac{1}{p!}\,\tau^{p-1} E_{p-1}\right] f_k. \end{cases}
$$

In connection with this representation of the difference scheme we remark that the analytical solution $\tilde{U}(t)$ satisfies a similar relation, i.e.

$$
(2.10)\qquad \tilde{U}(t+\tau) = e^{D\tau}\,\tilde{U}(t) + e^{D(t+\tau)} \int_t^{t+\tau} e^{-D\theta}\, F(\theta)\, d\theta.
$$

Here the operator $e^{D\tau}$ is defined by

$$
(2.11)\qquad e^{D\tau} = 1 + \tau D + \frac{1}{2!}\,(\tau D)^2 + \ldots = A_\infty(\tau D).
$$

Thus in (2.9) the polynomial operator $A_p(\tau D)$ is in fact an approximation of order p to the exponential operator $\exp(\tau D)$. This suggests that other approximations of $\exp(\tau D)$ may be used. For instance

$$
\exp(\tau D) \cong A_p(\tau D) + (\tau D)^{p+1}\, B_q(\tau D),
$$

where $B_q(\tau D)$ is a polynomial operator of degree q in τD. It will turn out that it is advantageous to employ approximations of this type.
For future reference we give the corresponding difference schemes in the form (2.8) and (2.9), respectively:

$$
(2.8')\qquad \begin{cases} u_0 = \tilde{U}_0, \\ u_{k+1} = u_k + \tau c_k^{(1)} + \frac{1}{2}\,\tau^2 c_k^{(2)} + \ldots + \frac{1}{p!}\,\tau^p c_k^{(p)} + \beta_{p+1}\,\tau^{p+1} c_k^{(p+1)} + \ldots + \beta_n\,\tau^n c_k^{(n)}, \\ c_k^{(1)} = Du_k + f_k, \\ c_k^{(2)} = Dc_k^{(1)} + \frac{d}{dt}\, f_k, \\ \ldots\ldots\ldots\ldots \\ c_k^{(n)} = Dc_k^{(n-1)} + \frac{d^{n-1}}{dt^{n-1}}\, f_k, \quad n = p + q + 1, \end{cases}
$$

$$
(2.9')\quad \begin{cases} u_0 = \tilde{U}_0, \\ u_{k+1} = P_n(\tau D)\, u_k + \tau g_k^{(n)}, \quad k = 0, 1, 2, \ldots, \\ P_n(\tau D) = A_p(\tau D) + (\tau D)^{p+1} B_q(\tau D), \quad n = p + q + 1, \\ B_q(\tau D) = \beta_{p+1} + \beta_{p+2} \tau D + \ldots + \beta_n (\tau D)^q, \\ g_k^{(n)} = \left[E_0 + \frac{1}{2!} \tau E_1 + \ldots + \frac{1}{p!} \tau^{p-1} E_{p+1} + \beta_{p+1} \tau^p E_p + \ldots + \beta_n \tau^{n-1} E_{n-1} \right] f_k. \end{cases}
$$

The parameters $\beta_{p+1}, \ldots, \beta_n$ are real parameters to be determined later. In formulae (2.8') and (2.9') it is assumed that $F^{(n-1)}$ exists.

2.2 The discretization error

We now discuss the error which is introduced in the k-th time step, i.e. the local discretization error $\rho_k(\tau)$. Let $\tilde{U}'$ denote the solution of the initial value problem (see figure 2.1)

$$
(2.12)\quad \begin{cases} \dfrac{d\tilde{U}'}{dt} = D\tilde{U}' + F, \quad t \geq t_k, \\ \tilde{U}' = u_k, \quad t = t_k . \end{cases}
$$

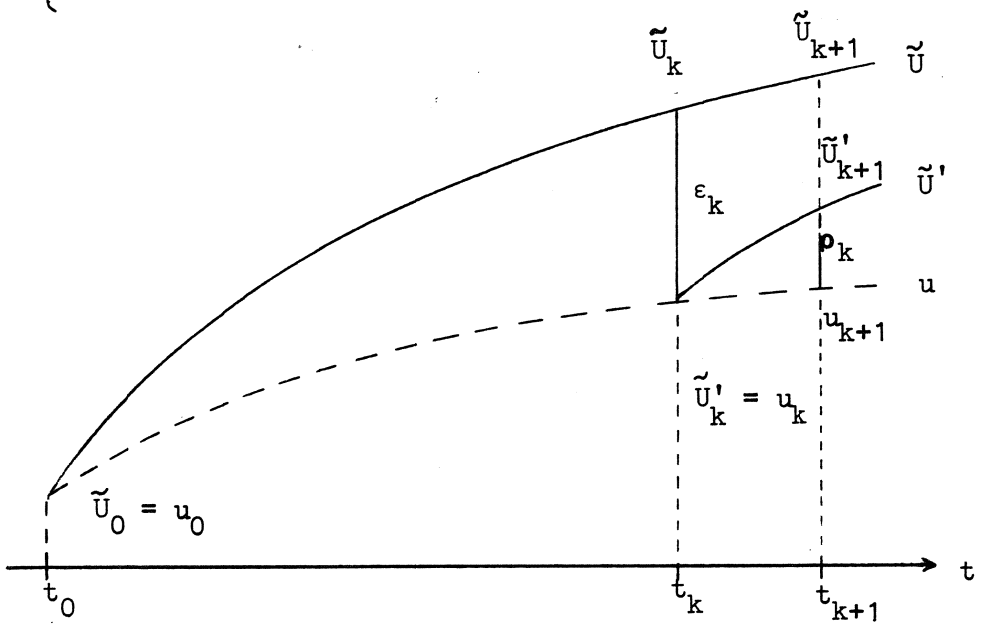

fig. 2.1 Local discretization error in a scalar case

Then the local discretization error is defined by

$$\rho_k(\tau) = \tilde{U}'_{k+1} - u_{k+1}. \tag{2.13}$$

By applying scheme (2.8') to (2.12) we find

$$\rho_k(\tau) \sim \left(\frac{1}{(p+1)!} - \beta_{p+1}\right) \tau^{p+1} c_k^{(p+1)} \text{ as } \tau \to 0. \tag{2.14}$$

The difference method is said to have an accuracy of order p as $\tau \to 0$.

The next step is to consider the development of the local discretization errors. These errors together, produce the total discretization error ε_k which is defined by

$$\varepsilon_k = \tilde{U}_k - u_k. \tag{2.15}$$

From (2.2), (2.4) and (2.9')it follows that ε_k satisfies the difference scheme

$$\varepsilon_{k+1} = P_n(\tau D)\, \varepsilon_k + \rho_k(\tau),\ k = 0, 1, 2, \ldots,\ \varepsilon_0(\tau) = 0. \tag{2.16}$$

Before this scheme is studied we consider another source of errors which may influence the difference solution, i.e. the effect of round-off errors.

2.3 The numerical error

In actual computation, we cannot obtain the solution of the difference scheme exactly, as one is faced with the phenomenon of round-off errors which give rise to a numerical solution u^* instead of the difference solution u. Suppose that u^* is the solution of the scheme

$$u^*_{k+1} = P_n(\tau D)\, u^*_k + \tau g_k^{(n)} - \rho^*_k,\ k = 0, 1, 2, \ldots, \tag{2.17}$$

where ρ^*_k is the local numerical error generated in the k-th time step. Then the (total) numerical error

(2.18) $$\varepsilon_k^* = u_k - u_k^*$$

satisfies the difference scheme

(2.19) $$\varepsilon_{k+1}^* = P_n(\tau D)\, \varepsilon_k^* + \rho_k^*, \quad k = 0, 1, 2, \ldots.$$

From (2.16) and (2.19) it follows that the total error

(2.20) $$e_k = \tilde{U}_k - u_k^* = \varepsilon_k + \varepsilon_k^*$$

satisfies the difference scheme

(2.21) $$e_{k+1} = P_n(\tau D)\, e_k + r_k, \quad k = 0, 1, 2, \ldots,$$

where r_k is the sum of the local discretization error and the local numerical error.

2.4 Convergence and stability

In the preceding sections the difference scheme is derived which determines the error propagation in actual computation. In this section conditions will be given to control the accumulation of errors.

Let $||\ ||_2$ denote the Euclidean norm in the space of level functions. With respect to this norm the following theorem holds:

Theorem 2.1

The error e_k satisfies the inequality

(2.22) $$||e_k||_2 \leq \left[1 + C(\tau) \sum_{j=1}^{k-1} j^{\gamma-1} \left[\sigma(P_n(\tau D))\right]^j\right] \operatorname*{Max}_{0 \leq j \leq k-1} ||r_j||_2,$$

where $C(\tau)$ is an uniformly bounded function as $\tau \to 0$, $\sigma(P_n(\tau D))$ is the spectral radius of $P_n(\tau D)$, and γ is the largest order of all diagonal submatrices J_m of the Jordan normal form J of $P_n(\tau D)$ with $\sigma(J_m) = \sigma(P_n(\tau D))$.

Proof. See reference [4], p. 57.

The sum corresponding in (2.22) may be estimated by the integral

$$\int_0^k x^{\gamma-1} \left[\sigma(P_n(\tau D))\right]^x \, dx.$$

A simple calculation yields

$$(2.23) \quad ||e_k||_2 \leq \{1+C(\tau) \{\frac{\sigma^k}{\ln^\gamma \sigma} [(k\ln\sigma)^{\gamma-1} - (\gamma-1)(k\ln\sigma)^{\gamma-2} + \ldots + (-1)^{\gamma-2}(\gamma-1)! \, k\ln\sigma + (-1)^{\gamma-1}(\gamma-1)!] - \frac{(-)^{\gamma-1}(\gamma-1)!}{\ln^\gamma \sigma}\}\} \max_{0\leq j\leq k-1} ||r_j||_2,$$

where σ denotes $\sigma(P_n(\tau D))$.
For large values of k and $\sigma \neq 1$ this expression reduces essentially to

$$(2.23') \quad ||e_k||_2 \leq \{1+C(\tau) \frac{\sigma^k (k\ln\sigma)^{\gamma-1} - (-)^{\gamma-1}(\gamma-1)!}{\ln^\gamma \sigma}\} \max_{0\leq j\leq k-1} ||r_j||_2$$

For $\sigma = 1$ we have

$$(2.23'') \quad ||e_k||_2 \leq \{1+C(\tau) \frac{k^\gamma}{\gamma}\} \max_{0\leq j\leq k-1} ||r_j||_2.$$

In the following the cases $\sigma > 1$, $\sigma = 1$ and $\sigma < 1$ will be discussed separately.

Case I: $\sigma > 1$. From (2.23') it follows that the error e_k may increase exponentially with k. Even when round-off errors are neglected, so that the local error r_k tends to zero like τ^{p+1} as $\tau \to 0$, we have an exponential increase, because the total number of steps in a given interval of integration $[0,T]$ is T/τ. The local errors do not vanish rapidly enough to compensate for the factor $[\sigma]^{T/\tau}$. Therefore, we may not expect convergence to the analytical solution. Further, the round-off errors may increase exponentially destroying the solution completely. This latter phenomenon is called instability.

Case II: $\sigma = 1$. Here we use inequality (2.23"). Neglecting round-off errors we see that

$$k^{\gamma}\,\tau^{p+1} \leq T^{\gamma}\,\tau^{p-\gamma+1}$$

determines the convergence of the difference solutions as $\tau \to 0$. Therefore, the condition

$$p > \gamma - 1 \tag{2.24}$$

is a sufficient condition for convergence in the case $\sigma = 1$.
For small values of τ, however, the numerical error ρ_k^* cannot be neglected with respect to ρ_k, so that e_k behaves like $\tau^{-\gamma}$ as $\tau \to 0$. Such a behaviour of the error is called stable by Forsythe and Wasow [1], p. 32, but unstable by Rjabenki and Filippov [3], p. 15. In most practical cases this behaviour is acceptable.

Case III: $\sigma < 1$. Again we use inequality (2.23'). As $\tau \to 0$ we have

$$||e_k||_2 \leq \text{const.} \max_{0 \leq j \leq k-1} ||r_j||_2 .$$

Evidently, the scheme is convergent and stable as well.

Our final conclusion is that a necessary condition for convergence is

$$\sigma(P_n(\tau D)) \leq 1. \tag{2.25}$$

Further, this condition guarantees a certain insensitivety for round-off errors. In literature, condition (2.25) is called the stability condition of the difference scheme.
In fact, (2.25) is a condition for the time step τ. To see this we define the numbers $\tau^{(j)}$ as the non-zero solutions of the equations $|P_n(\tau\delta_j)| = 1$, where δ_j represent the eigenvalues of the operator D (see figure 2.2). The minimum of all numbers $\tau^{(j)}$ obviously is an upper bound for τ, i.e.

$$\tau \leq \operatorname*{Minimum}_{|P_n(\tau^{(j)}\delta_j)|=1} \tau^{(j)}. \tag{2.25'}$$

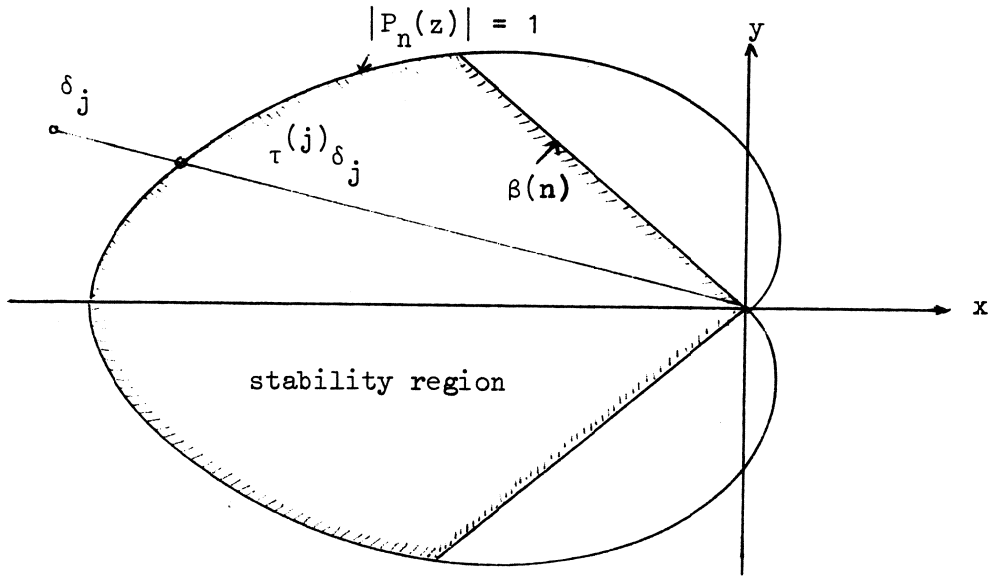

fig. 2.2 Stability region of the polynomial operator $P_n(\tau D)$ in the $(z=x+iy)$-plane.

A simple but rather rough stability condition can be obtained as follows. Let $\beta(n)$ be that point at the curve $|P_n(z)| = 1$ which is nearest to the origin and which lies in the sector in which all eigenvalues δ_j are situated. Then (2.25) is satisfied when

$$\tau \leq \frac{\beta(n)}{\sigma(D)}. \qquad (2.25'')$$

Conditions (2.25') and (2.25") are identical when the eigenvalues δ_j are situated at two lines which are conjugate complex; for instance, when the δ_j are real or purely imaginary.

It may be remarked that in many important applications the bound $\beta(n) / \sigma(D)$ is considerably smaller than the time step prescribed by accuracy considerations. In such cases one should construct difference schemes for which the number $\beta(n)$ is as great as possible. In view of this requirement we have introduced the polynomial operator $B_q(\tau D)$ in section 2.1.

2.5 Step-size control

Condition (2.25) controls the accumulation of the local errors r_k. The next step is to control the local errors itself. In this section the discretization error ρ_k is discussed. In section 2.6 the numerical error ρ^*_k is considered.

Suppose that it is required that ρ_k is bounded by some quantity η_k, i.e.

$$||\rho_k|| \leq \eta_k ,$$

where $||\ ||$ denotes some norm in the space of level functions. Then we derive from (2.14) the inequality

$$\tau \leq \left[\frac{(p+1)!\ \eta_k}{|1 - \beta_{p+1}(p+1)!|\ \ ||c_k^{(p+1)}||} \right]^{\frac{1}{p+1}} . \tag{2.26}$$

This condition prescribes at each level a new maximal time step, contrary to condition (2.25) which yields a uniform upperbound for τ. In fact, the right hand side in (2.26) may vary considerably with k. This will be illustrated by the following example. Given the homogeneous equation

$$\frac{d}{dt}\tilde{U} = D\tilde{U},$$

where D has negative eigenvalues δ_j, $\delta_m < \delta_{m-1} < \ldots < \delta_1 < 0$ with normalized eigenfunctions E_j. Let the initial condition be

$$\tilde{U} = \sum_{j=1}^{m} E_j \quad , \quad t = 0.$$

Then the analytical solution of the differential equation is given by

$$\tilde{U} = \sum_{j=1}^{m} \exp(\delta_j t) E_j ,$$

so that

$$c_k^{(p+1)} = D^{p+1} u_k \sim \sum_{j=1}^{m} \delta_j^{p+1} \exp(\delta_j t_k) E_j \quad \text{as} \quad \tau \to 0 .$$

For small values of t_k the terms with large values of $|\delta_j|$ are dominating, for large values of t_k the terms with small $|\delta_j|$ are dominating. To be more specific, let us take the special case in which $\delta_1 = -1$ and $\delta_2 = -1000$ (m=2). Then we have

$$||c_k^{(p+1)}|| \sim 1000^{p+1} \text{ for } t_k \sim 0, \quad ||c_k^{(p+1)}|| \sim 1 \text{ for } t_k \sim 1.$$

This means that for constant η_k the maximal allowed time step changes a factor 1000 over the interval $0 \leq t \leq 1$.

This example shows that it is desirable to employ a variable step τ_k which satisfies both (2.25) and (2.26), particularly when we deal with ill-conditioned matrices D, i.e. $|\delta_1| \; / \; |\delta_m| << 1$.

2.6 Numerical stability

The error ρ_k^* arises in the calculation of the successive correction terms $\beta_j \tau^j c_k^{(j)}$. Suppose that instead of calculating $c_k^{(j)}$ recursively we calculate the correction term itself by a recurrence formula. Let

$$v_k^{(j)} = \beta_j \tau^j c_k^{(j)} \quad , \quad j = 1, \ldots, n.$$

Then it follows from the recurrence relation for $c_k^{(j)}$ that

$$(2.27) \quad v_k^{(j+1)} = \frac{\beta_{j+1}}{\beta_j} \tau_k D \, v^{(j)} + \beta_{j+1} \tau_k^{j+1} \frac{d^j}{dt^j} F_k, \; j = 0, 1, \ldots, n-1 ,$$

where $v_k^{(0)} = u_k$ and $\beta_0 = 1$.
From the arguments in section 2.4 it follows that a necessary condition for the stability of this process is

$$(2.28) \quad \left|\frac{\beta_{j+1}}{\beta_j}\right| \tau_k \sigma(D) \leq 1 \; , \quad j = 0, 2, \ldots, n-1.$$

When this condition is satisfied the local error ρ_k^* will be small in general. For large values of n it is important to satisfy (2.28). We shall call the process numerically stable when ρ_k^* is small.

It may be remarked that the calculation of the $c_k^{(j)}$ according to scheme (2.8') may be very dangerous for large values of n.

3. Runge-Kutta methods

In this section we study difference schemes of the type

$$(3.1) \quad u_0 = \tilde{U}_0, \; u_{k+1} = A_p(\tau_k D) u_k + \tau_k \, g_k^{(p)}, \quad k = 0, 1, 2, \ldots,$$

where the polynomial $A_p(z)$ and $g_k^{(p)}$ are defined as in (2.9).
This scheme is related to the Runge-Kutta method of order p. We will show this relation for the case p = 2.
The second order Runge-Kutta Method (or Heun's method) for linear equations of type (2.1) is defined by (compare [2], p. 896)

$$(3.2)\quad \begin{cases} u_0 = \tilde{U}_0, \\ u_{k+1} = u_k + \frac{1}{2}(c_1+c_2), \quad k = 0, 1, 2, \ldots, \\ c_1 = \tau_k(Du_k + f_k) , \\ c_2 = \tau_k(Du_k + Dc_1 + f_{k+1}) . \end{cases}$$

By substituting c_1 and c_2 this scheme reduces to

$$\begin{cases} u_0 = \tilde{U}_0 , \\ u_{k+1} = (1 + \tau_k D + \frac{1}{2}\tau_k^2 D^2)u_k + \tau_k(\frac{1}{2}(f_k + f_{k+1}) + \frac{1}{2}\tau_k D f_k). \\ \qquad = A_2(\tau_k D)u_k + \tau_k(f_k + \frac{1}{2}\tau_k(D + \frac{d}{dt})f_k + \frac{1}{4}\tau_k^2 \frac{d^2}{dt^2} F(\bar{t}_k) \\ \qquad = A_2(\tau_k D)u_k + \tau_k g_k^{(2)} + \frac{1}{4}\tau_k^3 \frac{d^2}{dt^2} F(\bar{t}_k) , \end{cases}$$

where $\bar{t}_k = t_k + \theta\tau_k$, $0 \le \theta \le 1$.
This scheme resembles (3.1). However, it has a different local discretization error, namely

$$\rho_k(\tau_k) \sim \frac{1}{3!}\tau_k^3 \left[\frac{d^3}{dt^3}\tilde{U}(t_k) - \frac{3}{2}\frac{d^2}{dt^2}F(t_k) \right] \quad \text{as } \tau_k \to 0.$$

For numerical calculations they are, of course, equivalent, as both schemes have a local discretization error of order τ_k^3.

3.1 Regions of stability

The stability regions of the operator $A_p(\tau_k D)$ are defined by the curves $|A_p(z)| = 1$. In figure 3.1 these curves are given for p = 1, 2, 3 and 4. Since $|A_p(z)| = |\overline{A_p(\bar{z})}|$, the lower part of the curve $|A_p(z)| = 1$ is omitted.

This figure shows that we should require that the eigenvalues δ_j of D have non-positive real parts. This requirement is related to a similar condition one has to impose upon the ordinary differential equation (2.1) in order to guarantee stability in the sense of Lyapunov. In this and subsequent sections we shall assume that Re $\delta_j \leq 0$.
From figure 3.1 the following table is derived.

Table 3.1 Approximate values of $\beta(p)$ in the stability condition

$$\tau_k \leq \frac{\beta(p)}{\sigma(D)} \text{ of the polynomials } A_p(\tau_k D)$$

p	arbitrary δ_j		real δ_j		imaginary δ_j	
	$\beta(p)$	$\beta_{eff}(p)$	$\beta(p)$	$\beta_{eff}(p)$	$\beta(p)$	$\beta_{eff}(p)$
1	0	0	2	2	0	0
2	0	0	2	1	0	0
3	1.72	0.57	2.54	0.85	1.72	0.57
4	2.63	0.67	2.78	0.70	2.82	0.71

In this table the values of $\beta_{eff}(p) = \beta(p)/p$ have been added. These values take into account that a scheme of order p requires p times as much work per time step as a scheme with p = 1. Therefore, the effective time step τ_{eff}, defined by $\tau_{eff} = \tau/p$, satisfies the condition

$$\tau_{eff} \leq \frac{\beta_{eff}(p)}{\sigma(D)} = \frac{\beta(p)}{p\,\sigma(D)} . \tag{3.3}$$

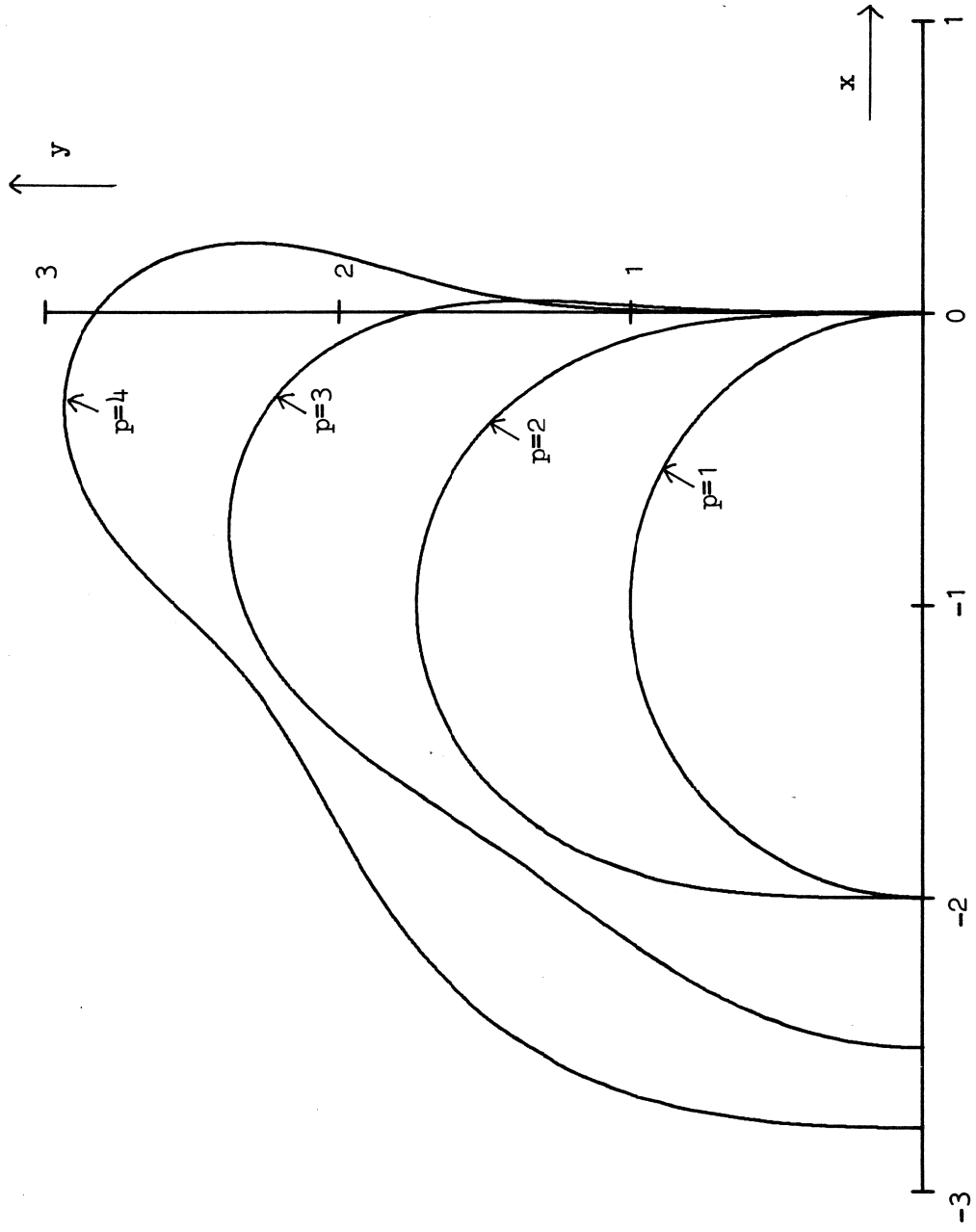

fig. 3.1 Stability regions of the polynomials $A_p(z)$ for p = 1, 2, 3, 4.

We are now in a position to evaluate the merits of the Runge-Kutta methods of order 1 to 4. In general, the first and second order methods (method of Euler and Heun, respectively) are not recommended, as their stability, although the differential equation itself is supposed to be stable, is not guaranteed for any time step. Moreover, the accuracy is considerably less than the accuracy of the third and fourth order method. Only when the solution is slowly varying Euler's and Heun's method may be advantageous (see the numerical examples in [7]). In practice, the fourth order method is most widely used.

4. The use of Chebyshev polynomials

The examples in [7] show that in the asymptotic region of the solution the time step is not prescribed by accuracy requirements but by stability requirements. Therefore, in this part of the integration interval it suffices to employ first order schemes like Euler's scheme. However, it is possible to construct first order schemes which have considerably less stringent stability conditions than Euler's scheme. In this section a class of first order schemes is considered which is appropriate in cases where the matrix D has real or "almost real" eigenvalues.

4.1 Construction of the difference scheme

The first order schemes which arise from (2.9') for p = 1 are of the type

$$
(4.1)\qquad \begin{cases} u_0 = \tilde{U}_0, \\ u_{k+1} = \left[A_1(\tau_k D) + (\tau_k D)^2 \, B_{n-2}(\tau_k D) \right] u_k + \tau_k \, g_k^{(n)} \\ \qquad = \left[1 + \tau_k D + \beta_2 \tau_k D^2 + \ldots + \beta_n \, \tau_k^n D^n \right] u_k + \tau_k \, g_k^{(n)}, \\ \qquad k = 0, 1, 2, \ldots . \end{cases}
$$

For a given polynomial $P_n(x)$, the number $\beta(n)$, as defined in section 2.4, is the largest number such that $P_n(x) = A_1(x) + x^2 B_{n-2}(x)$ has values in the interval $[-1, 1]$ for $-\beta(n) \leq x \leq 0$. Thus, we are lead to the problem to construct a polynomial $P_n(x)$ for which this number $\beta(n)$ is maximal.

Theorem 4.1

Of all polynomials $P_n(x)$ of the degree n in x, which satisfy the conditions

$$P_n(0) = 1, \qquad P_n'(0) = 1,$$

the polynomials $T_n(1 + n^{-2}x) = \cos[n \arccos(1 + n^{-2}x)]$ has the largest value for $\beta(n)$. This value equals $2n^2$.

Proof See reference [4], p. 38.

From this theorem it follows that the scheme

$$(4.2) \qquad \begin{cases} u_0 = \tilde{U}_0, \\ u_{k+1} = T(1 + n^{-2}\tau_k D)u_k + \tau_k g_k^{(n)}, \quad k = 0, 1, 2, \ldots \end{cases}$$

is the scheme we are looking for. The first four polynomials together with their $\beta(n)$ values are given by

$$(4.3) \qquad \begin{cases} P_1(x) = 1 + x, \quad \beta(1) = 2, \\ P_2(x) = 1 + x + \frac{1}{8}x^2, \quad \beta(2) = 8, \\ P_3(x) = 1 + x + \frac{4}{27}x^2 + \frac{4}{729}x^3, \quad \beta(3) = 18, \\ P_4(x) = 1 + x + \frac{5}{32}x^2 + \frac{1}{128}x^3 + \frac{1}{8192}x^4, \quad \beta(4) = 32. \end{cases}$$

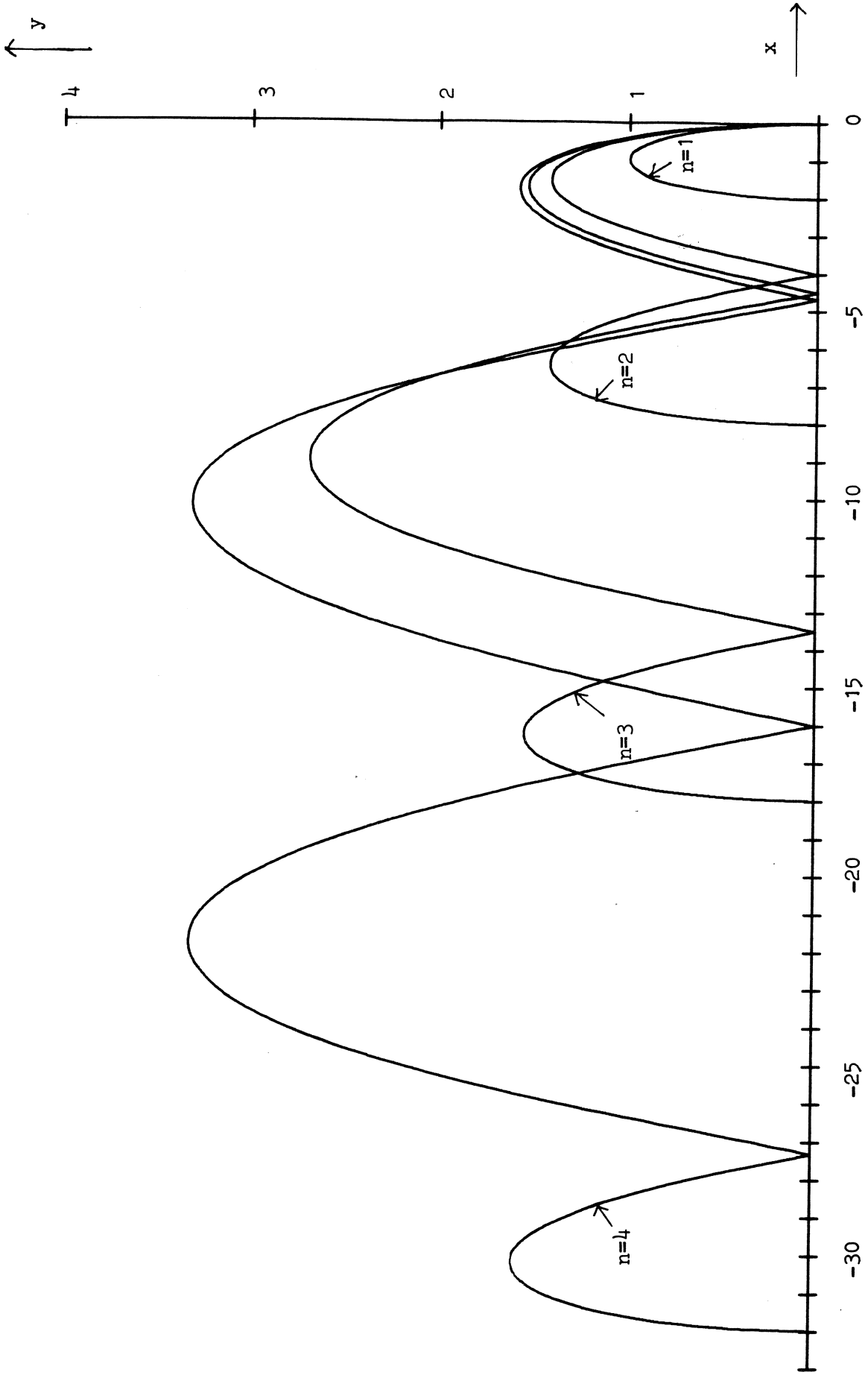

fig. 4.1 Stability regions of the polynomials $T_n(1 + n^{-2}z)$ for n = 1, 2, 3, 4.

4.2 Regions of stability

In the case of real eigenvalues we have from theorem 4.1 the stability condition

$$(4.4) \qquad \tau_k \leq \frac{2n^2}{\sigma(D)}, \qquad (\tau_k)_{eff} \leq \frac{2n}{\sigma(D)}.$$

Comparing this condition with table 3.1 we see that in the real case the use of Chebyshev polynomials allows us to take n times larger steps than Euler's scheme.

For complex eigenvalues Chebyshev polynomials are appropriate in those cases where the eigenvalues are within the stability regions of the polynomials $T_n(1 + n^{-2}z)$.

In figure 4.1 the curves $|T_n(1 + n^{-2}z)| = 1$ are given. From these curves it may be concluded that stability is expected if the eigenvalues are "almost real", i.e. if they are situated in a small strip along the (negative) real axis. In such cases it is recommended to use polynomials of different degree in succession. Then, for those points $\tau_k\delta_j$ which are outside the stability region of $P_{n_k}(z)$, we may hope that the points $\tau_{k+1}\delta_j$ are within the stability region of $P_{n_{k+1}}(z)$. In this manner the instabilities introduced in the k-th step are reduced in the (k+1)-st step.

5. The case of purely imaginary eigenvalues

We now consider schemes of type (4.1) in which D is an operator with purely imaginary eigenvalues $\delta = iy$. The number $\beta(n)$ is now defined as the largest number such that $P_n(iy) = A_1(iy) + (iy)^2 \beta_{n-2}(iy)$ has values on or within the unit circle $|z| = 1$. Again we are faced with the problem to construct a polynomial $P_n(iy)$ for which this number $\beta(n)$ is maximal.

5.1 A polynomial problem

For n = 2, 3, 4 the optimal polynomial $P_n(z)$ was derived in [4], p.45 For odd values of n a general expression of $P_n(z)$ is given in [5]. For future reference the first four polynomials and the general expression for odd values of n are given, together with the corresponding values of $\beta(n)$.

$$
(5.1)\quad \begin{cases}
P_1(z) = 1 + z, \ \beta(1) = 0, \\
P_2(z) = 1 + z + z^2, \ \beta(2) = 1, \\
P_3(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{4}z^3, \ \beta(3) = 2, \\
P_4(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4, \ \beta(4) = 2\sqrt{2}, \\
P_n(z) = T_{\frac{n-1}{2}}\left[\frac{(n-1)^2 + 2z^2}{(n-1)^2}\right] + 2z\,\frac{(n-1)^2 + z^2}{(n-1)^3}\,U_{\frac{n-3}{2}}\left[\frac{(n-1)^2 + 2z^2}{(n-1)^2}\right], \\
\qquad\qquad \beta(n) = n-1, \ n = 1, 3, 5, \ldots .
\end{cases}
$$

Here $U_m(y)$ denotes the Chebyshev polynomial of the second kind, i.e.

$$
(5.2)\qquad U_m(y) = \frac{\sin[(m+1)\arccos y]}{\sin \arccos y}.
$$

5.2 Regions of stability

In figure 5.1 the curves $|P_n(z)| = 1$, $P_n(z)$ defined by (5.1), are given.

For n = 2 we have stability if

$$
(5.3)\qquad \tau_k \leq \frac{1}{\sigma(D)} \quad , \quad (\tau_k)_{eff} \leq \frac{.5}{\sigma(D)} \quad .
$$

This condition does not hold only in the imaginary case, but for any set of eigenvalues with Re $\delta_j \leq 0$. Therefore, with respect to stability the operator $1 + \tau_k D + \tau_k^2 D^2$ is the best possible one of degree 2 in those cases where the only information about the eigenvalues of D is that they are in the non-positive half-plane. As regards the accuracy, however, we observe that $P_2(\tau_k D)$ has only first order accuracy.

The polynomial $P_3(\tau_k D)$ is second order exact and satisfies the condition

$$\tau_k \leq \frac{2}{\sigma(D)}, \quad (\tau_k)_{eff} \leq \frac{.67}{\sigma(D)}, \tag{5.4}$$

which is a slightly better than the third order Runge-Kutta process.

For $n = 4$ the polynomial $P_4(z)$ coincides with the polynomial $A_4(z)$. We recall (see table 3.1) that τ_k must satisfy the condition

$$\tau_k \leq \frac{2\sqrt{2}}{\sigma(D)}, \quad (\tau_k)_{eff} \leq \frac{.71}{\sigma(D)}. \tag{5.5}$$

The next polynomials $P_5(z)$, $P_7(z)$, ... yield second order exact schemes with a slowly increasing upper bound for the effective time step, i.e.

$$\tau_k \leq \frac{n-1}{\sigma(D)}, \quad (\tau_k)_{eff} \leq \frac{1-1/n}{\sigma(D)}, \quad n = 1, 3, 5, 7, \ldots . \tag{5.6}$$

Thus $P_4(\tau_k D)$ already has 70% of the maximal attainable stability, and it is therefore that we recommend the fourth order exact Runge-Kutta method in the case of imaginary eigenvalues.

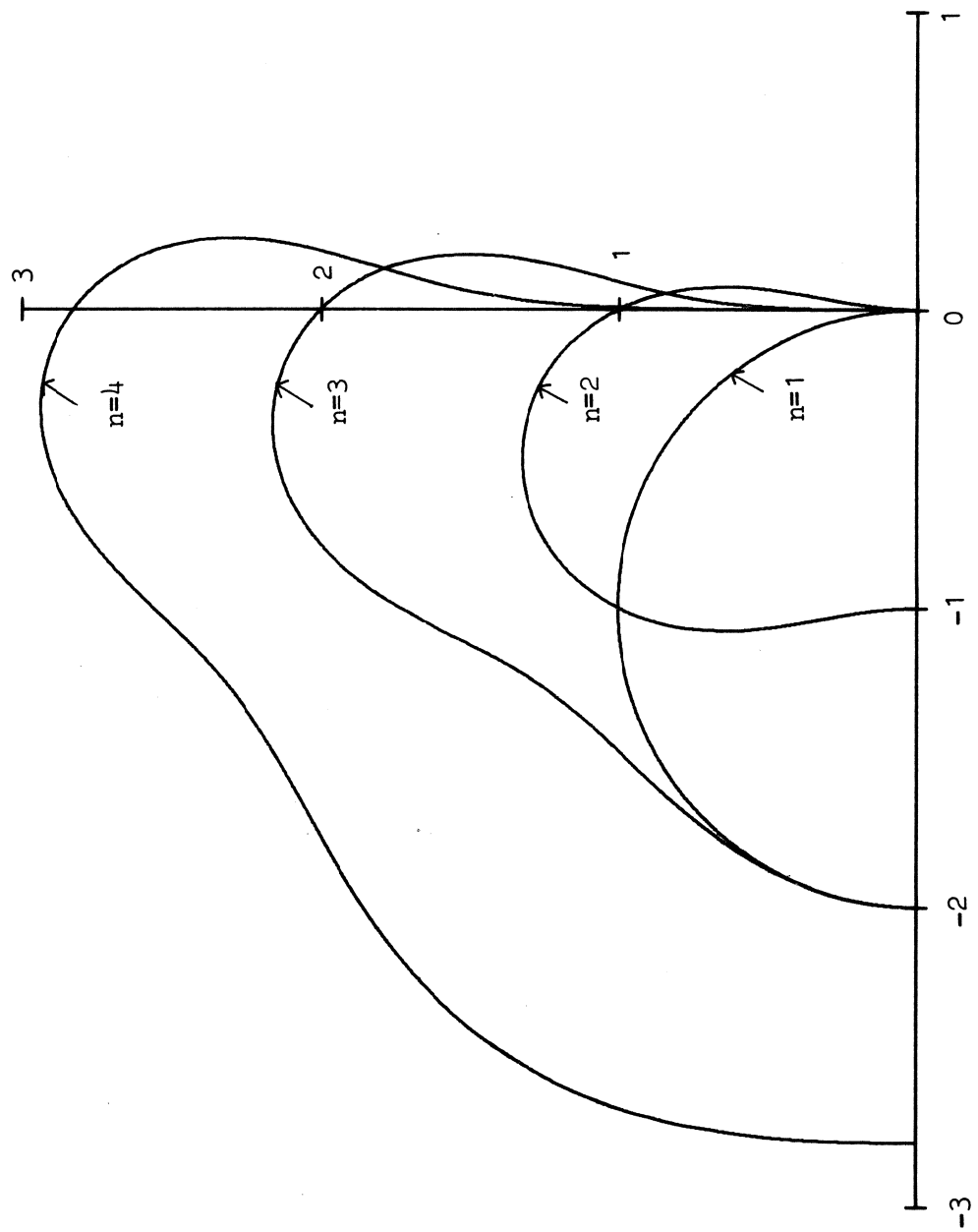

fig. 5.1 Stability regions of the polynomials $P_n(z)$ for n = 1, 2, 3, 4.

6. Stabilization of higher order schemes

In the preceding sections the problem is discussed to maximize the number $\beta(n)$ associated to polynomials of the type

$$P_n(z) = A_1(z) + z^2 B_{n-2}(z),$$

where z was real and imaginary, respectively. In the imaginary case, only a small improvement was obtained in comparison with the fourth order Runge-Kutta process. In the real case, a considerable improvement was obtained; however, these methods are only first order exact. Therefore, it is natural to try to maximize the number $\beta(n)$ associated with polynomials of the type

$$P_n(x) = A_p(z) + x^{p+1} B_q(x), \quad p + q + 1 = n, \quad p > 1. \tag{6.1}$$

Thus, instead of stabilizing first order schemes we now try to stabilize a p-th order scheme.

6.1 Properties of the polynomial $B_q(x)$

Let $P_n(x)$ have values between -1 and +1 for $-\beta(n) \leq x \leq 0$, then $B_q(x)$ satisfies the inequalities

$$\begin{cases} -x^{-p-1}(1 + A_p(x)) \leq B_q(x) \leq x^{-p-1}(1 - A_p(x)), \\ \qquad\qquad p \text{ odd}, \ -\beta(n) \leq x \leq 0, \\ x^{-p-1}(1 - A_p(x)) \leq B_q(x) \leq -x^{-p-1}(1 + A_p(x)), \\ \qquad\qquad p \text{ even}, \ -\beta(n) \leq x \leq 0. \end{cases} \tag{6.2}$$

Let $l_p(x)$ and $r_p(x)$ denote the left and right hand side of (6.2), respectively.

Then we have

Theorem 6.1

Of all polynomials of degree q in x the polynomial $B_q(x)$ yields the largest number $\beta(n)$ if it has at least q + 1 alternating tangent points with the boundary curves $l_p(x)$ and $r_p(x)$.

Proof

The maximization of $\beta(n)$ means that the curve $y = B_q(x)$ remains as long as possible in the region bounded by the curves $y = l_p(x)$, $y = r_p(x)$ and $x = 0$. In figure 6.1 the behaviour of the boundary curves $y = l_p(x)$ and $y = r_p(x)$ is illustrated. For $x \to 0$ they tend to $-\infty$ and $+\infty$, respectively, for $x \to -\infty$ they both converge to zero.

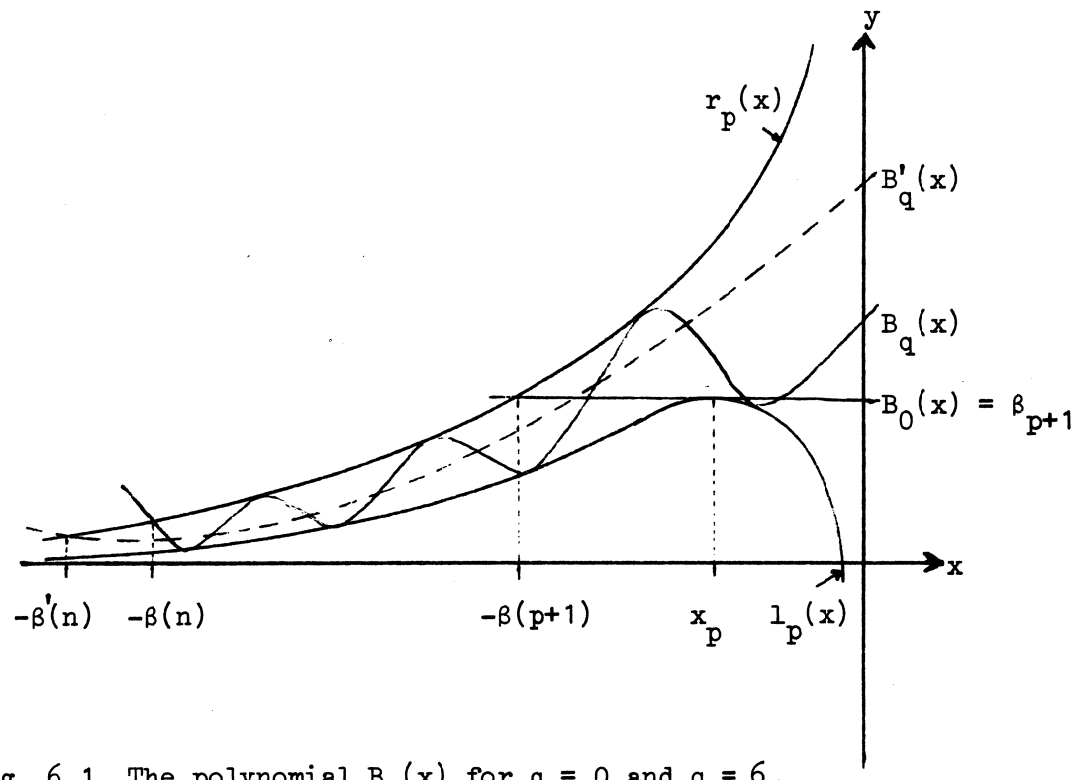

fig. 6.1 The polynomial $B_q(x)$ for q = 0 and q = 6.

Suppose that there exists a polynomial $B_q(x)$ with $q + 1$ different tangent points with $l_p(x)$ and $r_p(x)$, and a polynomial $B'_q(x)$ which satisfies inequality (6.2) over an interval $[-\beta'(n),0]$ where $\beta'(n) > \beta(n)$. Then the curve $y = B'_q(x)$ intersects the curve $y = B_q(x)$ at least at $q + 1$ points. Hence the polynomial $B'_q(x) - B_q(x)$, which is atmost of degree q in x, has at least $q + 1$ zeroes.
This contradiction proves the theorem.

Although this theorem does not guarantee the existence of a polynomial $B_q(x)$ with $q + 1$ tangent points, it may guide us in constructing the "best" polynomial.

6.2 Introduction of a single stability term

We consider the case $q = 0$, i.e.

$$B_q(x) = \beta_{p+1}. \tag{6.3}$$

From theorem 6.1 it follows that the line $y = \beta_{p+1}$ which touches the curve $y = l_p(x)$ defines the optimal value of β_{p+1} (see figure 6.1). Obviously, the tangent point is the point where $l_p(x)$ reaches its maximum. Let $x = x_p$ be this point, then the following relations hold:

$$\begin{cases} B_0(x_p) = \beta_{p+1} = l_p(x_p) , \\ r_p(-\beta(n)) = l_p(x_p) , \quad n = p + 1. \end{cases} \tag{6.4}$$

In table 6.1 the values of x_p, β_{p+1}, $\beta(n) = \beta(p+1)$ and $\beta_{eff}(p+1)$ are listed for $p = 1, 2, 3$ and 4.

Table 6.1 Parameter values for polynomials of the type

$$P_n(x) = A_p(x) + \beta_{p+1} x^{p+1}.$$

p	x_p	β_{p+1}	$\beta(p+1)$	$\beta_{eff}(p+1)$
1	- 4.00	. 1250000	8.00	4.00
2	- 4.00	. 0625000	6.27	2.09
3	- 4.39	. 0184557	6.00	1.50
4	- 4.69	. 0040869	6.05	1.21

The corresponding polynomials are given by

$$(6.5)\begin{cases} P_2(x) = 1 + x + .125\, x^2, \\ P_3(x) = 1 + x + .5\, x^2 + .0625\, x^3, \\ P_4(x) = 1 + x + .5\, x^2 + .1666667\, x^3 + .0184557\, x^4, \\ P_5(x) = 1 + x + .5\, x^2 + .1666667\, x^3 + .0416667\, x^4 + .0040869\, x^5. \end{cases}$$

By comparing the stability properties of these polynomials and the Runge-Kutta formulae for real eigenvalues (see table 3.1) it may be concluded that the introduction of just one stability term allows us to employ 70% larger time steps in the fourth order case.

6.3 The case of two stability terms

For q = 1 we have

$$(6.6) \qquad B_1(x) = \beta_{p+1} + \beta_{p+2} x.$$

From figure 6.2 it is clear that there actually exists a polynomial $B_1(x)$ which touches both $l_p(x)$ and $r_p(x)$.

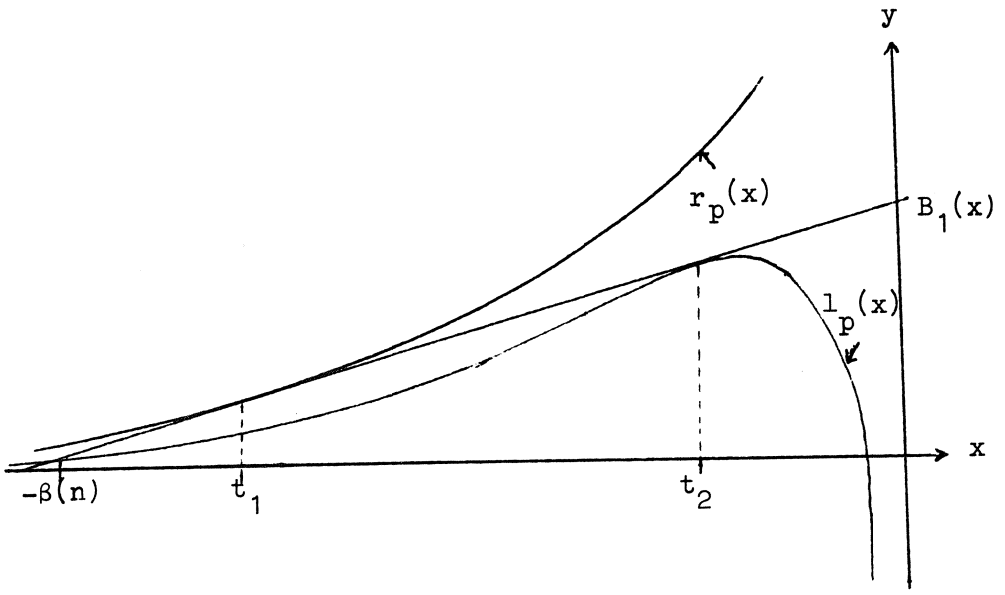

fig. 6.2

Let the line $y_1 = B_1(x)$ touch $y = l_p(x)$ and $y = r_p(x)$ in $x = t_2$ and $x = t_1$ respectively. Then the following equations hold for t_1 and t_2:

$$(6.7) \qquad \begin{cases} r'_p(t_1) = l'_p(t_2) \ , \\ r_p(t_1) - t_1\, r'(t_1) = l_p(t_2) - t_2\, l'(t_2) \ . \end{cases}$$

When these equations are solved for t_1 and t_2 the line $y = B_1(x)$ is defined by

$$(6.8) \qquad y = r'_p(t_1)x + (r_p(t_1) - r'_p(t_1)t_1) \ ,$$

thus

$$(6.9) \qquad \beta_{p+1} = r_p(t_1) - r'_p(t_1)t_1 \ , \ \beta_{p+2} = r'_p(t_1) \ .$$

Further, $\beta(n) = \beta(p+2)$ is defined by the equation

$$l_p(-\beta(p+2)) = B_1(-\beta(p+2)). \tag{6.10}$$

We now solve the equations (6.7) for $p = 2$. From the definition of $r_2(x)$ and $l_2(x)$ it follows that (6.7) may be written as

$$\begin{cases} t_2^{-2} - t_1^{-1} + 4(t_2^{-3} - t_1^{-3}) = 12\, t_1^{-4} , \\ t_2^{-1} - t_1^{-1} + 3(t_2^{-2} - t_1^{-2}) = 8\, t_1^{-3} , \end{cases} \tag{6.7'}$$

or equivalently

$$\begin{cases} 8\, t_2^{-2} - t_1^{-2} - t_1^{-1}\, t_2^{-1} + 2\, t_2^{-1} - t_1^{-1} = 0 , \\ 3\, t_2^{-2} + t_2^{-1} - t_1^{-1}(8\, t_1^{-2} + 3\, t_1^{-1} + 1) = 0 . \end{cases} \tag{6.7''}$$

Elimination of t_2 yields the following equation for t_1

$$(3t_1^{-1} + 2)^2 (96t_1^{-3} + 36t_1^{-2} + 12t_1^{-1} + 1) - (384t_1^{-3} + 126t_1^{-2} + 33t_1^{-1} + 2)^2 = 0.$$

A numerical calculation reveals that $t_1 \sim -10$ is an approximate zero of this equation. A corresponding value of t_2 is given by $t_2 \sim -4.8$. By formulae (6.8) - (6.10) we finally obtain

$$P_4(x) = 1 + x + .5\, x^2 + .078\, x^3 + .0036\, x^4 , \quad \beta(4) \sim 12 . \tag{6.11}$$

Hence we have gained a factor 3 over Heun's method (see table 3.1).

6.4 Polynomials $B_q(x)$ of higher degree

Analogous to the considerations in the preceding section polynomials $B_q(x)$ of higher degree can be constructed by setting up the equations for the tangent points $x = t_j$, $j = 1, 2, \ldots, q+1$. This leads to $q+1$ non-linear equations for $q+1$ unknowns. The solution of these equations is difficult to find, even when numerical methods are employed. Therefore, we look for other methods to construct $B_q(x)$. In the following, a method will be described which approximately determines the coefficients of $B_q(x)$. This method is based on the Taylor-expansion of the function

$$a_p(x) = \frac{1}{2}(r_p(x) - l_p(x)) \tag{6.13}$$

for large negative values of x (see figure 6.3).

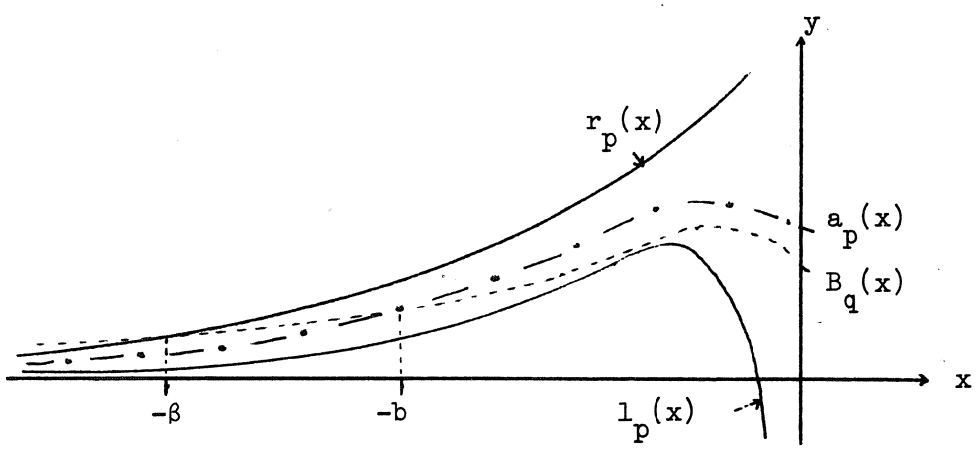

fig. 6.3 Taylor-expansion of $a_p(x)$

We shall give the analysis for the case p = 2.
Let x = -b be a point at the negative axis. Then the Taylor-expansion of $a_2(x)$ at x = -b is given by

$$a_2(x) = -\frac{A_2(x)}{x^3} = \sum_{j=0}^{\infty} \frac{b^2 - 2(j+1)\,b + (j+1)(j+2)}{2\,b^3} \left(\frac{x+b}{b}\right)^j . \tag{6.14}$$

We now define the polynomial $B_q(x)$ by

$$(6.15) \qquad B_q(x) = \sum_{j=0}^{q} \frac{b^2 - 2(j+1)\,b + (j+1)(j+2)}{2\,b^3} \left(\frac{x+b}{b}\right)^j .$$

The parameter b in this expression is determined by the condition that for $-b \le x \le 0$ the remainder $R_{q+1}(x) = a_2(x) - B_q(x)$ does not exceed in absolute value the "distance" between $a_2(x)$ and $r_2(x)$ or $l_2(x)$. Thus we require

$$(6.16) \qquad |R_{q+1}(x)| \le |a_2(x) - r_2(x)| = \left|\frac{1}{x^3}\right| , \quad -b \le x \le 0.$$

The remainder $R_{q+1}(x)$ may be written as

$$(6.17) \qquad R_{q+1}(x) = \frac{1}{2\,b^3} \left(\frac{x+b}{b}\right)^{q+1} \sum_{j=0}^{\infty} \left[b^2 - 2(q+2+j)b + (q+2+j)(q+3+j)\right] \left(\frac{x+b}{b}\right)^j .$$

An upper bound for $R_{q+1}(x)$, $-b \le x \le 0$, is easily seen to be

$$\frac{1}{2\,b^3} \left(\frac{x+b}{b}\right)^{q+1} \left[b^2 - 2(q+2)b + (q+2)(q+3)\right] . \frac{b}{-x} .$$

For small values of q (q = 1, 2, 3) this bound is rather rough, getting closer to the true value of $R_{q+1}(x)$ for larger values of q. By using this majorizing function for $R_{q+1}(x)$ condition (6.16) becomes

$$(6.16') \qquad x^2\,|x+b|^{q+1} \le \frac{2\,b^{q+3}}{b^2 - 2(q+2)b + (q+2)(q+3)} , \quad -b \le x \le 0 .$$

The maximal value of the left hand side in the interval $[-b, 0]$ is reached at $x = -2b/(q+3)$. Hence, the maximal value of b satisfies the equation

$$(6.18) \qquad \frac{4\,b^2}{(q+3)^2} \left|b - \frac{2b}{q+3}\right|^{q+1} = \frac{2\,b^{q+3}}{b^2 - 2(q+2)b + (q+2)(q+3)} .$$

From this equation it follows that

$$(6.19) \qquad b = q+2 + \sqrt{\frac{1}{2}\,\frac{(q+3)^{q+3}}{(q+1)^{q+1}} - (q+2)}\,.$$

The corresponding value of β is determined by the equation

$$(6.20) \qquad |R_{q+1}(-\beta)| = \frac{1}{\beta^3}\,.$$

For $x < -b$ the remainder $R_{q+1}(x)$ satisfies the condition

$$|R_{q+1}(x)| \leq \frac{b^2 - 2(q+2)b + (q+2)(q+3)}{2\,b^3}\,\left|\frac{x+b}{b}\right|^{q+1}\,.$$

Using this upper bound for $R_{q+1}(x)$ at $x = -\beta$ we obtain from (6.18) and (6.20) the relation

$$(6.21) \qquad \beta = b\left[1 + \left(4\,\frac{(q+1)^{q+1}}{(q+3)^{q+3}}\,\frac{b^3}{\beta^3}\right)^{\frac{1}{q+1}}\right]\,.$$

In table 6.2 numerical values of b, β/b, β, β_{eff} are listed.

Tabel 6.2 Parameter values for polynomials of the type $P_n(x) = A_2(x) + x^3 B_q(x)$.

$n = q+3$	$b(n)$	$\beta(n) / b(n)$	$\beta(n)$	$\beta_{eff}(n)$
3	5.391	1.109	5.979	1.993
4	8.385	1.192	9.995	2.499
5	11.339	1.259	14.276	2.855
6	14.280	1.314	18.764	3.127
7	17.215	1.360	23.412	3.345
8	20.145	1.399	28.183	3.523
9	23.074	1.434	33.088	3.676
10	26.001	1.464	38.065	3.807
$n\to\infty$	2.922 n	2	5.844 n	5.844

The results for $n \to \infty$ needs some explanation. From (6.19) we derive

$$b \sim q+2 + \frac{1}{2}(q+3)\, e\, \sqrt{2} \qquad \text{as } q \to \infty ,$$

where $e \sim 2.7182$. Substitution of this numerical value of e and $q = n-3$ yields the result given above. The result $\beta(\infty)/b(\infty) = 2$ follows directly from (6.21).

Our conclusion from table 6.2 is that, although the polynomials constructed above are not optimal, the effect for $q = 0,1$ is only slightly less than the corresponding optimal polynomials derived in the preceding sections. Hence, we may expect that the higher degree polynomials $B_q(x)$, $q > 1$, which are even better approximations, are nearly as good as the optimal ones.

Finally, we give in table 6.3 the coefficients $\beta_3, \beta_4, \ldots, \beta_7$ of the resulting polynomials.

Table 6.3 Coefficients $\beta_3, \ldots, \beta_7$ of the approximating polynomials $P_n(x) = A_2(x) + x^3 B_q(x)$

$n = q + 3$	$10^9\, \beta_3$	$10^{10}\, \beta_4$	$10^{11}\, \beta_5$	$10^{12}\, \beta_6$	$10^{14}\, \beta_7$
3	64720219				
4	83375271	43257975			
5	92476529	70859328	19346763		
6	97883479	89253815	38670047	6466748	
7	101469351	102270702	55104889	15282050	17209475

References

[1] Forsythe, G.E. and W.R. Wasow, Finite difference methods for partial differential equations, John Wiley & Sons, Inc., New York (1960).

[2] Handbook of Mathematical Functions, National Bureau of Standards, Applied Mathematics Series 55, edited by M. Abramowitz and J.A. Stegun (1964).

[3] Rjabenki, V.S. and A.F. Filippov, Über die Stabilität von Differenzengleichungen, Deutscher Verlag der Wissenschaften, Berlin (1960).

[4] Van der Houwen, P.J., Finite difference methods for solving partial differential equations, M.C. Tract 20, Mathematisch Centrum, Amsterdam (1968).

[5] Van der Houwen, P.J., Difference schemes with complex time steps, Report M.R. 105, Mathematisch Centrum, Amsterdam (1969).

[6] Van der Houwen, P.J., One-step methods for linear initial value problems II. Applications to stiff equations, T.W. Mathematisch Centrum, Amsterdam (to appear).

[7] Van der Houwen, P.J., One-step methods for linear initial value problems III. Numerical examples, Report T.W. Mathematisch Centrum, Amsterdam (to appear).