

*The Maximal Accuracy of Stable Difference  
Schemes for the Wave Equation*

Rolf Jeltsch

Seminar für Angewandte Mathematik  
ETH Zürich, Switzerland

Rosemary A. Renaut

Department of Mathematics  
Arizona State University, Tempe, AZ 85287-1804

Kosie J.H. Smit

Department of Mathematics  
University of Stellenbosch, Republic of South Africa

September 19th 1994

---

1980 Mathematics Subject classification (1985 Revision); 65M10.

Keywords: finite difference methods, wave equation, accuracy, stability,  
Padé approximants, order stars, Riemann surface.

## Abstract

We consider three time-level difference schemes, symmetric in time and space, for the solution of the wave equation,  $u_{tt} = c^2 u_{xx}$ , given by

$$\sum_{j=-S}^S b_j U_{n+1, m+j} + \sum_{j=-s}^s a_j U_{n, m+j} + \sum_{j=-S}^S b_j U_{n-1, m+j} = 0 .$$

It has already been proved that the maximal order of accuracy  $p$  of such schemes is given by  $p \leq 2(s + S)$ . In this paper we show that the requirement of stability does not reduce this maximal order, for any choice of the pair  $(s, S)$ . The result is proved by introducing an order star on the Riemann surface of the algebraic function associated with the scheme. Furthermore, Padé schemes, with  $S = 0, s > 0$  and  $s = 0, S > 0$  are proved to be stable for  $0 < \mu < 1$ , where  $\mu$  is the Courant number. These schemes can be implemented with high-order absorbing boundary conditions without reducing the range of  $\mu$  for which stable solutions are obtained.

## 1. Introduction

Consider an initial value problem for the wave equation

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2}(t, x) &= c^2 \frac{\partial^2 u}{\partial x^2}(t, x) ; t \geq 0, x \in \mathbb{R} \\ u(0, x), \frac{\partial u}{\partial t}(0, x) &\text{ given .} \end{aligned} \tag{1.1}$$

Let  $\Delta t$  and  $\Delta x$  denote step sizes in the time and space variables, respectively, and  $\mu = \frac{c\Delta t}{\Delta x}$  be the Courant number. An approximate solution of (1.1) is determined via the finite difference scheme

$$\sum_{j=-S}^S b_j U_{n+1, m+j} + \sum_{j=-s}^s a_j U_{n, m+j} + \sum_{j=-S}^S b_j U_{n-1, m+j} = 0 , \tag{1.2}$$

$$m = 0, \pm 1, \pm 2, \dots, \quad n = 1, 2, \dots$$

where  $U_{n, j}$  approximates the exact solution  $u(n\Delta t, j\Delta x)$ , and the coefficients  $a_j, b_j$  are functions of  $\mu$ . Further, because solutions of (1.1) are symmetric in both the time and space variables ((1.1) is invariant under the transformations  $x \rightarrow -\zeta, t \rightarrow -\tau$ ) the form of (1.2) has been chosen so that the same coefficients  $b_j$  are used at time levels  $t_{n-1}$  and  $t_{n+1}$  and it is assumed that the coefficients  $a_j$  and  $b_j$  are symmetric in space,  $a_{-j} = a_j, b_{-j} = b_j$ . The equal flow of information in both directions also imposes the conditions  $a_j(-\mu) = a_j(\mu), b_j(-\mu) = b_j(\mu)$ . Thus (1.2) is of the same type as already considered in [8] and [9].

In [8] Renaut showed that the order of accuracy,  $p$ , of (1.2) satisfies  $p \leq 2(s + S)$ . The proof for  $\mu = \frac{1}{2}$  is clear, but for arbitrary  $\mu$  relies on an assumption about the location of

the poles and zeros of a rational approximation associated with (1.2). Here we will present a proof of this assumption which thus verifies the result in [8]. Renaut [8] also showed that certain schemes (1.2), those for which the coefficients  $a_j, b_j$  are analytic as functions of  $\mu$  and have order  $p = 2(s + S)$ , are stable in the limit as  $\mu \rightarrow 0$ . But are these schemes stable for practical choices of  $\mu$ ? In [9] it was shown that for the “more explicit” schemes, i.e. those with  $s \geq S$ , the imposition of stability does not restrict  $p$  so that  $p$  is still bounded by  $2(s + S)$ . Here we investigate the accuracy of stable schemes satisfying  $S > s$ , the “more implicit” schemes. For the sake of completeness the results presented here cover all choices of the pairs  $(s, S)$ .

The method of proof relies on the introduction of an order star defined on the Riemann surface of the algebraic function associated with (1.2). In section 2 of this paper the notions of stability and order of accuracy of (1.2) are introduced and related to the characteristic function of (1.2). It is observed that these criteria can be investigated via consideration of the algebraic function approximating  $z^\mu$  in the neighbourhood of  $z = 1$ , which is also a root of the characteristic function of (1.2). The properties of this algebraic function are given in section 3 and the order star defined on its Riemann surface introduced in section 4. This order star has properties very similar to the one described in [9,5] but actually possesses a simpler geometric structure. The geometric structure is detailed in section 5 and used to prove an interlace property about the negative real zeros and poles of rational approximations to  $(\frac{z^\mu + z^{-\mu}}{2})$ .

From these results we prove that for stable schemes  $p$  is limited by  $p \leq 2(s + S)$  when  $0 < \mu \leq 1$ . Furthermore, except when  $S = 0$  the order must be reduced for  $\mu > 1$ . Finally, in section 6, we prove the stability of the Padé schemes with  $s = 0, S > 0$  and  $S = 0, s > 0$ , for  $0 < \mu < 1$ . In the latter case the proof corrects an error occurring in the proof in [8]. These schemes can also be used as interior schemes for the solution of the wave equation on an artificially bounded domain, without stability imposing any additional conditions on the boundary schemes beyond those already imposed by the standard second order explicit scheme [10].

## 2. Stability and order of accuracy

As is customary the stability and order of accuracy of (1.2) are studied via the application of a Fourier transform, see e.g. [7,9,3]. Thus associated with (1.2) are the characteristic functions

$$a(z) = \sum_{j=-s}^s a_j z^j, \quad b(z) = \sum_{j=-S}^S b_j z^j \quad (2.1)$$

and

$$\Phi(z, w) = b(z)w + a(z) + b(z)w^{-1},$$

where the variables  $z$  and  $w$  relate to space and time amplification, respectively. The functions  $a(z)$  and  $b(z)$  are assumed to have no common factor and it is also assumed that the implicit part  $b(z)$  of the scheme (1.2) is invertible. By the Wiener–Hopf criterion, [3],  $b(z)$  is invertible if it has  $S$  zeros inside and  $S$  zeros outside the unit circle in  $\mathbb{C}$ . But the

symmetry in the coefficients  $b_j, b_{-j} = b_j$ , means that  $b(z) = b(1/z)$  and hence invertibility translates to the requirement that  $b(z)$  is nonzero along the unit circle.

A scheme (1.2) is said to be stable if

$$\left. \begin{array}{l} \Phi(z, w) = 0 \\ |z| = 1 \end{array} \right\} \implies |w| \leq 1. \quad (2.2)$$

Note that this definition is relaxed slightly from that given in [8] and [9], where roots on the unit circle were required to be simple, unless  $z = 1$ , but is in agreement with that given by Strikwerda [13].

**Remark 2.1.** Observe from (2.1) that  $\Phi(z, w) = \Phi(z, w^{-1})$ . Hence if  $w$  is a root of  $\Phi$  then  $w^{-1}$  is also. Because  $\Phi$  is quadratic in  $w$  it has either two simple roots or one double root. But in order for  $w$  to be a double root we must have  $w = w^{-1}$  which implies that the only double roots are  $w = \pm 1$ . Therefore there are no double complex roots of (2.1) and hence the consequence of the relaxed definition is only that for  $|z| = 1$  double roots  $w = \pm 1$  are allowed. Furthermore, and, as we see later, more importantly, if  $w_1(z)$  is a complex root of (2.1) for  $|z| = 1$  and (1.2) is stable then

$$|w_1(z)| = |w_2(z)| = 1$$

and  $w_1$  and  $w_2$  are a complex conjugate pair of zeros of (2.1).

**Remark 2.2.** It is appropriate to assume that (1.2) is consistent, i.e. (1.2) gives the exact solution of (1.1) if  $u(t, x) = \text{constant}$ , which leads to the condition

$$\Phi(1, 1) = 2 \sum_{j=-S}^S b_j + \sum_{j=-s}^s a_j = 0. \quad (2.3)$$

Hence  $w = 1$  is necessarily a double root of  $\Phi(z, w) = 0$  at  $z = 1$  and the stability condition must allow for linear growth in time of the solution  $u(t, x)$ .

A scheme (1.2) is said to have order of accuracy  $p$  if for a sufficiently smooth solution  $u(t, x)$  of (1.1)

$$\begin{aligned} & \sum_{j=-S}^S b_j u(t + \Delta t, x + j\Delta x) + \sum_{j=-s}^s a_j u(t, x + j\Delta x) + \sum_{j=-S}^S b_j u(t - \Delta t, x + j\Delta x) \\ &= C \frac{\partial^{p+2}}{\partial x^{p+2}} u(t, x) (\Delta x)^{p+2} + O((\Delta x)^{p+3}), (\Delta x \rightarrow 0). \end{aligned} \quad (2.4)$$

Because the numerical values  $U_{n,m}$  computed by (1.2) are independent of whether all coefficients  $a_j, b_j$  are multiplied by some constant, the constant  $C$  is not a good measure

of accuracy. Instead it is more appropriate to introduce the “normalised error constant”  $\tilde{C}$ ,

$$\tilde{C} = \frac{C}{2\mu b(1)}, \quad (2.5)$$

where the motivation for this definition is provided in the next theorem.

**Theorem 2.1.** Assume that a scheme given by  $\Phi(z, w)$  and Courant number  $\mu$  satisfies  $\Phi(1, 1) = 0$ . Then the following three conditions are equivalent:

- (i) the scheme has order  $p = 2q$  and error constant  $C$ ;
- (ii)  $\Phi(z, z^\mu) = C(z-1)^{p+2} + O((z-1)^{p+3}), (z \rightarrow 1)$ ;
- (iii) the algebraic function  $w$  given by

$$\Phi(z, w(z)) \equiv 0$$

has exactly one branch  $w_1(z)$  which is analytic in a neighbourhood of  $z = 1$  and satisfies

$$z^\mu - w_1(z) = \tilde{C}(z-1)^{p+1} + O((z-1)^{p+2}), (z \rightarrow 1) \quad (2.6)$$

where  $\tilde{C} = \frac{C}{2\mu b(1)}$ . □

To prove the theorem we first prove the following Lemma.

**Lemma 2.1.** The method (1.2) is of order  $p = 2q$  if and only if

$$\sum_{j=-S}^S \{(j+\mu)^{2k} + (j-\mu)^{2k}\} b_j + \sum_{j=-s}^s a_j j^{2k} = 0, \quad (2.7)$$

$$k = 0, 1, \dots, q.$$

**Proof.** Observe that

$$\begin{aligned} & u(t+\tau, x) + u(t-\tau, x) \\ &= 2\left[u(t, x) + \frac{\tau^2}{2!}u_{tt}(t, x) + \frac{\tau^4}{4!}u_{tttt}(t, x) + \dots\right] \\ &= 2\left[u(t, x) + \frac{c^2\tau^2}{2!}u_{xx}(t, x) + \frac{c^4\tau^4}{4!}u_{xxxx}(t, x) + \dots\right] \\ &= u(t, x+c\tau) + u(t, x-c\tau), \end{aligned} \quad (2.8)$$

using  $u_{tt} = c^2u_{xx}$ ,  $u_{tttt} = c^4u_{xxxx}$ , ... Therefore (2.4) can be rewritten in the form

$$\begin{aligned} & \sum_{j=-S}^S b_j [u(t, x+j\Delta x + \mu\Delta x) + u(t, x+j\Delta x - \mu\Delta x)] \\ &+ \sum_{j=-s}^s a_j u(t, x+j\Delta x) = C \frac{\partial^{p+2}}{\partial x^{p+2}} u(t, x) (\Delta x)^{p+2} + O((\Delta x)^{p+4}), (\Delta x \rightarrow 0), \end{aligned} \quad (2.9)$$

here noting that, because of the symmetry in the coefficients  $a_j, b_j$ , the odd powers of  $\Delta x$  in the Taylor expansion cancel. In (2.9)  $u$  is evaluated throughout at the same time level and therefore can be replaced by a function  $y$  of one variable:

$$\begin{aligned} & \sum_{j=-S}^S b_j \left\{ y(x + (j + \mu)\Delta x) + y(x + (j - \mu)\Delta x) \right\} \\ & + \sum_{j=-s}^s a_j y(x + j\Delta x) = C y^{p+2}(x) (\Delta x)^{p+2} + O((\Delta x)^{p+4}), (\Delta x \rightarrow 0). \end{aligned} \quad (2.10)$$

The function  $y(x)$  on the left hand side in (2.10) can now be expanded using Taylor series to give

$$\begin{aligned} & \sum_{j=-S}^S b_j \left\{ \sum_{k=0}^{\infty} \frac{y^k(x) (j + \mu)^k (\Delta x)^k}{k!} + \sum_{k=0}^{\infty} \frac{y^k(x) (j - \mu)^k (\Delta x)^k}{k!} \right\} \\ & + \sum_{j=-s}^s a_j \sum_{k=0}^{\infty} \frac{y^k(x) j^k (\Delta x)^k}{k!}, \end{aligned}$$

which, using the symmetry in the coefficients  $a_j, b_j$  and  $[(-j - \mu)^{2k+1} + (j + \mu)^{2k+1}] = 0$ , replaces (2.10) by

$$\begin{aligned} & \sum_{j=-S}^S b_j \sum_{k=0}^{\infty} \frac{\{(j + \mu)^{2k} + (j - \mu)^{2k}\} (\Delta x)^{2k} y^{2k}(x)}{(2k)!} \\ & + \sum_{j=-s}^s a_j \sum_{k=0}^{\infty} \frac{j^{2k} (\Delta x)^{2k} y^{2k}(x)}{(2k)!} \\ & = C y^{p+2}(x) (\Delta x)^{p+2} + O((\Delta x)^{p+4}), (\Delta x \rightarrow 0). \end{aligned}$$

Thus (1.2) is of order  $p = 2q$  if and only if (2.7) holds.  $\square$

**Corollary.** If the method (1.2) has order  $p = 2q$  then the error constant  $C$  is given by

$$C = \frac{1}{(p+2)!} \left[ \sum_{j=-S}^S \{(j + \mu)^{p+2} + (j - \mu)^{p+2}\} b_j + \sum_{j=-s}^s j^{p+2} a_j \right] \neq 0. \quad (2.11)$$

$\square$

**Proof.** (Theorem 2.1)

(i)  $\iff$  (ii). In (2.1) let  $(z, w)$  be replaced by  $(z, z^\mu)$  with  $z = e^v$ , so that for  $v \rightarrow 0, z \rightarrow 1$ .

Then

$$\begin{aligned}
\Phi(z, z^\mu) &= \sum_{j=-S}^S b_j z^j (z^\mu + z^{-\mu}) + \sum_{j=-s}^s a_j z^j \\
&= \sum_{j=-S}^S b_j (e^{(j+\mu)v} + e^{(j-\mu)v}) + \sum_{j=-s}^s a_j e^{jv} \\
&= \sum_{j=-S}^S b_j \left\{ \sum_{k=0}^{\infty} \frac{((j+\mu)^k + (j-\mu)^k) v^k}{k!} \right\} + \sum_{j=-s}^s a_j \sum_{k=0}^{\infty} \frac{j^k v^k}{k!}.
\end{aligned}$$

But then because of the symmetry in the coefficients  $b_j = b_{-j}$ ,  $a_j = a_{-j}$  the odd powered terms cancel yielding

$$\begin{aligned}
\Phi(z, z^\mu) &= \sum_{k=0}^{\infty} \left\{ \sum_{j=-S}^S b_j \{((j+\mu)^{2k} + (j-\mu)^{2k})\} + \sum_{j=-s}^s j^{2k} a_j \right\} \frac{v^{2k}}{(2k)!} \\
&= \sum_{k=0}^{\infty} \left\{ \sum_{j=-S}^S b_j \{(j+\mu)^{2k} + (j-\mu)^{2k}\} + \sum_{j=-s}^s j^{2k} a_j \right\} \frac{(z-1)^{2k}}{(2k)!}.
\end{aligned}$$

Hence from Lemma (2.1) we conclude that method (1.2) is of order  $p = 2q$  if and only if

$$\Phi(z, z^\mu) = C(z-1)^{p+2} + O((z-1)^{p+4}), (z \rightarrow 1).$$

(ii)  $\iff$  (iii). Suppose that  $w = z^\mu + D(z-1)^\alpha + E(z-1)^{\alpha+1} + O((z-1)^{\alpha+2})$ , then

$$\begin{aligned}
w^{-1} &= \frac{1}{z^\mu + D(z-1)^\alpha + E(z-1)^{\alpha+1} + O((z-1)^{\alpha+2})} \\
&= \frac{z^{-\mu}}{1 + z^{-\mu} D(z-1)^\alpha + E(z-1)^{\alpha+1} z^{-\mu} + O((z-1)^{\alpha+2})} \\
&= z^{-\mu} \left( 1 - z^{-\mu} D(z-1)^\alpha - E(z-1)^{\alpha+1} z^{-\mu} + O((z-1)^{\alpha+2}) \right) \\
&= z^{-\mu} - z^{-2\mu} D(z-1)^\alpha - E z^{-2\mu} (z-1)^{\alpha+1} + O((z-1)^{\alpha+2}) \\
&= z^{-\mu} - D(z-1)^\alpha (1 - 2\mu(z-1)) - E(z-1)^{\alpha+1} + O((z-1)^{\alpha+2}),
\end{aligned}$$

where we have used  $z^k = (1+z-1)^k = 1 + k(z-1) + O((z-1)^2)$ . Thus

$$\begin{aligned}
\Phi(z, w(z)) &= (w + w^{-1})b(z) + a(z) \\
&= \left( z^\mu + D(z-1)^\alpha + E(z-1)^{\alpha+1} + O((z-1)^{\alpha+2}) \right. \\
&\quad \left. + z^{-\mu} - D(z-1)^\alpha - E(z-1)^{\alpha+1} + 2\mu D(z-1)^{\alpha+1} + O((z-1)^{\alpha+2}) \right) b(z) + a(z) \\
&= \Phi(z, z^\mu) + 2\mu D(z-1)^{\alpha+1} b(z) + O((z-1)^{\alpha+2}).
\end{aligned}$$

Let the pair  $(z, w(z))$  be such that  $\Phi(z, w(z)) = 0$  then  $\Phi(z, z^\mu) = -2\mu D(z-1)^{\alpha+1}b(z) + O((z-1)^{\alpha+2})$ . Hence  $\Phi(z, z^\mu) = C(z-1)^{p+2} + O((z-1)^{p+3})$ ,  $(z \rightarrow 1)$  if and only if  $\alpha = p+1$  and (2.6) holds. To show now that (2.6) holds for a unique branch of  $\Phi(z, w) = 0$  we observe that by consistency  $\Phi(1, w) = 0$  has two branches which satisfy  $w_\pm(1) = 1$ . But for these branches  $w_+w_- = 1$  also holds and hence only one can satisfy (2.6). Further  $C = -2\mu Db(1)$ , where now we see from (2.6) that  $D = -\tilde{C}$  so that  $\tilde{C} = \frac{C}{2\mu b(1)}$ , and we have used  $b(z) = b(1) + O((z-1)^2)$ ,  $(z \rightarrow 1)$ .  $\square$

**Remark 2.3.** If  $b(z) \neq 0$  (2.1) can be written in the form

$$\Phi_1(z, w) = w^2 + \frac{a(z)}{b(z)}w + 1. \quad (2.12)$$

Therefore by introducing the rational function  $h(z) := -\frac{a(z)}{2b(z)}$ , where for convenience the explicit dependance on  $\mu$  is not expressed, (2.12) can be further written in the form

$$\Phi_1(z, w) = w^2 - 2h(z)w + 1, \quad (2.13)$$

from which it is easily seen that the two roots  $w_\pm(z)$  of  $\Phi_1(z, w) = 0$  are given by

$$w_\pm(z) = h(z) \pm \sqrt{h(z)^2 - 1}. \quad (2.14)$$

The stability requirement can thus be stated in terms of the function  $h(e^{i\theta})$ , which because of the symmetry in  $a_j, b_j$  is real, as the requirement

$$-1 \leq h(e^{i\theta}) \leq 1. \quad (2.15)$$

Furthermore, consistency (2.3) imposes  $h(1) = 1$  and from statement (ii) Theorem 2.1 order of accuracy  $p = 2q$  translates to the expansion

$$h(z, \mu) = \frac{z^\mu + z^{-\mu}}{2} - \frac{C}{2b(1)}(z-1)^{p+2} + O((z-1)^{p+3}), (z \rightarrow 1).$$

Note that here it is convenient to indicate the dependance of  $h$  on  $\mu$ . With the introduction of the variable  $v$ , via  $z = e^v$ , we see that

$$h(e^v, \mu) = \cosh \mu v + C^* v^{p+2} + O((v)^{p+3}), (v \rightarrow 0), \quad (2.16)$$

where  $C^* = -C/2b(1)$ , and

$$h(e^{i\theta}, \mu) = \cos \mu \theta + O((\theta)^{p+2}), (\theta \rightarrow 0). \quad (2.17)$$

Hence the normalisation of  $h$  by the factor  $-\frac{1}{2}$  is explained by the form of the right hand sides in (2.16) and (2.17).

### 3. Properties of the algebraic function $w$

The algebraic function  $w$ , which satisfies  $\Phi(z, w(z)) = 0$ , or rather

$$w\Phi_1(z, w) = w^2 - 2h(z)w + 1 = 0 \quad (3.1)$$

is multiple valued, consisting in general for a given  $z$  of two values  $w_{\pm}(z)$ , from (2.14). Associated with  $w\Phi_1(z, w) = 0$  there is, therefore, a two-sheeted Riemann surface  $M$ , i.e.

$$M = \{(z, w) \in \bar{\mathbb{C}} \times \bar{\mathbb{C}} : w\Phi_1(z, w) = 0\} .$$

On  $M$  the algebraic function  $w$  is again single valued and the two sheets of  $M$  interact only at branch points  $z$ , where the functions  $w_{\pm}$  coincide. It is convenient to refer to the portion on  $M$  with  $|z| < 1$  as the unit disc  $\Delta$ , the set with  $|z| > 1$  as the outside of the unit disc, and the portion with  $|z| = 1$  as the unit circle. Strictly speaking, it is rather the projections of these sets onto the  $z$ -plane which we are actually referring to.

#### 3.1 Branch points of $w$

Branch points of  $w$  can occur where the two roots of  $w\Phi_1(z, w)$  are equal. By Remark 2.1 branch points therefore occur only if  $w_{\pm} = \pm 1$  and, by Remark 2.3 equation (2.14), at these places

$$h(z) = +1 \quad \text{or} \quad h(z) = -1 . \quad (3.2)$$

#### Remark 3.1

(i) The symmetry in the coefficients  $a_j, b_j$  implies that  $h(z) = h(1/z)$ . Therefore there are exactly the same number of branch points inside and outside the unit circle.

(ii) Although  $z = 1$  is a double root of the equation  $h(z) = 1$  the point  $z = 1$  is not a branch point of  $M$  because from Theorem 2.1

$$w_{\pm}(z) = 1 \pm \mu(z - 1) + O((z - 1)^2), (z \rightarrow 1) . \quad (3.3)$$

(iii) Because the coefficients of  $h(z)$  are real  $\overline{h(z)} = h(\bar{z})$ , where  $\bar{z}$  is the complex conjugate of  $z$ . Therefore the branch points of  $M$  are either real or they occur in complex conjugate pairs. Thus branch cuts of  $M$  can be taken to fall on the real axis or to be perpendicular and symmetric with respect to the real axis. Sometimes, however, it is simpler to “picture” some cuts as pairs of conjugate cuts.

(iv) If a stable scheme has a branch point on the unit circle at  $z = e^{i\theta_0}$ ,  $\frac{\partial h(e^{i\theta}, \mu)}{\partial \theta} |_{\theta=\theta_0} = 0$  and  $\theta_0$  is either a maximum or minimum point of the real function  $h(e^{i\theta}, \mu)$ .

#### 3.2 Poles of $w$

The poles of  $w$  occur where  $b(z) = 0$  and because  $b(z) = b(1/z)$  there are equal numbers of poles inside and outside  $\Delta$  away from  $z = 0$ . Furthermore, by invertibility, this number is  $S$ , and if  $S < s$  there will also be a pole at  $z = 0$ .

### 3.3 Zeros of $w$

Because the two roots  $w_{\pm}$  of  $w\Phi_1$  satisfy  $w_+ = \frac{1}{w_-}$  the algebraic function  $w$  will have zeros (of the corresponding multiplicities) at exactly the same  $z$ -values where it has poles, i.e. a pole on one sheet and a zero on the other for the same value of  $z$ .

### 3.4 Behaviour at $z = 0$ .

If  $S \geq s$ ,  $h(z)$  has the form

$$h(z) = z^{S-s} \frac{a^*(z)}{b^*(z)},$$

where  $a^*(z)$  and  $b^*(z)$  are polynomials of degree  $2s$  and  $2S$ , respectively, for which  $b^*(0) \neq 0$ . Therefore the point  $z = 0$  is an ordinary point of  $w$  and  $w_{\pm}$  have the Taylor expansions

$$w_{\pm}(z) = c_{0\pm} + c_{1\pm}z + c_{2\pm}z^2 + \dots,$$

at  $z = 0$  on the two sheets of  $M$ . Furthermore, for  $S > s$ ,  $h(0) = 0$  so that  $c_{0\pm} = \pm i$  and  $|w_{\pm}(0)| = 1$ .

If  $S < s$ ,  $h(z) = \frac{a^*(z)}{z^{s-S}b^*(z)}$  and  $w$  has a pole of multiplicity  $s - S$  on one sheet and a zero of multiplicity  $s - S$  on the other sheet,

$$\begin{aligned} w_+(z) &= z^{S-s}(c_0 + c_1z + c_2z^2 + \dots) \\ w_-(z) &= z^{s-S}(d_0 + d_1z + d_2z^2 + \dots). \end{aligned} \tag{3.4}$$

**Remark 3.2.** In what follows the sheet of  $M$  on which the root  $w_+$  occurs when  $z = 1$ , i.e. the root with expansion  $w_+(z) = 1 + \mu(z - 1) + O((z - 1)^2)$  at  $z = 1$ , will be called the principal sheet. Because the Riemann surface is connected, this notion is basically a local property in the neighbourhood of  $z = 1$ . Further, we use the convention that the principle sheet will refer to that part of  $M$  which can be connected to  $(1, w_+)$  without crossing a branch cut. The remaining part of  $M$  will be called the secondary sheet.

## 4. Order stars

In [8] the order of accuracy of schemes (1.2) was investigated by defining an order star on a portion of the complex plane using the approximation  $h(e^v, \mu) \approx \cosh \mu v$  given by (2.16). It was determined in [8], however, that this order star did not supply any information about the stability of schemes (1.2). Here, to obtain this information, we consider an order star on the Riemann surface  $M$  of the algebraic function  $w$ . Define the function  $\phi$  by

$$\phi(z, w) = z^{-\mu}w(z), \quad (z, w) \in M$$

and consider the Riemann surface  $M$  partitioned according to

$$\begin{aligned} \Omega &= \{(z, w) \in M : |\phi(z, w)| > 1\}, \\ \Omega^c &= \{(z, w) \in M : |\phi(z, w)| < 1\} \end{aligned} \tag{4.1}$$

and

$$\partial = \{(z, w) \in M : |\phi(z, w)| = 1\}.$$

Then, although the function  $\phi$  is multivalued on  $M$  because of the factor  $z^{-\mu}$ , the order star, defined by the set  $\Omega$ , is again single-valued on  $M$ , because it is defined via the modulus of  $\phi$  which is single-valued. Because the coefficients  $a_j, b_j$  are real, it is easily verified that  $\Omega$  (and  $\Omega^c$  and  $\partial$ ) are symmetric with respect to the real axis.

This order star has been considered by Jeltsch and Smit in [7], where three-time-level difference schemes for the linear advection equation  $u_t = cu_x$  were investigated. The properties of that order star also apply in the present situation, but because of the extra symmetry in the form of (1.2), compared with the schemes in [7], the order star here is considerably simplified. Furthermore, we will see that, whereas in [7] the geometry of  $\Omega$  dictates the order of the scheme, here it is simpler to look at  $\Omega^c$ . The results which coincide with those in [7] are presented without proof.

**Lemma 4.1.** “Stability”.

The order star  $\Omega$  of a stable scheme has the property

$$\Omega \cap \{(z, w) \in M : |z| = 1\} = \emptyset.$$

Furthermore  $\Omega^c \cap \{(z, w) \in M : |z| = 1\} = \emptyset$ . Equivalently  $\{(z, w) \in M : |z| = 1\} \subset \partial$ .  $\square$

**Lemma 4.2.** “Order of accuracy”.

A scheme (1.2) has order  $p$  if and only if at the point  $z = 1$  on the principal sheet of  $M$ ,  $\Omega$  consists of  $(p + 1)$  sectors of angle  $\pi/(p + 1)$ , separated by  $(p + 1)$  sectors of  $\Omega^c$ , each of the same angle.  $\square$

A subset  $\Omega_1$  (with boundary  $\partial\Omega$ ) of  $\Omega(\Omega^c)$  is said to be an  $\Omega(\Omega^c)$ -component if  $\partial\Omega_1 \subset \partial\Omega(\partial\Omega^c)$  and  $\Omega_1$  is connected. Further, an  $\Omega(\Omega^c)$ -component is said to have multiplicity  $m$  if it contains  $m$   $\Omega(\Omega^c)$ -sectors at  $z = 1$  on the principal sheet.

By Lemma 4.1 it is clear that there is a distinction between the portion of  $\Omega$  inside and the portion outside the unit disc  $\Delta$ . Defining a component  $\Omega_1$  (of  $\Omega$  or  $\Omega^c$ ) as bounded if  $\sup\{|z| : (z, w) \in \Omega_1\} < \infty$  it is then sufficient to consider only components of  $\Omega(\Omega^c)$  which are bounded, otherwise (1.2) is unstable.

Note that the symmetry of (1.2) with respect to  $\mu$  means that it is sufficient to consider  $\mu > 0$  alone. Thus in what follows  $\mu > 0$  is assumed.

**Lemma 4.3.** Inside  $\Delta$  the branch points of  $w$  occur inside  $\Omega$ -components while outside  $\Delta$  the branch points of  $w$  occur inside  $\Omega^c$ -components. On  $|z| = 1$  the branch points lie in  $\partial$ .

**Proof.** From Section 3.1  $|w_{\pm}(z_i)| = 1$  at any branch point  $z_i$ . Hence

$$|\phi(z_i, w)| \begin{cases} > \\ = \\ < \end{cases} 1 \quad \text{if} \quad \begin{cases} |z_i| < 1 \\ |z_i| = 1 \\ |z_i| > 1. \end{cases}$$

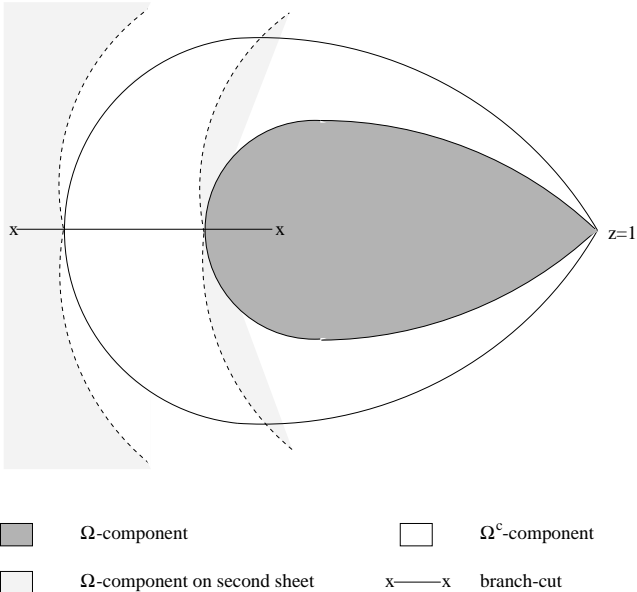
□

**Lemma 4.4.** For every  $z$  inside  $\Delta$  at least one of the points  $(z, w_+)$ ,  $(z, w_-)$  belongs to  $\Omega$ .

**Proof.** When  $b(z) = 0$  by (3.3) there is a pole on one sheet and a zero on the other, and thus the sheet with the pole belongs to  $\Omega$  at the pole. For  $b(z) \neq 0$  we have from (2.13) that  $w_+w_- = 1$  and hence for at least one of the points  $(z, w_+)$ ,  $(z, w_-)$

$$|\phi(z, w)| = |z^{-\mu}w| > 1 \quad \text{for} \quad |z| < 1. \quad \square$$

These two lemmas are crucial in the situation here compared with that in [7] for the advection equation because they induce a considerable simplification on the structure of  $\Omega^c$ . By lemma 4.3 it is clear that no  $\Omega^c$ -component inside  $\Delta$  can contain a branch cut of  $M$ . Therefore the only way that such a component can exist on more than one sheet of  $M$  is if it meets a branch cut of  $M$ . Suppose that it meets a cut on the real axis inside  $\Delta$ . Then by the symmetry of the order star both of the points  $(z, w_+)$  and  $(z, w_-)$  are contained in  $\Omega^c$  for a set of  $z$  values near the branch cut. This contradicts Lemma 4.4 and hence cannot happen, see Figure 4.1.



**Figure 4.1:**  
 $\Omega^c$ -component crossing a branch cut on the negative real axis

Furthermore, inside  $\Delta$ ,  $\partial$  can only intersect with a branch cut on the real axis at  $z = -1$ .

This structure contrasts with that in [7] and [5] where it is demonstrated that particularly the situation of the kind in Figure 4.1 leads to the maximum multiplicity of  $\Omega$ -components. Thus this maximum cannot be attained for schemes of type (1.2). In the next paragraphs we investigate the multiplicities that can occur and, because of the simplicity of the  $\Omega^c$ -components inside  $\Delta$ , we consider only  $\Omega^c$ -components.

## 4.2 Classification of $\Omega^c$ -components inside $\Delta$ .

It is well known in the order star literature, ([14,3,12,11]), that the multiplicity of a bounded  $\Omega$  (or  $\Omega^c$ ) component is related to the total multiplicity of the “poles” (“zeros”) inside the component. For the order star here, however, the multiplicities of the poles and zeros do not provide all the information needed to determine the multiplicity of the component. Hence it is appropriate to distinguish different classes of bounded components, the different properties arising due to the factor  $z^{-\mu}$  in the definition of  $\phi$ .

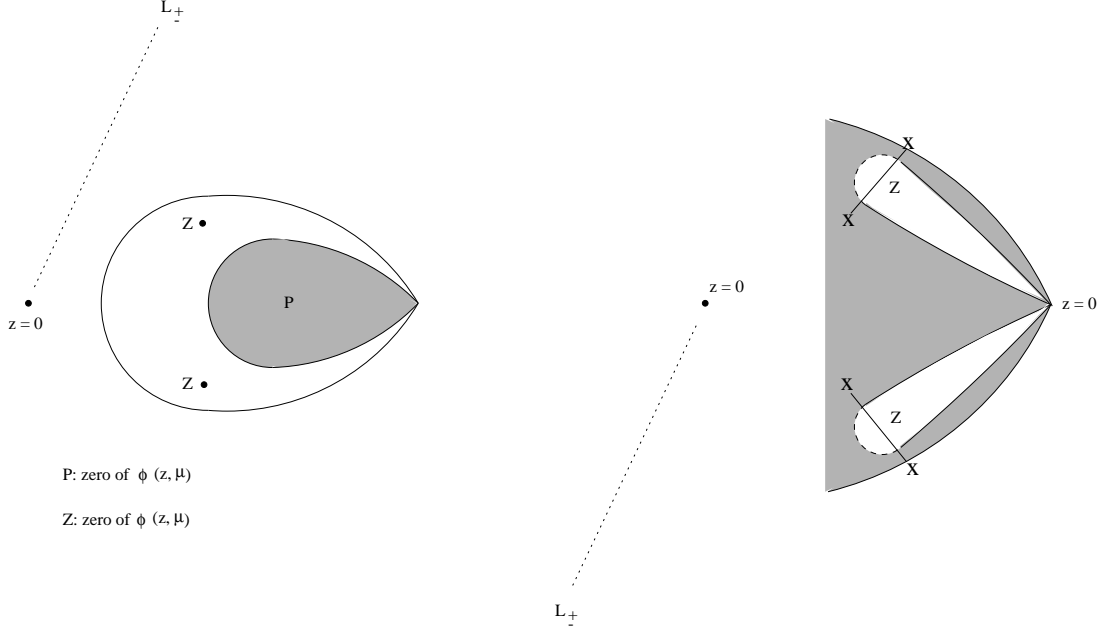
In order to establish the multiplicity of a given component the argument principle is applied with respect to the function  $\phi$  by integration along the boundary of the component. This cannot be carried out unless  $z^\mu$  has been defined uniquely on  $M$ . To achieve this branch cuts  $L_\pm$  emanating from the zero points  $(0, w_\pm(0))$  to the infinity points  $(\infty, w_\pm^\infty)$  are introduced, one on each sheet of  $M$ . These cuts are made according to the following rules, which are made to ensure that  $\log z$  and hence  $z^\mu$  are defined uniquely on  $M$ , (see [7]).

*Rule 1:* The cuts  $L_\pm$  have to be such that their projections onto the  $z$ -plane are identical, or “enclose” a sector of  $C$  which does not contain a branch cut of  $M$ .

*Rule 2:* The cuts  $L_\pm$  have to be such that each cut occurs on only one sheet of  $M$ , i.e.  $L_\pm$  do not cross branch cuts of  $M$ .

By adhering to rule 1,  $z^{-\mu}$  is defined uniquely on  $M$  even if the cuts  $L_\pm$  are allowed to cross branch cuts of  $M$ . It is, however, convenient to avoid this latter possibility and hence Rule 2 is also imposed. These two rules can then always be satisfied if the branch cuts  $L_\pm$  go along radial lines from the zero — to the infinity — points, such that their projections onto the  $z$ -plane are the same and do not pass through  $z = 1$ .

It is clear that for any bounded component  $\Omega_1$ , where  $\Omega_1$  is a component of either the order star or its complement, which does not either contain a branch cut of  $M$ , or “cross” a branch cut of  $M$ , and does not intersect  $L_\pm$  at any point, the multiplicity of  $\Omega_1$  is equal to the total multiplicities of the poles, respectively zeros, of  $\phi$  inside  $\Omega_1$ . Furthermore, even if the former does occur, so that some portion of  $\Omega_1$  exists on the secondary sheet, but still without intersecting  $L_\pm$  the multiplicity is still equal to the total multiplicities of the poles, respectively zeros, of  $\phi$  inside  $\Omega$ , see Figure 4.2.



**Figure 4.2:**

*$\Omega^c$ -components not involving the zero points*

Thus these components do not differ in any way to those normally exhibited by an order star defined on  $\mathbb{C}$ .

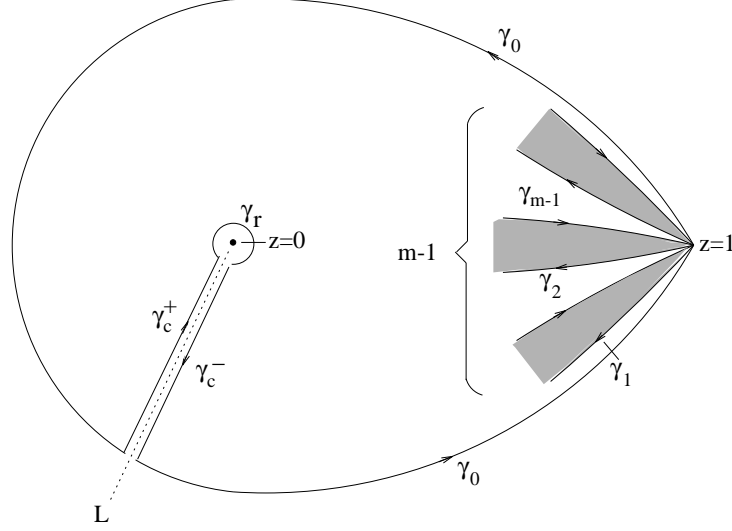
The second option to consider is a component which contains one of the zero points  $(0, w_{\pm}(0))$ . Because we are interested only in  $\Omega^c$ -components we know by Lemma 4.4 that no  $\Omega^c$ -component inside  $\Delta$  can contain both zero points, and hence it is really sufficient to assume that the component contains only one zero point, which can be on either sheet of  $M$ . Further, because the component contains a zero point, one of the cuts  $L_{\pm}$  must intersect the boundary  $\partial\Omega^c$  of the component and the curve  $\gamma$  along which the integration is carried out has to be extended to exclude the zero point from the integration so that  $\phi$  is well defined inside  $\gamma$ .

**Proposition 4.1.** Suppose that  $\Omega_1^c$  is a bounded component containing one zero point  $(0, w_+(0))$  or  $(0, w_-(0))$  and that the leading exponent of  $\phi$  at  $z = 0$  is  $\alpha$ . Furthermore, assume that  $\Omega_1^c$  contains  $Z$  zeros of  $\phi$  away from  $z = 0$ . Then the multiplicity  $m$  of  $\Omega_1^c$  satisfies

$$m \leq [\alpha] + 1 + Z,$$

where  $[\alpha]$  is defined to be the largest integer smaller than  $\alpha$ .

**Proof.** If  $\Omega_1^c$  is of multiplicity  $m$  then there are  $m - 1$   $\Omega$ -components emerging from  $z = 1$  on the principal sheet lying inside  $\Omega_1^c$ , for which we can assume there is no intersection of their boundaries with the cuts  $L_{\pm}$ . We denote the paths around these components by  $\gamma_i$ ,  $i = 1 : m - 1$ , see Figure 4.3. As can be seen from the figure the curve  $\gamma$  is then comprised of the union of the  $\gamma_i$  with the  $\gamma_0$  around  $\Omega_1^c$ , the paths  $\gamma_c^+$ ,  $\gamma_c^-$  along either side of the cut  $L$  (where  $L = L_+$  or  $L = L_-$ ) and the circular curve  $\gamma_r$  around  $z = 0$  centred at  $z = 0$ .



**Figure 4.3:**  
*Simple  $\Omega^c$ -component*

The proof follows by application of the argument principle for the integral of  $\frac{\phi'}{\phi}$  around  $\gamma$ . Because there are  $Z$  zeros inside  $\gamma$  we have

$$\begin{aligned} Z &= \frac{1}{2\pi i} \int_{\gamma} \frac{\phi'(z, w)}{\phi(z, w)} dz \\ &= \frac{1}{2\pi i} \left[ \int_{\gamma_0} \frac{\phi'(z, w)}{\phi(z, w)} dz + \int_{\gamma_c^+} \frac{\phi'(z, w)}{\phi(z, w)} dz \right. \\ &\quad \left. + \int_{\gamma_c^-} \frac{\phi'(z, w)}{\phi(z, w)} dz + \int_{\gamma_r} \frac{\phi'(z, w)}{\phi(z, w)} dz + \sum_{j=1}^{m-1} \int_{\gamma_j} \frac{\phi'(z, w)}{\phi(z, w)} dz \right]. \end{aligned}$$

By Proposition 4.3 [7] and the assumption that  $\phi(z, w) \approx z^\alpha$  at  $z = 0$

$$\frac{1}{2\pi i} \int_{\gamma_r} \frac{\phi'(z, w)}{\phi(z, w)} dz = -\alpha .$$

The integrals  $\int_{\gamma_c^+} \frac{\phi'(z, w)}{\phi(z, w)} dz$ , and  $\int_{\gamma_c^-} \frac{\phi'(z, w)}{\phi(z, w)} dz$  cancel because by Proposition 4.4 [7]

$$\int_{\gamma_c^+} \frac{\phi'(z, w)}{\phi(z, w)} dz = - \int_{\gamma_c^-} \frac{\phi'(z, w)}{\phi(z, w)} dz .$$

Furthermore, we also know that  $\arg \phi$  decreases monotonically along the boundary of an  $\Omega$  region and increases monotonically along the boundary of an  $\Omega^c$  region, when the path taken is oriented in the positive direction. Hence

$$\frac{1}{2\pi i} \int_{\gamma_j} \frac{\phi'(z, w)}{\phi(z, w)} dz \geq 1 \quad \text{and} \quad \frac{1}{2\pi i} \int_{\gamma} \frac{\phi'(z, w)}{\phi(z, w)} dz > 0 .$$

Therefore

$$Z > -\alpha + (m - 1)$$

and

$$m < Z + \alpha + 1 .$$

But  $m$  is integer and hence

$$m \leq [\alpha] + Z + 1 .$$

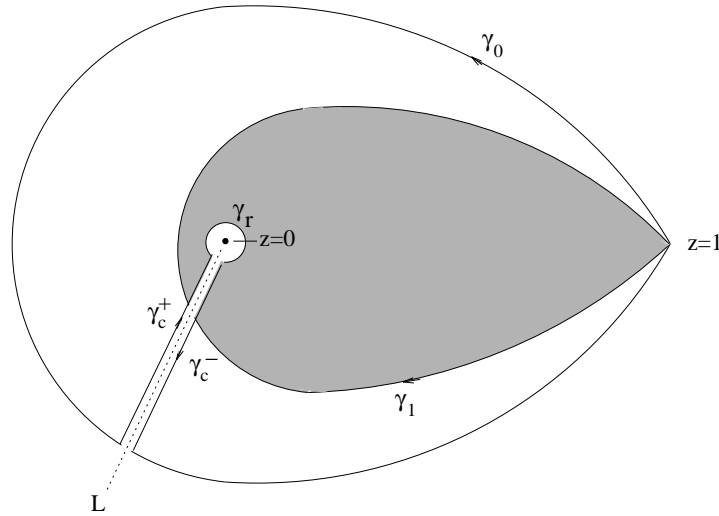
□

Note that this result is a direct corollary of Proposition 4.5 in [7].

Finally we consider the situation in which a zero of the function  $\phi$  can contribute to more than one sector of  $\Omega^c$  approaching the point  $z = 1$ . This can only happen if the zero points  $(0, w_{\pm}(0))$  are not contained in the component  $\Omega_1^c$ .

**Proposition 4.2.** Suppose that  $\Omega_1^c$  is a bounded component containing neither of the zero points  $(0, w_{\pm}(0))$ , but that  $\Omega_1^c$  surrounds one of the zero points in the sense that  $\Omega_1^c$  intersects the negative real axis. Then  $\Omega_1^c$  has multiplicity 2 and contains at least one zero of  $\phi$ , away from  $z = 0$ .

**Proof.** We consider a component of the type in Figure 4.4.



**Figure 4.4:**  
 $\Omega^c$ -component of multiplicity 2

The integration is performed along the paths  $\gamma_a = \gamma_0 \cup \gamma_c^+ \cup \gamma_c^- \cup \gamma_r$ , and  $\gamma_b = \gamma_1 \cup (-\gamma_c^-) \cup (-\gamma_c^+) \cup (-\gamma_r)$ . The path  $\gamma = \gamma_a \cup \gamma_b$  goes around the complete boundary of  $\Omega_1^c$ . Note that for  $\gamma_b$  we have written  $-\gamma_c^-$ ,  $-\gamma_c^+$  and  $-\gamma_r$  because the direction is opposite to that taken by  $\gamma_a$ . Also strictly speaking on  $\gamma_b$  the path taken along  $\gamma_c^+$ ,  $\gamma_c^-$  is shorter than that in  $\gamma_a$ , but because the two integrals will cancel, the length of  $\gamma_c^+$ ,  $\gamma_c^-$  is unimportant. Thus

$$\begin{aligned} \frac{1}{2\pi i} \int_{\gamma_a \cup \gamma_b} \frac{\phi'(z, w)}{\phi(z, w)} dz &= \frac{1}{2\pi i} \left[ \int_{\gamma_0} \frac{\phi'(z, w)}{\phi(z, w)} dz \right. \\ &\quad \left. + \int_{\gamma_1} \frac{\phi'(z, w)}{\phi(z, w)} dz + \int_{\gamma_r} \frac{\phi'(z, w)}{\phi(z, w)} dz + \int_{-\gamma_r} \frac{\phi'(z, w)}{\phi(z, w)} dz \right] . \end{aligned}$$

Suppose that the leading exponent of  $\phi$  at  $z = 0$  is  $-\alpha$  then by Proposition 4.3 [7],

$$\int_{\gamma_r} \frac{\phi'(z, w)}{\phi(z, w)} dz + \int_{-\gamma_r} \frac{\phi'(z, w)}{\phi(z, w)} dz = \alpha - \alpha = 0.$$

Therefore,

$$\frac{1}{2\pi i} \int_{\gamma} \frac{\phi'(z, w)}{\phi(z, w)} dz = \frac{1}{2\pi i} \sum_{j=0}^1 \int_{\gamma_j} \frac{\phi'(z, w)}{\phi(z, w)} dz.$$

But now, because  $\arg \phi$  increases monotonically along  $\partial\Omega_1^c$  and because after traversing both  $\gamma_0$  and  $\gamma_1$  the path has returned to the point where it started, we must have

$$\frac{1}{2\pi i} \sum_{j=0}^1 \int_{\gamma_j} \frac{\phi'(z, w)}{\phi(z, w)} dz \geq 1.$$

Note that after  $\gamma_a$  has been traversed the path has returned to the point  $z = 1$  but on the other side of the cut  $L$  and hence  $\arg \phi$  need not have increased by an integer multiple of  $2\pi i$ . Therefore,  $\Omega_1^c$  has multiplicity 2 and  $Z$ , the number of zeros inside  $\Omega_1^c$ , satisfies  $Z \geq 1$ .  $\square$

**Remark 4.1.** (i) The situation depicted in Figure 4.4 is given the name “binary loop” in [7], because the component has 2 sectors at  $z = 1$  for a minimum of one zero in its interior.

(ii) If there is just one zero inside the component then it must lie on the negative real axis, otherwise it has a complex conjugate zero, by symmetry also inside the component, and then  $\frac{1}{2\pi i} \int_{\gamma} \frac{\phi'(z, w)}{\phi(z, w)} dz = 2$ .

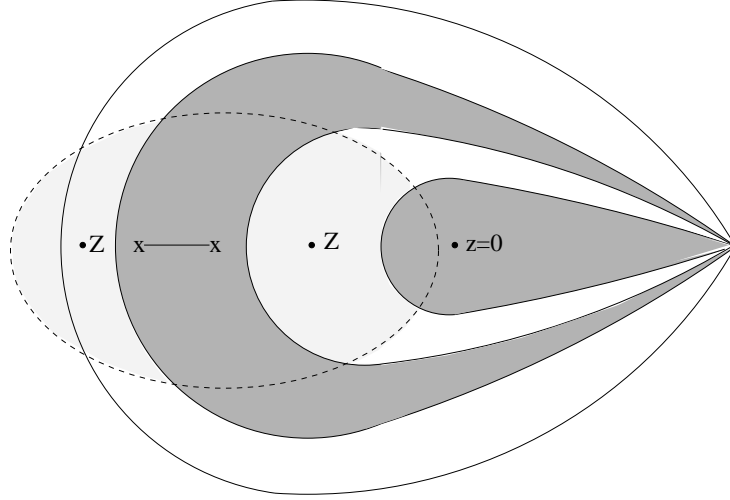
**Remark 4.2.** In the proof of Proposition 4.2 we did not assume, as in 4.1, that there could be many sectors of the component at  $z = 1$ . Suppose that we did make this assumption then either they would also be binary loops, or simple loops not crossing the real axis as in Figure 4.3. In the latter case they would each correspond to an increase in the argument of  $\phi$  of at least  $2\pi i$ , and would then be seen to require extra zeros inside  $\Omega_1^c$ , placed “inefficiently”. Otherwise, as binary loops, they would not be part of the same component  $\Omega_1^c$  because by Lemma 4.3 they could not be accessed by branch cuts within  $\Omega_1^c$ .

**Definition.** A zero of  $w(z, \mu)$  is said to be “efficient” if it lies on the negative real axis inside an  $\Omega^c$ -component on either sheet of  $M$ .

**Corollary 4.3.** An  $\Omega_1^c$ -component which contains neither zero point  $(0, w_{\pm}(0))$  has maximum multiplicity 2.  $\square$

**Remark 4.3.** A component  $\Omega_1^c$  for which Corollary 4.3 is valid will be called an “efficient” component. Furthermore, if we have two efficient components  $\Omega_1^c$  and  $\Omega_2^c$  then

between them there must be a component of the order star which also has multiplicity at least 2. The most efficient way for the multiplicity of the order star component to be attained is via a cut to the other sheet so that the point which acts as a zero of  $w(z, \mu)$  on one sheet acts as a pole of  $w(z, \mu)$  on the other sheet. Thus an efficient zero of  $w(z, \mu)$  contributes not only to the multiplicity of an  $\Omega^c$ -component but also of an  $\Omega$ -component, see Figure 4.5, via a branch cut.



**Figure 4.5:**  
*Efficient  $\Omega^c$ -component*

## 5. The maximal order of stable schemes

In section 4 we have outlined the structure of  $\Omega^c$ -components inside the unit disc. This is now used to prove the following result.

**Proposition 5.1.** Suppose we have a stable scheme (1.2) with  $S \leq s$  and  $0 < \mu \leq 1$ . Then the order  $p$  of the scheme is bounded by

$$p \leq 2(s + S).$$

Further, for  $s > S$  a stable scheme cannot achieve the maximum order  $p = 2(s + S)$  when  $\mu > 1$ .

**Proof.** The maximum order of a scheme is achieved when, by Corollary 4.3, all the zeros of  $w(z, \mu)$  away from  $z = 0$  are efficient. These zeros correspond to zeros of  $\phi$  with exactly the same orders as the corresponding zeros of  $w$ , which implies that inside  $\Delta$  their total order is  $S$ . Thus the maximum number of sectors of  $\Omega^c$  approaching  $z = 0$  due to these zeros is  $2S$ , and assumes that at least one of the zero points  $(0, w_{\pm}(0))$  lies inside an  $\Omega$ -component.

At  $z = 0$  we have, when  $S = s$ , ordinary points of  $w$ , so that  $\phi$  has expansion with leading term  $z^{-\mu}$  at  $z = 0$  on both sheets of  $M$ , and hence both zero points lie inside the order star and do not contribute to sectors of  $\Omega^c$  at  $z = 1$ .

When  $S < s$ , by (3.4),  $\phi$  has expansions

$$\begin{aligned} \phi(z, w_+) &= z^{S-s-\mu}(c_0 + c_1z + \cdots) \\ \text{and} & \\ \phi(z, w_-) &= z^{s-S-\mu}(d_0 + d_1z + \cdots), \end{aligned} \tag{5.1}$$

i.e., a “pole” of order  $S - s - \mu$  on one sheet and a zero, when  $s - S - \mu > 0$ , of order  $s - S - \mu$  on the second sheet. By Proposition 4.1 the  $\Omega^c$ -component containing the “zero” of order  $s - S - \mu$  has multiplicity bounded by

$$\lfloor s - S - \mu \rfloor + 1 = s - S \quad \text{for } 0 < \mu \leq 1. \tag{5.2}$$

The total number,  $m$ , of  $\Omega^c$ -sectors at  $z = 1$  from inside  $\Delta$  is thus bounded by

$$m \leq s - S + 2S \quad 0 < \mu \leq 1,$$

which implies

$$m \leq \begin{cases} s + S & \text{if } s > S \\ 2S = s + S & \text{if } s = S, \end{cases}$$

for  $0 < \mu \leq 1$ . Because by stability  $\partial\Delta$  is either the outer boundary of an  $\Omega$ -component or an  $\Omega^c$ -component inside  $\Delta$  there are either  $m + 1$  or  $m - 1$  sectors of  $\Omega^c$  reaching  $z = 1$  from outside  $\Delta$ . The maximum is achieved in the former case, and hence by Lemma 4.2

$$p + 1 \leq 2m + 1$$

implies  $p \leq 2(s + S)$ ,  $s = S$  or  $S < s$  and  $0 < \mu \leq 1$ . We see from (5.1) and hence (5.2) that when  $S < s$  and  $\mu > 1$  the order is necessarily reduced. Hence, schemes of the maximal order can only be stable for  $0 < \mu \leq 1$ . When  $\mu > 1$  the order is reduced.  $\square$

Note that in this proof it is essential to assume that the zero point lies in a component not containing zeros of  $\phi$ , away from  $z = 0$ . If the contrary were true then the zeros would be inefficient and the maximum order could not be achieved.

### 5.1 “More implicit” schemes, $S > s$ , and an interlace property.

From Theorem 4.1 it is easy to see that if we apply the same arguments to schemes which have  $S > s$  the maximum order  $p \leq 4S$  will result. Examples demonstrate that this overestimates  $p$ , particularly for  $S \gg s$ , and that actually the result should still be  $p \leq 2(s + S)$ . Thus, further information is needed to prove the correct bound, and hence limit the number of efficient zeros that can exist. In order to do this we use the rational function  $h(z)$ , originally introduced in Remark 2.3, and in terms of which the properties of  $w(z, \mu)$  were given in Section 3.

First we return to Remark 4.3 and Figure 4.5 from which it is clear that in order for zeros of  $w(z, \mu)$  to be efficient the order star has a particular structure, of the type in Figure 4.5. This means that if we just consider the negative real axis it consists of intervals

which lie inside  $\Omega$  separated by intervals of  $\Omega^c$  which each contain zeros of  $w(z, \mu)$ . Note that the intervals inside  $\Omega$  might also consist of a real branch cut, in order than a pole of  $w(z, \mu)$  is reached, but that this is not necessary because two or more conjugate branch points can also provide branch cuts to the other sheet. Furthermore, the “zeros” of  $w(z, \mu)$  of interest are actually poles of the rational function  $h(z, \mu)$ , and henceforth we will refer to the roots of  $b(z, \mu) = 0$ , away from  $z = 0$ , as poles of  $h(z, \mu)$ , which will be called “efficient” when they provide efficient zeros of  $w(z, \mu)$ .

In [8] it was postulated that between any two efficient poles of  $h(z, \mu)$  there must be a zero of  $h(z, \mu)$ , i.e., that the poles and zeros of  $h(z, \mu)$  interlace along the negative real axis, inside the unit circle. The proof of the statement for  $\mu = \frac{1}{2}$  is easily seen in [8] but the extension to other values of  $\mu$  was not explained. In fact, this interlace property can be proved via the order star in [8] but here it is more appropriate, and informative, to use the order star defined by (4.1). Furthermore, we will see that what we actually need is not just an interlace property but, more precisely, that associated with every efficient pole there is at least one zero of  $h(z, \mu)$ . The proof employs the argument principle and examines the change in the argument of  $\phi$  around an efficient  $\Omega^c$ -component. With this in mind we prove the following results:

**Lemma 5.1.** Suppose that an efficient  $\Omega^c$ -component,  $\Omega_1^c$ , inside  $\Delta$  intersects the negative real axis on the interval  $[x_1, x_2]$ , where  $-1 < x_1 < x_2 < 0$  and that  $b(x^*, \mu) = 0$  where  $x_1 < x^* < x_2$ . Then for  $0 < \mu \leq 1$ ,  $h(x_1) < 0$  and  $h(x_2) > 0$ .

**Proof.** Suppose that  $-1 < x_i < 0$  such that  $|\phi(x_i, w)| = 1$  for  $w = w_+$  or  $w_-$ . Because  $|x_i| < 1$ ,  $|x_i^{-\mu}| > 1$ ,  $|w| < 1$  and by (2.14) this means that  $|h(x_i)| > 1$ . Further, either  $x^*$  is a pole of even order of  $h$  and  $h$ , as a real function along the negative real axis, does not change sign on the interval  $[x_1, x_2]$ , or  $h$  changes sign. But if  $x^*$  is efficient as a pole of  $h(z, \mu)$  then it is simple and hence we have one of the two situations,

$$\begin{aligned} (i) \quad & h(x_1, \mu) > 1 > -1 > h(x_2, \mu) \\ (ii) \quad & h(x_1, \mu) < -1 < 1 < h(x_2, \mu) . \end{aligned} \tag{5.3}$$

Further, from (2.14)

$$\arg w(x_i) = \begin{cases} 0 & \text{if } h(x_i) > 0 \\ \pi & \text{if } h(x_i) < 0 \end{cases} \tag{5.4}$$

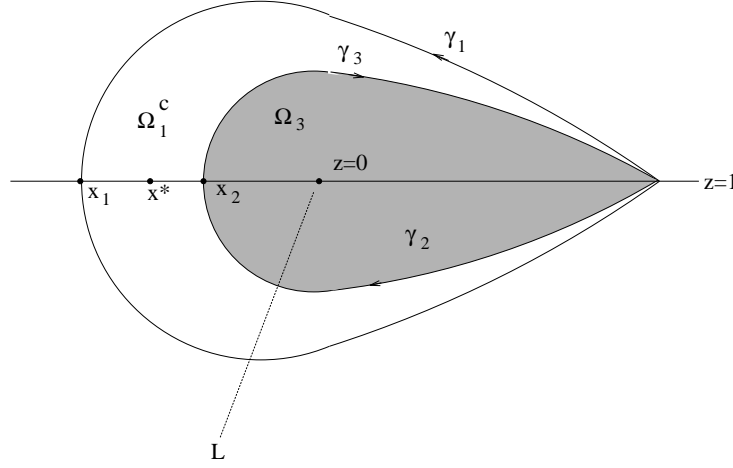
and by consistency (2.3)  $h(1) = 1$  so that

$$\arg w(z = 1) = 0 . \tag{5.5}$$

Now  $\arg \phi$  is monotonically increasing along the boundary of an  $\Omega^c$ -component traversed in the positively oriented direction. Thus the change in the argument of  $\phi$  along any portion of  $\partial\Omega^c$  is positive. In particular

$$\varliminf_{i=1,2,3} [\arg \phi] > 0 ,$$

where the  $\gamma_i$ ,  $i = 1, 2, 3$  are, as indicated in Figure 5.1, the portions of  $\partial\Omega^c$  traversed from  $z = 1$  to  $z = x_1$ ,  $z = x_1$  to  $z = x_2$ , and  $z = x_2$  to  $z = 1$ , respectively.



**Figure 5.1:**  
 *$\Omega^c$ -component crossing the negative real axis*

Here we can assume without loss of generality, that the cut  $L = L_+$  or  $L = L_-$  lies beneath the real axis. Further, if there are conjugate cuts so that the component  $\Omega_1^c$  crosses onto the lower sheet, the argument is unchanged at the cut. Because  $\phi = z^{-\mu}w(z, \mu)$

$$\arg \phi(z, \mu) = \arg(z^{-\mu}) + \arg(w(z, \mu)) , \quad (5.6)$$

and for  $z = re^{i\theta}$ ,  $\arg(z^{-\mu}) = -\theta\mu$ . Thus on any branch of the function  $z^{-\mu}$ ,  $-\theta\mu$  is monotonically decreasing as  $\theta$  increases. Therefore  $\arg w(z, \mu)$  must be monotonically increasing along  $\gamma_1$ , and thus by (5.3), (5.4) and (5.5)

$$\text{var}_{\gamma_1}[\arg w] = \begin{cases} 2k\pi \\ (2k-1)\pi \end{cases} \geq \begin{cases} 2\pi & \text{case (i)} \\ \pi & \text{case (ii)} \end{cases} , \quad (5.7)$$

where  $k \in \mathbb{N}$ .

Along the first portion of  $\gamma_2$ , up to  $z = 1$ ,  $\theta$  increases and then decreases by the same amount back to the point  $z = x_2$ . Although  $z^{-\mu}$  crosses to a different branch when  $\gamma_2$  meets  $L$ , suppose at  $z = re^{i\theta_0}$ , it crosses back again on the second portion of  $\gamma_2$  and, because  $L$  is taken to be radial from  $z = 0$ , it crosses back to the initial branch at a new point  $z = r_2e^{i\theta_0}$  with the same argument. For the portions of  $\gamma_2$  on the different branches, therefore, the change in the argument of  $z^{-\mu}$  is zero. Hence, also on this segment of  $\partial\Omega_1^c$  we must have  $\text{var}_{\gamma_2}[\arg w] > 0$ , and thus from (5.3) and (5.4)

$$\text{var}_{\gamma_2}[\arg w] \geq \pi . \quad (5.8)$$

Thus

$$\text{var}_{\gamma}[\arg \phi] = \sum_{i=1}^3 \text{var}_{\gamma_i}[\arg \phi]$$

$$\geq \begin{cases} (2\pi - \pi\mu) + \pi + \delta_1 & \text{case (i)} \\ (\pi - \pi\mu) + \pi + \delta_2 & \text{case (ii)} \end{cases}, \quad (5.9)$$

where  $\delta_j = \text{var}_{\gamma_3}[\arg \phi] > 0$ ,  $j = 1, 2$  for cases (i) and (ii), respectively. Therefore for  $0 < \mu \leq 1$

$$V = \text{var}_{\gamma}[\arg \phi] > \begin{cases} 2\pi & \text{case (i)} \\ \pi & \text{case (ii)} \end{cases}. \quad (5.10)$$

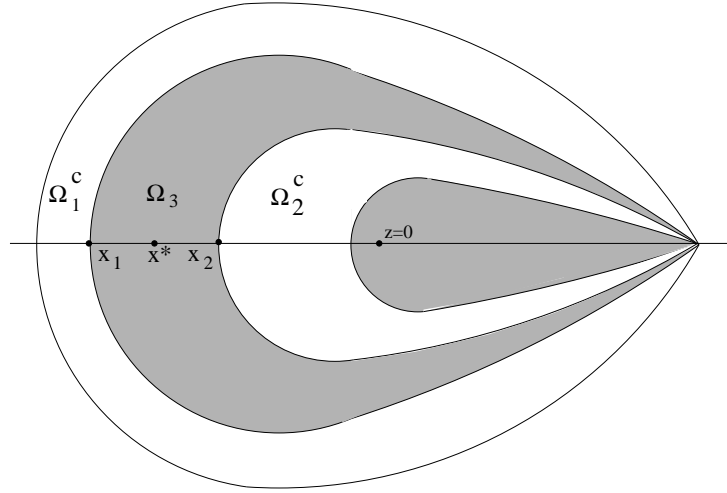
But by the argument principle  $\frac{V}{2\pi}$  is integer and hence

$$Z - P \geq \begin{cases} 2 & \text{case (i)} \\ 1 & \text{case (ii)} \end{cases},$$

where  $Z$  and  $P$  are the number of zeros, and poles, respectively, of  $\phi$  inside the component. But, because we assume efficiency of the component,  $Z = 1$ , and hence case (ii) must be the correct choice.  $\square$

**Corollary 5.1.** Suppose that  $0 < \mu \leq 1$  and that an efficient  $\Omega$ -component inside  $\Delta$  intersects the negative real axis on the interval  $[x_1, x_2]$ , where  $-1 \leq x_1 < x_2 < 0$  and either lies between two efficient  $\Omega^c$ -components, if  $x_1 \neq -1$ , or surrounds such a component if  $x_1 = -1$ . Then (i) if  $x_1 \neq -1$ ,  $h(x_1) > 0$ ,  $h(x_2) < 0$  and there exists at least one zero of  $h(z, \mu)$  at some point  $x^*$ , such that  $x_1 < x^* < x_2$  or (ii)  $x_1 = -1$  and either  $h(x^*, \mu) = 0$  for some  $-1 < x^* < x_2$  or  $h(e^{i\theta_0}, \mu) = 0$  for some  $0 < \theta_0 < \pi$ .

**Proof.** Consider the first situation as depicted in Figure 5.2.



**Figure 5.2:**  
*Efficient  $\Omega^c$ -component*

Here  $\Omega_1^c$  and  $\Omega_2^c$  are efficient components and so  $h(x_1, \mu) > 0$  and  $h(x_2, \mu) < 0$ . Therefore  $h$  changes sign on the interval  $[x_1, x_2]$  and because  $\Omega_3$  is efficient  $h$  can only change sign at a zero of  $h$ . Thus there exists  $x^*$  such that  $h(x^*, \mu) = 0$ ,  $x_1 < x^* < x_2$ .

In the second case  $\partial\Delta$  coincides with the outer boundary of  $\Omega_3$  and we know that  $h(x_2, \mu) < 0$  by Lemma 5.1. Further  $h(e^0) = 1 > 0$ , and  $h(e^{i\theta})$  is a real function, as is

$h(x, \mu)$ . Thus  $h$  changes sign either along the unit circle at some point  $e^{i\theta_0}$ , or  $h(-1) > 0$  and  $h(x^*, \mu) = 0$  for  $-1 < x^* < x_2$ .  $\square$

The implication of this corollary is not only that between any two efficient zeros of  $\phi(z, \mu)$ , zeros of  $b(z, \mu)$ , there is at least one zero of  $h(z, \mu)$ , i.e.  $a(x, \mu) = 0$ , but also that, if we have a stable scheme with  $\partial\Delta$  coinciding with the outer boundary of some component  $\Omega_1$ , and efficient  $\Omega^c$  components inside  $\Delta$ , then associated with the component  $\Omega_1$  is also at least one zero of  $a(z, \mu)$ . Furthermore, we know that for  $S > s$  the zero points lie inside  $\Omega$ -components so that each efficient  $\Omega^c$ -component is sandwiched between  $\Omega$ -components, as in Figure 5.2. Also, because  $a(z, \mu) = a(1/z, \mu)$  there are at most  $s$  zeros of  $a(z, \mu)$  with  $z = x$  and  $|x| < 1$ .

**Proposition 5.2.** For a stable scheme and  $0 < \mu < 1$  the number of efficient zeros,  $E$ , of  $\phi(z, \mu)$  is limited by

$$E \leq \min\{S, s\}, \quad (5.11)$$

when  $\partial\Delta$  coincides with the outer boundary of an  $\Omega$ -component and

$$E \leq \min\{S, s + 1\}, \quad (5.12)$$

when  $\partial\Delta$  coincides with the outer boundary of an  $\Omega^c$  component.  $\square$

In fact we can also prove in a manner similar to that in the proof of Lemma 5.1 that

$$E \leq \min\{S, s\}$$

in either case. But when an  $\Omega^c$ -component has outer boundary coincident with  $\partial\Delta$  it is sufficient to have (5.12) to obtain the desired result.

**Proposition 5.3.** Suppose we have a stable scheme (1.2) with  $S \geq s$ . Then for  $0 < \mu \leq 1$  the order  $p$  of the scheme is bounded by

$$p \leq 2(S + s). \quad \square$$

**Proof.** The proof follows that of Proposition 5.1 with the adjustment that the maximum number,  $m$ , of sectors of  $\Omega^c$  approaching  $z = 0$  inside  $\Delta$  due to zeros of  $b(z, \mu)$  is given by

$$m \leq 2E + (S - E), \quad (5.13)$$

where  $E$  is the number of zeros which are efficient. When  $\partial\Delta$  coincides with the outer boundary of an  $\Omega$ -component, by Proposition 5.2,  $E = s$  and

$$m \leq S + s,$$

so that  $p \leq 2(s + S)$  follows.

When  $\partial\Delta$  coincides with the outer boundary of an  $\Omega^c$ -component,  $E = s + 1$  and

$$m \leq S + s + 1. \quad (5.14)$$

But now there are only  $(m - 1)$  sectors of  $\Omega^c$  reaching  $z = 1$  from outside  $\Delta$  and hence

$$p + 1 \leq 2m - 1 \leq 2(S + s) + 1. \quad (5.15)$$

□

**Corollary 5.2.** For a stable scheme of maximum order with  $S \geq s$  and  $0 < \mu \leq 1$ ,  $\partial\Delta = \partial\Omega$ . □

**Proof.** This follows by the observation that (5.11) holds also when  $\partial\Delta$  coincides with the outer boundary of an  $\Omega$ -component. Hence (5.15) is replaced by  $p + 1 \leq 2m - 1 \leq 2(S + s) - 1$ , and maximum order cannot be achieved.

**Corollary 5.3.** For the schemes with  $S \geq s$ ,  $s \neq 0$ , the maximal order  $p = 2(s + S)$  cannot be achieved for  $\mu > 1$ . □

**Proof.** We look at the proof of lemma 5.1 and see that when  $1 < \mu \leq 2$  (5.7) becomes

$$\text{var}_{\gamma_1}[\arg \omega] = \begin{cases} 2k\pi \\ (2k + 1)\pi \end{cases} \geq \begin{cases} 2\pi & \text{case (i)} \\ 3\pi & \text{case (ii)} \end{cases},$$

if  $\text{var}_{\gamma_1}[\arg \phi] > 0$  is to be satisfied. Thus (5.10) is replaced by

$$V = \text{var}_{\gamma}[\arg \phi] > \begin{cases} 2\pi & \text{case (i)} \\ 2\pi & \text{case (ii)} \end{cases}$$

and hence  $Z \geq 2$ , which means that the zero of  $\phi$  is a double zero and so  $h$  does not change sign on the interval  $[x_1, x_2]$ . But then we actually have that

$$V \geq \begin{cases} (2\pi - \pi\mu) + 2\pi + \delta_1 & \text{if } h(x_1) > 0 \\ (3\pi - \pi\mu) + 2\pi + \delta_2 & \text{if } h(x_1) < 0 \end{cases}$$

and hence

$$V > \begin{cases} 2\pi & \text{if } h(x_1) > 0 \\ 3\pi & \text{if } h(x_1) < 0. \end{cases}$$

Thus  $h(x_1) > 0$ , and  $h(x_2) > 0$ , and the interlace property is lost. But each zero of  $\phi$  can now only contribute to one sector of  $\Omega^c$  at the origin and so

$$m \leq S \quad \text{and} \quad p \leq 2S.$$

For larger  $\mu$  the proof follows in a similar manner. □

**Theorem 5.1.** The maximal order of stable schemes (1.2) satisfying  $0 < \mu \leq 1$  is given by

$$p \leq 2(s + S).$$

Further, for  $s \neq 0$  either order reduction occurs when  $\mu > 1$  or the schemes become unstable. □

## 6. Padé schemes.

The results of the preceding section demonstrate that the maximal order of a stable scheme (1.2) is given by  $p \leq 2(s + S)$ . This does not tell us, however, whether schemes of the maximal order are stable, merely that they may be stable. In order to investigate the stability of these schemes we need more information about the schemes themselves. From expression (2.16) we see that schemes which attain the maximal order satisfy

$$h(e^v, \mu) = \cosh \mu v + C^* v^{2(s+S+1)} + o(v^{2(s+S+2)}), (v \rightarrow 0). \quad (6.1)$$

We shall adopt the usual convention and call those schemes with coefficients obtained from the Padé approximations described by (6.1), Padé schemes. By Lemma (2.1) these coefficients can be obtained by solving the system of equations

$$\begin{aligned} \sum_{j=1}^s a_j j^{2k} + \sum_{j=1}^S \{(j + \mu)^{2k} + (j - \mu)^{2k}\} b_j &= -\mu^{2k}, k = 1 : s + S \\ 2 \sum_{j=1}^S b_j + \sum_{j=1}^s a_j + \frac{a_0}{2} &= -1. \end{aligned} \quad (6.2)$$

Here the normalisation  $b_0(\mu) = 1$  is assumed.

System (6.2) is Vandermonde when  $s = 0$  and, as such, is easily solved (see Golub and van Loan [2]).

**Theorem 6.1.** The choice of coefficients

$$a_j(\mu) = \frac{2(-1)^{s-j+1} \mu(\mu - s)_{2s+1}}{(s-j)!(s+j)!(\mu^2 - j^2)} \quad 0 \leq j \leq s \quad (6.3)$$

solves the Vandermonde system (6.2) for  $S = 0$ . The error constant of the method is given by

$$C_s^* = \frac{-\mu(\mu - s)_{2s+1}}{(2s + 2)!},$$

so that the normalised error constant is

$$\tilde{C}_S = \frac{(\mu - s)_{2s+1}}{(2s + 2)!}. \quad (6.4)$$

Here

$$(\alpha)_k = \begin{cases} \alpha(\alpha + 1)(\alpha + 2) \dots & (\alpha + k - 1) \\ 1 & k = 0 \end{cases}$$

is the Pochhammer symbol. □

In [8] it was claimed that the schemes given by (6.3) are stable for  $0 < \mu \leq 1$ . This claim is valid but the proof is not quite correct, although the technique is appropriate,

and can also be employed to demonstrate the stability of the fully implicit Padé schemes,  $s = 0, S > 0$ .

**Theorem 6.2.** The explicit Padé schemes which have  $S = 0, s > 0$  are stable for  $0 < \mu \leq 1$ .

**Proof.** The first part of the proof follows the form taken in [8]. Consider the Taylor expansion of  $\cos(\mu\theta)$  in  $\omega := 2(1 - \cos \theta)$

$$\cos(\mu\theta) = \sum_{k=0}^{\infty} d_k(\mu)\omega^k. \quad (6.5)$$

Clearly  $d_0(\mu) = 1$  and in [8] it is proved that the coefficients  $d_k(\mu), k \geq 1$  are given by

$$d_{k+1} = \frac{(k^2 - \mu^2)}{(2k+2)(2k+1)} d_k(\mu), \quad k \geq 1.$$

By induction the solution is

$$d_k(\mu) = \frac{1}{(2k)!} \prod_{j=0}^{k-1} (j^2 - \mu^2), \quad k \geq 1. \quad (6.6)$$

Now, by (2.17), the explicit schemes satisfy

$$a(e^{i\theta}, \mu) = -2 \cos(\mu\theta) + 0((\theta)^{p+2}), \quad (\theta \rightarrow 0). \quad (6.7)$$

But  $\cos j\theta$  is just the Chebyshev polynomial  $T_j(x)$ ,  $x := \cos \theta$  and hence each  $\cos j\theta$  can be expanded in powers of  $\cos \theta$ , and hence  $w$ . Therefore

$$\begin{aligned} a(e^{i\theta}, \mu) &= \sum_{j=0}^s a_j^*(\mu) (2(1 - \cos \theta))^j \\ &= \sum_{j=0}^s a_j^*(\mu) \omega^j. \end{aligned} \quad (6.8)$$

Thus, by (6.7),

$$\sum_{j=0}^s a_j^*(\mu) \omega^j = -2 \sum_{j=0}^{\infty} d_j(\mu) \omega^j + 0((\omega)^{s+1}), \quad (\omega \rightarrow 0),$$

and hence

$$a_j^*(\mu) = -2d_j(\mu), \quad 0 \leq j \leq s. \quad (6.9)$$

From consistency  $a(1, \mu) = -2$  and  $\tilde{a} = -\frac{a}{2}$  satisfies  $\frac{\partial \tilde{a}}{\partial \theta} > 0$  near  $\theta = 0$ . Therefore,  $|\tilde{a}(e^{i\theta}, \mu)|$  can become greater than 1 only if either  $\tilde{a}(e^{i\theta}, \mu) < -1$  for some  $\theta$  and  $\mu$  or there

is a turning point and  $\tilde{a}(e^{i\theta}, \mu) > 1$  for some  $\theta$  and  $\mu$ . But  $\frac{\partial \tilde{a}}{\partial \theta} = -\sin \theta \sum_{j=1}^s j a_j^*(\mu) \omega^{j-1}$  only has zeros at  $\theta = 0$  and  $\pi$  for  $0 < \mu < 1$ , because  $\omega \geq 0$  and  $d_k(\mu) < 0$  for  $0 < \mu < 1$ ,  $k \geq 1$ . Hence  $|\tilde{a}(e^{i\theta}, \mu)| > 1$  only if  $\tilde{a}(e^{i\theta_0}, \mu_0) < -1$  for some  $0 < \theta_0 < \pi$  and  $0 < \mu_0 < 1$ . This then also implies  $\tilde{a}(-1, \mu_0) < -1$ , and  $a(-1, \mu_0) > 2$ , because  $\frac{\partial \tilde{a}}{\partial \theta} \neq 0$  for  $\theta \in (0, \pi)$  and  $0 < \mu < 1$ .

In particular, consider  $a(-1, \mu)$  for  $\mu$  integer

$$a(-1, \mu) = a_0 + 2 \sum_{j=1}^s (-1)^j a_j(\mu) .$$

For integer  $\mu$ ,  $\mu = n$ ,  $n \leq s$  we have

$$a(-1, \mu) = a_0(n) + 2 \sum_{j=1}^s (-1)^j a_j(n) . \quad (6.10)$$

But because  $a_j(\mu)$  is known it is easily shown that

$$\begin{aligned} a_j(n) &= -\delta_{jn} , \quad j \neq 0 \\ a_0(n) &= -2\delta_{0n} . \end{aligned}$$

Therefore

$$a(-1, n) = 2(-1)^{n+1} \quad (6.11)$$

and further, because  $a(e^{i\theta}, \mu) = a(e^{i\theta}, -\mu)$ ,  $a(-1, -n) = 2(-1)^{n+1}$ , also. Thus, on the interval  $\mu \in [-s, s]$ ,  $a(-1, \mu)$  as a function of  $\mu$  changes sign at least  $2s$  times and must have at least  $2s - 1$  turning points on the interval. Note that  $a(-1, \mu)$  is clearly bounded on this interval. Therefore  $\frac{\partial a(-1, \mu)}{\partial \mu}$  has at least  $2s - 1$  zeros on  $[-s, s]$ . But  $a(-1, \mu)$  is a function in  $\mu$  of degree  $2s$  so that  $\frac{\partial a(-1, \mu)}{\partial \mu}$  has degree at most  $2s - 1$  in  $\mu$ , and thus at most  $2s - 1$  zeros on any interval. Hence  $\frac{\partial a(-1, \mu)}{\partial \mu}$  has exactly  $2s - 1$  zeros on  $\mu \in [-s, s]$  and thus cannot be zero more than once on the interval  $\mu \in (0, 1)$ .

Now  $a(-1, 0) = -2 < 0$  and  $a(-1, \mu) = -2(1 + \sum_{j=1}^s d_j(\mu)4^j) > -2$  for  $0 < \mu < 1$  because  $d_j(\mu) < 0$  for  $0 < \mu < 1$ . Hence  $a(-1, \mu)$  is increasing at  $\mu = 0$ . Now consider  $\frac{\partial a(-1, \mu)}{\partial \mu}$  at  $\mu = 1$ . Differentiating (6.3)

$$\frac{\partial a_j}{\partial \mu} = \frac{4\mu(-1)^{s-j+1}}{(s-j)!(s+j)!} \sum_{\substack{i=0 \\ i \neq j}}^s \prod_{\substack{k=0 \\ k \neq i \\ k \neq j}}^s (\mu^2 - k^2) . \quad (6.12)$$

At  $\mu = 1$

$$\frac{\partial a_1}{\partial \mu} = -\left(\frac{1}{2} + \frac{1}{s} + \frac{1}{s+1}\right)$$

and

$$\frac{\partial a_j}{\partial \mu} = \frac{2(-1)^j (s-1)!(s+1)!}{(s-j)!(s+j)!(1-j)(1+j)} . \quad (6.13)$$

Therefore by (6.12) and (6.13)

$$\begin{aligned} \frac{1}{2} \frac{\partial a(-1, 1)}{\partial \mu} &= \left[ \frac{s+1}{s} \right] + \left[ \frac{1}{2} + \frac{1}{s} + \frac{1}{s+1} \right] \\ &+ 2(s-1)!(s+1)! \sum_{j=2}^s \frac{1}{(s-j)!(s+j)!(1-j)(1+j)}. \end{aligned} \quad (6.14)$$

With some manipulation (6.14) can be simplified to give

$$\frac{1}{2} \frac{\partial a(-1, 1)}{\partial \mu} = \left( \frac{1}{s(s+1)} + \frac{2(s-1)!(s+1)!}{(2s+2)!} (2s+3+s2^{2s+2}) \right) > 0. \quad (6.15)$$

Therefore  $a(-1, \mu)$  is increasing at  $\mu = 1$  and thus the explicit schemes are stable for  $0 < \mu < 1$ , because by (6.11)  $a(-1, 1) = 2$  and  $a(-1, \mu)$  is monotonically increasing on the interval  $0 < \mu < 1$ .

Note that when  $\mu = 1$ ,  $|a(e^{i\theta}, \mu)| = |-2 \cos \theta| \leq 2$ , with  $|a(e^{i\theta}, \mu)| = 2$  for  $\theta = 0, \pi$ , and  $|w(e^{i\theta})| = 1$  with double roots at  $\theta = 0$  and  $\theta = \pi$ , neither of which is a branch point of  $w$ , because

$$a(z, 1) = -(z^1 + z^{-1})$$

implies that the roots  $w_{\pm}$  of  $\Phi(z, w) = 0$  are given by  $w_+ = z$  and  $w_- = z^{-1}$ .  $\square$

We now consider numerical schemes for the artificially-bounded domain wave equation problem, *global absorbing boundary schemes*. These schemes consist of a difference scheme for the wave equation on the interior of the domain, in conjunction with a high order boundary scheme applied at the artificial boundaries. In [10] we see that in order to ascertain the global stability of such schemes it is necessary to investigate the effect of the interior scheme (1.2) on oscillatory solutions of the boundary condition. A necessary condition for stability is that incoming solutions from the boundary are not propagated by the interior scheme. For the high order absorbing boundary conditions in [10] it is shown that in conjunction with interior schemes for which the oscillatory modes satisfy  $\arg(w) = -\arg(z)$  and  $\arg(w) = \arg(z)$  for incoming and outgoing modes, respectively, necessary conditions for stability in terms of the coefficients of the absorbing boundary equation can be obtained. With the explicit schemes considered here it can be shown from Theorem 6.2 that the oscillatory modes are of the appropriate type, i.e. that for incoming (outgoing) modes

$$\arg(w) = -\arg(z), \quad (\arg(w) = \arg(z)).$$

Hence Corollary 6.2 follows.

### Corollary 6.2

For the global absorbing boundary scheme, comprised of an explicit Padé scheme on the interior of the domain and a high order boundary scheme [10] at the boundaries, and

$0 < \mu \leq 1$ , necessary conditions for stability depend only on the coefficients of the high order boundary scheme, and are as given in Theorem 1, [10].  $\square$

**Proof**

By evaluating the group velocity  $C_x$  component in the  $x$ -direction it can be seen that

$$C_x = -K \frac{\sin \theta}{\sin \phi}, \quad \text{where } \theta = \arg(z), \quad \phi = \arg(w)$$

and  $K > 0$  for all  $\theta$ , and  $0 < \mu \leq 1$ .  $\square$

**Theorem 6.3.** The Padé schemes which have  $s = 0$ ,  $S > 0$  are stable for  $0 < \mu < 1$ .  $\square$

**Proof.** We consider the Padé schemes which satisfy (6.1);

$$h(e^{i\theta}, \mu) = \frac{a_0}{1 + 2 \sum_{j=1}^S b_j \cos j\theta} = \cos \mu\theta + O((\theta)^{2S+2}), \quad \theta \rightarrow 0. \quad (6.16)$$

For stability we have to show  $|h(e^{i\theta}, \mu)| \leq 1$ . As in the proof of Theorem 6.2 we write

$$1 + 2 \sum_{j=1}^S b_j \cos j\theta = \sum_{j=0}^S b_j^* \omega^j$$

for coefficients  $b_j^*(\mu)$ ,  $0 \leq j \leq S$ . Then by (6.5)

$$\frac{a_0}{\sum_{j=0}^S b_j^* \omega^j} = \sum_{k=0}^{\infty} d_k(\mu) \omega^k + O((\omega)^{s+1}), \quad (\omega \rightarrow 0)$$

which means that

$$a_0 = \left( \sum_{j=0}^S b_j^* \omega^j \right) \left( \sum_{k=0}^{\infty} d_k(\mu) \omega^k + O((\omega)^{s+1}) \right),$$

and thus

$$a_0 = d_0 b_0^* = b_0^*$$

and

$$0 = \sum_{i=0}^j d_i b_{j-i}^*, \quad j = 1 : S.$$

Therefore  $b_j^* = -\sum_{i=1}^j d_i b_{j-i}^*$ ,  $j = 1 : S$ , and because  $d_i(\mu) < 0$ ,  $i \geq 1$  and  $0 < \mu < 1$  it can be shown inductively that the  $b_j^*$  all have the same sign for  $0 < \mu < 1$ . Thus

$$h(e^{i\theta}, \mu) = \frac{a_0}{\sum_{j=0}^S b_j^* \omega^j} = \frac{b_0^*}{\sum_{j=0}^S b_j^* \omega^j} = \frac{1}{1 + \sum_{j=1}^S (b_j^*/b_0^*) \omega^j} < 1,$$

because  $b_j^*/b_0^* > 0$  for all  $j$ . Thus the schemes satisfying (6.16) are stable for  $0 < \mu < 1$ .  $\square$

**Corollary 6.3.** For the global absorbing boundary scheme, comprised of a fully implicit,  $s = 0, S > 0$ , Padé scheme on the interior of the domain and a high order boundary scheme [10] at the boundaries, and  $0 < \mu \leq 1$ , necessary conditions for stability depend only on the coefficients of the high order boundary scheme, and are as given in Theorem 1, [10].  $\square$

**Proof.** It can be seen that the group velocity component  $C_x$  can be expressed as in the proof of corollary 6.2.  $\square$

The system (6.2) is, in general, not Vandermonde and hence we cannot find explicit solutions of (6.2) for any pair  $(s, S)$ . But we can use the uniqueness of Padé approximations to rewrite (6.1) as

$$\frac{1}{h(\epsilon^v, \mu)} = \operatorname{sech} \mu v - C^* v^{2p+2} + o((v)^{2p+3}), \quad (6.17)$$

so that for  $s = 0$  (6.17) becomes

$$1 + 2 \sum_{j=1}^S b_j \cosh jv = a_0 \operatorname{sech} \mu v - a_0 C^* v^{2S+2} + o((v)^{2S+3}), \quad (v \rightarrow 0). \quad (6.18)$$

Therefore the coefficients  $b_j$  satisfy

$$2 \sum_{j=1}^S j^{2k} b_j = a_0 \mu^{2k} e_{2k}, \quad 1 \leq k \leq S, \quad (6.19)$$

where

$$e_k = \left( \frac{d^k}{dv^k} \operatorname{sech} v \right)_{v=0}$$

and

$$a_0 = 1 + 2 \sum_{j=1}^S b_j.$$

The system (6.19) is Vandermonde and given  $e_{2k}$  a solution can be found. Now  $e_k = 0$  for  $k$  odd but for  $k$  even the first few values are given by

$$e_2 = -1, \quad e_4 = 5, \quad e_6 = -61, \quad e_8 = 1385,$$

and we realise, of course, that  $e_{2k} = E_{2k}$  where  $E_{2n}$ ,  $n = 0, 1, 2, \dots$  are the Euler numbers (see p. 810 in Abramowitz and Stegun [1]). The explicit form of the coefficients in the case  $s = 0, S > 0$  can therefore also be found by solving a Vandermonde system.

**Theorem 6.4.** The choice of coefficients

$$\begin{aligned} a_0(\mu) &= -2\tilde{a}_0 \\ b_j(\mu) &= \begin{cases} (\tilde{a}_0 \tilde{b}_j(\mu))/2 & 1 \leq j \leq S \\ 1 & j = 0 \end{cases}, \end{aligned}$$

where

$$\begin{aligned}\tilde{b}_j(\mu) &= \frac{2 \left| e_{SS}^j \right| (-1)^j}{(s+j)!(s-j)!}, \quad 1 \leq j \leq S, \\ \tilde{a}_0 &= \left( 1 - \sum_{j=1}^S \tilde{b}_j(\mu) \right)^{-1}, \\ e_{SS}^j &= \sum_{k=0}^{S-1} \mu^{2(s-k)} E_{2(s-k)} \alpha_k^j, \quad 1 \leq j \leq S,\end{aligned}\tag{6.20}$$

and

$$\alpha_k^j = \begin{cases} (-1)^k \sum_{I=I^1}^{I^2} X_I, & k \geq 1, \\ 1, & k = 0, \end{cases},$$

for which

$$\begin{aligned}I &= \text{set of indices} = \{i_1, i_2, i_3, \dots, i_k\} \\ I^1 &= \begin{cases} \{k, k-1, \dots, 1\} & j \geq k \\ \{k+1, k, \dots, j+1, j-1, \dots, 1\} & j < k \end{cases} \\ I^2 &= \{s, i_1 - 1, i_2 - 1, \dots, i_{k-1} - 1\}\end{aligned}$$

and

$$X_I = \prod_{\ell=1}^k (i_\ell)^2,$$

solves the Vandermonde system (6.19) for  $s = 0$ . Here the expression for the coefficient  $\alpha_k^j$  represents the sum over all the indices in  $I$  for the limits from  $I^1$  to  $I^2$ , with the exception that no index assumes the value  $j$ . The ordering in each set is assumed to be the same, so that for example when  $j \geq k$  the index  $i_1$  is summed from  $k$  to  $s+1$ . The error constant of the scheme is given by

$$C_S^* = \frac{\mu^{2(S+1)} E_{2(S+1)} - \sum_{j=1}^S j^{2(S+1)} \tilde{b}_j}{(2S+2)!},\tag{6.21}$$

so that the normalised error constant is given by

$$\tilde{C}_S = \frac{-C_S^*}{\mu} . \quad \square$$

Examples indicate that these schemes are stable for any  $\mu$ , although this is only proved for  $0 < \mu \leq 1$ . From (6.20) it is clear that the  $\tilde{b}_j$  are polynomials in  $\mu$  of degree at most  $2S$ , but from (6.21) the error constant is a polynomial in  $\mu$  of order  $2(S+1)$ . Hence the normalised error constant is a polynomial in  $\mu$  of order  $2S+1$  with leading coefficient  $E_{2(S+1)}$ . Therefore, for  $\mu > 1$ , the leading term dominates, so it is easily seen that  $\left| \tilde{C}_S \right|$  increases rapidly for  $\mu > 1$ . Thus despite the apparent unconditional stability of these “fully implicit” schemes the occurrence of large moduli error constants indicates that in

practice these schemes are not useful for  $\mu > 1$ . When  $\mu < 1$  the error constants satisfy  $|\tilde{C}_S| < 1$  but then  $|\tilde{C}_s| < 1$  also for the explicit schemes. The latter schemes are easily programmed in comparison to the implicit schemes which require the solution of a system of equations at each time level. Therefore there can be no reason to use these implicit schemes for any choice of  $\mu$ , and Theorem 6.4 is presented to validate the above statements, rather than with expectation that it will be used.

**Acknowledgements:** The work of the second author was supported by an NSF US–Switzerland cooperative research grant INT9123314 and funding from the Forschungsinstitut für Mathematik, ETH Zürich. The work of the first and third authors was supported under project Nr. 21-33551.92 of the Schweizerische Nationalfonds. Travel funds for the third author were provided by the University of Stellenbosch.

## References

- 1 M. Abramowitz and I.A. Stegun: Handbook of Mathematical Functions; Dover Publications, New York, 1968.
- 2 G.H. Golub and C.F. van Loan: Matrix computations; The John Hopkins University Press, Baltimore, MD, 1983.
- 3 Iserles A., Strang G., The optimal accuracy of difference schemes, Trans. Amer. Math. Soc. **277**, 779-803, (1983).
- 4 Jeltsch R., Stability and accuracy of difference schemes for hyperbolic problems, J. Comput. Appl. Math. **12 & 13**, 91-108, (1985).
- 5 R. Jeltsch, R.A. Renaut, and J.H. Smit: An accuracy barrier for stable three-time-level difference schemes for hyperbolic equations; Research Report 95–01, Seminar für Angewandte Mathematik, ETH Zürich.
- 6 Jeltsch R., Smit J. H., Accuracy barriers of difference schemes for hyperbolic equations, SIAM J. Numer. Anal. **24**, 1-11, (1987).
- 7 Jeltsch R., Smit J. H., Accuracy barriers of three-time-level difference schemes for hyperbolic equations, Ann. University of Stellenbosch, 1992/2, 1-34, (1992).
- 8 Renaut R. A., Full discretizations of  $u_{tt} = u_{xx}$  and rational approximations to  $\cosh \mu z$ , SIAM J. Numer. Anal. **26**, (1989).
- 9 Renaut R. A., Smit J. H., Order stars and the maximal accuracy of stable difference schemes for the wave equation, Quaestiones Math. **15**, 307-323, (1992).
- 10 R.A. Renaut: Absorbing boundary conditions, difference operators and stability; J. Comp. Phys. **102** (1992), 236–251.
- 11 Smit J. H., Order stars and the optimal accuracy of stable, explicit difference schemes, Quaestiones Math. **8**, 167-188, (1985).
- 12 Strang G., Iserles A., Barriers to stability, SIAM J. Numer. Anal. **20**, 1251-1257, (1983).

- 13 J.C. Strikwerda: Finite difference schemes and partial differential equations; Wadsworth and Brooks, Pacific Grove, California, 1989.
- 14 Wanner G., Hairer E., Nørsett S. P., Order stars and stability theorems, BIT 18, 475-489, (1978).