Chapter 4
Complex Spectra

Abstract Recognition of Lyapunov and Sylvester matrix equations as model ADI problems stimulated generalization to complex spectra. Rouchés theorem replaces the alternating extreme property in this analysis. Complex spectra are embedded in "elliptic function regions" for which ADI iteration parameters are generated. Jordan's spectral alignment is generalized for application to Sylvester equations.

4.1 Introduction

Generalization of Chebyshev minimax theory into the complex plane has been considered by many researchers, and much of the relevant theory may be found in Smirnov (1968). A concise review of some of this theory was given by Rivlin (1980). Application to the ADI minimax problem by Starke (1989) to obtain asymptotically optimal parameters motivated recent work by Istace and Thiran (1993) in which nonlinear optimization numerical techniques were devised to determine optimal parameters. The following brief chronology was extracted from the Istace–Thiran paper: [Gonchar, 1969] characterized the general minimax problem and showed how asymptotically optimal parameters could be obtained with generalized Léja or Fejér points. [Starke, 1989] subsequently applied this theory to the ADI minimax problem. [Gutknecht, 1983] obtained a necessary Kolmogorov optimality condition for the general problem and [Ruttan, 1985] found an alternative which was refined in [Istace and Thiran, 1993]. The latter then implemented this theory for the ADI iteration problem with iterative solution of the relevant nonlinear optimization problem, starting with a complex version of an [Osborne-Watson, 1978] exchange algorithm and then switching to a Newton iteration to achieve the desired accuracy.

My analysis from 1982 to 1994 [Wachspress 1988c, 1990, 1991] was directed toward practical determination of efficient ADI iteration parameters for spectra anticipated in practice. My goal was to generalize the elliptic-function theory into the complex domain in much the same manner as the Chebyshev-polynomial theory had been generalized for polynomial approximation in the complex plane.

E. Wachspress, *The ADI Model Problem*, DOI 10.1007/978-1-4614-5122-8_4,

Although the relevant spectra were somewhat restricted, it was found in practice that many problems could be solved expediently by embedding actual spectra in these "elliptic-function" regions. My analysis led to Starke's investigations which in turn stimulated the development of Istace and Thiran. This marriage of the more erudite and general development with my limited studies was most gratifying. One may now choose among the various approaches to find effective ADI iteration parameters for specific problems.

It was observed in Chap. 3 that my analysis of ADI iteration in the presence of complex spectra was motivated by application to solution of Lyapunov and Sylvester matrix equations. The commutation property required for application to the Dirichlet problem is not restrictive in this new application. However, complex spectra to which the ADI theory described in the previous chapters does not apply are now encountered. Complex spectra also arise in boundary-value problems containing odd order derivatives like, for example, convection diffusion equations. Theory of elliptic functions plays a prominent role in the analysis.

The coefficient matrices of the linear systems to be solved are assumed to be real and N-stable. This yields spectra in the positive-real half plane which are symmetric about the real axis. (We will designate these as PRS spectra.) It may be shown that the set of optimum parameters for any PRS spectrum must also be PRS. Unlike the corresponding polynomial approximation to zero, the rational ADI approximation is bounded in absolute value by unity for any choice of PRS parameters. It is thus possible to partition the spectral region into subregions for each of which parameters may then be selected to yield the prescribed error reduction. This was in fact the approach used by Douglas and Rachford (1956) for real spectra when ADI iteration was first introduced.

Nonsymmetric systems create other problems. Convergence of ADI iteration is retarded by deficient eigenvector spaces. Although means for handling such deficiencies may be addressed, we consider primarily problems with complete eigenvector spaces. This is assumed unless specified otherwise in the ensuing analysis. When dealing with nonsymmetric systems, the error norm reduction is not bounded by the spectral radius of the iteration matrix. If Λ is the diagonal matrix of eigenvalues of the matrix A then $A = G\Lambda G^{-1}$, where G is the matrix whose columns are the right eigenvectors of A. The condition of matrix G is $\kappa(G) = \|G\| \, \|G^{-1}\|$. If ρ is the spectral radius of an iteration matrix that commutes with A, then the norm of the error reduction is bounded by $\kappa\rho$. Nachtigal, Reddy and Trefethen (1990) considered more subtle problems associated with nonsymmetry and demonstrated for polynomial approximation to zero that one should choose parameters which minimize the spectral radius of the iteration matrix for a spectrum chosen to be slightly larger than the actual spectral region. Thus far, ADI iteration convergence with parameters based on the actual spectra has been quite satisfactory.

Just as Wilkinson's "backward error analysis" proved fruitful in studies of numerical stability, "backward spectrum analysis" has proved to be useful for ADI iteration. One chooses convenient sets of iteration parameters and determines families of spectral regions for which these sets are optimal. One then embeds a given spectrum in an approximating member of the families of spectra generated

in this manner. For example, the ADI parameters determined for real spectra are also optimal for a class of complex spectra which we denote as "elliptic-function regions." One may embed a given spectral region in an elliptic-function region for which optimum parameters and the resulting error reduction are known.

Complex iteration parameters enter in conjugate pairs. By combining the two iterations with a conjugate pair, one can perform the iteration in real arithmetic with essentially no increase in computation time over that required for two iterations with real parameters.

The elliptic-function region theory has been verified with numerical solution of the Lyapunov matrix equation by ADI iteration. Some of the early studies were reported by Saltzman (1987), and later studies were reported by Lu and Wachspress (1991).

The more general development in [Starke, 1989] describes how Remez-type arguments may be applied to yield asymptotically optimal iteration parameters in similar fashion to schemes used for polynomial approximation to zero with Fejér and Léja parameters on the spectral boundary. For example, if the spectrum is bounded by a circle with real diameter the interval $[a, b]$, then the optimum parameter set is repeated use of $\sqrt{ab}$. The Fejér and Léja parameters for J iterations are J equally spaced points on the boundary. As J increases, the error reduction with these points approaches from above that obtained with the single optimum parameter repeated J times. Optimal parameters for more general regions are not found readily, but the asymptotically optimal Léja parameters can be approximated quite well by a Remez-type algorithm.

The theory for Chebyshev approximation over complex domains differs from the minimax theory for real domains. However, this complex theory applies equally as well to polynomial and rational approximation to zero. Initial analysis was for polynomials, and we lay the groundwork for the ADI iteration theory by first describing the polynomial theory.

4.2 Chebyshev Approximation with Polynomials

The general theory for rational Chebyshev approximation over complex domains is not needed for this development. We require only application of Rouché's theorem [Copson, 1935]:

ROUCHÉ'S THEOREM. *If functions f and g have no essential singularities in a region bounded by a simple closed curve on which fg is bounded and $|f|$ is everywhere greater than $|g|$, then f and $f - g$ have the same number of zeros minus poles, counting multiplicities, in the region bounded by the curve.*

This result follows directly from the Cauchy residue theorem. In our application, there are no poles within the region which contains the spectrum over which the spectral radius of the iteration matrix is to be minimized. In this section the functions are polynomials. When we treat the ADI iteration problem, the functions are rational

with nonvanishing denominators over the spectral region. Suppose the iteration function f is a polynomial of degree J normalized to unity at a fixed point outside the spectral region and that it has constant absolute value H on the spectral boundary and all J of its zeros within the spectral region. If we assume the existence of a polynomial g of maximal degree J with a smaller maximum absolute value over the entire boundary, also normalized to unity at the fixed point outside the spectral region, we may apply Rouché's Theorem to arrive at a contradiction. The difference polynomial $f - g$ has J zeros inside the boundary and is zero at the normalization point. The polynomial is analytic over the spectral region and hence attains its maximum absolute value on the boundary.

One might think this result too restrictive to be of practical value since one cannot hope except in very special cases for a polynomial of this type to exist for a given spectrum. That is where the backward analysis enters. Any polynomial of degree J is optimal for a family of regions defined by its absolute level contours. The simplest example is z^J with circular contours around the origin. This leads to the important result that successive overrelaxation (SOR) applied to the usual Dirichlet-type systems with optimal extrapolation cannot be accelerated by linear combination of the SOR iterates. The eigenvalues of the SOR iteration matrix all lie on a circle. The result of J iterations is the polynomial z^J where z lies on the circle of radius equal to the spectral radius of the SOR iteration matrix. This is the unique polynomial of maximal degree J which has the least maximum absolute value over the disk bounded by the circle.

We now consider the family of regions for which Chebyshev polynomials are optimal. In my book on iterative solution of elliptic systems [Wachspress, 1966], I investigated the application of Chebyshev polynomials to spectral regions bounded by ellipses with real major axes and real normalization point. Convergence and adaptive updating of parameters was considered. There was no discussion of optimality, and the relevance of Rouché had not yet been disclosed. The first published account of application of Chebyshev polynomials to complex spectra was [Clayton, 1963] and this analysis was first applied to acceleration of Jacobi iteration in [Wrigley, 1963]. A more recent and thorough treatment of this problem was presented in [Manteuffel, 1977], who appears to be the first to discuss the relevance of Rouché's Theorem to this problem. The paper [Opfer and Schober, 1984] clarified some of the problems associated with Chebyshev approximation in the complex plane. They showed, for example, that the translated and rotated Chebyshev polynomial is not optimal for the line spectrum equal to the interval $[(1 - i), (1 + i)]$ perpendicular to the real axis, a result implicit in Manteuffel's work. (The Chebyshev parameters are asymptotically optimal for this case.) A real affine transformation normalizes a spectrum bounded by an ellipse to the region bounded by

$$\Gamma = \left\{ z = \cos(\phi + i\,\psi) \mid 0 \le \phi \le 2\pi,\ \psi = \operatorname{arc\,tanh} \frac{b}{a} \right\}, \tag{1}$$

where $2\cosh b$ is the length of the minor imaginary axis and $2\cosh a$ is the length of the major real axis of the bounding ellipse. The Chebyshev polynomial

$$C_J(z) = \cos(J \arccos z) = \cos(J\theta) \tag{2}$$

varies along Γ as

$$\cos[J(\phi + i\psi)] = \cos J\phi \cosh J\psi - i \sin J\phi \sinh J\psi. \tag{3}$$

This function has absolute value which varies between a maximum of $\cosh J\psi$ and a minimum of $\sinh J\psi$ on the boundary. The ellipse may be enclosed in a scalloped region on which $|\cos J\theta| = \cosh J\psi$. The Chebyshev polynomial is optimal over this extended region. As J increases the ratio of $\sinh J\psi$ to $\cosh J\psi$ approaches unity and the enclosing scalloped curve approaches the ellipse. Thus, the Chebyshev polynomial is asymptotically optimal.

Having established that the Chebyshev polynomials are optimal for regions which close in on an ellipse as J increases, we consider the possibility that these parameters are optimal for this ellipse for all J. I thought Tom Manteuffel had proved optimality and he responded to my query by sending me a reprint of a definitive paper by Fischer and Freund (1991) in which an error in Clayton's 1963 proof of optimality was disclosed. Fischer and Freund proved that when $J \leq 4$ the Chebyshev polynomial is optimal, but that for $J > 4$ it is not optimal when the normalization point is sufficiently close to $\cosh b$. They gave precise rules for establishing optimality. As J increases, the interval of nonoptimality decreases (as it must in view of the Rouché result). We note that as the normalization point approaches $\cosh b$ the number of iterations for significant error reduction increases. Thus, for significant error reduction J increases as the normalization point approaches $\cosh b$ and the Chebyshev polynomial approaches optimality. Moreover, the Chebyshev polynomial is optimal for all J when the ellipse is either a circle or degenerates to the real axis. Fischer and Freund were unable to detect any analytic representation of the truly optimal polynomials. In general, little is lost in practice by use of the Chebyshev polynomials. The ease with which the extrapolation parameters may be found and the associated error reduction predicted makes them well suited for this application. Their use for spectral regions other than elliptic must be examined more carefully. One must evaluate loss in convergence when the actual region is embedded in an elliptic region for selection of parameters.

There is a rather general procedure for obtaining asymptotically optimal parameters. One variant of this approach will now be outlined for a polygonal boundary. One starts with a polynomial that vanishes at each vertex. One then chooses a set of sampling points uniformly spaced on the boundary and computes the value of the iteration function at these points. If the maximum absolute value of the iteration function at sampling points between two consecutive points already selected is greater than the prescribed error reduction, then one introduces an additional parameter at the sampling point at which the function attains this maximum absolute value. This is continued until the maximum absolute value among all sampling

points is less than the prescribed error reduction. These are the Léja [Léja, 1957] parameters. The scalloped boundary extension on which the constant maximum absolute value is retained defines a larger region for which the chosen parameters are optimal. This scalloped boundary approaches the actual boundary as the number of parameters is increased.

Some insight into the validity of this approximation is gained by considering the unit disk with the iteration function normalized to unity at $z = 2$. The optimum parameter is repeated use of zero, in which case the normalized iteration function for n iterations is $(\frac{z}{2})^n$ and the error reduction is 2^{-n}. The Léja points are the n roots of unity and the iteration function is $\frac{z^n-1}{2^n-1}$. Its maximum absolute value on the unit circle is $\frac{2}{2^n-1}$. As n increases this approaches $2^{(1-n)}$. Although this is twice the optimal value, the asymptotic convergence rate is the n-th root, which is $1/2$ for both the optimal and the Léja points. Thus, the Léja points are not truly optimal even asymptotically. The rate of convergence approaches the optimal rate of convergence asymptotically. The Léja points are easily generated, and the associated error reduction is found during their generation.

Relative advantages of Léja points and embedding a given spectrum in an ellipse or other region for which analytic optimization is possible must be weighed. When relatively few iterations are needed to attain the prescribed accuracy, the Léja points may be poor. On the other hand, there may be no convenient embedding of the actual region which yields efficient parameters. Then again, one must consider the advantage of having analytic error reduction bounds as a function of embedded spectral parameters. In practice, details of the spectrum are often not known and it is common to enclose known values like minimum and maximum real eigenvalues and maximum imaginary component in a conservative ellipse.

One must also consider the use of more sophisticated algorithms [Istace and Thiran] to compute optimal parameters when neither embedding in an ellipse nor using asymptotically optimal parameters is efficient. One must take care that time spent in computing parameters does not outweigh the convergence improvement derived therefrom. Significant computer programming is required for some of these schemes.

4.3 Early ADI Analysis for Complex Spectra

The earliest reported analysis of complex spectra was in [Saltzman, 1987] which applied primarily to spectra with relatively small imaginary components. A review of this analysis provides a springboard for study of more general spectra. Extensive use is made of the theory of elliptic functions and in particular Jacobi elliptic functions. My 1995 monograph drew heavily on [Abramowitz and Stegun, 1964]. The recent update of this work, [NIST Handbook of Mathematical Functions, 2010] is referenced here with the notation “N-xx.x” denoting formula or tables in Chap. 22 of the NIST handbook.

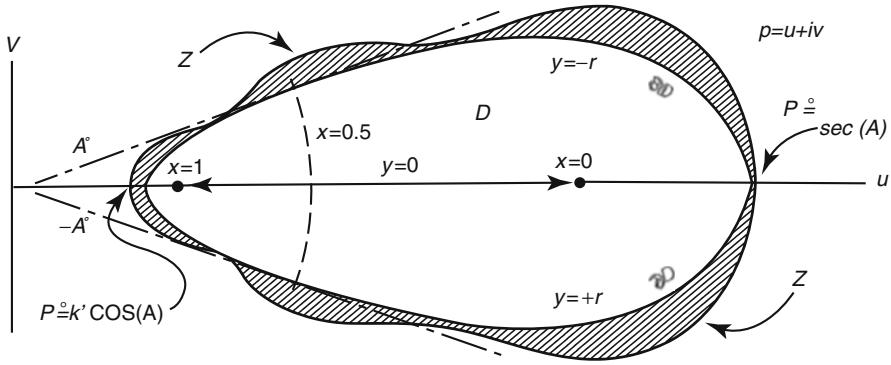

Fig. 4.1 An elliptic-function region

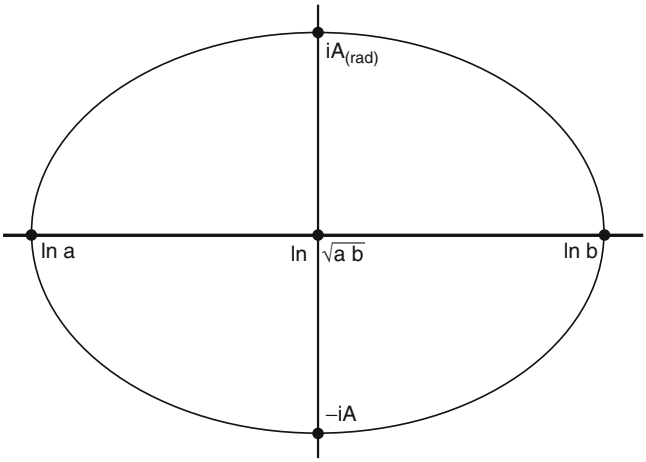

Fig. 4.2 Logarithm of an elliptic-function region

Let the eigenvalues p over which the error reduction is to be minimized for J iterations be enclosed by the elliptic-function region D:

$$D = \{p = dn(zK,\ k) \mid z = x + iy,\ \ 0 \le x \le 1 \text{ and } |y| \le r\}. \tag{4}$$

When $1-k << 1$, D is an egg-shaped region with parameters displayed in Fig. 4.1.

The logarithm of an elliptic-function region, normalized so that the product of the endpoints of its real intercept is unity, is symmetric about both the real and imaginary axes (Fig. 4.2).

The logarithmic spectra are analogous to the elliptic spectra for polynomial approximation. The complementary modulus $k' = (1 - k^2)^{1/2}$ is less than unity. The complete elliptic integral for modulus k is approximated well by $K \doteq \ln(4/k')$

and for modulus k' by $K' \doteq \pi/2$. Region D is tangent at $\sqrt{k'}\exp[iKr]$ to the ray from the origin at angle $A = rK \doteq rln(4/k')$. This tangent point in D corresponds to $x = 0.5$ and $y = -r$. It should be noted that the inversion of D in the circle of radius $\sqrt{k'}$ is D. If the ADI parameters are computed for the real interval $[k', 1]$, the iteration function for J iterations is

$$g(z) = k(J)^{1/2} sn[(1 + 2Jz)K(J),\ k(J)], \tag{5}$$

where $k(J)$ may be computed from k and J. As J increases, $k(J)$ goes to zero, and a desired error reduction is attained by suitable choice of J. We assume henceforth that J is sufficiently large that $k(J) << 1$. It can then be shown by approximating elliptic functions of small modulus with trigonometric functions and those of modulus close to unity with hyperbolic functions that for all z on a closed curve Z (shown in Fig. 4.1) which is close to the boundary of D, $|g(z)|$ has the constant value of

$$R(Z) \doteq 2\exp\left[-\frac{\pi^2 J}{2\ln(\frac{4}{k'})}\right]\cosh\left[\frac{J\pi A}{\ln(\frac{4}{k'})}\right]. \tag{6}$$

Rouché's theorem may be used to prove [Saltzman, 1987] that $R(Z)$ is the least possible value for R attainable with J parameters for the domain bounded by curve Z. The proof is slightly more complicated for the ADI rational function. The difference between the two rational functions is a rational function with numerator of degree $2J - 1$. Only J zeros are identified by Rouché's theorem within the spectral region. However, the numerator is an odd function of z so that there must be an additional negative J zeros. This together with the normalization to unity at $z = 0$ identifies $2J + 1$ zeros for the numerator. This contradiction establishes that the parameters are optimal for the region bounded by the scalloped extension of the elliptic-function region. That this is the unique solution may be proved by more subtle arguments [Stephenson and Sundberg, 1985].

We note that $|g(z)| \leq R$ of Eq. 6 for all p in D. As J increases, Z approaches the boundary of D. Note that if we define the x-intercepts as a and b, then the modulus of the elliptic function satisfies

$$k' \doteq \frac{a}{b}\sec^2 A. \tag{7}$$

When J is large enough that $\cosh[\cdot]$ can be approximated well by $0.5\exp[\cdot]$, the value of R may be approximated by

$$R(Z) \doteq \exp\left[-\frac{\pi J(\pi - 2A)}{2\ln\frac{4}{k'}}\right]. \tag{8}$$

Thus, it is seen that as A approaches $\pi/2$, R approaches unity. This is the correct limit for eigenvalues on the imaginary axis. This form allows one to compute the loss in convergence as a function of A. For example, when $A = \pi/4$, approximately twice as many iterations are required as when A is close to zero. However, before one can use these results on problems where A is not small one must review some of the assumptions leading to Eq. 6. In particular, the value of k' in Eq. 7 is valid only for a particular range of a/b and A.

It is easily proved [Saltzman] that when the spectrum is bounded by a circle the optimum parameters are repeated use of the single value $w(j) = \sqrt{ab}$. Let $z = c + d \exp i\theta$ on the circular boundary, where $c - d > 0$. Then $a = c - d$ and $b = c + d$ and

$$\left|\frac{\sqrt{ab} - z}{\sqrt{ab} + z}\right|^2 = \frac{c - \sqrt{c^2 - d^2}}{c + \sqrt{c^2 - d^2}} = \frac{\frac{a+b}{2} - \sqrt{ab}}{\frac{a+b}{2} + \sqrt{ab}}. \tag{9}$$

The rational iteration function has constant absolute value on the circle and is, therefore, optimal. This corresponds to $k' = 1$ and $A = \arccos \frac{2\sqrt{a/b}}{1+(a/b)}$. Clearly, Eq. 7 is not valid in this case. Eq. 7 would give $k' = [(1 + a/b)/2]^2$ instead of 1. In the next section, theory will be developed for the entire range of elliptic-function domains varying from the real line to a disk and from the disk to a circle arc.

4.4 The Family of Elliptic-Function Domains

We retain the domain of Eq. 4 but drop the approximations based on $k' << 1$. The ratio of the real intercepts is obtained from Tables N-4.3 and N-6.1 as

$$\frac{a}{b} = \frac{dn[K(1 + ri), k]}{dn[Kri, k]} = \frac{k' cn^2(Kr, k')}{dn^2(Kr, k')}. \tag{10}$$

By formula N-6.1, $cn^2(\text{mod } k') = (dn^2 - k^2)/k'^2$, and Eq. 10 may be solved for dn^2:

$$dn^2(Kr, k') = \frac{1 - k'^2}{1 - \frac{ak'}{b}} \tag{11}$$

We then obtain

$$cn^2(Kr, k') = \frac{\frac{a}{b}(1 - k'^2)}{k'[1 - \frac{ak'}{b}]}, \tag{12}$$

and since $sn^2 = 1 - cn^2$, we have

$$sn^2(Kr, k') = \frac{1 - \frac{a}{bk'}}{1 - \frac{ak'}{b}}. \tag{13}$$

Note that when $r = 0, k' = a/b$ is the appropriate value for this real domain and this yields $dn = 1, cn = 1$, and $sn = 0$ in the above equations.

The maximum angle is attained when $x = 1/2$ and $| y | = r$ as in the previous analysis. However, it is crucial that we not assume $k' << 1$ in evaluating this angle. As the boundary becomes more circular, k' approaches unity. Formula N-8.3 and Tables N-5.1-2 yield

$$\tan^2 A = (1 - k')^2 \frac{sn^2(Kr, k')}{cn^2(Kr, k') dn^2(Kr, k')}. \tag{14}$$

Substitution of Eqs. 11–13 into Eq. 14 results in

$$\tan^2 A = \frac{(k' - \frac{a}{b})(1 - \frac{ak'}{b})}{\frac{a}{b}(1 + k')^2}. \tag{15}$$

Given a domain with angle A and real intercept ratio a/b, we may solve Eq. 15 for k'. We define

$$\cos^2 B = \frac{2}{1 + \frac{1}{2}(\frac{a}{b} + \frac{b}{a})} \tag{16}$$

and

$$m = \frac{2 \cos^2 A}{\cos^2 B} - 1. \tag{17}$$

If $A < B$, then $m > 1$ and we obtain from Eq. 15

$$k' = \frac{1}{m + \sqrt{m^2 - 1}} \quad \text{in } (0, 1]; \tag{18.1}$$

$$w(j) = \sqrt{\frac{ab}{k'}} dn \left[\frac{(2j - 1)K}{2J}, \; k \right].$$

$$j = 1, 2, \ldots, J \tag{18.2}$$

Two limiting cases are of interest. Let $p \equiv k'b/a$. When $m >> 1$, $k' = pa/b << 1$ and from Eq. 15, $\tan^2 A \doteq p - 1$ or $p \doteq \sec^2 A$ as in Eq. 7.

Next let $m = 1$, the smallest value for which k' remains real. In this limit $k' = 1$ and Eq. 15 yields $\tan A = \frac{(1 - \frac{a}{b})}{2\sqrt{a/b}}$. Hence, $\cos A = \frac{2\sqrt{a/b}}{(1 + \frac{a}{b})}$ as was previously derived for the disk. It is thus established that these new relationships provide a transition between the real line and the disk. It should be noted that in this limit $Kr \to K'$ which becomes infinite at $k' = 1$ and that the elliptic-function region does approach a disk rather than a point.

To illustrate a spectrum in the transition region, let $a/b = 0.1$ and $A = 45°$. Then $m = 2.025$ and $k' = 0.264$. Note that $(a/b) \sec^2 A = 0.2$. The larger value of k' here reflects the contraction of the parameters toward $\sqrt{ab}$ as the domain moves from the real line to the disk. When $m \geq 1$ the optimum parameters are real. If $m < 1$ all the parameters lie on an arc of the circle of radius $\sqrt{ab}$.

We preempt analysis with elliptic functions of complex moduli by defining a dual spectrum. To motivate the dual spectrum technique, we consider optimum parameters for the spectrum consisting of the arc of the unit circle between $-A$ and $+A$. The folding $z' = \frac{(z+\frac{1}{z})}{2}$ transforms the arc into the real interval $[\cos A, 1]$. The dual interval $[a, 1/a]$ folds into $[1, \sec A]$ when $a = \tan(\pi/4 - A/2)$. Hence, if we first compute the optimum parameters over $[a, 1/a]$ for J iterations, these parameters will fold into parameters over the interval $[1, \sec A]$ which give the proper Chebyshev alternating extremes property. The reciprocal of these parameters will retain this property over the interval $[\cos A, 1]$. The inverse transformation back to the arc will then yield the optimum parameters over the arc! When J is odd, the real parameter at angle $A(j) = 0$ is used only once and all the other parameters on $[\cos A, 1]$ transform back into the $\pm$ angles on the arc.

The recipe for computing the optimum $A(j)$ for $j = 1, \ldots, J$ derived on this basis is

$$k' = \tan^2(\pi/4 - A/2), \tag{19.1}$$

$$z(j) = \frac{1}{\sqrt{k'}} dn\left[\frac{(2j-1)K}{2J}, k\right], \tag{19.2}$$

$$w(j) = \frac{1}{2}\left[z(j) + \frac{1}{z(j)}\right],$$

$$j = 1, 2, \ldots, \text{ integer part of } \frac{1+J}{2}, \tag{19.3}$$

$$A(2j-1) = \arccos\frac{1}{w(j)}, \tag{19.4}$$

$$A(2j) = -A(2j-1), \tag{19.5}$$

$$w(j) = \exp[iA(j)]. \tag{19.6}$$

When J is odd, the value $A[(1 + J)/2] = 0$ is not repeated. This technique generalizes to elliptic-function spectra. The actual and dual spectra fold into reciprocal spectra with m' for the dual spectrum > 1 when m for the actual spectrum is < 1. The algebra is not trivial, but the resulting equations are easily verified. The duality relationships are remarkable. They highlight an elegant application of classical analysis to a crucial problem of numerical analysis.

The elliptic spectrum is defined by the triplet $\{a, b, A\}$. The dual elliptic spectrum is defined by the triplet $\{a', 1/a', A'\}$, with $A' = B$ of Eq. 16 and

$$a' = \tan\left(\frac{\pi}{4} - \frac{A}{2}\right). \tag{20}$$

Substituting this value for a' into Eq. 16, we find that

$$B' = A \tag{21}$$

and therefore

$$m' = \frac{2\cos^2 B}{\cos^2 A} - 1 \tag{22}$$

must be greater than 1 when $m < 1$. We use m' in place of m in Eq. 18.1 and compute the optimum real parameters $\{w'(j)\}$ for the dual problem. The corresponding parameters for the actual elliptic spectrum may then be computed from

$$\cos A(j) = \frac{2}{w'(j) + \frac{1}{w'(j)}}$$

$$\text{for } j = 1, 2, \ldots, \text{integer part of } \frac{1+J}{2}, \tag{23}$$

$$w(2j-1) = \sqrt{ab}\ \exp[iA(j)], \tag{24}$$

and

$$w(2j) = \sqrt{ab}\ \exp[-iA(j)]. \tag{25}$$

When J is odd, Eq. 25 is dropped for $2j = J + 1$ and the last value computed by Eq. 24 is $w(J) = \sqrt{ab}$.

This is illustrated with $\{a, b, A\} = \{0.1, 1.0, 60°\}$. We compute $\cos^2 B = 2/(1.0 + 10.1/2) = 0.3306$ and since this is greater than $\cos^2 60° = 0.25, m = 0.5125 < 1$. For the dual problem, $m' = 2(.3306)/0.25 - 1.0 = 1.645$ and $k' = 0.3389$. For $J = 2$ we compute $w'(1) = 0.685114$ and $w'(2) = 1.45961$, so $2/[w'(1) + 1/w'(1)] = 0.93252$ and $A(1) = \arccos(0.93252) = 0.3695$ rad. The optimum parameters are therefore $w(1) = 0.316\exp(0.3695i)$ and $w(2) = 0.316\exp(-0.3695i)$.

The range of elliptic-function domains for which we have now developed a theory for computing nearly optimum parameters is displayed in Fig. 4.3.

Rouché's theorem establishes that our elliptic-function parameters are optimal for a scalloped region which approaches the elliptic-function region as J increases. Thus, the parameters are "nearly optimal" in the same sense that Chebyshev parameters are nearly optimal for polynomial extrapolation. We now consider the possibility that the elliptic-function parameters may be optimal even when J is small. We have already described how Chebyshev polynomials are not optimal for polynomial approximation to zero over elliptic spectra in certain cases. Does this result carry over into rational approximation with elliptic functions over elliptic domains? Although it may be true that there are cases where the elliptic functions are suboptimal, it is easily shown that the polynomial result does not apply.

Theorem 12 (Optimality of elliptic parameters when $J=2^n$). *When $J = 2^n$, the elliptic-function parameters are optimal over elliptic-function spectral domains.*

Proof. The algorithm for obtaining optimal parameters over a real spectral interval when $J = 2^n$ applies to elliptic-function spectra. Successive Landen transformations (N-7) and renormalizations reduce the number of parameters to one, and this optimal parameter is the geometric mean of the endpoints of the real intercept. The back transformations yield the usual elliptic-function parameters. □

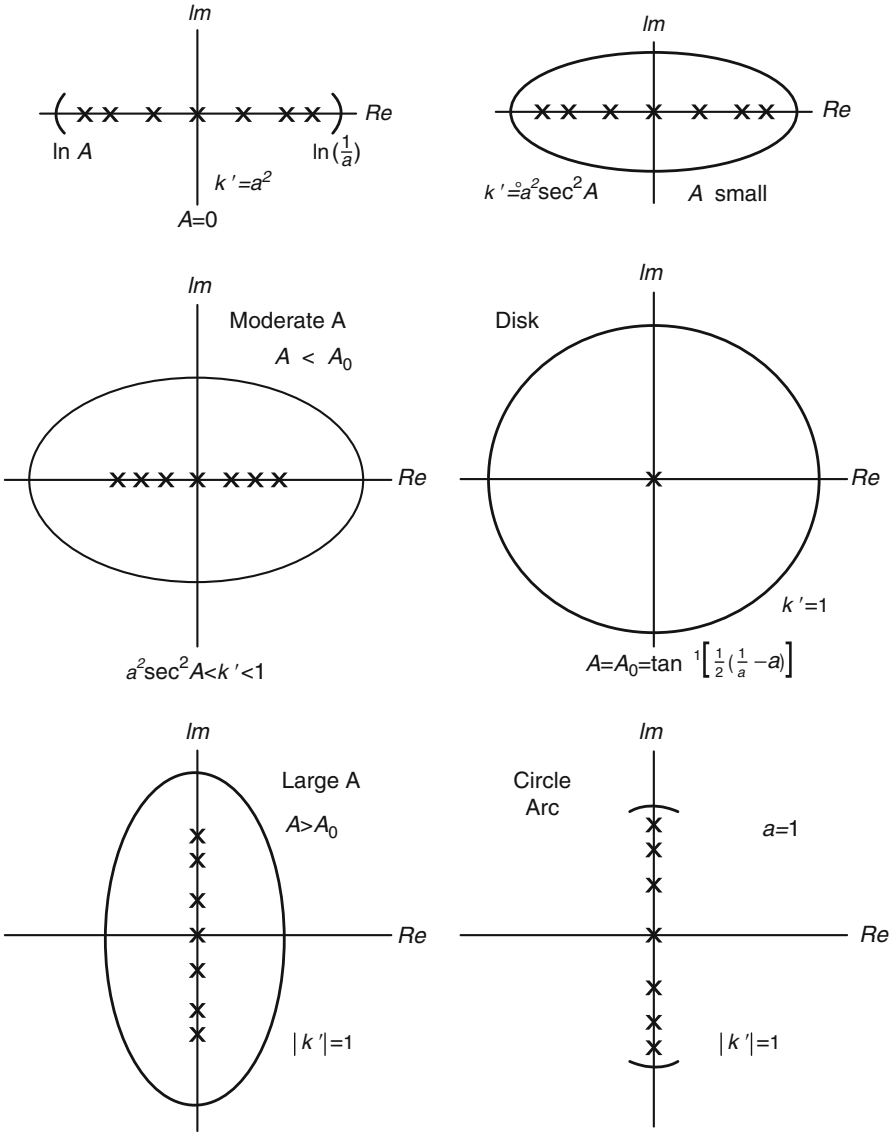

Fig. 4.3 Log of elliptic-function spectra ($0 \leq A \leq \pi/2$) x = locations of optimum iteration parameters for $J = 7$

Theorem 12 follows from the logarithmic symmetry of the rational approximation. There is no corresponding symmetry argument for the polynomial approximation. Note that the Chebyshev polynomial is optimal when the normalization point is far enough away from the spectrum. The logarithm of the ADI normalization point is at infinity. Although it may be true that the elliptic-function parameters are optimal for all J, this has yet to be proved.

4.5 Error Reduction

Several approximations were made in the early analysis of error reduction associated with optimal parameters for complex elliptic-region spectra. We now consider more precise estimates. When k' is complex, we consider the dual problem with the same rate of convergence and real k' in $(0, 1]$. Referring to Eq. 5 and noting that the maximum absolute value of the error function occurs when $z = 1 + ir$, we find that the precise bound is R^2 where

$$R = |\sqrt{k(J)}sn\,\{[1 + 2J(1 + ri)]K(J),\, k(J)\}|. \tag{26}$$

By Table N-4.3, this reduces to

$$R = \sqrt{k(J)}cd\,[2JriK(J),\, k(J)], \tag{27}$$

and Jacobi's imaginary transformation in Table N-6.1 yields

$$R = \sqrt{k(J)}nd\,[2JrK(J),\, k'(J)]. \tag{28}$$

We recall that

$$\frac{K'(J)}{K(J)} = 4J\,\frac{K'}{K}. \tag{29}$$

Hence,

$$R = \sqrt{k(J)}nd\left[\frac{rK}{2K'}K'(J),\, k'(J)\right]. \tag{30}$$

To evaluate R, we must first evaluate rK/K'. By Table N-4.3,

$$sn\,[(K' - rK),\, k'] \equiv cd\,(rK,\, k') \equiv \frac{cn\,(rK,\, k')}{dn\,(rK,\, k')}. \tag{31}$$

By Eqs. 11 and 12,

$$cd\,(rK,\, k') = \sqrt{\frac{a}{bk'}}, \tag{32}$$

and we may compute $K' - rK$ with the AGM$(1, k)$ algorithm in Sect. 1.6, starting with $\phi_0 = \arcsin\sqrt{\frac{a}{bk'}}$ and using Eq. 51 of Chap. 1to compute ϕ_N. We have

$$K' - rK = \frac{\phi_N}{2^N a'_N} \text{ with } \phi \text{ in radians}$$
$$= \frac{\phi_N \pi}{2^N (180) a'_N} \text{ with } \phi \text{ in degrees.}$$

Since $K' = \pi/2a'_N$, we obtain when ϕ is expressed in degrees

$$v \equiv 1 - \frac{rK}{K'} = \frac{\phi_N}{90 \cdot 2^N}. \tag{33}$$

We now define

$$y \equiv \frac{1}{2}(1 - v) = \frac{rK}{2K'} \tag{34}$$

and

$$w \equiv q(J) = q^{4J}. \tag{35}$$

We may compute the nd-function in Eq. 30 with the AGM algorithm to any desired accuracy. However, there is a simple approximation to R which seems adequate for virtually all applications. We substitute the approximation to the dn-function given in Eq. 56 of Chap. 1 into Eq. 30 to obtain

$$R \doteq w^{\frac{1-2y}{4}} \frac{1 + w^y + w^{2-y}}{1 + w^{1-y} + w^{1+y}}. \tag{36}$$

The value for q may be approximated as described in Sect. 1.6. Equations 50.1–50.3 of Chap. 1 are often suitable for this purpose. We recall that the error reduction is equal to R^2. As J increases, R approaches q^{vJ} and the asymptotic rate of convergence is $\rho_\infty = q^{2v}$.

The procedure will now be illustrated for the spectrum $\{a, b, A\} = \{0.1, 1.0, 45°\}$, for which we have already determined that $k' = 0.26414$. We compute $k = 0.96448$ and $\phi_0 = \arcsin(\sqrt{\frac{a}{bk'}}) = 37.973°$. The AGM algorithm converges to five digit accuracy after two steps:

Table 4.1 An AGM table

n	$a(n)$	$b(n)$	$c(n)$	$b(n)/a(n)$	ϕ_n
0	1.0	0.96448	0.26414	0.96448	37.973°
1	0.98224	0.98208	0.01776	0.999837	74.946°
2	0.98216	0.98216	0.00008	1.0000	149.89°

Thus, $v = 149.89/360 = 0.416361$ and $y = (1 - v)/2 = 0.29182$. The approximations in Eqs. 50 of Chap. 1 yield

$$z \doteq \frac{1}{2}\frac{1-\sqrt{k}}{1+\sqrt{k}} = 0.00452,$$
$$q' \doteq z(1+z^4) \doteq 0.00452$$

and Eq. 48 of Chap. 1 yields $q = \exp[\frac{\pi^2}{\ln q'}] = 0.16074$. Setting $w = q^{4J}$, we compute R with Eq. 36 for $J = 1, 2, 4$ and compare with truth determined with the optimal parameters and the actual rational function evaluated at the point $x = 0.1$.

Table 4.2 Convergence rate estimates

J	1	2	4
R(Truth)	0.5195	0.2213	0.04763
R(Eq. 36)	0.4987	0.2208	0.04763
q^{vJ}	0.46715	0.2182	0.04763

The last row indicates how rapidly the asymptotic convergence rate is attained when k' is not close to zero. This row was computed with $q^v = 0.46715$ (Table 4.2).

It is instructive to examine some limiting cases. When $\frac{a}{b} << 1$ and angle A is small, we set $k' = p\frac{a}{b}$ in Eq. 15 and find that $p = \sec^2 A$ as in the early result in Eq. 7. It follows that $\sqrt{\frac{a}{bk'}} = \cos A$. Moreover, when $k' << 1$, $sn(z, k')$ may be approximated (N-10.4) by $\sin(z)$. From Eqs. 31 and 32, we have

$$sn\,[(K' - rK),\, k'] = \sqrt{\frac{a}{bk'}} \doteq \cos A \doteq \sin(K' - rK). \tag{37}$$

In this case, $K' \doteq \frac{\pi}{2}$ and it follows that

$$rK \doteq A. \tag{38}$$

The asymptotic convergence rate is now obtained by allowing J to increase until $k(J) << 1$. By N-10.9, $nd(z) < \cosh(z) < e^z$. Equation 28 yields

$$R(J) < 2\exp\left[-\frac{\pi JK'}{K}\left(1 - \frac{2A}{\pi}\right)\right]. \tag{39}$$

Although K' remains close to $\pi/2$ as A increases slightly from 0, the value for K decreases from $\ln\frac{4b}{a}$ to $\ln\frac{4b\cos^2 A}{a}$. Thus, as A increases the number of iterations required for prescribed error reduction increases by a factor of

$$\frac{J(A)}{J(0)} \doteq \frac{1 + \frac{2\ln(\cos A)}{\ln(4/k')}}{1 - \frac{2A}{\pi}}. \tag{40}$$

For small k', the decrease in K is insignificant and the ratio is approximately $(1 - 2A/\pi)^{-1}$. Even though the approximation applies to small A, we observe that an increase by a factor of two occurs when A is near $\pi/4$. Relatively large complex

components do not seem to be as detrimental to ADI iteration as corresponding components are to polynomial approximation.

We may also consider convergence as the elliptic-function region approaches a disk, in which case k' approaches unity and both K and $K(J)$ are close to $\pi/2$. By N-10.7, $sn(z, k')$ may be approximated well by $\tanh(z)$, and we obtain the approximation

$$K' - rK = \operatorname{arc\,tanh} \sqrt{\frac{a}{bk'}}. \tag{41}$$

Now even for small J, $k'(J)$ is close to unity and

$$nd\,[2rJK(J),\, k'(J)] \doteq \cosh[2rJK(J)] \doteq \cosh(rJ\pi). \tag{42}$$

For J sufficiently large,

$$\sqrt{k(J)} \doteq 2\exp\left(-\frac{\pi JK'}{K}\right) \tag{43}$$

and the approximation $\cosh[\cdot] \doteq \exp[\cdot]/2$ yields

$$R(J) \doteq \exp\left[-\frac{\pi J}{K}(K' - rK)\right] \doteq \exp[-2J(K' - rK)]. \tag{44}$$

Substituting Eq. 41 into Eq. 44 and using the identity

$$\operatorname{arc\,tanh} u = \ln\left(\frac{1+u}{1-u}\right)^{\frac{1}{2}}, \tag{45}$$

we obtain

$$R(J) = \left[\frac{1 - \sqrt{\frac{a}{bk'}}}{1 + \sqrt{\frac{a}{bk'}}}\right]^J. \tag{46}$$

The example with $(a, b, A) = (0.1, 1.0, 45°)$ is not far from a disk, and Eq. 46 yields an approximate asymptotic convergence rate of $\rho_\infty = 0.23816$ and $R(J) \doteq (0.488)^J$, which may be compared with the correct asymptotic value of $(0.46715)^J$.

When $k' = 1$, the region is a disk and $R(J)$ attains the correct limit of

$$R(J) = \left[\frac{1 - \sqrt{\frac{a}{b}}}{1 + \sqrt{\frac{a}{b}}}\right]^J. \tag{47}$$

Added in 2012: The maximum for $|R|$ on the boundary of the elliptic-function region occurs at $w = 1 + ir$ as in Eq. 26. The minimum absolute value on the boundary occurs when $w = 0.5 + ir$, in which case Table N-4.3 yields $R(u = 0.5) = \sqrt{k(J)}sc[2JrK(J), k'(J)]$. By Rouché's theorem, the optimum

parameters for the elliptic-function region cannot result in a lower value for $|R|$. For significant error reduction $k(J)$ is small and $k'(J)$ is close to unity. The elliptic-function region may be approximated by hyperbolic functions (cf. N-10.7-9) and the ratio of the minimum to maximum over the boundary is $\tanh[2rJK(J)]$. Here $K(J) \doteq \pi/2$ and the ratio is $\tanh(rJ\pi)$. When $k' << 1$, by Eq. 38 $r = \frac{A}{K}$ and the greatest loss in error reduction through use of the elliptic rather than optimum parameters is bounded by

$$L \equiv \left| \frac{R^2(u = .5)}{R^2(u = 1)} \right| = \tanh^2 \left(\frac{\pi A J}{K} \right).$$

By Eq. 8, the error reduction is $\varepsilon = \exp[-\frac{\pi J(\pi - 2A)}{K}]$, from which we find that

$$\frac{\pi J}{K} = \frac{\ln \frac{1}{\varepsilon}}{\pi(1 - \frac{2A}{\pi})}.$$

When A is small this analysis is not relevant since $A = 0$ and $L = 0$. However, the elliptic parameters are optimum for this real spectrum. When A is sufficiently large we define

$$\zeta = \frac{\frac{2A}{\pi}}{(1 - \frac{2A}{\pi})}$$

and observe that for significant error reduction $L = 1 - 2\varepsilon^{\zeta}$. For example, when $A = \pi/6, \zeta = 1/2$ and $L = 1 - 2\sqrt{\varepsilon}$ are close to unity. The elliptic parameters are close enough to optimum to justify embedding an actual spectrum in an elliptic-function region whenever feasible without undue enlargement of the spectrum. Algorithms for generating truly optimum parameters cannot yield significant improvement for such regions and may be quite complicated and time consuming. In practice the spectrum may not be known well enough to preclude embedding in a conservative elliptic-function region.

4.6 The Two-Variable Problem

4.6.1 Generalized Spectral Alignment

A need for treating complex spectra first arose with the discovery that the Lyapunov matrix equation $AX + XA^T = C$ is a model ADI problem when matrix C is SPD and the eigenvalues of matrix A are in the positive-real half plane. For this application the two spectral regions are the same. We have just exposed theory for treating this case. The Sylvester matrix equation is $AX + XB = C$, where A is a given $m \times m$ real matrix, B is a given $n \times n$ real matrix, C is a given $m \times n$ real matrix, and the $m \times n$ real matrix X is to be determined. This is

an ADI model problem when the eigenvalues λ and γ of matrices A and B satisfy $\min[Re\lambda(A)] + \min[Re\gamma(B)]>0$. In general, the eigenvalues of A and B are complex and the spectra of these matrices may differ widely. Our goal is to generalize the real transformation in order to align two complex spectra of this type. We first apply a WBJ transformation to align the real intercepts at $[k', 1]$ and normalize to $[\sqrt{k'}, 1/\sqrt{k'}] \equiv [e, 1/e]$. We may then treat eigenvalues subtending large angles (e.g., greater than 1 rad) discretely, thus removing them from the spectra to be aligned. The maximum angles of the remaining spectra may then be found as θ_1 and θ_2 in $S_1 = (e, 1/e, \theta_1)$ and $S_2 = (e, 1/e, \theta_2)$. If optimal parameters can be found for these spectra, they may be transformed back to parameters for the actual spectra. By Eqs. 49.1 and 49.2, further alignment is needed when the angles differ.

The optimum parameters for J iterations over spectrum $S = (e, 1/e, \theta)$ are given in Eq. 18.2 as

$$w_j = \frac{1}{\sqrt{k'}} dn\left[\frac{2j-1}{2J}K, k\right], \quad j = 1, 2, \ldots, J, \tag{48}$$

where k' depends on e and θ as shown in Eqs. 16–18.1. Useful parameters for this analysis are defined as

$$f_e \equiv \frac{1}{2}\left(e + \frac{1}{e}\right), \tag{49.1}$$

$$\zeta \equiv f_e \cos\theta. \tag{49.2}$$

Then it is easily shown that k' may be computed with Eqs. 18.1 and 18.2 from ζ, when one observes that

$$m = 2\zeta^2 - 1, \tag{49.3}$$

$$k' = \frac{1}{m + \sqrt{m^2 - 1}}. \tag{49.4}$$

Having aligned the real intercepts of our two spectra, we observe that when $\theta_1 \neq \theta_2$ they do not share the same optimum parameters. The ratio $\cos\theta_1/\cos\theta_2$ is a measure of the disparity of the two spectra. If this ratio is close to unity, we may choose parameters for the larger of the two angles. Each parameter w_j may then be transformed back to yield the corresponding values for p_j and q_j. In general, the two angles will differ. We seek another transformation to align the spectra. To this end, we first establish a relationship between s and t such that there is a Jordan-type transformation which maps $S_1 = (e, 1/e, \theta_1)$ onto $S_1(s) = (s, 1/s, \psi_1)$ and $S_2 = (e, 1/e, \theta_2)$ onto $S_2(t) = (t, 1/t, \psi_2)$. In an attempt to accomplish this objective, we repeat the analysis in Chap. 2 with a few simple modifications.

4.6.2 A One-Parameter Family of Spectral Pairs

We first define

$$K = \begin{bmatrix} s & 0 \\ 0 & 1/s \end{bmatrix}, \quad L = \begin{bmatrix} t & 0 \\ 0 & 1/t \end{bmatrix}, \quad A = \begin{bmatrix} 1 & -e \\ 1 & -1/e \end{bmatrix}, \text{ and } F = \begin{bmatrix} 1 & e \\ 1 & 1/e \end{bmatrix}. \tag{50}$$

Then the matrix C in Eqs. 2–10 is replaced by

$$C = \begin{bmatrix} KA & A \\ LF & -F \end{bmatrix} = \begin{bmatrix} KAF^{-1} & 0 \\ L & -I \end{bmatrix} \begin{bmatrix} F & FA^{-1}K^{-1}A \\ 0 & (F + LFA^{-1}K^{-1}A) \end{bmatrix}. \tag{51}$$

Now we define $G \equiv FA^{-1}K + LFA^{-1}$ and note that G can be singular only when

$$\det \begin{bmatrix} (s+t)(e+1/e) & -2e(t+1/s) \\ 2(s+1/t)/e & -(e+1/e)(1/s+1/t) \end{bmatrix} = 0. \tag{52}$$

We determine that Eq. 52 is satisfied when

$$t = \frac{1 - sf_e}{f_e - s}. \tag{53}$$

Algebra identical to that used in Chap. 2 now yields corresponding values for the transformation parameters of

$$\alpha' = \delta' = 1 - es, \qquad \beta' = \gamma' = e - s, \tag{54}$$

Since in this application $\alpha' > 0$, we may normalize to $\alpha = 1$. We identify S_1 as the spectrum with the smaller angle. Then $s > e$ and we define the normalized positive $\beta \equiv -\beta'/(1 - es) = (s - e)/(1 - es)$ to obtain the transformations

$$z_1 = \frac{w_1 - \beta}{1 - \beta w_1} \text{ and } z_2 = \frac{w_2 + \beta}{1 + \beta w_2}. \tag{55}$$

These transformations map the unit circle into the unit circle.

We now seek a value for β such that

$$\zeta_1 \equiv f_s \cos \psi_1 \tag{56}$$

and

$$\zeta_2 \equiv f_t \cos \psi_2 \tag{57}$$

with $\zeta_1 = \zeta_2$. If this is possible, then the transformed spectra are aligned and optimal parameters for these spectra may be transformed back.

4.6.3 Transformation from [e/1/e] to [s,1/s] and [t,1/t]

The inverses of the transformations in Eq. 55 are

$$w_1 = \frac{z_1 + \beta}{1 + \beta z_1} \quad \text{on } S_1 \quad and \qquad w_2 = \frac{z_2 - \beta}{1 - \beta z_2} \quad \text{on } S_2. \tag{58}$$

We recall that these transformations leave the unit circle invariant. We now prove that the transformed spectra remain elliptic-function regions.

Theorem 13. *If*

$$z(u + iv) \equiv \frac{1}{\sqrt{k_0}} dn\,[(u + iv)K_0,\, k_0] \text{ and}$$

$$w(u + iv) \equiv \frac{z + \beta}{1 + \beta z},\ \text{then}$$

$$w(u + iv) = f(u + iv) \equiv \frac{1}{\sqrt{k}} dn\,[(u + iv)K,\, k],$$

where k is uniquely determined by β and k_0.

Proof. We first observe that $f(1/2) = w(1/2) = 1$. The real and imaginary periods of f and w are the same. If we can choose k so that they have the same zeros and poles, the functions are the same. We have $w = 0$ when $u = 1$ and $v = r_0$, where r_0 is determined from $dn[(1 + ir_0)K_0,\, k_0] = nd(ir_0K_0,\, k_0) = cd(r_0K_0,\, k'_0) = -\beta\sqrt{k_0}$. If $\beta = 0, ir_0 = \tau_0$ and $r_0 = K'_0/K_0$. If $\beta > 0, ir_0 > \tau_0$, and if $\beta < 0, ir_0 < \tau_0$. The value for k is chosen so that $dn[(1 + ir_0)K,\, k] = 0$. This is true when $ir_0 = \tau$. Now $dn(ir_0K,\, k) = dn(iK',\, k) = \infty$ so that ir_0 is a pole of $f(u + iv)$. We now note that

$$dn(ir_0K_0,\, k_0) = \frac{k_0}{dn[(1 + ir_0)K_0,\, k_0]} = -\frac{\sqrt{k_0}}{\beta},$$

so that $z(ir_0) = -1/\beta$ and

$$w(ir_0) = \frac{\beta - \frac{1}{\beta}}{1 - \beta\frac{1}{\beta}} = \infty.$$

Thus w and f have the same poles. Both functions are elliptic, their ratio is unity at one point, they have the same periods, and they have the same zeros and poles. It follows that they are equal. □

Having established that the transformed region remains elliptic, we need not determine r_0 to evaluate k. We need only compute the angle subtended by the unit circle and compute k' from Eq. 49.

Inspection of Eq. 58 reveals that

$$p \equiv (1 + st)/(s + t) \tag{59}$$

is an invariant of these transformations.

The transformation to real intervals $[s, 1/s]$ and $[t, 1/t]$ is accomplished with

$$\beta = (s - e)/(1 - se). \tag{60}$$

The angles for the transformed spectra are determined as

$$\theta_s = \arccos\left[\frac{(1 + \beta^2)\cos\theta_1 + 2\beta}{(1 + \beta^2) + 2\beta\cos\theta_1}\right] \tag{61}$$

and

$$\theta_t = \arccos\left[\frac{(1 + \beta^2)\cos\theta_2 - 2\beta}{(1 + \beta^2) - 2\beta\cos\theta_2}\right]. \tag{62}$$

We now define

$$f_s = \frac{1}{2}\left(s + \frac{1}{s}\right) \qquad \text{and} \qquad f_t = \frac{1}{2}\left(t + \frac{1}{t}\right). \tag{63}$$

Then the values for ζ are

$$\zeta_s = f_s \cos\theta_s \qquad \text{and} \qquad \zeta_t = f_t \cos\theta_t. \tag{64}$$

The spectra are aligned when $\zeta_s = \zeta_t$. We associate S_1 with the spectrum for which ζ is larger and seek a value for β which will yield $\zeta_s = \zeta_t$.

4.6.4 *Alignment When $\zeta_s > 1$ and $\zeta_t > 1$*

We attempt alignment by increasing s. Invariance of p in Eq. 59 establishes that t decreases according to

$$t = (1 - sp)/(p - s). \tag{65}$$

Thus, as s approaches $1/p$, t approaches zero. It appears that we may decrease t until the spectra are aligned, but this is not always possible. Since $\beta > 0$, the numerator in the expression for θ_t in Eq. 62 decreases as s increases so that the decrease in $\cos\theta_t$ can dominate the increase in f_t. We first show that the spectra may be aligned when $\zeta_1 > \zeta_2 > 1$

Theorem 14. *If $\zeta_2 > 1$, then θ_t is bounded away from $\pi/2$ for $0 < t < t_0$.*

Proof. As s increases, θ_t increases to its maximum at $s = 1/p$ at which point $\beta = t_0$. Substituting this value for β into Eq. 62, we find that

$$\cos\theta_{t=0} = \frac{(1+t_0^2)\cos\theta_2 - 2t_0}{(1+t_0^2) - 2t_0\cos\theta_2} = \frac{\zeta_2\cos\theta_2 - 1}{\zeta_2 - \cos\theta_2},$$

which is in the interval $(0, 1]$. Hence,

$$\theta_t < \theta_{t=0} < \frac{\pi}{2}, \qquad 0 < t < t_0. \tag{66}$$

□

It follows that as s is increased, $\zeta_s = f_s\cos\theta_s < f_s < f_{s_0}$ and $\zeta_t = f_t\cos\theta_t > f_t\cos\theta_{t=0}$. Thus, as t approaches zero f_t increases until at some point in $(0, t_0)$ $\zeta_t = \zeta_s$. The spectra can always be aligned when ζ_1 and ζ_2 are both greater than one. Let

$$\tau \equiv \frac{\beta + \frac{1}{\beta}}{2}. \tag{67}$$

Then

$$\begin{aligned} s + \frac{1}{s} &= \frac{e+\beta}{1+e\beta} + \frac{1+e\beta}{e+\beta} \\ &= \frac{(1+e^2)(1+\beta^2) + 4e\beta}{(1+e^2)\beta + e(1+\beta^2)} \\ &= \frac{2(f_e\tau + 1)}{\tau + f_e}. \end{aligned}$$

A similar expression applies to $(t + 1/t)$ with τ replaced by $-\tau$. We have shown that

$$f_1 = \frac{(f_e\tau + 1)}{\tau + f_e}, \tag{68.1}$$

$$f_2 = \frac{(f_e\tau - 1)}{\tau - f_e}. \tag{68.2}$$

If we define $c_1 = \cos(A_1)$ and $c_2 = \cos(A_2)$, then Eq. 62 yield

$$\cos(B_1) = \frac{c_1\tau + 1}{\tau + c_1}, \tag{69.1}$$

$$\cos(B_2) = \frac{c_2\tau - 1}{\tau - c_2}. \tag{69.2}$$

Hence, $\zeta_1 = \zeta_2$ when

$$(f_e\tau + 1)(\tau - f_e)(c_1\tau + 1)(\tau - c_2) = (f_e\tau - 1)(\tau + f_e)(c_2\tau - 1)(\tau + c_1). \quad (70)$$

We define

$$\phi \equiv \frac{\frac{c_1+c_2}{2}(f_e - \frac{1}{f_e}) - (1 - c_1c_2)}{c_1 - c_2}. \quad (71)$$

Then Eq. 70 reduces to

$$\tau^4 - 2\phi(\tau^3 - \tau) - 1 = 0 \quad (72)$$

or

$$(\tau^2 - 1)(\tau^2 - 2\phi\tau + 1) = 0. \quad (73)$$

We seek a value of β which is less than e. By Eq. 67, τ must be greater than unity. Thus, the roots $\tau = \pm 1$ are extraneous. We next demonstrate that $\phi > 1$: Let $c_1 = (1 + r)c_2,\ r > 0$, and $f_e c_2 = 1 + p,\ p > 0$. Then

$$\phi = \frac{(1 + \frac{r}{2})(1 + p - \frac{c_2^2}{1+p}) - 1 + (1 + r)c_2^2}{rc_2}$$

$$> \frac{p + \frac{r}{2} + \frac{r}{2}c_2^2}{rc_2} > \frac{1}{2}\left(c_2 + \frac{1}{c_2}\right) > 1.$$

As p approaches zero, ϕ approaches $\frac{1}{2}(c_2 + \frac{1}{c_2})$. The only root of Eq. 73 greater than unity is thus

$$\tau = \phi + \sqrt{\phi^2 - 1}. \quad (74)$$

From Eq. 67, we obtain

$$\beta = (\tau + \sqrt{\tau^2 - 1})^{-1}. \quad (75)$$

By way of illustration, consider $S_1(0.1, 10, 0°)$ and $S_2(0.1, 10, \cos^{-1}\frac{1}{f_{0.1}})$, where $f_{0.1} = (0.1 + 10)/2 = 5.05$. The optimum single parameter is $w = 1$ with associated error reduction of $R_1 = R_2 = \frac{1-0.1}{1+0.1} = \frac{0.9}{1.1}$. Hence, $R = R_1R_2 = \frac{0.81}{1.21}$ is the error reduction when $J = 1$. The transformation with $\beta = 0.1$ yields $s = \frac{0.1+0.1}{1+0.01} = \frac{0.2}{1.01}$ with $S_1(s, 1/s, 0°)$. The error reduction over the infinite disc is $R_2 = 1$ and the error reduction over the transformed region 1 is $R_1 = \frac{1-s}{1+s} = \frac{0.81}{1.21}$. The back transformation of $w = 1$ remains at $w = 1$ and it is no surprise that this optimum single parameter is invariant. Although R_1 and R_2 change, the product is invariant when we back transform.

When $J = 1$ there is no need for further alignment. For $J > 1$, however, choice of optimum parameters is facilitated by the transformation. Consider the case of $J = 2$. Repeated use of $w = 1$ squares the reduction to $R = 0.448$. The optimum two parameters over region S_1 yield $R_1 = 0.384$ (determined as described in Chap. 1) while R_2 is greater than the value obtained for the disc with its optimum parameters of $w_1 = w_2 = 1$ which is 0.669 so that $R > 0.669 \times 0.384 = 0.257$,

with a possible improvement over $w_1 = w_2 = 1$. Improvement is guaranteed by transforming to determine the truly optimum elliptic-function parameters. The transformed interval for S_1 is (0.198,5.05) for which the optimum two parameters yield $R_1 = 0.2366$. Although $R_2 = 1$, the product is now $R = 0.2366$. As the cycle length J increases, greater improvement is achieved through use of the optimum cycle in the transformed space.

In general, the product $R = R_1 R_2$ remains fixed after the back transformation even though the individual values change. Error reduction in the presence of complex spectra is analyzed in Sect. 4.5. Two parameters play a crucial role. One is the nome, q, of the elliptic-function region. This is a function of ζ only and is the same for both spectra after alignment. The other, $v \in (0, 1]$, depends on both the real interval and the angle. The number of iterations needed to yield a prescribed error reduction varies inversely as v, which is a measure of the retardation in convergence as the angle increases. An approximate value is $v \simeq (1 - A°/90°)$, and the precise value is computed as described on pp.77–78. Asymptotic error reduction per iteration (as J increases) varies as

$$R^{1/J} \sim q^{(v_1+v_2)}. \tag{76}$$

A prescribed error reduction ε is obtained by choosing

$$J > \frac{\ln \frac{\varepsilon}{4}}{(v_1 + v_2) \ln q}. \tag{77}$$

For this J, a more accurate estimate of the error reduction is given by $R = R_1 R_2$ where

$$R_k = q^{v_k J} \frac{1 + q^{2J(1-v_k)}}{1 + q^{2J(1+v_k)}}. \tag{78}$$

When the spectra can be embedded in elliptic-function regions the parameters determined by this method are close to optimum. Comparison with parameters determined by methods described by [Istace] and [Starke] should be of great interest.

4.6.5 Alignment When One Spectrum Is a Disk

We now address the case where either value is equal to unity. The corresponding spectrum is a disk. It is known that the transformations in Eq. 58 retain a disk spectrum. We will prove this while establishing properties useful in alignment.

Theorem 15. *The transformation in Eq. 58 transforms a disk, which is characterized by $\zeta = 1$, into a disk. Thus, $\zeta = 1$ is invariant.*

Proof. The theorem may be established with either the transformation or its inverse. We choose $s > s_0$ so that $\beta > 0$, and replacing $\cos\theta$ by ζ/f in Eq. 62, we have

$$\cos\theta_s = \frac{f + \frac{\zeta}{2}(\beta + 1/\beta)}{\zeta + \frac{f}{2}(\beta + 1/\beta)}, \tag{79}$$

where $f = f_{s_0}$ and $\zeta = \zeta_{s_0}$. Hence,

$$\zeta_s = f_s \cos\theta_s = f_s \frac{f + \frac{\zeta}{2}(\beta + 1/\beta)}{\zeta + \frac{f}{2}(\beta + 1/\beta)}.$$

One can apply Eq. 61 to establish the identity

$$\frac{1}{2}(\beta + 1/\beta) = \frac{f_s f - 1}{f - f_s}.$$

It follows that

$$\zeta_s = f_s \left[\frac{f + \zeta\left(\frac{f_s f - 1}{f - f_s}\right)}{\zeta + f\left(\frac{f_s f - 1}{f - f_s}\right)} \right], \tag{80}$$

and when $\zeta = 1$, $\zeta_s = f_s \frac{f^2 - 1}{f_s f^2 - f_s} = 1$. □

When either spectrum is a disk, the two-variable problem is reduced to one variable by a simple but elegant method. As s approaches $1/p$, the radius of the disk increases. In the limit, when we choose optimal parameters for $S_1(1/p)$, the error function has absolute value unity on the boundary of the infinite disk. Hence, the back transformation will have constant absolute value on the boundary of S_2 and theory establishes this as sufficient for simultaneous parameter optimization for the two spectra. When $s = 1/p$, $\beta = t_0$ and angle θ_s is determined with Eq. 62.

It is instructive to consider the alignment equations when $p = 0$. From Eq. 73, $\phi = (\tau + \frac{1}{\tau})/2$, $\tau > 1$. Hence, $\tau = \frac{1}{c_2} = f_e$. From Eq. 69.2, $B_2 = 90°$. Also, from Eq. 75, $\beta = (f_e + \sqrt{f_e^2 - 1})^{-1} = (\frac{e+1/e}{2} + \sqrt{(\frac{1/e-e}{2})^2})^{-1} = e$.

When both spectra are disks, the transformation leaves $S_1(1/p)$ as a disk with optimal iteration parameters all equal to unity. The back transformation yields optimal values for **p** and **q** for the two disks.

Careless application of the theory can lead to erroneous conclusions. One pitfall will now be illustrated by example. Let S_1 be the line $\cos\theta + i \sin\phi$, $|\phi| \leq \theta$. Let S_2 be a disk with real intercept $[\cos\alpha, \sec\alpha]$. The optimum single parameter for both spectra is unity. The corresponding error reduction is

$$R = \left(\frac{1 - \cos\alpha}{1 + \cos\alpha}\right) \tan\frac{\theta}{2}.$$

Note that as θ approaches zero the line shrinks to a point and R approaches zero. We may subtract $\cos\theta$ from S_1 and add $\cos\theta$ to S_2. Then for any real parameter the absolute value of the error reduction is unity along the shifted line which now falls on the imaginary axis. If we choose as the parameter the square root of the endpoints of the real intercept of the translated disk, then the error reduction has a constant absolute value on the boundary of the disk. As θ approaches zero, the translation approaches unity and the error reduction over the disk approaches

$$R = \left(\frac{1-\sqrt{\cos\alpha}}{1+\sqrt{\cos\alpha}}\right).$$

This is not zero and certainly not optimal for these spectra. Yet the error reduction has constant absolute value along both boundaries. The flaw in the argument is that the roots of the error function do not lie within spectrum S_1. Rouché's theorem cannot be applied. The line in this example is not an elliptic-function region. The analysis in this section applies only to spectra embedded in elliptic-function regions.

When θ approaches $\pi/2$, error reduction along the line is slight and the shift of S_1 to the imaginary axis yields nearly optimal parameters. This works for any S_2 which may be embedded in an elliptic-function region after the shift. In general, if one region dominates the other in error reduction, one may shift the weaker region until the smallest real component of its shifted eigenvalues is zero and compute parameters which are optimal for the shifted dominant region.

When $(\zeta_1 - 1)(\zeta_2 - 1) < 0$, the spectra of $S_1(s)$ and $S_2(t)$ are always separated by a disk and cannot be aligned with Eq. 58. The simplest resolution is to not align and just use parameters for the spectrum with $\zeta < 1$.

4.6.6 Illustrative Examples

We now illustrate some of the algorithms with a few simple examples. The real intervals $S_1 = [0.1, 10]$ and $S_2 = [10, 100]$ transform as in Part 1 into the interval $[0.19967, 1]$ for which the optimum single parameter is $w_1 = 0.44684$. If $\theta_1 = 78.58°$ and $\theta_2 = 54.9°$, the complex regions are disks and repeated use of parameters back transformed from w_1 is optimal. We compute these values as $q_1 = 4.1$ and $p_1 = 20.74$. The error reduction per iteration assumes its largest magnitude at the interval endpoints and is $\varepsilon = 0.1462$. If one were to arbitrarily try the geometric means of the intervals ($q_1 = 1$ and $p_1 = 10^{3/2}$) as parameters, one would find that the error at the point $z = 10$ is equal to $\varepsilon = 0.4287$. This may not even be the bound, but it is certainly not nearly as small as the optimal value.

We now illustrate our algorithms for the case where one of the spectra is a disk. We obtain optimum parameters for $J = 2$ by transforming the disk to infinite radius. Let $S_1 = (1/4, 4, 0°)$ and $S_2 = (1/2, 2, 36.87°)$. We compute $\zeta_2 = 1$, thereby establishing that S_2 is a disk. By Eq. 59, $p = 1.5$, and we realign at $s = 1/p = 2/3$. The transformation from $s_0 = 0.25$ to $s = 2/3$ is attained with $\beta = 1/2$.

(The inverse is with $\beta = -1/2$.) The best parameter for $J = 1$ is unity, which transforms back into $p_1 = q_1 = 1$ with corresponding error reduction of $R = 0.2$. This was known from the given spectra, and required no transformation. The best two parameters for $S_1(2/3) = (2/3, 3/2, 0°)$ are $w_1 = 0.7522$ and its reciprocal $w_2 = 1.3295$. We compute $q_1 = (w_1 - 0.5)/(1 - 0.5w_1) = 0.4042$ and $q_2 = 1/q_1 = 2.474$. We compute $p_1 = (w_1 + 0.5)/(1 + 0.5w_1) = 0.90995$ and $p_2 = 1/p_1 = 1.099$. We compute the error reduction at various points on the boundaries of S_1 and S_2 and ascertain that the reduction is indeed a constant value of $R = 0.02$.