1991

A.N. Malyshev

Parallel aspects of some spectral problems in linear algebra

Department of Numerical Mathematics Report NM-R9113 July

CWI, nationaal instituut voor onderzoek op het gebied van wiskunde en informatica

CWI is the research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a non-profit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands organization for scientific research (NWO).

Parallel Aspects of some Spectral Problems in Linear Algebra

Alexander N. Malyshev

Institute of Mathematics
Siberian Branch of the USSR Academy of Sciences
Universitetsky pr., 4
Novosibirsk, 630090,
USSR

Abstract

The generalized eigenvalue problem $Ax = \lambda Bx$ is central to some important control theory applications, for example, the Lyapunov and the Riccati equations. In this report an algorithm is being studied for computing deflating subspaces of the regular linear matrix pencil $\lambda B - A$ on parallel computing systems.

More precisely, the algorithm is intended to compute the projector matrices P and $I - P$ onto deflating subspaces of the matrix pencil corresponding to the eigenvalues inside and outside the unit circle. The algorithm is based upon orthogonal transformations. It possesses a quadratic convergence rate and simultaneously with the projectors it computes the condition number of the problem. We study the extent to which this algorithm can be parallelized.

The algorithm has been implemented in a portable way suitable for shared memory parallel computers. Applications of the developed code are made to the solution of the Lyapunov and the Riccati equations. Timing results are provided for the Alliant FX/4.

1980 Mathematics Subject Classification (1985 revision): 65F15, 65W05.
Key Words & Phrases: Generalized eigenvalue problem, invariant subspaces, deflating subspaces, Lyapunov equation, Riccati equation, block Householder transformations.

Note: This work was carried out in the period December 1990 - May 1991, while the author was an ERCIM - fellow at CWI.

Report NM-R9113
CWI
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

Contents

Introduction

In 1988 the Soviet academic publishing house "Nauka" issued the monograph [12] "Guaranteed accuracy of the solution to systems of linear equations in Euclidean spaces" by S.K. Godunov et al. The book states theoretical background and practical aspects of some algorithms in linear algebra which are connected by the common idea to compute a solution with guaranteed accuracy. The list of the problems considered in the book involves:

- evaluation of eigenvalues for symmetric matrices;
- solution of a well-posed system of linear equations;
- calculation of an orthonormal basis of eigenvectors for a symmetric matrix;
- calculation of the singular value decomposition;
- solution of the linear least squares problem.

All the algorithms are provided for the exploration of resolvability of the problem during computation and for the evaluation of efficient (realistic) error bounds for the computed solution.

The approach used to treat the algorithms is based upon the backward error analysis and an appropriate perturbation theory. The same approach was successfully used earlier by J. Wilkinson in his classical treatise "The algebraic eigenvalue problem". In [12] this method was advanced to absolute mathematical strictness.

The present work is devoted to the investigation of unsymmetric eigenvalue problems with guaranteed accuracy. This kind of problems is much more complicated than the symmetric ones. The complexity is due to the fact that small perturbations of input data may cause tremendous perturbations of eigenvalues. It is worth to remark here that as a rule the input data are the results of some measurements or previous calculations and contain rounding errors of these measurements or calculations.

A standard example of such instability to small perturbations is the stability analysis (after Lyapunov) for a lower bidiagonal 20-by-20 matrix A with -1 on the diagonal and 10 on the subdiagonal. Replace the element $A_{1,20}$ at the upper right corner of matrix A by ϵ of small magnitude. Then the determinant of the matrix $A - \lambda I$, where I is the identity 20-by-20 matrix, equals $(-1-\lambda)^{20} - 10^{19}\epsilon$. For $\epsilon = 10^{-18}$ which is less than the relative rounding error for the majority of modern computers one can find that there is a root of the polynomial $\det(A - \lambda I)$ equal to $\lambda = \sqrt[20]{10} - 1 > 0.1$. Thus the stable matrix A with all eigenvalues being equal to -1 becomes unstable owing to very small perturbation of its elements. These circumstances demand to regard the matrix A as an "ill stable" or "practically unstable" matrix.

A numerical criterion for the quality of the matrix stability which can be efficiently evaluated on a computer was suggested in [5, 13]. Further development of these researches resulted in the criteria of a dichotomy quality of the matrix spectrum with respect to the imaginary axis and the unit circle [11, 7, 19].

By the spectrum dichotomy problem with respect to a closed contour γ in the complex plane we mean the following:

1. To find out whether there are some eigenvalues on the contour γ or in a small neigbourhood of γ.

2. If there are no such eigenvalues then invariant (deflating) subspaces corresponding to the parts of the spectrum inside and outside the contour γ have to be computed.

Criteria of the dichotomy quality are the numbers which characterize the stability of these invariant (deflating) subspaces for small perturbations of input data.

Under the spectrum dichotomy approach to eigenvalue problems the method of investigation of the spectrum consists of:

a) partition of the complex plane by circles and straight lines into some parts;

b) evaluation of the dichotomy quality parameters for each circle and straight line;

c) computation of the deflating (invariant) subspaces associated with the parts of the spectrum inside and outside each circle or halfplane provided that the corresponding dichotomy parameter is not too large;

d) computation of deflating (invariant) subspaces in the intersections of some circles and/or halfplanes.

In fact, this method is quite similar to the well-known bisection method for symmetric tridiagonal matrices:

Bisection procedure	**Dichotomy procedure**
Partition of *the real axis* by *a point* into two half-axes.	Partition of *the complex plane* by *a circle* or *straight line* into two parts.
Computation of *a number of eigenvalues* within each half-axis by means of *the Sturm sequences.*	Computation of *a number of eigenvalues* within each part by means of *the spectrum dichotomy problem solver.*

In both procedures the neigbourhoods of isolated parts of the spectrum are computed instead of individual eigenvalues.

The spectrum dichotomy problem method is much reinforced by the dichotomy parameter which can be efficiently evaluated on a computer. Thanks to this parameter we are able to develop an effective perturbation theory for some unsymmetric eigenvalue problems and to investigate the rate of convergence and stability of several effective numerical procedures. One may consider the dichotomy parameter to be *a condition number* for the spectrum dichotomy problems.

The dichotomy parameter is a positive real number which characterizes the stability of invariant subspaces of a matrix associated with two parts of the spectrum separated by a given closed contour. If the value of this parameter increases then the corresponding invariant subspaces become less stable. The stability deterioriates, for instance, when some eigenvalues come nearer to the contour or when the angle between the invariant subspaces diminishes.

When the computations are executed with rounding errors one can argue that the given contour does not "practically" separate the spectrum when the dichotomy parameter exceeds some quite large number. The magnitude of this bound depends mostly on the relative rounding error of the arithmetical operations. As usual, it is chosen in accordance with the requirements of guaranteed accuracy and applications.

Classical eigenelements of a square matrix are the set of eigenvalues and a vector basis composed of Jordan chains for a given matrix. In the dichotomy spectrum approach the main eigenelements are the dichotomy parameter with respect to a given closed contour and, if this parameter is finite, the invariant subspaces and the generalized Lyapunov function associated with the parts of the spectrum inside and outside the contour.

There are some alternative approaches to guaranteed accuracy for unsymmetric eigenvalue problems. The most important one is developed in [24, 25, 28, 8, 9]. Let us refer to it as the Schur method. This method has already been existing quite long and its algorithmical aspects are developed to a great extent [31, 22, 26, 27, 3].

Here I want to discuss briefly some theoretical difficulties in the Schur approach. Given an N-by-N unsymmetric matrix A. The crucial idea of the Schur method is to reduce the matrix

A to upper triangular form by means of similarity transformations [30]. As usual the orthogonal transformations are preferable and QR-like methods are used. As a result we have the Schur decomposition $A = Q^{-1}RQ$, where Q is an orthogonal matrix. The matrix R is a "nearly upper triangular" matrix, i.e. all the elements of R below the diagonal are of small magnitude but not all of them are zeros.

The first problem is that the methods under consideration have no global convergence for any matrix A [30]. Moreover, they can behave chaotically and the set of matrices on which chaotic behavior occurs has positive Lebesgue measure in the space $\mathbf{C}^{n\times n}$ [30].

The second difficulty is the justification of the fact that the matrix R can be used afterwards as if it is upper triangular. The above example with the 20-by-20 matrix A clearly demonstrates the possible danger of neglecting the lower triangle. In order to justify the utilization of R as a triangular matrix in some situations the following procedure was proposed [24]. Let

$$R = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}$$

with k-by-k matrix R_{11} and $(N-k)$-by-$(N-k)$ matrix R_{22}. Assume that the matrix R_{21} has a small norm. Then we can try to reduce the matrix R to upper block triangular form by means of the following procedure: Find a matrix X of small norm such that

$$RU = U \begin{pmatrix} \tilde{R}_{11} & \tilde{R}_{12} \\ 0 & \tilde{R}_{22} \end{pmatrix}$$

with the orthogonal matrix U of the form

$$U = \begin{pmatrix} I & -X^T \\ X & I \end{pmatrix} \begin{pmatrix} (I+X^TX)^{-1/2} & 0 \\ 0 & (I+XX^T)^{-1/2} \end{pmatrix}.$$

It follows that X has to satisfy the equation

$$XR_{11} - R_{22}X = R_{21} - XR_{12}X$$

which is solvable if the spectrum of R_{11} has no intersection with the spectrum of R_{22} and the norm of R_{21} is small enough.

Summarizing, we need to be able to compute the condition number of the so called Silvester equation $XR_{11} - R_{22}X = R_{21}$ which is, in fact, a $(N-k)k$-by-$(N-k)k$ system of linear equations [2, 15]. This may be quite expensive to compute [16].

Chapter 1

Eigenelements of a regular linear matrix pencil

Given real N-by-N matrices A and B. Let us refer to the matrix pencil $\lambda B - A$ as *regular with respect to the unit circle,* if $\det(\lambda B - A) \neq 0$ for every complex number λ on the unit circle $|\lambda| = 1$. All matrix pencils being considered in this chapter are assumed to be regular with respect to the unit circle. Throughout this paper we are interested only in right-hand side eigenelements for matrix pencils.

1.1 Deflating subspaces of a matrix pencil associated with the spectrum parts inside and outside the unit circle

At first I remind the definitions of eigenvalues and associated Jordan chains for the regular linear matrix pencil $\lambda B - A$.

The complex numbers λ and $1/\mu$ (if $\mu = 0$ then $1/\mu = \infty$) are *the eigenvalues of the pencil* $\lambda B - A$, if $\det(\lambda B - A) = 0$ or $\det(B - \mu A) = 0$ respectively.

The vectors $x_1, x_2, \ldots, x_t$ *form a Jordan chain of the pencil* $\lambda B - A$ *corresponding to an eigenvalue* λ if $(\lambda B - A)x_1 = 0$, $(\lambda B - A)x_l + Bx_{l-1} = 0$, $l = 2, \ldots, t$, provided the linear system $(\lambda B - A)z + Bx_t = 0$ has no solution z. Similarly, the vectors $x_1, x_2, \ldots, x_t$ form a Jordan chain of the pencil $\lambda B - A$ corresponding to an eigenvalue $1/\mu$ if $(B - \mu A)x_1 = 0$, $(B - \mu A)x_l + Ax_{l-1} = 0$, $l = 2, \ldots, t$, provided the linear system $(B - \mu A)z + Ax_t = 0$ has no solution z.

The linear span $\mathcal{L}_0$ of all the Jordan chains associated with all eigenvalues λ satisfying the condition $|\lambda| < 1$ is called *the deflating subspace of the matrix pencil* $\lambda B - A$ corresponding to the eigenvalues inside the unit circle. The linear span $\mathcal{L}_\infty$ of all the Jordan chains associated with all eigenvalues $1/\mu$ satisfying the condition $|\mu| < 1$ is called *the deflating subspace of the matrix pencil* $\lambda B - A$ corresponding to the eigenvalues outside the unit circle. The deflating subspaces are a suitable generalization of invariant subspaces used in the matrix case.

The main properties of the eigenelements introduced above are revealed in the following theorem on the Kronecker form of a linear matrix pencil regular with respect to the unit circle.

Theorem 1 *If* $\det(\lambda B - A) \neq 0$ *for all complex* λ *satisfying the condition* $|\lambda| = 1$ *then there exist nonsingular* N*-by-*N *matrices* Q_l *and* Q_r *such that*

$$Q_l(\lambda B - A)Q_r = \textit{block diag}\{\lambda I - J_0, \lambda J_\infty - I\}, \tag{1.1}$$

$$J_0 = \textit{block diag}\{J_{i_1}(\lambda_1), \ldots, J_{i_m}(\lambda_m)\}, \quad |\lambda_i| < 1,$$

$$J_\infty = \textit{block diag}\{J_{j_1}(\mu_1), \ldots, J_{j_n}(\mu_n)\}, \quad |\mu_j| < 1,$$

where I is the identity matrix of appropriate size and

$$J_t(\nu) = \begin{bmatrix} \nu & 1 & & O \\ & \ddots & \ddots & \\ & & \nu & 1 \\ O & & & \nu \end{bmatrix}$$

is a Jordan t-by-t block.

PROOF. As a matter of fact the proof of the Theorem 1 consists of the appropriate use of the Jordan canonical form. Let us rewrite the matrix pencil $\lambda B - A$ in the following way:

$$\lambda B - A = (B - A)[(\lambda - 1)(B - A)^{-1}B + I].$$

There exists a nonsingular matrix F such that

$$\lambda B - A = (B - A)F[(\lambda - 1)J + I]F^{-1},$$

where J is the Jordan canonical form of the matrix $(B - A)^{-1}B$.

Now we divide the matrix J into two diagonal blocks: $J = \text{block diag}\{J_1, J_2\}$, where J_1 is the block containing all Jordan blocks associated with the eigenvalues of the pencil $\lambda J + I$ inside the unit circle and J_2 is associated with the eigenvalues outside the unit circle. It is easy to make sure that the matrices J_1 and $J_2 - I$ are invertible.

The pencil $\lambda J + I - J$ can be decomposed in the following way:

$$\lambda J - (J - I) = \begin{pmatrix} J_1 & 0 \\ 0 & J_2 - I \end{pmatrix} \left[\lambda \begin{pmatrix} I & 0 \\ 0 & (J_2 - I)^{-1}J_2 \end{pmatrix} - \begin{pmatrix} J_1^{-1}(J_1 - I) & 0 \\ 0 & I \end{pmatrix} \right].$$

Reducing the matrices $J_1^{-1}(J_1 - I)$ and $(J_2 - I)^{-1}J_2$ to Jordan forms $F_1 J_0 F_1^{-1}$ and $F_2 J_\infty F_2^{-1}$ we obtain

$$\lambda B - A = (B - A)F \begin{pmatrix} J_1 F_1 & 0 \\ 0 & (J_2 - I)F_2 \end{pmatrix} \begin{pmatrix} \lambda I - J_0 & 0 \\ 0 & \lambda J_\infty - I \end{pmatrix} \begin{pmatrix} F_1^{-1} & 0 \\ 0 & F_2^{-1} \end{pmatrix} F^{-1}.$$

■

Let us rewrite the matrix Q_r in block form

$$Q_r = [Q_1 \dots Q_m Q_{m+1} \dots Q_{m+n}],$$

where the partition into block columns conforms to the sizes of the diagonal blocks in the decomposition (1.1). Then evidently the columns of each block Q_i form a Jordan chain.

One can show that the deflating subspace $\mathcal{L}_0$ of the pencil $\lambda B - A$ corresponding to the eigenvalues inside the unit circle is spanned by the columns of blocks $Q_1, Q_2, \dots, Q_m$, and that the deflating subspace $\mathcal{L}_\infty$ corresponding to the eigenvalues outside the unit circle is spanned by the columns of the blocks $Q_{m+1}, Q_{m+2}, \dots, Q_{m+n}$. Therefore, the intersection of the subspaces $\mathcal{L}_0$ and $\mathcal{L}_\infty$ is equal to the null space and the sum of these subspaces is equal to the whole space R^N.

Vector subspaces are conveniently defined by means of projectors onto the subspaces. Let N_0 be the dimension of the subspace $\mathcal{L}_0$ and let $N_\infty = N - N_0$ be the dimension of the subspace $\mathcal{L}_\infty$. Consider the matrices P_0 and P_∞ designed with the help of decomposition (1.1):

$$P_0 = Q_r \begin{pmatrix} I_{N_0} & 0 \\ 0 & 0 \end{pmatrix} Q_r^{-1}, \quad P_\infty = Q_r \begin{pmatrix} 0 & 0 \\ 0 & I_{N_\infty} \end{pmatrix} Q_r^{-1}. \tag{1.2}$$

It is not difficult to see that P_0 is a projector onto the deflating subspace $\mathcal{L}_0$ of the pencil $\lambda B - A$, i.e. $P_0^2 = P_0$ and the equality $P_0 x = x$ is equivalent to the fact that the vector x belongs to to this subspace. Accordingly, the matrix P_∞ is a projector onto the deflating subspace $\mathcal{L}_\infty$.

Observe that $P_0 + P_\infty = I$. This identity means that P_0 and P_∞ is a pair of projectors onto $\mathcal{L}_0$ and $\mathcal{L}_\infty$.

Apart from the projectors P_0 and P_∞ it is sometimes useful to deal with orthogonal projectors Π_0 and Π_∞ onto the invariant subspaces of the matrix pencil $\lambda B - A$. We deduce the formulas that express the matrices Π_0 and Π_∞ in terms of the matrices P_0 and P_∞. At first premultiply the identity $P_0^2 = P_0$ from the left-hand side by P_0^H as well as the identity $P_\infty P_0 = 0$ by P_∞^H. Adding the results and grouping the left-hand side we obtain the identity

$$(P_0^H P_0 + P_\infty^H P_\infty)P_0 = P_0^H P_0.$$

From this one can deduce the equalities

$$P_0 = (P_0^H P_0 + P_\infty^H P_\infty)^{-1} P_0^H P_0, \quad P_0 = P_0 (P_0^H P_0 + P_\infty^H P_\infty)^{-1} P_0^H P_0.$$

The latter of these allows to check that the matrix

$$\Pi_0 = P_0 (P_0^H P_0 + P_\infty^H P_\infty)^{-1} P_0^H \tag{1.3}$$

satisfies the system of matrix equations

$$\Pi_0^2 = \Pi_0 = \Pi_0^H,$$

$$P_0 \Pi_0 = \Pi_0,$$

$$\Pi_0 P_0 = P_0,$$

which uniquely defines the orthogonal projector Π_0 onto the deflating subspace $\mathcal{L}_0$ of the pencil $\lambda B - A$. In a similar way the formula

$$\Pi_\infty = P_\infty (P_0^H P_0 + P_\infty^H P_\infty)^{-1} P_\infty^* \tag{1.4}$$

can be derived.

Below we also use the following identities:

$$I - \Pi_0 = P_\infty^H (P_0 P_0^H + P_\infty P_\infty^H)^{-1} P_\infty, \quad I - \Pi_\infty = P_0^H (P_0 P_0^H + P_\infty P_\infty^H)^{-1} P_0. \tag{1.5}$$

Finally, we reveal how norms of the projectors are connected with the angle between nontrivial deflating subspaces $\mathcal{L}_0$ and $\mathcal{L}_\infty$. Here nontrivial means that $\mathcal{L}_0 \neq 0$ and $\mathcal{L}_\infty \neq 0$.

The angle φ, $0 < \varphi \leq \pi/2$, between the deflating subspaces with projectors P_0 and P_∞ is defined by means of the formula

$$\cos\varphi = \max_{\substack{P_0 x \neq 0 \\ P_\infty y \neq 0}} \frac{|(P_0 x, P_\infty y)|}{\|P_0 x\| \, \|P_\infty y\|}. \tag{1.6}$$

It can be shown that

$$\sin\varphi = \frac{1}{\|P_0\|} = \frac{1}{\|P_\infty\|} = \sigma_{\min}(\Pi_0 - \Pi_\infty), \tag{1.7}$$

where $\sigma_{\min}(\Pi_0 - \Pi_\infty)$ is the least singular value of matrix $\Pi_0 - \Pi_\infty$.

1.2 The canonical form of a linear matrix pencil which is regular with respect to the unit circle

Let us use the notations of Theorem 1 in order to define the sequence of matrices

$$P_{+0} = P_0 = Q_r \begin{pmatrix} I_{N_0} & 0 \\ 0 & 0 \end{pmatrix} Q_r^{-1}, \quad P_{-0} = P_\infty = Q_r \begin{pmatrix} 0 & 0 \\ 0 & I_{N_\infty} \end{pmatrix} Q_r^{-1},$$

$$P_k = Q_r \begin{pmatrix} J_0^k & 0 \\ 0 & 0 \end{pmatrix} Q_r^{-1} \text{ for integer } k \geq 1, \tag{1.8}$$

$$P_{-k} = Q_r \begin{pmatrix} 0 & 0 \\ 0 & J_\infty^{-k} \end{pmatrix} Q_r^{-1} \text{ for integer } k \geq 1.$$

The following identities are satisfied for these matrices P_k:

$$\begin{aligned} P_k P_l &= P_{k+l}, \text{ if } k \text{ and } l \text{ have the same sign,} \\ P_k P_l &= 0, \text{ if } k \text{ and } l \text{ have different signs.} \end{aligned} \tag{1.9}$$

The index $+0$ has the sign plus but the index -0 has the sign minus.

Assuming that $T = Q_r Q_l$ we obtain from Theorem 1

Corollary. *If* $\det(\lambda B - A) \neq 0$ *for each* λ, $|\lambda| = 1$, *then the pencil* $\lambda B - A$ *can be presented in the canonical form*

$$\lambda B - A = T^{-1}[\lambda(P_{+0} + P_{-1}) - (P_{-0} + P_{+1})] \tag{1.10}$$

with nonsingular N*-by-*N *matrix* T *and* N*-by-*N *matrices* P_{+0}, P_{-1}, P_{-0}, P_{+1}, *which are defined by 1.8.*

Note that the matrix T is equal to $P_{+0} + P_{-1}$ when $B = I$.

The canonical form (1.10) may be defined independently of the decomposition (1.1) from Theorem 1. Such a characterization is contained in the following theorem.

Theorem 2 *Given*

$$\lambda B - A = T^{-1}[\lambda(P_{+0} + P_{-1}) - (P_{-0} + P_{+1})], \ \det T \neq 0.$$

$$\begin{aligned} P_{+0}^2 - P_{+0} = 0, \ \ P_{-0} = I - P_{+0}, \\ P_{-0}P_{+1} = P_{+1}P_{-0} = P_{+0}P_{-1} = P_{-1}P_{+0} = 0 \end{aligned} \tag{1.11}$$

and all the eigenvalues of the matrix $P_{-1} + P_{+1}$ *lie inside the unit circle. Then* $\det(\lambda B - A) \neq 0$ *for each complex* λ *on the unit circle* $|\lambda| = 1$ *and*

$$T = \frac{1}{2\pi} \int_0^{2\pi} (B - e^{i\phi} A)^{-1} (1 + e^{i\phi}) \, d\phi,$$

$$P_{+0} = \left[\frac{1}{2\pi} \int_0^{2\pi} (B - e^{i\phi} A)^{-1} d\phi \right] T^{-1}, \tag{1.12}$$

$$P_{-0} = -\left[\frac{1}{2\pi} \int_0^{2\pi} (B - e^{i\phi} A)^{-1} e^{i\phi} d\phi \right] T^{-1}.$$

PROOF. We carry out all our arguments in a vector basis where the matrices P_{+0} and P_{-0} have the diagonal form. Then the system (1.11) implies that

$$P_{+0} = \begin{pmatrix} I_{N_0} & 0 \\ 0 & 0 \end{pmatrix}, \ P_{-0} = \begin{pmatrix} 0 & 0 \\ 0 & I_{N_\infty} \end{pmatrix},$$

$$P_{+1} = \begin{pmatrix} K_0 & 0 \\ 0 & 0 \end{pmatrix}, \ P_{-1} = \begin{pmatrix} 0 & 0 \\ 0 & K_\infty \end{pmatrix},$$

where all the eigenvalues of K_0 and K_∞ lie inside the unit circle. Therefore, $\det(B - e^{i\phi} A) \neq 0$ for each real ϕ.

In order to derive the formulas (1.12) let us find out the relations between the sequence of matrices P_k and the sequence of Green matrices G_k for the finite difference equation $Bx_n - Ax_{n-1} = f_n$. Expand the periodic matrix function $\mathcal{D}(\phi) = (B - e^{i\phi} A)^{-1}$ into a Fourier series with respect

to the basis $e^{ik\phi}$: $\mathcal{D}(\phi) = \sum_{k=-\infty}^{\infty} Z_k e^{ik\phi}$. The coefficients of the Fourier series are evaluated by means of the integrals

$$Z_k = \frac{1}{2\pi} \int_0^{2\pi} (B - e^{i\phi}A)^{-1} e^{-ik\phi} d\phi. \tag{1.13}$$

Hence the uniform boundness of the sequence $\|Z_k\|$ follows easily for all integers k.

From the identity $(B - e^{i\phi}A)\mathcal{D}(\phi) = I$ the system of equations is deduced which connects the matrices Z_k:

$$BZ_k - AZ_{k-1} = \begin{cases} I, & k = 0, \\ 0, & k \neq 0. \end{cases} \tag{1.14}$$

Therefore the sequence of matrices Z_k is the sequence of Green matrices G_k, i.d. $Z_k = G_k$.

Properties of the matrices P_k allow us explicitly to find a solution to the infinite system (1.14) with the values $\|Z_k\|$ bounded for all integers k:

$$\begin{gathered} G_0 = Z_0 = P_{+0}T, \quad G_k = Z_k = P_k T \quad \text{for } k \geq 1, \\ G_{-1} = Z_{-1} = -P_{-0}T, \quad G_k = Z_k = -P_{k+1}T \quad \text{for } k \leq -2. \end{gathered} \tag{1.15}$$

Hence, $T = G_0 - G_{-1}$.

Finally, the formulas (1.12) are obtained from (1.13) and (1.15). ■

1.3 The parameter ω—a criterion to determine whether a regular linear matrix pencil has no eigenvalues on the unit circle and within a small neighbourhood of it

Let $\det(B - e^{i\phi}A) \neq 0$ for all real ϕ, i.e. the pencil $\lambda B - A$ is regular with respect to the unit circle. Consider the Hermitian matrix

$$H = \frac{1}{2\pi} \int_0^{2\pi} (B - e^{i\phi}A)^{-1}(AA^H + BB^H)(B - e^{i\phi}A)^{-H} d\phi. \tag{1.16}$$

The spectral norm, i.e. the largest singular value, of the matrix H is denoted by $\omega = \|H\|$ and is referred to as *the criterion of absence of the eigenvalues of pencil $\lambda B - A$ on the unit circle and within a small neighbourhood of it.*

Note that if the pencil $\lambda B - A$ is premultiplied by any nonsingular matrix from the left-hand side then the Hermitian matrix H is not changed. Let a nonsingular matrix L satisfy the equation

$$LL^H = AA^H + BB^H. \tag{1.17}$$

Introducing the matrices $A_0 = L^{-1}A$, $B_0 = L^{-1}B$, we obtain the identity

$$H = \frac{1}{2\pi} \int_0^{2\pi} (B_0 - e^{i\phi}A_0)^{-1}(AA^H + BB^H)(B_0 - e^{i\phi}A_0)^{-H} d\phi. \tag{1.18}$$

The pencil $\lambda B_0 - A_0$ satisfies the equation

$$A_0 A_0^H + B_0 B_0^H = I. \tag{1.19}$$

We shall refer to such pencils $\lambda B_0 - A_0$ as *orthonormalized pencils* which are obtained with the help of orthonormalization of pencil $\lambda B - A$ by the condition (1.19). Evidently, all orthornomalized pencils being obtained from the same pencil $\lambda B - A$ differ by left-hand side matrix orthogonal multiples only.

The previous remarks prompt the idea that the parameter ω reflects the spectral properties of the orthonormalized pencil $\lambda B_0 - A_0$. Actually, $\sigma_{\min}(B_0 - e^{i\phi}A_0)$ can be estimated by means of ω.

Theorem 3 *Let* $A_0A_0^H + B_0B_0^H = I$ *and* $\det(B_0 - e^{i\phi}A_0) \neq 0$ *for each real* ϕ. *Then*

$$\max_{\phi} \|(B_0 - e^{i\phi}A_0)^{-1}\| < 14\omega. \tag{1.20}$$

Thus the eigenvalues of the orthonormalized matrix pencil $\lambda B_0 - A_0$, and of the given pencil $\lambda B - A$ equivalently, being regular with respect to the unit circle are separated from the unit circle by the distance not less than $1/14\omega$.

As an illustration to Theorem 1.20 we consider one example. Let $B = I$ and A be a symmetric matrix. We make use of the diagonal form of the matrix A: $A = U^H DU$, where $U^H U = I$. Since in this case

$$\frac{1}{2\pi}\int_0^{2\pi}(B - e^{i\phi}A)^{-1}(AA^H + BB^H)(B - e^{i\phi}A)^{-H}d\phi =$$

$$= U^H\left[\frac{1}{2\pi}\int_0^{2\pi}(I - e^{i\phi}D)^{-1}(I + D^2)(I - e^{-i\phi}D)^{-1}d\phi\right]U,$$

then

$$\omega = \max_i\left(\left|\frac{1}{2\pi}\int_0^{2\pi}\frac{1 + d_i^2}{1 + d_i^2 - 2d_i\cos\phi}d\phi\right|\right) = \max_i\left|\frac{1 + d_i^2}{1 - d_i^2}\right|,$$

where d_i are the diagonal elements of matrix D. One of the orthonormalized pencils $\lambda B_0 - A_0$ is equal to

$$U^H\frac{\lambda I - D}{\sqrt{I + D^2}}U,$$

and, consequently,

$$\max_{\phi}\|(B_0 - e^{i\phi}A_0)^{-1}\| = \max_i\left|\frac{\sqrt{1 + d_i^2}}{1 - |d_i|}\right|.$$

So the following estimate holds for this example:

$$\max_{\phi}\|(B_0 - e^{i\phi}A_0)^{-1}\| \leq \max_{x\geq 0}\frac{1 + x}{\sqrt{1 + x^2}} = \sqrt{2}\omega.$$

The above example demonstrates that the exponent of ω in the estimate (1.20) is precise.

In numerical mathematics the question is extremely important about the stability of the computed solution under small perturbations of the input data. As remarked above the parameter ω reflects the spectral properties of an orthonormalized matrix pencil $\lambda B_0 - A_0$. Therefore the problem about estimation of stability of ω is posed as follows:

Let $\tilde{A}_0$ *and* $\tilde{B}_0$ *be perturbations of the matrices* A_0, B_0 *and*

$$A_0A_0^H + B_0B_0^H = I, \quad \|(\tilde{A}_0\ \tilde{B}_0) - (A_0\ B_0)\| \leq \delta,$$

$$\omega = \left\|\frac{1}{2\pi}\int_0^{2\pi}(B_0 - e^{i\phi}A_0)^{-1}(B_0 - e^{i\phi}A_0)^{-H}d\phi\right\|,$$

$$\tilde{\omega} = \left\|\frac{1}{2\pi}\int_0^{2\pi}(\tilde{B}_0 - e^{i\phi}\tilde{A}_0)^{-1}(\tilde{A}_0\tilde{A}_0^H + \tilde{B}_0\tilde{B}_0^H)(\tilde{B}_0 - e^{i\phi}\tilde{A}_0)^{-H}d\phi\right\|.$$

Then we have to evaluate the quantity $|\tilde{\omega} - \omega|$.

In order to derive this estimate let us take advantage of the identity $(K + M)^{-1} = K^{-1} - (K + M)^{-1}MK^{-1}$. Since

$$(\tilde{B}_0 - e^{i\phi}\tilde{A}_0)^{-1} = (B_0 - e^{i\phi}A_0)^{-1} -$$

$$-(B_0 - e^{i\phi}A_0)^{-1}[(\tilde{B}_0 - B_0) - e^{i\phi}(\tilde{A}_0 - A_0)](\tilde{B}_0 - e^{i\phi}\tilde{A}_0)^{-1},$$

the formula for $\tilde{\omega}$ can be rewritten in the way

$$\tilde{\omega} = \left\| \frac{1}{2\pi} \int_0^{2\pi} (B_0 - e^{i\phi}A_0)^{-1}\{I - [(\tilde{B}_0 - B_0) - e^{i\phi}(\tilde{A}_0 - A_0)](\tilde{B}_0 - e^{i\phi}\tilde{A}_0)^{-1}\}\times \right.$$

$$\left. \times(\tilde{A}_0\tilde{A}_0^H + \tilde{B}_0\tilde{B}_0^H)\{I - [(\tilde{B}_0 - B_0) - e^{i\phi}(\tilde{A}_0 - A_0)]\,(\tilde{B}_0 - e^{i\phi}\tilde{A}_0)^{-1}\}^H (B_0 - e^{i\phi}A_0)^{-H} d\phi \right\|.$$

Hence,

$$|\tilde{\omega} - \omega| \le \left\| \frac{1}{2\pi} \int_0^{2\pi} (B_0 - e^{i\phi}A_0)^{-1}\Delta(B_0 - e^{i\phi}A_0)^{-H} d\phi \right\|,$$

where

$$\Delta = \{I - [(\tilde{B}_0 - B_0) - e^{i\phi}(\tilde{A}_0 - A_0)](\tilde{B}_0 - e^{i\phi}\tilde{A}_0)^{-1}\}\times$$

$$\times(\tilde{A}_0\tilde{A}_0^H + \tilde{B}_0\tilde{B}_0^H)\{I - [(\tilde{B}_0 - B_0) - e^{i\phi}(\tilde{A}_0 - A_0)](\tilde{B}_0 - e^{i\phi}\tilde{A}_0)^{-1}\}^H - I.$$

Then

$$\left\| \frac{1}{2\pi} \int_0^{2\pi} (B_0 - e^{i\phi}A_0)^{-1}\Delta(B_0 - e^{i\phi}A_0)^{-H} d\phi \right\| =$$

$$= \sup_{\|x\|=1} \left| \left(\left[\frac{1}{2\pi} \int_0^{2\pi} (B_0 - e^{i\phi}A_0)^{-1}\Delta(B_0 - e^{i\phi}A_0)^{-H} d\phi \right] x, x \right) \right| =$$

$$= \sup_{\|x\|=1} \left| \frac{1}{2\pi} \int_0^{2\pi} (\Delta(B_0 - e^{i\phi}A_0)^{-H}x, (B_0 - e^{i\phi}A_0)^{-H}x) d\phi \right| \le$$

$$\le \|\Delta\| \sup_{\|x\|=1} \left\| \frac{1}{2\pi} \int_0^{2\pi} \|(B_0 - e^{i\phi}A_0)^{-H}x\|^2 d\phi \right\| = \|\Delta\|\omega.$$

Estimate $\|\Delta\|$:

$$\|(\tilde{B}_0 - B_0) - e^{i\phi}(\tilde{A}_0 - A_0)\| \le \delta \left\| \begin{pmatrix} I \\ -e^{i\phi}I \end{pmatrix} \right\| = \delta\sqrt{2},$$

$$\|(\tilde{B}_0 - e^{i\phi}\tilde{A}_0)^{-1}\| \le \frac{\|(B_0 - e^{i\phi}A_0)^{-1}\|}{1 - \sqrt{2}\delta\|(B_0 - e^{i\phi}A_0)^{-1}\|},$$

$$\|\Delta\| \le \left(1 + \frac{\sqrt{2}\delta\|(B_0 - e^{i\phi}A_0)^{-1}\|}{1 - \sqrt{2}\delta\|(B_0 - e^{i\phi}A_0)^{-1}\|} \right)^2 (1+\delta)^2 - 1 \le$$

$$\le \frac{1}{1 - 2\delta - 2\sqrt{2}\delta\|(B_0 - e^{i\phi}A_0)^{-1}\|} - 1 \le \frac{2\delta + 40\delta\omega}{1 - 2\delta - 40\delta\omega}.$$

As a result

$$|\tilde{\omega} - \omega| \le \omega \frac{2\delta + 40\delta\omega}{1 - 2\delta - 40\delta\omega}. \tag{1.21}$$

Of course, this estimate is valid only when $2\delta + 40\delta\omega < 1$. Taking into account the inequality $\omega \ge 1$ which will be proved below, the estimate (1.21) may be somewhat simplified:

$$\left| \frac{\tilde{\omega} - \omega}{\omega} \right| \le \frac{42\omega\delta}{1 - 42\omega\delta}. \tag{1.22}$$

What is following now is aimed to find out the structure of the matrix H in terms of the matrices P_k. Recall that

$$H = \frac{1}{2\pi} \int_0^{2\pi} (B_0 - e^{i\phi}A_0)^{-1}(B_0 - e^{i\phi}A_0)^{-H} d\phi,$$

where $A_0 = L^{-1}A$, $B_0 = L^{-1}B$, $LL^H = AA^H + BB^H$. Due to the equality $A_0A_0^H + B_0B_0^H = I$ the matrix T_0 from the canonical form of the pencil

$$\lambda B_0 - A_0 = \lambda T_0^{-1}(P_{+0} + P_{-1}) - T_0^{-1}(P_{-0} + P_{+1})$$

satisfies the identity

$$T_0T_0^H = (P_{-0} + P_{+1})(P_{-0}^H + P_{+1}^H) + (P_{+0} + P_{-1})(P_{+0}^H + P_{-1}^H). \tag{1.23}$$

The relations (1.15) imply that

$$(B_0 - e^{i\phi}A_0)^{-1} = \sum_{j=-\infty}^{\infty} P_jT_0e^{i\phi}. \tag{1.24}$$

With the help of Parseval's equality for the function $(B_0 - e^{i\phi}A_0)^{-1}$ we obtain the formula

$$H = \sum_{j=-\infty}^{\infty} P_jT_0T_0^HP_j^H. \tag{1.25}$$

Substituting (1.23) into (1.25) and reducing parentheses by means of (1.9) we deduce the following important identity:

$$H = 2\sum_{j=-\infty}^{\infty} P_jP_j^H - P_{+0}P_{+0}^H - P_{-0}P_{-0}^H. \tag{1.26}$$

In particular this identity allows to derive the inequality

$$\omega \geq 1. \tag{1.27}$$

In fact,

$$H \geq P_{+0}P_{+0}^H + P_{-0}P_0^H, \tag{1.28}$$

$$\omega = \max_{\|x\|=1} |(Hx, x)| \geq \max_{\|x\|=1} |([P_{+0}P_{+0}^H + P_{-0}P_0^H]x, x)| = \max_{\|x\|=1} (\|P_{+0}^Hx\|^2 + \|P_{-0}^Hx\|^2) \geq 1.$$

In the latter inequality it is sufficient to choose x satisfying the equality $P_{+0}^Hx = x$ if $P_{+0} \neq 0$ or $P_{-0}^Hx = x$ if $P_0 \neq 0$.

Of great importance is an effective estimate of the matrices P_k with the help of the parameter ω. It allows to justify many numerical procedures for exploration of the spectrum of unsymmetric problems. Now we start to derive this important estimate.

For $k \geq 1$ the chain of inequalities is valid:

$$(P_0HP_0^HP_k^Hf, P_k^Hf) = (P_1HP_1^HP_{k-1}^Hf, P_{k-1}^Hf) = (P_0HP_0^HP_{k-1}^Hf, P_{k-1}^Hf) -$$

$$-([P_0P_0^H + P_1P_1^H]P_{k-1}^Hf, P_{k-1}^Hf) = (P_0HP_0^HP_{k-1}^Hf, P_{k-1}^Hf) - (P_{k-1}^Hf, P_{k-1}^Hf) - (P_k^Hf, P_k^Hf),$$

$$(P_k^Hf, P_k^Hf) + (P_0HP_0^HP_k^Hf, P_k^Hf) = (P_{k-1}^Hf, P_{k-1}^Hf) + (P_0HP_0^HP_{k-1}^Hf, P_{k-1}^Hf) -$$

$$-2(P_{k-1}^Hf, P_{k-1}^Hf) \leq \left(1 - \frac{2}{1+\|H\|}\right)[(P_{k-1}^Hf, P_{k-1}^Hf) + (P_0HP_0^HP_{k-1}^Hf, P_{k-1}^Hf)],$$

$$(P_k^Hf, P_k^Hf) + (P_0HP_0^HP_k^Hf, P_k^Hf) \leq$$

$$\leq \left(1 - \frac{2}{1+\|H\|}\right)^k [(P_0^Hf, P_0^Hf) + (P_0HP_0^Hf, f)] \leq 2\left(1 - \frac{2}{1+\|H\|}\right)^k \|H\|\|f\|^2.$$

Hence for $k \geq 0$ the inequality

$$(P_k^Hf, P_k^Hf) = (P_0^HP_k^Hf, P_0^HP_k^Hf) = (P_0P_0^HP_k^Hf, P_k^Hf) \leq (P_0HP_0^HP_k^Hf, P_k^Hf),$$

is valid, which yields the following estimates:

$$\|P_k\|^2 \leq \|H\| \left(1 - \frac{2}{1+\|H\|}\right)^k,$$

$$\|P_k\| \leq \sqrt{\omega} \left(1 - \frac{1}{1+\omega}\right)^k \leq \sqrt{\omega} e^{-k/(1+\omega)}, \; k \geq 1.$$

The case of $k \leq 0$ is treated in a similar way. Thus we proved the following theorem which is extremely important for the researches of convergence of power methods used for the calculation of the eigenelements.

Theorem 4 *If the linear matrix pencil $\lambda B - A$ is regular with respect to the unit circle then*

$$\|P_k\| \leq \sqrt{\omega} \left(1 - \frac{1}{1+\omega}\right)^{|k|} \leq \sqrt{\omega} e^{-|k|/(1+\omega)}. \tag{1.29}$$

Chapter 2

An algorithm to compute the projectors onto the deflating subspaces of a linear matrix pencil which is regular with respect the unit circle

2.1 Description of the algorithm

We start with some heuristical motives which lead to the scheme of our algorithm. The system of difference equations (1.14) for infinite sequence of matrices G_k may be replaced by the similar system for the matrices

$$\ldots, P_{-2}, P_{-1}, P_{-0}, P_{+0}, P_{+1}, P_{+2}, \ldots.$$

The new system has the form

$$\begin{aligned} BP_k - AP_{k-1} &= 0, \quad k \leq -0, \\ P_{+0} + P_{-0} &= I, \\ BP_k - AP_{k-1} &= 0, \quad k \geq 1. \end{aligned} \tag{2.1}$$

The system (2.1) is not suitable for numerical treatment because of the infinite number of equations. In order to overcome this problem a modification of the system (2.1) is suggested which is based on the fact that the norm of the matrices P_k decays rapidly when $|k| \to \infty$.

Define the matrices $P_k^{(n)} = \sum_{p=-\infty}^{\infty} P_{k+np}$ for integers k, where n is a sufficiently large natural even number. Note that for $k = np$ there are two matrices $P_{k-0}^{(n)}$ and $P_{k+0}^{(n)}$. Applying Theorem 4 one can show that for $-n/2 \leq k \leq n/2$ the following inequality is valid

$$\|P_k - P_k^{(n)}\| \leq 2\sqrt{\omega}\frac{e^{-n/(2(1+\omega))}}{1 - e^{-n/(1+\omega)}},$$

which means that the matrix $P_k^{(n)}$ approximates the matrix P_k for $|k| \leq n/2$ when n is quite large.

The sequence of matrices $P_k^{(n)}$ is periodical with period n and satisfies the system of matrix equations

$$\begin{aligned} BP_k^{(n)} - AP_{k-1}^{(n)} &= 0, \quad k \neq np + 0, \\ P_{k+0}^{(n)} + P_{k-0}^{(n)} &= I, \quad k = np + 0. \end{aligned}$$

Let us consider this system within the period $[+0, n-0]$:

$$\begin{aligned} &Z_0 + Z_n = I, \\ &BZ_k - AZ_{k-1} = 0, \quad 1 \le k \le n, \end{aligned} \tag{2.2}$$

where $Z_k = P_k^{(n)}$. When n is large enough the matrices $Z_0, Z_1, \ldots, Z_{n/2}$ are good approximations for the matrices $P_{+0}, P_1, \ldots, P_{n/2}$ and the matrices $Z_{n/2}, \ldots, Z_n$ are the approximations for $P_{-n/2}, \ldots, P_{-0}$.

We only need the projectors $P_{+0} = P_0$ and $P_{-0} = P_\infty$. For this it turns out that one may carry out a fast orthogonal procedure to compute the approximations to these projectors, the matrices Z_0 and Z_n. Let $n = 2^{m_0}$ and for some m, $0 \le m < m_0$, the following finite difference identities holds:

$$B_m Z_k - A_m Z_{k-2^m} = 0. \tag{2.3}$$

Consider the equation (2.3) for the matrices Z_{k+2^m} and Z_k:

$$B_m Z_{k+2^m} - A_m Z_k = 0. \tag{2.4}$$

Premultiply the equation (2.3) by a matrix X_m as well as the equation (2.4) by matrix Y_m and add the results:

$$Y_m B_m Z_{k+2^m} + (X_m B_m - Y_m A_m) Z_k - X_m A_m Z_{k-2^m} = 0. \tag{2.5}$$

For the matrices X_m and Y_m let us choose N-by-N matrices satisfying the system of matrix equations

$$\begin{cases} X_m X_m^H + Y_m Y_m^H = I, \\ X_m B_m - Y_m A_m = 0. \end{cases} \tag{2.6}$$

A method of calculation of such matrices will be described below.

Using the notations $A_{m+1} = X_m A_m$, $B_{m+1} = Y_m B_m$, the equation (2.5) is rewritten as:

$$B_{m+1} Z_k - A_{m+1} Z_{k-2^{m+1}} = 0.$$

At the end of the iterative process of orthogonal elimination we obtain the linear system which connects the matrices Z_0 and $Z_{2^{m_0}}$

$$\begin{aligned} &Z_0 + Z_{2^{m_0}} = I, \\ &B_{m_0} Z_{2^{m_0}} - A_{m_0} Z_0 = 0. \end{aligned} \tag{2.7}$$

It is easy to verify that a solution of the system is equal to the matrices $Z_0 = (A_{m_0} + B_{m_0})^{-1} B_{m_0}$, $Z_{2^{m_0}} = (A_{m_0} + B_{m_0})^{-1} A_{m_0}$. Hence for large m_0 the matrix $(A_{m_0} + B_{m_0})^{-1} B_{m_0}$ comes close to the projector P_0 and the matrix $(A_{m_0} + B_{m_0})^{-1} A_{m_0}$ approximates P_∞.

Summing up, we have formulated the main stages of the numerical procedure to compute the projectors P_0 and P_∞ onto the deflating subspaces of the matrix pencil $\lambda B - A$ which is regular with respect to the unit circle corresponding to the eigenvalues inside and outside the unit circle.

Stage I. The pencil $\lambda B - A$ is normalized in an appropriate way, i.e. the matrices B_0 and A_0 should be computed such that $\lambda B - A = L(\lambda B_0 - A_0)$ with a nonsingular matrix L. As a rule it is the orthonormalization by the condition $A_0 A_0^H + B_0 B_0^H = I$.

Stage II. The iterations to get the pair of matrices A_{m+1}, B_{m+1} from the pair of matrices A_m, B_m for $0 \le m < m_0$ are executed:

$$A_{m+1} = X_m A_m, \quad B_{m+1} = Y_m B_m,$$

where X_m and Y_m satisfy the system (2.6).

Stage III. Choosing some large enough m_0 solve the linear system

$$Z_0 + Z_\infty = I, \quad A_{m_0} Z_0 - B_{m_0} Z_\infty = 0 \tag{2.8}$$

with the unknown matrices Z_0 and Z_∞. When $m_0 \to \infty$ the matrices Z_0 and Z_∞ will be accurate approximations to the projectors P_0 and P_∞ of the matrix pencil $\lambda B - A$.

If you need to compute the orthogonal projectors Π_0 and Π_∞ onto the deflating subspaces $\mathcal{L}_0$ and $\mathcal{L}_\infty$ of the pencil $\lambda B - A$ then instead of preliminary computation of the projectors P_0 and P_∞ and subsequent using of the formulas (1.3),(1.4) one can do the following. At first compute the orthonomalized matrix pencil $\lambda \overline{B}_{m_0} - \overline{A}_{m_0}$ from the pencil $\lambda B_{m_0} - A_{m_0}$ with $\overline{A}_{m_0}\overline{A}_{m_0}^H + \overline{B}_{m_0}\overline{B}_{m_0}^H = I$. Then define the matrices $S_0 = I - \overline{A}_{m_0}^H \overline{A}_{m_0}$, $S_\infty = I - \overline{B}_{m_0}^H \overline{B}_{m_0}$. When $m_0 \to \infty$ the matrices S_0 and S_∞ converge to Π_0 and Π_∞ respectively.

2.2 Proof of a convergence of the algorithm

We begin with the investigation of the second, main, stage of the algorithm. Remember that before this stage a pair of matrices A_0 and B_0 has the canonical form $A_0 = T_0^{-1}(P_{-0} + P_{+1})$, $B_0 = T_0^{-1}(P_{+0} + P_{-1})$ with a nonsingular N-by-N matrix T_0. We prove now that for the matrices A_m, B_m the following representation is valid:

$$A_m = T_m^{-1}(P_{-0} + P_{2^m}), \; B_m = T_m^{-1}(P_{+0} + P_{-2^m}). \tag{2.9}$$

In order to show this let us rewrite the equation $X_m B_m - Y_m A_m = 0$, taking advantage of the decomposition (1.1):

$$(X_m T_m^{-1})(P_{+0} + P_{-2^m}) - (Y_m T_m^{-1})(P_{-0} + P_{2^m}) = 0,$$

$$\left[(X_m T_m^{-1} Q_r)\begin{pmatrix} I_{N_0} & 0 \\ 0 & J_\infty^{2^m} \end{pmatrix} - (Y_m T_m^{-1} Q_r)\begin{pmatrix} J_0^{2^m} & 0 \\ 0 & I_{N_\infty} \end{pmatrix}\right] Q_r^{-1} = 0.$$

It follows that all the solutions of the equation $X_m B_m - Y_m A_m = 0$ have the form

$$X_m = F_{m+1}\begin{pmatrix} J_0^{2^m} & 0 \\ 0 & I_{N_\infty} \end{pmatrix} Q_r^{-1} T_m, \; Y_m = F_{m+1}\begin{pmatrix} I_{N_0} & 0 \\ 0 & J_\infty^{2^m} \end{pmatrix} Q_r^{-1} T_m$$

with arbitrary matrix F_{m+1}.

Since the matrices X_m and Y_m must satisfy the equation $X_m X_m^H + Y_m Y_m^H = I$, the matrix F_{m+1} should be nonsingular. Therefore, a solution of the system (2.6) is representable in the form

$$X_m = T_{m+1}^{-1}(P_{-0} + P_{2^m}) T_m, \; Y_m = T_{m+1}^{-1}(P_{+0} + P_{-2^m}) T_m, \tag{2.10}$$

where $T_{m+1} = Q_r F_{m+1}^{-1}$.

Substitute (2.10) in the equality $X_m X_m^H + Y_m Y_m^H = I$, and after simple transformations we obtain the recurrent formula for the matrices T_m:

$$T_{m+1} T_{m+1}^H = (P_{-0} + P_{2^m}) T_m T_m^H (P_{-0}^H + P_{2^m}^H) + (P_{+0} + P_{-2^m}) T_m T_m^H (P_{+0}^H + P_{-2^m}^H). \tag{2.11}$$

From the formula (2.11) it follows that the solution of the system (2.6) is unique to an orthogonal left-hand side premultiplier, i.e. all the solutions of the system (2.6) have the form $X_m = U_m \overline{X}_m$, $Y_m = U_m \overline{Y}_m$, where U_m is an arbitrary unitary matrix but $\overline{X}_m$, $\overline{Y}_m$ present some partial solution of the system (2.6).

Now investigate the third stage of the algorithm. Taking into account (2.9) we rewrite the system (2.8) as:

$$Z_0 + Z_\infty = I, \; T_{m_0}^{-1}[(P_{-0} + P_{2^{m_0}}) Z_0 - (P_{+0} + P_{-2^{m_0}}) Z_\infty] = 0. \tag{2.12}$$

Let the matrix ξ be such that $Z_0 = P_{+0} + \xi$ and $Z_\infty = P_{-0} - \xi$. Then from the second equation of the system (2.12) we obtain

$$(P_{-0} + P_{2^{m_0}})(P_{+0} + \xi) - (P_{+0} + P_{-2^{m_0}})(P_{-0} - \xi) = 0,$$

$$(I + P_{2^{m_0}} + P_{-2^{m_0}})\xi = P_{-2^{m_0}} - P_{2^{m_0}}.$$

Theorem 4 guarantees the estimate

$$\max\{\|P_{2^{m_0}}\|, \|P_{-2^{m_0}}\|\} \le \delta = \sqrt{\omega} e^{-2^{m_0}/(1+\omega)}.$$

Assume that m_0 is large enough and the estimate $2\delta < 1$ holds. Then

$$\|P_{-2^{m_0}} - P_{2^{m_0}}\| \le 2\delta, \quad \sigma_{\min}(I + P_{2^{m_0}} + P_{-2^{m_0}}) \ge 1 - 2\delta, \quad \|\xi\| \le \frac{2\delta}{1 - 2\delta}.$$

Finally, we obtain an expression for the convergence rate of the algorithm:

$$\max\{\|Z_0 - P_0\|, \|Z_\infty - P_\infty\|\} \le \frac{2\sqrt{\omega} e^{-2^{m_0}/(1+\omega)}}{1 - 2\sqrt{\omega} e^{-2^{m_0}/(1+\omega)}}. \tag{2.13}$$

This estimate is valid for such values of m_0 for which the denominator in (2.13) is positive.

Apart from the computation of the projectors onto the deflating subspaces the algorithm possesses one more unexpected property: in the third stage the matrix H is easily computed.

Let us investigate the behaviour of the matrices T_m. By induction one can easily derive from (2.11) the formula

$$T_k T_k^H = \sum_{j=0}^{2^k - 1} (P_j + P_{j+1-2^k}) T_0 T_0^H (P_j^H + P_{j+1-2^k}^H), \tag{2.14}$$

where $P_{j+1-2^k} = P_{-0}$ if $j = 2^k - 1$. Compute the limit of (2.14) for $k \to \infty$. In order to do that we estimate the difference

$$\|T_k T_k^H - \sum_{j=-\infty}^{\infty} P_j T_0 T_0^H P_j^H\| \le \|\sum_{|j| \ge 2^k} P_j T_0 T_0^H P_j^H\| + 2\|\sum_{j=0}^{2^k-1} P_j T_0 T_0^H P_{j+1-2^k}^H\| \le$$

$$\le 2\|T_0\|^2 \left[\frac{\omega e^{-\frac{2^k+1}{1+\omega}}}{1 - e^{-\frac{2}{1+\omega}}} + \omega \sum_{j=0}^{2^k-1} e^{-\frac{2^k-1}{1+\omega}} \right] = 2\omega\|T_0\|^2 \left[\frac{e^{-\frac{2^k+1}{1+\omega}}}{1 - e^{-\frac{2}{1+\omega}}} + 2^k e^{-\frac{2^k-1}{1+\omega}} \right].$$

The latter estimate allows us to assert that

$$T_k T_k^H \to \sum_{j=-\infty}^{\infty} P_j T_0 T_0^H P_j^H \tag{2.15}$$

when $k \to \infty$. The decomposition (2.9) implies the formula

$$(A_{m_0} + B_{m_0})^{-1} = (I + P_{2^{m_0}} + P_{-2^{m_0}})^{-1} T_{m_0}.$$

For $m_0 \to \infty$ it follows that

$$(A_{m_0} + B_{m_0})^{-1} (A_{m_0}^H + B_{m_0}^H)^{-1} \to \sum_{j=-\infty}^{\infty} P_j T_0 T_0^H P_j^H. \tag{2.16}$$

Since $A_0 A_0^H + B_0 B_0^H = I$, (1.25) implies that

$$(A_{m_0} + B_{m_0})^{-1} (A_{m_0}^H + B_{m_0}^H)^{-1} \to$$

$$\rightarrow H = \frac{1}{2\pi}\int_0^{2\pi}(B - e^{i\phi}A)^{-1}(AA^H + BB^H)(B^H - e^{-i\phi}A^H)^{-1}d\phi. \tag{2.17}$$

One can obtain the following more precise rate of convergence in (2.17):

$$\|H - (A_k + B_k)^{-1}(A_k^H + B_k^H)^{-1}\| \leq$$

$$\leq \frac{4\omega^{3/2}e^{-2^k/(1+\omega)} + 2\omega^2 e^{-2^{k+1}/(1+\omega)} + 2^{k+2}\omega e^{-2^k/(1+\omega)}}{1 - 4\sqrt{\omega}e^{-2^k/(1+\omega)}}. \tag{2.18}$$

This estimate holds, of course, when $1 - 4\sqrt{\omega}e^{-\frac{2^k}{1+\omega}} > 0$.

Finally, consider the question on avoiding of a failure during the computations for the third stage of the algorithm. Recall once again that a solution to the system (2.8) is of the form $Z_0 = (A_{m_0} + B_{m_0})^{-1}B_{m_0}$, $Z_\infty = (A_{m_0} + B_{m_0})^{-1}A_{m_0}$. Therefore, one has to find out about the condition number of the matrix $A_{m_0} + B_{m_0}$. As $\|A_{m+1}\| \leq \|A_m\|$, $\|B_{m+1}\| \leq \|B_m\|$, we have

$$\|A_{m_0} + B_{m_0}\| \leq \|A_0\| + \|B_0\| \leq 2. \tag{2.19}$$

If m_0 is large enough then

$$\|(A_{m_0} + B_{m_0})^{-1}\| \leq \sqrt{\omega} + 1. \tag{2.20}$$

2.3 The stopping criteria for the algorithm

In practice the necessity to detect the convergence of the algorithm occurs. The maximal number of iterations may be extracted from the inequality (2.13). If we limit the range of ω by the interval $[1, \omega_{\max}]$ then the maximal number of iterations m_0 can be equal, for example, to

$$m_0 = \text{integer part of } \{1 + \log[(1+\omega)\log(2\sqrt{\omega}/\epsilon)]/\log 2\}, \tag{2.21}$$

where ϵ is the relative rounding error for arithmetic operations.

In order to stop before m_0 iterations we make use of the matrix identity $P^2 - P = 0$ which means that N-by-N matrix P is a projector.

Suppose now that we have N-by-N matrices A_∞ and B_∞ which are suspected to be a result of the convergence of the algorithm. Compute the matrices

$$P_\infty = (A_\infty + B_\infty)^{-1}B_\infty \text{ and } R = [P_\infty(A_\infty + B_\infty)^{-1}, (I - P_\infty)(A_\infty + B_\infty)^{-1}].$$

Now we estimate the difference

$$D = \|\frac{1}{2\pi}\int_0^{2\pi}(B_\infty - A_\infty e^{i\phi})^{-1}(B_\infty - A_\infty e^{i\phi})^{-H}d\phi - RR^H\|$$

in terms of the quantities

$$\delta = \|P_\infty^2 - P_\infty\| \text{ and } r = \|R\|.$$

Theorem 5 *If $\delta < 1/4$ then there exist a matrix $\hat{P}$ such that $\hat{P}^2 = \hat{P}$ and $\|\hat{P} - P_\infty\| \leq \delta/(1-2\delta)$.*

PROOF. Exploit the Schur form of the matrix P_∞:

$$P_\infty = U^H \begin{pmatrix} \alpha_1 & & * \\ & \alpha_2 & \\ & & \ddots \\ O & & \alpha_N \end{pmatrix} U, \quad U^H U = I.$$

We have $|\alpha_i^2 - \alpha_i| \leq \delta$, therefore if $\delta < 1/4$ either $|\alpha_i| \leq \delta/(1-2\delta)$ or $|\alpha_i - 1| \leq \delta/(1-2\delta)$.

Thus the eigenvalues of the matrix P_∞ are separated into two groups: the eigenvalues in the neighbourhood of zero and the eigenvalues in the neighbourhood of one. Choose the Schur form of P_∞ taking into account this structure:

$$P_\infty = U^H \begin{pmatrix} K_0 & K \\ 0 & K_1 \end{pmatrix} U,$$

where K_0 has all eigenvalues from the neighbourhood of zero, but K_1 has all eigenvalues from the neighbourhood of one. Then $P_\infty^2 - P_\infty = \Phi$ implies that

$$\begin{pmatrix} K_0^2 - K_0 & K_0K + KK_1 - K \\ 0 & K_1^2 - K_1 \end{pmatrix} = U\Phi U^* = \begin{pmatrix} \Phi_0 & \Phi \\ 0 & \Phi_1 \end{pmatrix}.$$

Therefore, we have two matrix equations:

$$K_0^2 - K_0 = \Phi_0, \quad K_1^2 - K_1 = \Phi_1.$$

A solution to $K_0^2 - K_0 = \Phi_0$ can be written in the form $K_0 = \frac{1}{2}(I - \sqrt{I + 4\Phi_0})$, where the square root from the matrix is taken by the principal value, i.e.

$$\sqrt{I + 4\Phi_0} = I + \frac{1}{2}(4\Phi_0) + \frac{\frac{1}{2}(\frac{1}{2} - 1)}{2!}(4\Phi_0)^2 + \ldots.$$

Hence,

$$\|K_0\| \le \frac{1}{2}(1 - \sqrt{1 - 4\|\Phi_0\|}) \le \frac{\|\Phi_0\|}{1 - 2\|\Phi_0\|} \le \frac{\|\Phi\|}{1 - 2\|\Phi\|} = \frac{\delta}{1 - 2\delta}.$$

In a similar way

$$\|K_1 - I\| \le \frac{\delta}{1 - 2\delta}.$$

Let us construct the matrix $\hat{P}$ by the formula

$$\hat{P} = U^H \begin{pmatrix} 0 & K \\ 0 & I \end{pmatrix} U,$$

then, evidently, $\hat{P}^2 - \hat{P} = 0$ and

$$\|\hat{P} - P_\infty\| = \left\| \begin{pmatrix} K_0 & 0 \\ 0 & K_1 - I \end{pmatrix} \right\| \le \frac{\delta}{1 - 2\delta}.$$

■

Introduce the following notations:

$$\Delta = \hat{P} - P_\infty, \quad \hat{B} = (A_\infty + B_\infty)\hat{P}, \quad \hat{A} = (A_\infty + B_\infty)(I - \hat{P}),$$

$$\hat{\omega} = \|\frac{1}{2\pi}\int_0^{2\pi} (\hat{B} - \hat{A}e^{i\phi})^{-1}(\hat{B} - \hat{A}e^{i\phi})^{-H} d\phi\|.$$

Then $\hat{A} + \hat{B} = A_\infty + B_\infty$, $\hat{A} - A_\infty = (A_\infty + B_\infty)(-\Delta)$, $\hat{B} - B_\infty = (A_\infty + B_\infty)\Delta$,

$$\hat{D} = \frac{1}{2\pi}\int_0^{2\pi} (B_\infty - A_\infty e^{i\phi})^{-1}(B_\infty - A_\infty e^{i\phi})^{-H} d\phi - \frac{1}{2\pi}\int_0^{2\pi} (\hat{B} - \hat{A}e^{i\phi})^{-1}(\hat{B} - \hat{A}e^{i\phi})^{-H} d\phi =$$

$$= \frac{1}{2\pi}\int_0^{2\pi} (\hat{B} - \hat{A}e^{i\phi})^{-1}\langle\{I - [(B_\infty - \hat{B}) - (A_\infty - \hat{A})e^{i\phi}](B_\infty - A_\infty e^{i\phi})^{-1}\}\times$$

$$\times\{I - [(B_\infty - \hat{B}) - (A_\infty - \hat{A})e^{i\phi}](B_\infty - A_\infty e^{i\phi})^{-1}\}^H - I\rangle(\hat{B} - \hat{A}e^{i\phi})^{-H} d\phi,$$

$$\|\hat{D}\| \le \hat{\omega}[(1 + 2\|\Delta\|\|A_\infty + B_\infty\|\|(B_\infty - A_\infty e^{i\phi})^{-1}\|)^2 - 1].$$

As $(\hat{B}-\hat{A}e^{i\phi})^{-1}=\hat{P}(\hat{A}+\hat{B})^{-1}+(I-\hat{P})(\hat{A}+\hat{B})^{-1}e^{-i\phi}$, then $\|(\hat{B}-\hat{A}e^{i\phi})^{-1}\|\le\sqrt{2\hat{\omega}}$,

$$\|(B_\infty-A_\infty e^{i\phi})^{-1}\|\le\|(\hat{B}-\hat{A}e^{i\phi})^{-1}\|/(1-2\|\Delta\|\|A_\infty+B_\infty\|\|(\hat{B}-\hat{A}e^{i\phi})^{-1}\|)\le$$

$$\le\sqrt{2\hat{\omega}}/(1-2\|\Delta\|\|A_\infty+B_\infty\|\sqrt{2\hat{\omega}}),$$

$$\|\hat{D}\|\le\hat{\omega}\frac{4\|\Delta\|\|A_\infty+B_\infty\|\sqrt{2\hat{\omega}}}{1-4\|\Delta\|\|A_\infty+B_\infty\|\sqrt{2\hat{\omega}}}\le\hat{\omega}\frac{4\delta\|A_\infty+B_\infty\|\sqrt{2\hat{\omega}}}{1-2\delta-4\delta\|A_\infty+B_\infty\|\sqrt{2\hat{\omega}}}.$$

Let $\hat{R}=[\hat{P}(\hat{A}+\hat{B})^{-1},(I-\hat{P})(\hat{A}+\hat{B})^{-1}]$, then

$$\|\hat{R}-R\|=\|[\Delta(A_\infty+B_\infty)^{-1},-\Delta(A_\infty+B_\infty)^{-1}]\|\le\sqrt{2}\|\Delta\|\|(A_\infty+B_\infty)^{-1}\|,$$

$$\|(A_\infty+B_\infty)^{-1}\|\le\sqrt{2}\|R\|,\ \ \|\hat{R}-R\|\le2\|\Delta\|\|R\|.$$

Therefore, $|\sqrt{\hat{\omega}}-r|\le2\|\Delta\|r$, $D\le\hat{D}+\|\hat{R}\hat{R}^H-RR^H\|\le$

$$\le\hat{D}+\|(\hat{R}-R)(\hat{R}-R)^H\|+\|R(\hat{R}^H-R)\|+\|(\hat{R}-R)R^H\|\le$$

$$\le\hat{D}+r^2(4\|\Delta\|+4\|\Delta\|^2)\le r^2\frac{4\sqrt{2}\delta\|A_\infty+B_\infty\|r}{1-8\delta-4\sqrt{2}\delta\|A_\infty+B_\infty\|r}+r^2\frac{4\delta}{1-4\delta},$$

$$D\le r^2\frac{4\delta(1+\sqrt{2}\|A_\infty+B_\infty\|r)}{1-8\delta-4\sqrt{2}\delta\|A_\infty+B_\infty\|r}. \tag{2.22}$$

Finally, some words about the relation of R to

$$H=\frac{1}{2\pi}\int_0^{2\pi}(B_0-A_0e^{i\phi})^{-1}(B_0-A_0e^{i\phi})^{-H}d\phi.$$

One can show that the following invariance property holds

$$\begin{array}{l}\frac{1}{2\pi}\int_0^{2\pi}(B_{m+1}-A_{m+1}e^{i\phi})^{-1}(B_{m+1}-A_{m+1}e^{i\phi})^{-H}d\phi=\\ =\frac{1}{2\pi}\int_0^{2\pi}(B_m-A_me^{i\phi})^{-1}(B_m-A_me^{i\phi})^{-H}d\phi.\end{array} \tag{2.23}$$

Therefore,

$$RR^H=H. \tag{2.24}$$

It is worth to remark here that (2.24) is valid for floating point calculations only if r is not very large.

Chapter 3

The problem of separation of the matrix spectrum by the imaginary axis

The developed method of computing the spectrum dichotomy problem for a linear matrix pencil which is regular with respect to the unit circle can be successfully applied to the spectrum dichotomy problem for a single matrix with respect to the imaginary axis.

3.1 Eigenelements of a matrix which has no eigenvalues on the imaginary axis

The main results of this section are obtained by S.K. Godunov and A.Ja. Bulgakov [11, 6].

Given an N-by-N matrix A which has no eigenvalues on the imaginary axis. Differentiable everywhere, except zero, the bounded matrix function $G(t)$ satisfying the differential equation

$$\frac{d}{dt}G(t) - AG(t) = \delta(t)I \tag{3.1}$$

and the condition at $t = 0$

$$G(+0) - G(-0) = I, \tag{3.2}$$

is called *the Green function of the differential operator* $I\frac{d}{dt} - A$. By $\delta(t)$ we denoted the delta-function with the centre at zero.

According to the Jordan theorem there exists a nonsingular matrix Q which reduces the matrix A to the canonical form

$$Q^{-1}AQ = \text{block diag}\{J_+, J_-\}, \tag{3.3}$$

where the matrix J_+ contains all the Jordan blocks with the eigenvalues within the left-hand side halfplane while the matrix J_- contains all the Jordan blocks with the eigenvalues within the right-hand side halfplane.

With the help of (3.3) one can easily verify that the Green function is unique and equal to

$$G(t) = \begin{cases} Q\begin{pmatrix} e^{tJ_+} & 0 \\ 0 & 0 \end{pmatrix}Q^{-1}, & t \geq 0, \\ Q\begin{pmatrix} 0 & 0 \\ 0 & e^{-tJ_-} \end{pmatrix}Q^{-1}, & t \leq 0. \end{cases} \tag{3.4}$$

From here it follows, firstly, that $G(+0)$ is the projector onto the invariant subspace of matrix A corresponding to the eigenvalues inside the left-hand halfplane but $-G(-0)$ is the projector

onto the invariant subspace of A corresponding to the eigenvalues inside the right-hand halfplane. Secondly, the following equalities hold:

$$\begin{aligned} G(t)G(s) &= G(t+s), \quad ts \geq 0, \\ G(t)G(s) &= 0, \quad ts \leq 0. \end{aligned} \tag{3.5}$$

Apply Fourier transformation to the equation (3.1):

$$(i\xi I - A)\hat{G}(\xi) = I. \tag{3.6}$$

The Parseval's equality gives the matrix

$$H_A = \int_{-\infty}^{\infty} G^H(t)G(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} (i\xi I - A)^{-H}(i\xi I - A)^{-1}d\xi. \tag{3.7}$$

Let us refer to the number

$$\kappa(A) = 2\|A\|\|H_A\| \geq 1 \tag{3.8}$$

as *the dichotomy parameter of matrix A*. This parameter is the criterion of absence of the eigenvalues of matrix A from the imaginary axis and a small neighbourhood of it.

The parameter $\kappa(A)$ allows to estimate the magnitude of $\sigma_{\min}(i\xi I - A)$:

$$\max_{\xi}\{2\|A\|\|(i\xi I - A)^{-1}\|\} < 14\kappa(A). \tag{3.9}$$

Let $\tilde{A}$ be a perturbation of the matrix A and $\|\tilde{A} - A\| \leq \delta\|A\|$ with quite small tolerance δ. Then

$$\left|\frac{\kappa(\tilde{A}) - \kappa(A)}{\kappa(A)}\right| \leq \delta\frac{1 + 14\kappa(A)}{1 - 14\delta\kappa(A)} \leq \frac{15\delta\kappa(A)}{1 - 14\delta\kappa(A)}. \tag{3.10}$$

One can prove a fundamental theorem (analogue of Theorem 4) about an estimate of the Green matrix $G(t)$ in terms of $\kappa(A)$.

Theorem 6 *If* $\det(i\xi I - A) \neq 0$ *for all real* ξ, *then the following estimate holds*

$$\|G(t)\| \leq \sqrt{\kappa}e^{-|t|\|A\|/\kappa}. \tag{3.11}$$

3.2 Modification of the algorithm

The algorithm to compute the projectors P_0 and P_∞ together with the generalized Lyapunov matrix can be slightly modified in order to compute the projectors $G(+0)$, $-G(-0)$ and the matrix H_A for the matrix A having no eigenvalues on the imaginary axis. The modified algorithm also consists of three stages.

Stage I. At first, compute the matrix exponential of the matrix $B = A^H/(2\|A\|)$. The invariant subspaces of the matrix A^H associated with the eigenvalues on the left-hand (right-hand) side from the imaginary axis obviously coincide with the deflating subspaces of the matrix pencil $\lambda I - e^B$ associated with the eigenvalues inside (outside) the unit circle.

Compute the matrix L satisfying the equation

$$LL^H = \int_0^1 e^{tB}e^{tB^H}dt \tag{3.12}$$

and define the matrix pencil

$$\lambda B_0 - A_0 = L^{-1}(\lambda I - e^B). \tag{3.13}$$

Stage II. This stage remains without any changes. Recall that here we carry out iterations to get the matrix pair A_m, B_m from A_{m-1}, B_{m-1} for $m = 1, 2, \ldots, m_0$. For example, we can apply to a block matrix

$$\begin{pmatrix} -A_{m-1} & B_{m-1} & 0 \\ 0 & \boxed{-A_{m-1}} & B_{m-1} \end{pmatrix}$$

the orthogonal transformations from the left-hand side which annihilate the block in the frame. As a result we obtain the matrix

$$\begin{pmatrix} C_m & R_m & D_m \\ -A_m & 0 & B_m \end{pmatrix}$$

containing the necessary matrices A_m, B_m.

Stage III. Compute the matrices

$$H_{m_0} = (A_{m_0} + B_{m_0})^{-1}(A_{m_0}^H + B_{m_0}^H)^{-1}, \quad G_{m_0} = [(A_{m_0} + B_{m_0})^{-1} B_{m_0}]^H. \tag{3.14}$$

For $m_0 \to \infty$ we have $H_{m_0} \to 2\|A\| H_A$, $G_{m_0} \to G(+0)$.

Consider the question about the rate of convergence for the proposed modified algorithm. It is not difficult to show by means of the estimate (3.11) and the formula

$$P_k = G^H\left(\frac{k}{2\|A\|}\right), \tag{3.15}$$

that the inequality (2.13) is transformed into the following:

$$\|(A_{m_0} + B_{m_0})^{-1} B_{m_0} - G^H(+0)\| \le \frac{2\sqrt{\kappa} e^{-2^{m_0-1}/\kappa}}{1 - 2\sqrt{\kappa} e^{-2^{m_0-1}/\kappa}}. \tag{3.16}$$

One can also deduce the rate of convergence for the generalized Lyapunov matrix:

$$\|2\|A\| H_A - (A_{m_0} + B_{m_0})^{-1}(A_{m_0}^H + B_{m_0}^H)^{-1}\| \le$$

$$\le \frac{2\kappa^2 e^{-2^{m_0}/\kappa} + 4\kappa^{3/2} e^{-2^{m_0-1}/\kappa} + 2^{m_0+1}\kappa e^{-2^{m_0-1}/\kappa}}{1 - 4\sqrt{\kappa} e^{-2^{m_0-1}/\kappa}}. \tag{3.17}$$

Chapter 4

On the implementation of the algorithm

An implementation of the algorithm described in chapters 2 and 3 has been developed for parallel computing systems with shared memory. In order to keep portability we have used BLAS routines for the elementary matrix operations [10]. The parallel processing capabilities are assumed to be exploited mostly inside the BLAS kernels. So we do not explicitly deal with the problems of synchronization between processors and leave them for developers of compilers and the BLAS routines. The programming technique used is clearly stated in [10] and in chapter 1 of [14].

The developed implementation is a part of library LINA which contains solvers for basic linear algebra problems with guaranteed accuracy. In the design of LINA the LAPACK programming style was kept [1], so LINA can be used as a complement to LAPACK with a few small modifications. It may also be used independently. In LINA we exploit a few modified LAPACK routines in order to maintain this independency. These modified routines pertain to the Householder transformations.

One can see below that the most time consuming parts of the algorithm are the matrix-matrix products and the Householder transformations. The first one is a Level 3 BLAS operation. The Householder transformations are used to reduce matrices to upper triangular form (QR factorisation) and the Householder transformations computed in the QR factorisation are applied to some matrices. We make use of the blocked versions of the Householder method as in the papers [4, 23].

4.1 The second stage of the algorithm

Recall that at this stage the N-by-N matrix pencil $\lambda B_{m+1} - A_{m+1}$ is calculated from the pencil $\lambda B_m - A_m$ by means of the formulas $A_{m+1} = X_m A_m$ and $B_{m+1} = Y_m B_m$, where N-by-N matrices X_m, Y_m satisfy the system $X_m X_m^H + Y_m Y_m^H = I$, $X_m B_m - Y_m A_m = 0$.

To compute the matrices X_m and Y_m we use the following procedure:

1. At first, the QR factorization for the $2N$-by-N matrix $\begin{pmatrix} B_m \\ -A_m \end{pmatrix}$ is computed, i.e. an $2N$-by-$2N$ orthogonal matrix Q_m and an N-by-N upper triangular matrix R_m are calculated such that

$$Q_m \begin{pmatrix} B_m \\ -A_m \end{pmatrix} = \begin{pmatrix} R_m \\ 0 \end{pmatrix}.$$

The matrix Q_m is constructed as the product of the Householder transformations

$$Q_m = \prod_{j=1}^{N} (I - u_j u_j^T), \text{ where } u_j^T u_j = 2. \tag{4.1}$$

The algorithm to compute the QR factorization is described in almost every handbook on numerical linear algebra, particularly, in [14].

2. Then we compute the matrix

$$\begin{pmatrix} X_m^T \\ Y_m^T \end{pmatrix} = Q_m^T \begin{pmatrix} 0 \\ I \end{pmatrix}.$$

Let us look at this procedure more attentively. Note that the matrix Q_m is not explicitly formed and it is available in a product form, since the vectors u_j are simply recorded in place of the columns they have been used to annihilate. Therefore, one has to optimize the calculation of

$$\prod_{j=1}^{K}(I - u_j u_j^T)C, \tag{4.2}$$

where C is an arbitrary matrix with $2N$ rows and $K \leq N$. If we do it consecutively then the procedure may be coded in terms of the Level 2 BLAS:

$$\begin{aligned} &y_j = C_j^T u_j \text{ (matrix} \times \text{ vector)}, \\ &C_{j+1} = C_j - u_j y_j^T \text{ (rank-one modification)}, \end{aligned} \tag{4.3}$$

where we denote $\prod_{i=1}^{j}(I - u_j u_j^T)C$ by C_j.

We make use of the blocked version of the Householder method suggested in [23]. The product (4.2) is calculated by means of accumulation of several Householder transformations into a single block transformation:

$$\prod_{j=1}^{K}(I - u_j u_j^T)C = [\prod_{i=1}^{K/k} \prod_{j=(i-1)*k+1}^{i*k} (I - u_j u_j^T)]C.$$

One can show that the product of k Householder transformations may be written as

$$\prod_{j=1}^{k}(I - u_j u_j^T) = I - UTU^T,$$

where $U = (u_1, u_2, \ldots, u_k)$ and T is an upper unit triangular matrix. This fact can be easily verified by induction. If $V = I - UTU^T$, then

$$V(I - uu^T) = I - (U, u)\begin{pmatrix} T & h \\ 0 & 1 \end{pmatrix}(U, u)^T$$

where $h = -TU^T u$. The computation of

$$(I - UTU^T)C = C - UT(U^T C)$$

is rich in matrix-matrix operations. The timing results showing the performance rate of such version of the Householder method can be found in [10].

Thus, the second stage of the algorithm may be implemented either in terms of the Level 2 BLAS or in terms of the Level 3 BLAS.

4.2 The third stage of the algorithm

At the third stage we need to compute the inverse of the matrix $A_\infty + B_\infty$, a few matrix-matrix products and some singular values.

Our general strategy to compute matrix inverses and the singular values consists of a reduction of the problem for dense matrices to the problem for bidiagonal matrices with the help of the

Householder transformations. For example, let A be a given nonsingular N-by-N matrix. Applying the Householder transformations to A from the left and the right hand sides [14] we obtain the bidiagonal factorization

$$PAQ = B, \tag{4.4}$$

where P and Q are the products of the Householder transformations and B is an upper bidiagonal matrix. Then,

$$A^{-1} = QB^{-1}P. \tag{4.5}$$

As to (4.5), the same technique like in the previous section is suitable for an implementation of applying the Householder transformations in Q and P.

Evidently, the singular values of the matrix A coincide with the singular values of the bidiagonal matrix B. Denoting the bidiagonal matrix as

$$B = \begin{pmatrix} d_1 & e_2 & & & \mathrm{O} \\ & d_2 & e_3 & & \\ & & \ddots & \ddots & \\ & & & d_{N-1} & e_N \\ \mathrm{O} & & & & d_N \end{pmatrix},$$

one can show that the singular values of B coincide with the N largest eigenvalues of the symmetric tridiagonal $2N$-by-$2N$ matrix

$$T = \begin{bmatrix} 0 & d_1 & & & & & & \\ d_1 & 0 & e_2 & & & & & \\ & e_2 & 0 & d_2 & & & \mathrm{O} & \\ & & d_2 & 0 & e_3 & & & \\ & & & e_3 & 0 & d_3 & & \\ & \mathrm{O} & & & \ddots & \ddots & \ddots & \\ & & & & e_N & 0 & d_N & \\ & & & & & d_N & 0 & \end{bmatrix}.$$

An excellent strategy to compute a few eigenvalues of a symmetric tridiagonal matrix on parallel computers is suggested in [18]. This method is based on the Sturm sequences approach.

Now let us discuss the parallel aspects of the computation of (4.4). The standard implementations of the bidiagonalization procedure are easily coded in terms of Level 2 BLAS. If

$$P = \prod_{i=1}^{N-1} (I - u_i u_i^T), \quad u_i^T u_i = 2,$$

$$Q = \prod_{i=1}^{N-1} (I - v_i v_i^T), \quad v_i^T v_i = 2, \quad v_{N-1} = 0,$$

$$A_0 = A, \quad A_j = [\prod_{i=1}^{j} (I - u_i u_i^T)] A [\prod_{i=1}^{j} (I - v_i v_i^T)],$$

then the j-th stage of the procedure consists of two steps:

1. calculate the vector u_j from the vector $A_j[j:N, j:j]$ and the matrix $\tilde{A}_j = (I - u_j u_j^T) A_{j-1}$;
2. calculate the vector v_j from the vector $\tilde{A}_j[j:j, j+1:N]$ and the matrix $A_j = \tilde{A}_j (I - v_j v_j^T)$.

Here we denote by $M[i_1 : i_2, j_1 : j_2]$ the submatrix of a matrix M consisting of the rows $i_1, i_1 + 1, \ldots, i_2$ and of the columns $j_1, j_1 + 1, \ldots, j_2$. These two steps are coded by means of Level 2 BLAS similar to (4.3).

Our block version of the bidiagonalization differs from the compact block Householder method described in [23]. The block version of the bidiagonalization is a further development of the block Householder method from [4]. The crucial idea is to use the following representation:

$$A_{jl} = \prod_{i=1}^{j}(I - u_i u_i^T)A\prod_{i=1}^{l}(I - v_i v_i^T) = A - U_j X_j^T - Y_l V_l^T, \tag{4.6}$$

where $l = j-1$ or $l = j$,

$$U_j = [u_1, u_2, \ldots, u_j], \quad V_l = [v_1, v_2, \ldots, v_l],$$

X_j and Y_l are j-by-N and l-by-N matrices respectively. The vectors u_i defining the Householder transformations are placed into the vector-columns $A[i:N, i:i]$ in the lower triangle. Accordingly, the vectors v_i are placed into the vector-rows $A[i:i, i+1:N]$ in the strict upper triangle.

In order to compute the j-th column of X_j, x_j, we compute the j-th column of $A_{j-1,j-1}$, a_j, with the help of the formula $A - U_{j-1}X_{j-1}^T - Y_{j-1}V_{j-1}^T$. Then the vector u_j is constructed from a_j by the usual procedure

$$u_j = \frac{1}{\sqrt{2(\|a_j\|^2 + \|a_j\|\|a_{jj}|)}}(a_j + \mathrm{sign}(a_{jj})\|a_j\|e_j), \tag{4.7}$$

where a_{jj} is the first component of $a_j = (a_{jj}, \ldots, a_{Nj})^T$, the sign function satisfies $\mathrm{sign}(0) = 1$ and $e_j = (1, 0, \ldots, 0)^T$. Therefore,

$$A - (U_{j-1}, u_j)(X_{j-1}, x_j)^T - Y_{j-1}V_{j-1}^T = (I - u_j u_j^T)(A - U_{j-1}X_{j-1}^T - Y_{j-1}V_{j-1}^T) =$$

$$= (A - U_{j-1}X_{j-1}^T - Y_{j-1}V_{j-1}^T) - u_j u_j^T(A - U_{j-1}X_{j-1}^T - Y_{j-1}V_{j-1}^T),$$

$$x_j = (A - U_{j-1}X_{j-1}^T - Y_{j-1}V_{j-1}^T)^T u_j. \tag{4.8}$$

In order to compute the j-th column of Y_j, y_j, we compute the j-th row of $A_{j,j-1}$, $\tilde{a}_j$, with the help of the formula $A - U_j X_j^T - Y_{j-1}V_{j-1}^T$. Then the vector v_j is constructed from $\tilde{a}_j$ in a similar way as in (4.7). Therefore,

$$y_j = (A - U_j X_j^T - Y_{j-1}V_{j-1}^T)v_j. \tag{4.9}$$

So far we described the strategy to accumulate several Householder transformations during the bidiagonalization of matrix A. This may be coded in terms of Level 2 BLAS. When we have accumulated k left and right hand side Householder transformations in the matrices U_k, X_k, V_k, Y_k, then the submatrix $A[k+1:N, k+1:N]$ can be updated by Level 3 BLAS operations: $A - U_k X_k^T - Y_k V_k^T$.

Note that these bidiagonalization algorithms are easily extended to the case of an arbitrary rectangular matrix A. Timing results to compare Level 2 BLAS and Level 3 BLAS implementations will be presented.

4.3 The first stage of the algorithm

We have two different procedures for the first stage of the algorithm: the first one for the spectrum dichotomy problem with respect to the unit circle and the second one for the spectrum dichotomy problem with respect to the imaginary axis. Let us discuss them in this order.

Given the N-by-N matrix pencil $\lambda B - A$. We compute an orthonormalized pencil $\lambda B_0 - A_0$ which satisfies the matrix system

$$\begin{cases} A_0 A_0^H + B_0 B_0^H = I, \\ \lambda B - A = L(\lambda B_0 - A_0), \ \det L \neq 0. \end{cases}$$

We compute the matrices A_0 and B_0 as follows:

1. Compute the bidiagonal factorization of the N-by-$2N$ matrix $(A\ B)$:

$$P(A\ B)Q = (\Sigma\ 0),$$

where P is an N-by-N orthogonal matrix, Q is a $2N$-by-$2N$ orthogonal matrix and Σ is an N-by-N lower bidiagonal matrix.

2. If the condition number of Σ is not large enough, compute

$$(A_0\ B_0) = (I\ 0)Q^{-1}.$$

The condition number of Σ is equal to $\mathrm{cond}(\Sigma) = \sigma_{\max}(\Sigma)/\sigma_{\min}(\Sigma)$, where $\sigma_{\max}$ and $\sigma_{\min}$ are the largest and the smallest singular values respectively. In fact, the parameter $\mathrm{cond}(\Sigma)$ is the condition number of the whole first stage of the algorithm. For the implementation we make use of the same technique for the Householder transformations like in the previous sections. For example, for the bidiagonalization we use noncompact accumulation of the Householder transformations of the form (4.6). However, the compact accumulation is used to apply them to a matrix.

Now consider the first stage of the modified algorithm, i.e. the algorithm for the spectrum dichotomy problem with respect to the imaginary axis. It is more sophisticated than the previous one. At first, we need to compute the norm of a given N-by-N matrix A. This should be done by bidiagonalization and subsequently by the computation of the largest singular value of the bidiagonal matrix. The possibilities for parallelism in the procedure are discussed above.

Afterwards we have to compute the matrix $B = A^T/(2\|A\|)$ the 2-norm of which is equal to $1/2$. Now one has to compute the matrices

$$e^B = \sum_{j=0}^{\infty} \frac{B^j}{j!} = I + B + \frac{B^2}{2!} + \ldots + \frac{B^j}{j!} + \ldots,$$

$$C = \int_0^1 e^{tB} e^{tB^T}\, dt.$$

To compute e^B we make use of two methods: Taylor series and Padé approximation [21]. These methods are both suitable for our case because the matrix B has a small norm.

When using Taylor series the polynomial of a matrix B has to be computed:

$$T_k(B) = I + B + \ldots + \frac{B^j}{j!} + \ldots + \frac{B^k}{k!}.$$

The number of terms k depends on the relative rounding error ϵ of the computer used. We find k from the relation $\|B\|^{k+1}/(k+1)! \approx \epsilon$. When $\epsilon = 2.22 \times 10^{-16}$, for example, $k = 13$ fits. The following Horner scheme is used to evaluate $T_k(B)$:

$$M_k = I + B/k,\ \ M_j = I + M_{j+1}B/j,\ \ j = k-1, \ldots, 1,\ \ T_k(B) = M_1.$$

Thus, this procedure requires $k-1$ matrix-matrix products which should be implemented in terms of Level 3 BLAS.

When using Padé approximation the rational function $R_{qq}(B)$ of a matrix B has to be computed, where

$$R_{qq}(z) = \frac{\sum_{j=0}^{q} c_j z^j}{\sum_{j=0}^{q} c_j (-z)^j},\quad c_j = \frac{(2q-j)!q!}{(2q)!j!(q-j)!}.$$

One can show [21] that $R_{qq}(B) = e^{B+E}$ with $BE = EB$ and $\|E\| \le \delta = 2^{2-2q}\frac{(q!)^2}{(2q)!(2q+1)!}$. The relation $\delta \approx \epsilon$ determines q ($q = 6$ fits).

In the course of forming $R_{qq}(B)$, we compute the matrices

$$N_{qq}(B) = \sum_{j=0}^{q} c_j B^j,\ \ D_{qq}(B) = \sum_{j=0}^{q} c_j (-1)^j B^j,\ \ R_{qq}(B) = [D_{qq}(B)]^{-1} N_{qq}(B).$$

One way to carry this out is to compute the sequence of powers B, B^2, ..., B^q, next the polynomials $N_{qq}(B)$ and $D_{qq}(B)$ and, finally, the matrix $[D_{qq}(B)]^{-1}N_{qq}(B)$. It requires $q-1$ matrix-matrix products and a solution of N linear systems with the same matrix. In this case the matrix $[D_{qq}(B)]^{-1}N_{qq}(B)$ can be calculated with the help of Gaussian elimination without pivoting [29].

Hence, the Padé approximation requires approximately twice as less floating point operations asthe truncated Taylor series approximant.

In order to compute the integral $C = \int_0^1 e^{tB}e^{tB^T}dt$ involving the exponential of a matrix we also used two methods:

Taylor series This method is based on the representation;

$$\int_0^1 e^{tB}Qe^{tB^T}dt = \sum_{j=1}^{\infty}\frac{\mathcal{L}^j(Q)}{j!}, \tag{4.10}$$

where $Q = Q^T$, $\mathcal{L}(Q) = BQ + QB^T$ is the Lyapunov operator for the matrix B^T and $\mathcal{L}^j(Q) = \mathcal{L}(\ldots(\mathcal{L}(Q))\ldots)$ is a recursive application of $\mathcal{L}$. In fact, the representation (4.10) is the Taylor series for $e^{\mathcal{L}}Q$. Thus, we can approximate C by the truncated series:

$$C_k = \sum_{j=1}^{k}\frac{\mathcal{L}^j(I)}{j!}.$$

The parameter k is defined by means of the relation $1/(k+1)! \approx \epsilon$. For $\epsilon = 2.22 \times 10^{-16}$ we can choose $k = 18$. This procedure may be coded using only $k-2$ matrix-matrix products because $\mathcal{L}(I) = B + B^T$.

Padé approximation A general technique to compute diverse integrals involving the exponential of a matrix is suggested in [29]. This technique combined with the diagonal Padé approximations for the exponential of a matrix was implemented for the spectrum dichotomy problem by Margreet Louter-Nool. Now we sketch the basic idea of this method.

For our case we need to compute the exponential

$$\exp\left(\begin{bmatrix} -A^T & Q \\ 0 & A \end{bmatrix}\right) = \begin{bmatrix} F_2 & G_2 \\ 0 & F_3 \end{bmatrix}. \tag{4.11}$$

It can be shown that $\int_0^1 e^{tB}Qe^{tB^T}dt = F_3^T G_2$. To compute the exponential (4.11), the above stated method of Padé approximants is used. Taking into account a particular structure of the matrices one can verify that this method requires $2q$ matrix-matrix products and a solution of two simultaneous systems of linear equations for N-by-N matrices. As a result of this method the exponential of B and the integral C are computed simultaneously. Note that the previous method computes only C. A very detailed description of this method is contained in [29].

Summing up, for computing e^B and $\int_0^1 e^{tB}e^{tB^T}dt$ the Taylor series method requires 28 matrix-matrix products (12 for computing the exponential and 16 for computing the integral) whereas the Padé approximation method requires 9 matrix-matrix products and the solution of two simultaneous linear systems of the form $AX = B$ with N-by-N matrices A and B. For the solution of these systems the Gaussian elimination without pivoting is suitable. The estimates for floating point operations are valid for the relative precision of arithmetic operations equal to 2.22×10^{-16}. So, the second method should be twice as fast as the first.

At the end, having the matrices e^B and $C = \int_0^1 e^{tB}e^{tB^T}dt$ we need to calculate the matrix pencil $\lambda L^{-1} - L^{-1}e^B$, where the N-by-N matrix satisfies the equation: $LL^T = C$. The matrix L can be computed by any parallel implementation for the Cholesky decomposition [10]. The inverse of the lower triangular matrix L may be computed by a Level 3 BLAS subroutine.

4.4 Timing results

In this section some timing results are presented. The algorithm for the spectrum dichotomy problem has been implemented exploiting BLAS routines. We deal with DOUBLE PRECISION floating point numbers. The main technique used for parallelism is the block version of the Householder method discussed above in detail.

For numerical experiments the ALLIANT FX/4 was used which is installed at CWI, Amsterdam. This computer has four parallel vector processors working with shared memory. We used the BLAS library supplied by the vendor and the optimizing FORTRAN compiler with automatic optimization (vector and concurrent) of FORTRAN loops.

The theoretical peak performance of the ALLIANT FX/4 is rated at 11.8 Mflops for a single processor and 47.2 Mflops for a complex of four concurrent processors. However, the performance rate of DGEMM routine from Level 3 BLAS achieves about 18-20 Mflops at 4 processors.

Now let us consider the timing results for the bidiagonalization procedure. The library LINA contains a routine named DGEBDB which executes both Level 2 BLAS as well as Level 3 BLAS codes for bidiagonalization. Theoretical aspects of these procedures are discussed in Section 4.2. Below (Table 4.1) we give the CPU time in seconds for execution on four concurrent processors. When executing the routine on a single processor the CPU time is approximately three times less than in the case of 4 processors. As a test matrix, an N-by-N matrix A with elements of the form $A_{ij} = i + (j-1) * N$ was taken.

N	BLAS 2	Blocksize for BLAS 3 implementation					
		4	8	16	32	64	128
25	0.03	0.07	0.08	0.08	—	—	—
50	0.08	0.17	0.18	0.19	0.21	—	—
100	0.34	0.50	0.53	0.57	0.65	0.73	—
150	1.08	1.20	1.22	1.31	1.46	1.72	2.14
200	2.63	2.40	2.44	2.57	2.82	3.36	4.26
250	5.53	4.47	4.38	4.64	5.15	5.90	7.40
300	10.1	7.44	7.26	7.66	8.11	9.37	11.8
400	26.2	17.2	16.6	17.8	18.3	20.8	26.3
500	52.8	32.9	31.9	34.0	34.4	39.0	49.5

Table 4.1: The timing results for the bidiagonalization by means of the Householder transformations

In order to estimate the performance of DGEBDB we make use of an "effective" number of floating point operations, i.e. the number of flops in the non-block algorithm of the bidiagonalization. One can verify that the procedure requires about $\frac{8}{3}N^3$ floating point operations. Thus, when $N = 500$ the "effective" performance rate achieves 6.3 Mflops for Level 2 BLAS implementation and 10.4 Mflops for Level 3 BLAS implementation.

It is interesting to compare the performance of the non-block version and the compact block version of the QR factorization. In fact, this procedure is the most time-consuming part of the spectrum dichotomy problem solver. In [10] one can find the timing results for the CRAY computers. Here (Table 4.2) we present some timing results for the ALLIANT FX/4 on four concurrent processors. Again we used the N-by-N matrix A with the elements $A_{ij} = i+(j-1)*N$. The CPU time is measured in seconds.

An "effective" number of floating point operations is about $\frac{4}{3}N^3$ for the QR factorization. Therefore, the "effective" performance rate achieves 6 Mflops for Level 2 BLAS implementation and 15.3 Mflops for Level 3 BLAS implementation. The performance rate on a single processor is about three times less.

N	BLAS 2	Blocksize for BLAS 3 implementation					
		4	8	16	32	64	128
25	0.01	0.03	0.03	0.03	—	—	—
50	0.04	0.06	0.06	0.07	0.08	—	—
100	0.19	0.18	0.18	0.20	0.24	0.30	—
150	0.59	0.45	0.43	0.46	0.53	0.68	0.94
200	1.41	0.93	0.84	0.88	1.04	1.30	1.70
250	2.90	1.71	1.53	1.54	1.75	2.12	2.85
300	5.31	2.85	2.54	2.62	2.86	3.32	4.35
400	13.6	6.47	5.74	5.95	6.10	7.03	8.90
500	27.6	12.4	10.9	11.4	11.3	12.5	15.1

Table 4.2: The timing results for the QR factorization by means of the Householder transformations

Note that the optimal blocksize for the block Householder method is equal to 8 for all our experiments.

It remains to compare the procedures for the calculation of the exponential of a matrix and the integral $C = \int_0^1 e^{tB} e^{tB^T} dt$. We obtained the following results:

N	Taylor method		Padé method	
	time	Mflops	time	Mflops
100	3.49	17.9	1.71	16.0
200	26.6	18.7	12.5	17.6
300	88.2	19.0	40.3	18.4

Table 4.3: The timing results for the Taylor method and for the Padé method

Finally, I present the timing results for the entire implementation of the spectrum dichotomy problem solver. When executing the solver for the spectrum dichotomy problem with respect to the imaginary axis the total CPU time for four concurrent processors was 35.8 seconds. When executing the solver for the spectrum dichotomy problem with respect to the unit circle the total CPU time on 4 processors was 35.7 seconds. These times were obtained for 100-by-100 matrices. The stopping criterion we used is (2.21) from Section 2.3, so the number of iterations for the second stage of the algorithm was equal to 33. This number is sufficient to let the algorithm converge for sufficiently ill-conditioned problems with the condition number less than or equal to 10^8.

The number of floating point operations for the spectrum dichotomy problem solver with respect to the imaginary axis equals about $402N^3$ and the number of floating point operations for the spectrum dichotomy problem with respect to the unit circle comprises about $360N^3$. Therefore, the performance rate of these two implementations is equal to 11.2 Mflops and 10.1 Mflops, respectively.

Numerical experiments with 300-by-300 matrices show a performance rate of the spectrum dichotomy problem solver of about 13 Mflops when the blocksize equal to 8.

Chapter 5

Some applications

The first problem being considered is the problem to compute an orthonormal vector basis of some deflating or invariant subspaces. For example, given an N-by-N matrix pencil $\lambda B - A$ and the circle of radius r with center a. If one need to calculate the right deflating subspace of the pencil $\lambda B - A$ corresponding to the eigenvalues inside this circle then the problem is reduced to the spectrum dichotomy problem with respect to the unit circle for the pencil $\mu(rB) - (A - aB)$.

Let A be an N-by-N matrix. It is easy to prove that the spectrum dichotomy problem for the matrix A with respect to the straight line $a + zt$, $-\infty < t < \infty$, is reduced to the spectrum dichotomy problem for the matrix $\frac{i|z|}{z}(A - aI)$ with respect to the imaginary axis. Thus, in order to solve the spectrum dichotomy problem with respect to arbitrary circles and straight lines we must be able to solve the spectrum dichotomy problem for complex matrices and complex matrix pencils.

Consider now the problem of how to calculate an orthonormal vector basis of a subspace when having computed a projector matrix onto this subspace. Let P be a projector matrix, i.e. $P^2 - P = 0$. One can verify that the subspace associated with P is the right null subspace of the matrix $I - P$, that is the subspace of the form $\{v \mid (I - P)v = 0\}$. The latter problem, the computation of the null subspace of a matrix, can be solved by means of SVD, the singular value decomposition.

The singular value decomposition for dense matrices is computed by reduction to the bidiagonal matrix case. This reduction is carried out by means of the Householder method. The parallel aspects of the bidiagonalization have already been discussed. Analogous to the SVD for bidiagonal matrices, we only need a few singular vectors, therefore the method from [18] fits.

5.1 The Riccati equation

In this section we apply the spectrum dichotomy problem solver to the solution of the Riccati equation arising in control theory. The problem of solving this equation is very attractive and many authors contributed to it [27].

We consider the algebraic Riccati equation of the form

$$Q + A^T\Lambda + \Lambda A - \Lambda B R^{-1} B^T \Lambda = 0, \tag{5.1}$$

where the matrix pair (A, Q) is detectable, the matrix pair (A, B) is stabilizable, $Q = Q^T \geq 0$, $R = R^T > 0$. For the definitions of "detectable" and "stabilizable" we refer to the literature, for example, [17].

Let us rewrite (5.1) as follows:

$$\begin{pmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{pmatrix} \begin{bmatrix} I \\ \Lambda \end{bmatrix} = \begin{bmatrix} I \\ \Lambda \end{bmatrix} (A - BR^{-1}B^T\Lambda). \tag{5.2}$$

Therefore, the Hamiltonian matrix

$$H = \begin{pmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{pmatrix} \tag{5.3}$$

has an invariant subspace spanned by the columns of the matrix $\begin{bmatrix} I \\ \Lambda \end{bmatrix}$ if and only if the Riccati equation (5.1) has a solution Λ.

One can show that the $2N$-by-$2N$ matrix H has N stable eigenvalues (i.e. $\mathrm{Re}(\lambda) < 0$) and N eigenvalues in the right hand side halfplane. Moreover, in this case equation (5.1) has the unique nonnegative definite solution Λ and the columns of $\begin{bmatrix} I \\ \Lambda \end{bmatrix}$ span the invariant subspace of H corresponding to the stable eigenvalues. Thus, to solve (5.1) is equivalent to the computation of the invariant subspace of H.

We discuss now an application of the spectrum dichotomy problem solver to the Riccati equation. For this case one has to make use of the solver for the spectrum dichotomy problem with respect to the imaginary axis. Recall that the matrix pencil $\lambda B_0 - A_0$ is formed in order to start the iterations, where $B_0 = L^{-1}$, $A_0 = L^{-1}e^K$, $K = H^T/(2\|H\|)$, $LL^T = \int_0^1 e^{tK}e^{tK^T}dt$. By means of the iterations of the algorithm we compute a convergent matrix pencil $\lambda B_\infty - A_\infty$. Since $A_\infty = T_\infty(I - P)$ with the matrix P being the projector onto the invariant subspace associated to the stable eigenvalues of H, the right null subspace of the matrix A_∞ is the invariant subspace of H which we need to compute. Let us partition the matrix A_∞ as follows:

$$A_\infty = [A_\infty^{(I)}, A_\infty^{(\Lambda)}].$$

Then a solution of the least squares problem

$$A_\infty^{(\Lambda)}\Lambda = -A_\infty^{(I)}$$

gives us the solution of the equation (5.1). Thus, we have no necessity to calculate explicitly the projector matrix P in order to compute the matrix λ.

All stages of the algorithm calculating the solution Λ of (5.1) are provided with the condition numbers. Hence, an entire error analysis is available for the algorithm. By the way, we also use the Householder method to solve the linear least squares problem, so the block versions of the method are applied in this case.

Now we present timing results for an implementation of the algorithm of a solution to the Riccati equation. The numerical experiments were executed on the ALLIANT FX/4. For $N = 50$ it required 39 seconds to find a solution. This CPU time corresponds to the worst case of convergence.

5.2 The Lyapunov equation

The Lyapunov equation is a particular case of the Riccati equation:

$$A^T\Lambda + \Lambda A + Q = 0. \tag{5.4}$$

Here the matrix A is stable and $Q = Q^T \geq 0$. In fact, the modified algorithm, which is intended for the solution of the spectrum dichotomy problem with respect to the imaginary axis, computes a solution of the Lyapunov equation with $Q = 2\|A\|I$. This special case of the right hand side is chosen in order to introduce the condition number for the matrix stability problem as $\kappa(A) = \|\Lambda\|$, where $A^T\Lambda + \Lambda A = -2\|A\|I$.

But if one still wants to solve (5.4) for an arbitrary matrix Q then the first stage of the modified algorithm should be slightly modified. At first, we compute the matrix $B = A^T/(2\|A\|)$ and then we have to compute $\int_0^1 e^{tB}Qe^{tB^T}dt$. Decomposing this integral as $LL^T = \int_0^1 e^{tB}Qe^{tB^T}dt$ and calculating the inverse of L^{-1} (if possible !) we form the matrix pencil $\lambda B_0 - A_0 = \lambda L^{-1} - l^{-1}e^B$.

Now one can carry out the iterations of the second stage of the spectrum dichotomy problem solver. After convergence, we obtain the solution of (5.4) in the form

$$\Lambda = (A_\infty + B_\infty)^{-1}(A_\infty + B_\infty)^{-T}.$$

There exists a more efficient way to compute a solution of (5.4) which is suitable for any matrix Q provided the matrix A is stable. The formulas of the method are extremely simple. At first, one has to calculate the matrices

$$E_1 = exp(\tau A) \text{ and } H_1 = \int_0^\tau e^{tA^T} Q e^{tA} dt$$

for $\tau = 1/(2\|A\|)$. Then the iterative process is executed:

$$H_i = H_{i-1} + E_{i-1}^T H_{i-1} E_{i-1}, \quad E_i = E_{i-1}^2. \tag{5.5}$$

If the matrix A is stable and the condition number $\kappa(A)$ is not very large then the iterative process (5.5) converges very rapidly. The method (5.5) takes a number of the floating point operations which is about three times less than that of the spectrum dichotomy problem solver; it may be coded in terms of the Level 3 BLAS routine DGEMM. Therefore, this latter solver for the Lyapunov equation seems to be the most efficient parallel method among all the existing methods.

Conclusions

In this report the technique for solving some eigenvalue problems has been discussed which is an alternative to the Schur method. When executing on sequential computers our method may be 5-10 times more expensive than the Schur method. On parallel computers, however, these two methods give approximately the same time-to-solution because our method may be effectively coded in terms of Level 3 BLAS with the performance rate not far from the performance rate for DGEMM whereas the Schur method does not reach such a high performance for parallel computers (about 1 Mflops on ALLIANT FX/4).

In addition, our method is equipped with the strict mathematical theory for the convergence rate and for the rounding errors [20]. For the Schur method it is very expensive to obtain the rounding error bounds.

Acknowledgements. I am very grateful to Margreet Louter-Nool for her help and implementation of the Padé approximation routine and the Lyapunov equation solver. I also wish to thank Herman te Riele and Margreet Louter-Nool for stimulating discussions.

Bibliography

[1] E. Anderson and J. Dongarra. LAPACK working note 18: Implementation guide for LAPACK. Technical Report CS-90-101, University of Tennessee, April 1990.

[2] R.H. Bartels and G.W. Stewart. Solution of the equation $AX + XB = C$. *Comm. ACM*, 15:820–826, 1972.

[3] T. Beelen and P. Van Dooren. An improved algorithm for the computation of Kroneker's canonical form of a singular pencil. *Linear Algebra Appl.*, 105:9–65, 1988.

[4] C.H. Bischof and Ch. Van Loan. The WY representation for products of Householder matrices. *SIAM J. Sci. Statist. Comput.*, 8:s2–s13, 1987.

[5] A.Ja. Bulgakov. An efficiently calculable parameter for the stability property of a system of linear differential equations with constant coefficients. *Siberian Math. J.*, 21(3):339–347, 1980.

[6] A.Ja. Bulgakov. An estimate of the Green matrix and the continuity of the dichotomy parameter. *Siberian Math. J.*, 30(1):139–142, 1989.

[7] A.Ja. Bulgakov and S.K. Godunov. Circular dichotomy of the spectrum of a matrix. *Siberian Math. J.*, 29(5):734–744, 1988.

[8] J. Demmel. Computing stable eigendecompositions of matrices. *Linear Algebra Appl.*, 79:163–193, 1986.

[9] J.W. Demmel and B. Kågström. Computing stable eigendecompositions of matrix pencils. *Linear Algebra Appl.*, 88/89:139–186, 1987.

[10] J.J. Dongarra, I.S. Duff, D.C. Sorensen, and H.A. Van der Vorst. *Solving Linear Systems on Vector and Shared Memory Computers.* SIAM Publications, Philadelphia, 1991.

[11] S.K. Godunov. The problem of dichotomy of the spectrum of a matrix. *Siberian Math. J.*, 27(5):649–660, 1986.

[12] S.K. Godunov, A.G. Antonov, O.P. Kiriluk, and V.I. Kostin. *Guaranteed Accuracy of the Solution to Systems of Linear Equations in Euclidean Spaces.* Nauka, Novosibirsk, 1988. (in Russian).

[13] S.K. Godunov and A.Ja. Bulgakov. Difficultés calculatives dans le problème de Hurwitz et méthodes à les surmonter. In *Proceedings Fifth International Conference on Analysis and Optimization of Systems, Versailles*, Lecture Notes in Control and Information 44, pages 845–851. Springer-Verlag, 1982. (in French).

[14] G.K. Golub and Ch.F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, Baltimore, Maryland, 2nd edition, 1989.

[15] B. Kågström, L. Nyström, and P. Poromaa. Parallel shared memory algorithms for solving the triangular Sylvester equation. In J. Dongarra, I. Duff, P. Gaffney, and S. McKee, editors, *Vector and Parallel Computing: Issues in Applied Research and Development*, pages 229–240. Ellis Horwood Limited, England, 1989.

[16] B. Kågström and L. Westin. GSYLV- Fortran routines for the generalized Schur method with dif^{-1} estimators for solving the generalized Sylvester equation. Technical Report UMINF-132.86, Information Processing, University of Umeå, S-901 87 Umeå, Sweden, 1987.

[17] H. Kwakernaak and R. Sivan. *Linear Optimal Control Systems.* Wiley-Interscience, New York, 1972.

[18] Sy-Shin Lo, B. Philippe, and A.H. Sameh. A multiprocessor algorithm for the symmetric tridiagonal eigenvalue problem. *SIAM J. Sci. Statist. Comput.*, 8(2):155–165, 1987.

[19] A.N. Malyshev. Computing invariant subspaces of a regular linear pencil of matrices. *Siberian Math. J.*, 30(4):559–567, 1989.

[20] A.N. Malyshev. Guaranteed accuracy for the eigenvalue problems in linear algebra. In S.K. Godunov, editor, *Computational Methods of Linear Algebra*, Proceedings of the Institute of Mathematics of the Siberian Division of the USSR Academy of Sciences, volume 17, pages 19–104. Nauka, Novosibirsk, 1990. (in Russian).

[21] C.B. Moler and Ch.F. Van Loan. Nineteen dubious ways to compute the exponential of a matrix. *SIAM Rev.*, 20(4):801–836, 1978.

[22] C.B. Moler and G.W. Stewart. An algorithm for generalized matrix eigenvalue problems. *SIAM J. Numer. Anal.*, 10:241–256, 1973.

[23] R. Schreiber and Ch. Van Loan. A storage efficient WY representation for products of Householder transformations. *SIAM J. Sci. Statist. Comput.*, 10(1):53–57, 1989.

[24] G.W. Stewart. On the sensitivity of the eigenvalue problem $Ax = \lambda Bx$. *SIAM J. Numer. Anal.*, 9:669–686, 1972.

[25] G.W. Stewart. Error and perturbation bounds for subspaces associated with certain eigenvalue problems. *SIAM Rev.*, 15:727–764, 1973.

[26] P. Van Dooren. The computation of Kronecker's canonical form of a singular pencil. *Linear Algebra Appl.*, 27:103–140, 1979.

[27] P. Van Dooren. A generalized eigenvalue approach for solving Riccati equations. *SIAM J. Sci. Statist. Comput.*, 2:121–135, 1981.

[28] P. Van Dooren. Reducing subspaces: definitions, properties and algorithms. In B. Kågström and A. Ruhe, editors, *Matrix Pencils*, pages 58–73. Springer-Verlag, New York, 1983.

[29] Ch.F. Van Loan. Computing integrals involving the matrix exponential. *IEEE Trans. Automat. Control*, AC-23(3):395–404, 1978.

[30] D.S. Watkins and L. Elsner. Convergence of algorithms of decomposition type for the eigenvalue problem. *Linear Algebra Appl.*, 143:19–47, 1991.

[31] J.H. Wilkinson. *The Algebraic Eigenvalue Problem.* Claredon Press, Oxford, England, 1965.