CWI Tracts

Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

The CWI is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

CWI Tract 52

Continuous decoupling transformations for linear boundary value problems

P.M. van Loon

Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

1980 Mathematics Subject Classification: 65Lxx, 34Bxx.
ISBN 90 6196 353 2
NUGI-code: 811

Contents

Preface

In this tract systems of linear boundary value problems (BVPs) for ordinary differential equations are studied. Especially are examined solution methods for BVPs based on the integration of initial value problems (IVPs), like (multiple) shooting techniques. The main problem to evade arises from the fact that a BVP may have both (fast) increasing and (fast) decaying modes. This implies that straightforward integration of the differential equation will be expensive (small stepsizes) and/or the results will be inaccurate (the decaying modes will be dominated by the increasing modes).

Some of these problems can be circumvented by a continuous decoupling of the (fast) increasing and the (fast) decaying modes. Such a decoupling can be interpreted as a transformation that actually accounts for the direction of the fastest increasing (called: dominant) modes. Hereafter, the growth of these modes can be determined separately. For this technique we generally have to solve only well-conditioned IVPs, the solutions of which are slowly varying or fast decaying. The price to be paid is that some of these IVPs are non-linear. One of the best known decoupling transformations is the (classical) Riccati transformation, for which an IVP, involving quadratic terms, has to be solved. One of the main difficulties is that the quadratic Riccati differential equation generally will not have a solution on the whole interval. This problem can be overcome in two ways: generalizing the Riccati transformation to an orthogonal transformation or rearranging the basis of $\mathbb{R}^n$ before the solution blows up.

In principle a decoupling transformation only determines the direction of the dominant modes. To obtain growth factors and the direction of the other modes various techniques are available. An obvious way is to integrate in the opposite direction. In that case, however, one has to store and interpolate intermediate results or to solve another non-linear IVP. Both options are not very attractive. Therefore a generalized invariant imbedding technique is suggested, which is not

necessarily related to the Riccati (decoupling) transformation. The resulting technique, which involves forward integration only, turns out to be similar to a backward integration, after the dominant modes have been decoupled.

The combination of a Riccati transformation with invariant imbedding is called the Riccati method. It turns out that this Riccati method is a rather robust method, which can solve both simple and complicated problems (such as stiff problems, turning point problems and singular problems)

In Chapter 1 some more or less elementary concepts from linear algebra that are fundamental in the discussion on differential equations (DEs) are treated. Important concepts are the gap and the distance of linear subspaces. Also a definition of the rotational activity of a solution (sub)space of a linear DE is given, which will be used throughout this thesis.

In Chapter 2 some recently developed fundamental concepts for linear BVPs are reviewed, like conditioning. A central role is played by the (exponential) dichotomy of the solution space. A dichotomic solution space can be partitioned in (at least) two parts, which solutions differ both in direction and in growth. An important result is formulated in Theorem 2.19, which shows that with straightforward integration and for almost any initial value the subspace of dominant modes will be obtained.

In Chapter 3 general decoupling transformations are studied. A special class is formed by continuous orthonormalization methods, where the decoupling transformation is orthogonal. In this chapter also the generalized invariant imbedding technique is discussed. In Theorem 3.14 we prove that this technique delivers the direction of modes, which are dominant in opposite direction.

In Chapter 4 decoupling via the Riccati transformation is considered. By its simple form (block lower diagonal) some reduction is obtained in number and complexity of the resulting IVPs. Of course one has to pay for this reduction: the transformation does not necessarily exist on the whole interval of interest. It will turn out, however, that under mild conditions this difficulty can be overcome quite satisfactorily. The theory results in the Riccati method, by which both stiff and non-stiff linear BVPs can be solved rather efficiently.

In Chapter 5 the Riccati method is examined in more detail when applied to so-called stiff problems. To this end two kind of stiff problems are considered: with and without turning points. If no turning points are present then some reduction in the number of DEs can be obtained, which is of interest for large

problems. When turning points are present then the super-stability property of integration methods based on the Backward Differentiation Formulas may cause inaccurate results. Some techniques to circumvent this problem are proposed.

In Chapter 6 the decoupling technique is examined when applied to BVPs with a singularity of the first kind. It turns out that by this technique one has to solve only singular IVPs that obtain (unique) analytic solutions. Using the first terms of the power series expansions of these solutions one is able to move away from the singularity, whereafter a regular problem remains. This regular problem can be solved by one of the techniques discussed in the Chapters 3 and 4.

Note for reading: the equation numbering does not contain the number of the chapter in which it appears. A reference to a numbered equation is always within the same chapter, unlike otherwise stated.

Chapter 1

Preliminaries

In this chapter we shall first discuss some elementary results from linear algebra. Therefore we use, where possible, the same notation as in ([20]). In the Sections 1.2 - 1.4 we consider some concepts that play a fundamental role in the discussion of differential equations. Some of them are already known for a long time, while others, like the rotational activity of a time-varying linear subspace, are quite new.

1.1 Elementary definitions and results

• span, range and rank

By the *span* of a given set of vectors $\{a_i\}_{i=1}^k$ in $\mathbb{R}^n$ we mean the linear subspace spanned by $a_1, \ldots, a_k$. This span is denoted by

$$\text{span}\{a_1, \ldots, a_k\} . \tag{1}$$

By $\left[a_1 \mid \cdots \mid a_k \right]$ we mean the $n \times k$ matrix whose i-th column is equal to the vector $a_i (i = 1, \ldots, k)$. Let $A = \left[a_1 \mid \cdots \mid a_k \right]$ be a matrix in $\mathbb{R}^{n \times k}$. The *range* of A is defined by

$$\mathcal{R}(A) = \text{span}\{a_1, \ldots, a_k\} . \tag{2a}$$

The *kernel* or *nullspace* of A is defined by

$$\ker(A) = \{ x \in \mathbb{R}^k \mid A x = 0 \} . \tag{2b}$$

The *rank* of a matrix is equal to the maximal number of independent columns. Hence,

$$\text{rank}\,(A) = \dim\,(\mathcal{R}(A)) \ . \tag{3}$$

If $\text{rank}\,(A) = \min\{n, k\}$, then A has *full rank.*

• **symmetry**

An $n \times n$ matrix A is called *symmetric* if $A = A^T$, and *skew-symmetric* if $A = -A^T$. Any matrix A can be written as the sum of a symmetric matrix and a skew-symmetric matrix:

$$A = \frac{1}{2}(A + A^T) \ + \ \frac{1}{2}(A - A^T) \ . \tag{4}$$

The *symmetric part* of A is defined by

$$\text{symm}\,(A) = \frac{1}{2}(A + A^T) \ . \tag{5}$$

• **orthogonality**

The matrix $A = \left[a_1 \mid \cdots \mid a_k \right] \epsilon \, \mathbb{R}^{n \times k}$ is *column orthogonal* if the columns of A are mutually orthonormal, i.e.,

$$a_i{}^T a_j = \delta_{ij}, \qquad (i, j = 1, \ldots, k) \ .$$

If $k = n$ then we just say that A is *orthogonal.* If $k > n$ and A^T is column orthogonal, then we call the matrix A *row orthogonal.*

For any k-dimensional subspace $\mathcal{S}_1 \subset \mathbb{R}^n$ there exists a column orthogonal matrix Q_1 such that $\mathcal{S}_1 = \mathcal{R}(Q_1)$. This is a result of the following:
Let $A_1 \epsilon \, \mathbb{R}^{n \times k}$, with $k \leq n$. Then there exist a column orthogonal matrix $Q_1 \epsilon \, \mathbb{R}^{n \times k}$ and an upper triangular matrix $R_{11} \epsilon \, \mathbb{R}^{k \times k}$ such that

$$A_1 = Q_1 R_{11} \ . \tag{6a}$$

This is called a *QR-decomposition* of A_1.
If A_1 has full rank and we moreover require that the diagonal elements of R_{11} are positive, then Q_1 and R_{11} are uniquely determined.
If $A_1 \epsilon \, \mathbb{R}^{k \times n}$, with $k < n$, then a QR-decomposition of A_1 is obtained by

$$A_1 \, Q = \left[0 \ \ R_{11} \right] \ , \tag{6b}$$

where $Q \epsilon \, \mathbb{R}^{n \times n}$ is orthogonal and $R_{11} \epsilon \, \mathbb{R}^{k \times k}$ is upper triangular.

It is always possible to extend a column orthogonal matrix Q_1 to a full $n \times n$ orthogonal matrix $Q = \left[\underset{\longleftrightarrow}{Q_1}{}_{k} \ \ \underset{\longleftrightarrow}{Q_2}{}_{n-k} \right]$. The orthogonal complement of $\mathcal{S}_1 = \mathcal{R}(Q_1)$ is then given by

$$\mathcal{S}_1^{\perp} = \mathcal{R}(Q_2) \ . \tag{7}$$

• norm and condition number

Throughout this thesis we shall mainly use the *Euclidean norm* (2-norm). Hence, if $x \in \mathbb{R}^n$ then $\| x \| = \sqrt{x^T x}$. For a matrix $A \in \mathbb{R}^{n \times k}$ we use the induced matrix norm

$$\| A \| = \max_{x \in \mathbb{R}^k} \{ \frac{\| A x \|}{\| x \|} \mid x \neq 0 \} \ . \tag{8}$$

This quantity will sometimes be denoted by $\mathrm{lub}(A)$. Similarly we define

$$\mathrm{glb}(A) = \min_{x \in \mathbb{R}^k} \{ \frac{\| A x \|}{\| x \|} \mid x \neq 0 \} \ . \tag{9}$$

Property 1.1

Let $A \in \mathbb{R}^{n \times k}$ *and* $B \in \mathbb{R}^{k \times k}$. *Then* $\mathrm{lub}(A B) \geq \mathrm{lub}(A)\, \mathrm{glb}(B)$.

■

Incidentally, we shall also use the following norm

$$\left| A \right| = \max_{i,j} \{ \mid a_{ij} \mid \Big| \ a_{ij} = (i,j)^{\text{th}} \text{ element of } A \} \ . \tag{10}$$

Observe that this norm is not an induced matrix norm, nor it is submultiplicative.

A matrix $A \in \mathbb{R}^{n \times k}$ $(k \leq n)$ has full rank if and only if $\mathrm{glb}(A) > 0$ and then the *condition number* of A is defined by

$$\kappa(A) = \frac{\mathrm{lub}(A)}{\mathrm{glb}(A)} \ . \tag{11}$$

If $k > n$ then there exists an $x \neq 0$ such that $A x = 0$. So, $\mathrm{glb}(A) = 0$. However, if $\mathrm{rank}(A) = n$ then we define the condition number of A by

$$\kappa(A) = \frac{\mathrm{lub}(A^T)}{\mathrm{glb}(A^T)} \ . \tag{12}$$

Observe that any (column/row) orthogonal matrix Q is perfectly conditioned, i.e., $\kappa(Q) = 1$.

An easily verifiable result is the following:

Lemma 1.2

Let $A,\ B \in \mathbb{R}^{n \times n}$. *Then* $\| \left[A \mid B \right] \| \leq \sqrt{2} \max \{ \| A \|, \| B \| \}$.

■

• singular value decomposition (SVD)

If $A \in \mathbb{R}^{n\times k}$ then there exist orthogonal matrices $U = \left[u_1 \mid \cdots \mid u_n\right] \in \mathbb{R}^{n\times n}$ and $V = \left[v_1 \mid \cdots \mid v_k\right] \in \mathbb{R}^{k\times k}$ and a diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{n\times k}$ such that

$$A = U \Sigma V^T \ , \tag{13}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ $(p = \min\{n, k\})$.
The σ_i $(i = 1, \ldots, p)$ are the *singular values* of A. It will be convenient to have the following notation:

$$\begin{aligned} \sigma_{\max}(A) &= \text{the largest singular value of } A\,, \\ \sigma_{\min}(A) &= \text{the smallest singular value of } A\,. \end{aligned}$$

• inverse and pseudo-inverse

If $A \in \mathbb{R}^{n\times n}$ is non-singular, then there exists a non-singular $B \in \mathbb{R}^{n\times n}$ such that $AB = BA = I_n$, the $n \times n$ identity matrix. The matrix B is called the *inverse* of A and is denoted by A^{-1}. For non-square matrices we make the following generalization. Let $A \in \mathbb{R}^{n\times k}$ $(k < n)$ have full rank. Then the *pseudo-inverse* of A is defined by

$$A^+ = (A^T A)^{-1} A^T \ . \tag{14}$$

For A^+ we obtain that $A^+ A = I_k$. Moreover, $A A^+$ describes the orthogonal projection onto $\mathcal{R}(A)$.

• eigenvalues and eigenvectors

The *eigenvalues* of a matrix $A \in \mathbb{R}^{n\times n}$ are the n (possibly complex) roots of its characteristic polynomial $\mathrm{p}(z) = \det(zI_n - A)$. The set of these roots is called the *spectrum* and is denoted by $\lambda(A)$. If $\lambda \in \lambda(A)$, then any non-zero vector $x \in \mathbb{R}^n$ that satisfies $Ax = \lambda x$ is referred to as an *eigenvector*. More generally, a subspace $\mathcal{S} \subset \mathbb{R}^n$ with the property that $x \in \mathcal{S} \Rightarrow Ax \in \mathcal{S}$ is said to be an *invariant subspace* for A.
Corresponding to the eigenvalues of A we shall use the definitions:

$$\begin{aligned} \lambda_{\max}(A) &= \max\{\,\mathrm{Re}(\lambda) \mid \lambda \in \lambda(A)\,\} \\ \lambda_{\min}(A) &= \min\{\,\mathrm{Re}(\lambda) \mid \lambda \in \lambda(A)\,\}\,. \end{aligned}$$

We have the following result:

$$\| A \|^2 = \lambda_{\max}(A^T A)\,. \tag{15}$$

• Schur-decomposition

A matrix $A \in \mathbb{R}^{n\times n}$ is *block upper triangular* if it can be partitioned in the form

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1l} \\ 0 & A_{22} & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & A_{ll} \end{bmatrix} ,$$

where each diagonal block A_{ii} is square. If each diagonal block is of order at most two, then A is said to be in *quasi-triangular form*. For the spectrum of A we have

$$\lambda(A) = \bigcup_{i=1}^{l} \lambda(A_{ii}) . \tag{16}$$

Theorem 1.3
Let $A \in \mathbb{R}^{n\times n}$. Then there exists an orthogonal matrix $U \in \mathbb{R}^{n\times n}$ such that $U^T A U$ is quasi-triangular. Moreover, U may be chosen such that any 2×2 diagonal block of $U^T A U$ has only complex eigenvalues (which therefore must be conjugates). The diagonal blocks may appear in any order along the diagonal.

■

If A is symmetric, then $\lambda(A) \subset \mathbb{R}$, which implies that $D = U^T A U$ is a diagonal matrix. If, moreover, $\lambda_{\min}(A) > 0$ then the *square root* of A is defined by

$$A^{\frac{1}{2}} = U^T \sqrt{D}\, U , \tag{17}$$

where $\sqrt{D} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$. In that case the matrix $(A^{\frac{1}{2}})^{-1}$ is denoted by $A^{-\frac{1}{2}}$.

• matrix product
Let $\{A_i\}_{i=1}^m$ be a given set of matrices in $\mathbb{R}^{n\times n}$. Then we shall use the following notation ($k \leq m$):

$$\prod_{i=1}^{k} A_i = A_k A_{k-1} \ldots A_1 . \tag{18}$$

If $k = 0$ then we identify this (empty) product with I_n.

1.2 Separation and measure

A linear equation that will return throughout this thesis is the so-called *Sylvester equation* ([20, p.242]):

$$B X - X C = D \,, \tag{19}$$

where $B \in \mathbb{R}^{k\times k}, C \in \mathbb{R}^{l\times l}$ and $D \in \mathbb{R}^{k\times l}$ are given matrices and $X \in \mathbb{R}^{k\times l}$ is to be determined.

Lemma 1.4 ([20, Lemma 7.1-4])
The Sylvester equation (19) *has a unique solution for every D, if and only if*

$$\lambda(B) \cap \lambda(C) = \emptyset \,.$$

■

Definition 1.5
The separation *of the matrices* $B \in \mathbb{R}^{k\times k}$ *and* $C \in \mathbb{R}^{l\times l}$ *is defined by*

$$\mathrm{sep}(B,\, C) = \min_{X \in \mathbb{R}^{k\times l}} \left\{ \frac{\| BX - XC \|}{\| X \|} \;\middle|\; X \neq 0 \right\}.$$

■

Property 1.6 ([57])
Let $B, E \in \mathbb{R}^{k\times k}$ *and* $C, F \in \mathbb{R}^{l\times l}$. *Then*

$$\mathrm{sep}(B + E, C + F) \;\geq\; \mathrm{sep}(B, C) - \| E \| - \| F \| \;.$$

■

Property 1.7 ([57])
The separation of $B \in \mathbb{R}^{k\times k}$ *and* $C \in \mathbb{R}^{l\times l}$ *satisfies the inequality*

$$\mathrm{sep}(B, C) \;\leq\; \min \{\, | \beta - \gamma | \,\big|\, \beta \in \lambda(B),\, \gamma \in \lambda(C) \,\} \,.$$

■

In the analysis of linear differential equations we sometimes need a (positive) lowerbound for the separation of matrices. To this end we consider, for $A \in \mathbb{R}^{n\times n}$, the initial value problem

$$\frac{d}{dt} X = AX \,, \qquad t \geq 0 \,, \tag{20}$$

subject to

$$X(0) = I_n \tag{21}$$

(X is an $n \times n$ matrix function). This initial value problem has the unique solution

$$X(t) = e^{tA}, \qquad t \geq 0 . \tag{22}$$

An important role in the dicussion of the stability of the differential equation (20) is played by the *measure* of the matrix A, which, with respect to the 2-norm for matrices, is defined as (cf. [59])

$$\mu(A) = \lambda_{\max}\left(\frac{1}{2}(A + A^T)\right) . \tag{23}$$

Property 1.8 ([59])
Let $A \in \mathbb{R}^{n\times n}$, *then* $\lambda_{\max}(A) \leq \mu(A)$ *and* $\| e^{tA} \| \leq e^{t\mu(A)}$, $t \geq 0$.

■

Theorem 1.9
Let $B \in \mathbb{R}^{k\times k}$ *and* $C \in \mathbb{R}^{l\times l}$ *be such that*

$$\mu(B) + \mu(-C) \; < \; 0 . \tag{24}$$

Then

$$\mathrm{sep}(B, C) \; \geq \; -\Big(\mu(B) + \mu(-C)\Big) .$$

Proof:
In the first place we observe that the condition (24) implies that the Sylvester equation

$$BX - XC = D \tag{25}$$

has a unique solution, for any $D \in \mathbb{R}^{k\times l}$. Moreover, from (25) it follows that

$$\frac{d}{dt}\left(e^{tB} X e^{-tC}\right) = e^{tB} D e^{-tC} .$$

Hence,

$$X = -\int_0^\infty e^{tB} D e^{-tC} \, dt ,$$

which exists by the condition (24). Therefore,

$$\| X \| \leq \; -\frac{\| D \|}{\Big(\mu(B) + \mu(-C)\Big)} .$$

Since this is true for any matrix D and corresponding solution X, we obtain

$$\mathrm{sep}(B, C) \; = \; \min\left\{\frac{\| BX - XC \|}{\| X \|}\right\}$$

$$\geq -\Big(\mu(B)+\mu(-C)\Big) .$$

■

1.3 Gap and distance

In chapter 2 we shall consider the (asymptotic) behaviour of time-dependent linear subspaces. An important tool for that discussion is the *gap* between two linear subspaces $\mathcal{A}$ and $\mathcal{B}$ $\Big(\mathrm{GAP}(\mathcal{A},\mathcal{B})\Big)$. This is defined as the sine of the smallest possible angle between a vector from $\mathcal{A}$ and a vector from $\mathcal{B}$. More formally (cf. [20, p.428]):

Definition 1.10
Let $\mathcal{A}$ and $\mathcal{B}$ be linear subspaces of $\mathbb{R}^n$. Define $\theta \in [0, \pi/2]$ such that

$$\cos\theta = \max_{u\in\mathcal{A}} \max_{v\in\mathcal{B}} \{ u^T v \mid \| u \| = 1 , \| v \| = 1 \} .$$

Then

$$\mathrm{GAP}(\mathcal{A},\mathcal{B}) = \sin\theta .$$

■

Example 1.11
Let $\mathcal{A} = \mathrm{span}\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\}$ and $\mathcal{B} = \mathrm{span}\left\{ \begin{pmatrix} \sqrt{1-\varepsilon^2} \\ \varepsilon \end{pmatrix} \right\}$ $(|\varepsilon| \leq 1)$. Then $\mathrm{GAP}(\mathcal{A},\mathcal{B}) = |\varepsilon|$.

■

The gap between two subspaces can be regarded as a measure for the separation of the two subspaces. Similarly we can define a measure for the distance of two (equidimensional) subspaces.

Definition 1.12
Let $\mathcal{A}$ and $\mathcal{B}$ be equidimensional subspaces of $\mathbb{R}^n$. The distance *between $\mathcal{A}$ and $\mathcal{B}$ is defined by*

$$\mathrm{DIST}(\mathcal{A},\mathcal{B}) = \max_{a\in\mathcal{A}} \Big\{ \mathrm{GAP}(\mathrm{span}\{a\},\mathcal{B}) \Big\} .$$

■

Remark 1.13
The equidimensionality requirement of $\mathcal{A}$ and $\mathcal{B}$ yields a symmetry in the definition of distance, i.e., $\mathrm{DIST}(\mathcal{A},\mathcal{B}) = \mathrm{DIST}(\mathcal{B},\mathcal{A})$.

■

The same definitions and notations of gap and distance will be used for matrices A and B. In that case we mean by $\mathrm{DIST}(A,B)$ the distance between $\mathcal{R}(A)$ and $\mathcal{R}(B)$. Similarly, by $\mathrm{GAP}(A,B)$ the gap between $\mathcal{R}(A)$ and $\mathcal{R}(B)$ is meant.

When we are dealing with the gap and the distance of linear subspaces the following more or less straightforward results will often be used.

Property 1.14
Let $\mathcal{A}$ and $\mathcal{B}$ be linear subspaces with $\mathcal{A} \oplus \mathcal{B} = \mathbb{R}^n$. Then

$$\mathrm{GAP}^2(\mathcal{A},\mathcal{B}) + \mathrm{DIST}^2(\mathcal{A}^\perp,\mathcal{B}) = 1 \ . \tag{26}$$

Moreover, let Q_2 be an orthogonal basis of $\mathcal{A}^\perp$ and $\tilde{Q}_2$ an orthogonal basis of $\mathcal{B}$. Then

$$\mathrm{GAP}(\mathcal{A},\mathcal{B}) = \| \left(Q_2{}^T \tilde{Q}_2\right)^{-1} \|^{-1} = \sigma_{\min}(Q_2{}^T \tilde{Q}_2) \ , \tag{27}$$

Similarly, with Q_1 an orthogonal basis of $\mathcal{A}$, we obtain

$$\begin{aligned} \mathrm{DIST}(\mathcal{A}^\perp,\mathcal{B}) &= \mathrm{DIST}(\mathcal{A},\mathcal{B}^\perp) \\ &= \| \tilde{Q}_2{}^T Q_1 \| = \sigma_{\max}(\tilde{Q}_2{}^T Q_1) \ . \end{aligned}$$

■

Property 1.15
Assume $A_1 \in \mathbb{R}^{n\times k} (k < n)$ with $\mathrm{rank}(A_1) = k$, $Q = \left[\underset{k}{\underset{\longleftrightarrow}{Q_1}} \ \underset{n-k}{\underset{\longleftrightarrow}{Q_2}} \right]$ an orthogonal matrix and $K_1 = \begin{bmatrix} K_{11} \\ K_{21} \end{bmatrix} \begin{matrix} \updownarrow k \\ \updownarrow n-k \end{matrix}$ such that $A_1 = Q_1 K_{11} + Q_2 K_{21}$. Then

$$\mathrm{DIST}(A_1,Q_1) = \begin{cases} 1 & \text{if } K_{11} \text{ singular} \\ \dfrac{\| K_{21} K_{11}{}^{-1} \|}{\sqrt{1 + \| K_{21} K_{11}{}^{-1} \|^2}} & \text{if } K_{11} \text{ non-singular} \end{cases} \ .$$

■

Less straightforward are the following two results.

Property 1.16
Assume $A_1 \in \mathbb{R}^{n\times k}(k < n)$ *with* $\text{rank}(A_1) = k$, $\tilde{Q}_1 \in \mathbb{R}^{n\times k}$ *and* $\tilde{Q}_2 \in \mathbb{R}^{n\times(n-k)}$ *column orthogonal matrices for which* $\tilde{Q} = \left[\tilde{Q}_1\ \tilde{Q}_2\right]$ *is non-singular and* $\tilde{K}_1 = \begin{bmatrix} \tilde{K}_{11} \\ \tilde{K}_{21} \end{bmatrix} \begin{matrix} \updownarrow k \\ \updownarrow n-k \end{matrix}$ *with* $\tilde{K}_{11}$ *non-singular are such that* $A_1 = \tilde{Q}_1\tilde{K}_{11} + \tilde{Q}_2\tilde{K}_{21}$. *Then we have:*

if $\| \tilde{K}_{21}\tilde{K}_{11}^{-1} \| < 1$ *then* $\text{DIST}(A_1, \tilde{Q}_1) \leq \dfrac{\| \tilde{K}_{21}\tilde{K}_{11}^{-1} \|}{1 - \| \tilde{K}_{21}\tilde{K}_{11}^{-1} \|}$.

Proof:
Let $Q = \left[Q_1\ Q_2\right] \in \mathbb{R}^{n\times n}$ be an orthogonal matrix with $Q_1 = \tilde{Q}_1$. Then

$$\tilde{Q} = Q \begin{bmatrix} I_k & Q_1{}^T\tilde{Q}_2 \\ 0 & Q_2{}^T\tilde{Q}_2 \end{bmatrix}$$

and

$$A_1 = \begin{bmatrix} \tilde{Q}_1 & \tilde{Q}_2 \end{bmatrix} \begin{bmatrix} \tilde{K}_{11} \\ \tilde{K}_{21} \end{bmatrix} = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} \tilde{K}_{11} + Q_1{}^T\tilde{Q}_2\tilde{K}_{21} \\ Q_2{}^T\tilde{Q}_2\tilde{K}_{21} \end{bmatrix}.$$

If $\| \tilde{K}_{21}\tilde{K}_{11}^{-1} \| < 1$ then $\tilde{K}_{11} + Q_1{}^T\tilde{Q}_2\tilde{K}_{21}$ is non-singular and

$$\| (I_k + Q_1{}^T\tilde{Q}_2\tilde{K}_{21}\tilde{K}_{11}^{-1})^{-1} \| \leq \frac{1}{1 - \| \tilde{K}_{21}\tilde{K}_{11}^{-1} \|}.$$

Therefore,

$$\begin{aligned}
\text{DIST}(A_1, \tilde{Q}_1) &= \text{DIST}(A_1, Q_1) \\
&= \frac{\| Q_2{}^T\tilde{Q}_2\tilde{K}_{21}(\tilde{K}_{11} + Q_1{}^T\tilde{Q}_2\tilde{K}_{21})^{-1} \|}{\sqrt{1 + \| Q_2{}^T\tilde{Q}_2\tilde{K}_{21}(\tilde{K}_{11} + Q_1{}^T\tilde{Q}_2\tilde{K}_{21})^{-1} \|^2}} \\
&\leq \| Q_2{}^T\tilde{Q}_2\tilde{K}_{21}\tilde{K}_{11}^{-1}(I_k + Q_1{}^T\tilde{Q}_2\tilde{K}_{21}\tilde{K}_{11}^{-1})^{-1} \| \\
&\leq \frac{\| Q_2{}^T\tilde{Q}_2\tilde{K}_{21}\tilde{K}_{11}^{-1} \|}{1 - \| \tilde{K}_{21}\tilde{K}_{11}^{-1} \|} \\
&\leq \frac{\| \tilde{K}_{21}\tilde{K}_{11}^{-1} \|}{1 - \| \tilde{K}_{21}\tilde{K}_{11}^{-1} \|}.
\end{aligned}$$

■

Lemma 1.17
Let $Q_1 \in \mathbb{R}^{n\times k}$ *and* $Q_2 \in \mathbb{R}^{n\times(n-k)}$ *be column orthogonal matrices with*

$\mathrm{GAP}(Q_1, Q_2) > 0$. *Then*

$$\text{(i)} \quad \| \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \|^2 = 1 + \sqrt{1 - \mathrm{GAP}^2(Q_1, Q_2)} \leq 2 .$$

$$\text{(ii)} \quad \| \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}^{-1} \|^2 = \frac{1}{1 - \sqrt{1 - \mathrm{GAP}^2(Q_1, Q_2)}} \leq \frac{2}{\mathrm{GAP}^2(Q_1, Q_2)} .$$

Proof:
Let $Q = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}$ and $Z_{21} = {Q_2}^T Q_1$. Then

$$Q^T Q = \begin{bmatrix} I_k & {Z_{21}}^T \\ Z_{21} & I_{n-k} \end{bmatrix} = I_n + \begin{bmatrix} 0 & {Z_{21}}^T \\ Z_{21} & 0 \end{bmatrix} .$$

From the singular value decomposition $Z_{21} = U \Sigma V^T$ we obtain

$$\begin{bmatrix} 0 & {Z_{21}}^T \\ Z_{21} & 0 \end{bmatrix} \begin{bmatrix} V & V \\ U & -U \end{bmatrix} = \begin{bmatrix} V & V \\ U & -U \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix} .$$

Hence,

$$\lambda_{\max}(Q^T Q) = 1 + \lambda_{\max}\left(\begin{bmatrix} 0 & {Z_{21}}^T \\ Z_{21} & 0 \end{bmatrix} \right) = 1 + \sigma_{\max}(Z_{21}) .$$

From Property 1.14 we obtain

$$\sigma_{\max}(Z_{21}) = \sqrt{1 - \mathrm{GAP}^2(Q_1, Q_2)} ,$$

by which (i) follows.
To see the second proposition note that

$$\| Q^{-1} \|^2 = 1/\lambda_{\min}(Q^T Q) = 1/(1 - \sigma_{\max}(Z_{21}))$$

and observe that

$$\frac{1}{1 - \sqrt{1 - t}} \leq \frac{2}{t} , \qquad \text{for all } t \in (0, 1].$$

■

1.4 Rotational activity

In this section we shall give definitions and derive properties concerning time-dependent matrices and subspaces. Let $X_1 = [x_1 | \cdots | x_k]$ be a continuously differentiable $n \times k$ $(k < n)$ matrix function on some interval $\mathcal{I}$. Assume

that $X_1(\tau)$ has full rank, for all $\tau \epsilon \mathcal{I}$. Decompose X_1 into a *direction matrix* $T_1 = [t_1|\cdots|t_k]$ and a positive *size matrix* $D_{11} = \mathrm{diag}(d_1, \ldots, d_k)$, (cf. [56])

$$X_1 = T_1 D_{11} , \tag{28}$$

where $\| t_i(\tau) \| = 1$ $(i = 1, \ldots, k)$, for all $\tau \epsilon \mathcal{I}$. We say that t_i is the *direction* of x_i and d_i is the *growth* of x_i $(i = 1, \ldots, k)$.

The following properties of T_1 and D_{11} directly follow from the requirements on X_1 and the differentiability of the 2-norm.

Property 1.18
The matrix functions T_1 and D_{11} have full rank and are continuously differentiable.

■

Now we can define the rotational activity of a matrix function.

Definition 1.19
Let $X_1: \mathcal{I} \rightarrow \mathbb{R}^{n \times k} (k \leq n)$ be continuously differentiable. Assume $X_1(\tau)$ has full rank, for all $\tau \epsilon \mathcal{I}$. The rotational activity *of the matrix function X_1 at the time τ is defined by $\| \frac{d}{d\tau} T_1(\tau) \|$, where T_1 is the direction matrix corresponding to X_1* (see (28)).

■

Property 1.20
The rotational activity of a matrix function X_1 is invariant under permutations of the columns of X_1.

■

For our purposes the concept of rotational activity of a matrix function may be too restrictive in some situations. Sometimes we are interested in the rotational activity of a linear subspace. In order to obtain a workable definition of the rotational activity of a linear subspace we first look at orthogonal bases of that subspace. Let $Q_1: \mathcal{I} \rightarrow \mathbb{R}^{n \times k}$ be continuously differentiable and column orthogonal. Then, for all $\tau \epsilon \mathcal{I}$, we have the inequalities

$$\begin{aligned} \| (I_n - Q_1 {Q_1}^T) \frac{dQ_1}{d\tau} \|^2 &\leq \| \frac{dQ_1}{d\tau} \|^2 \\ &\leq \| {Q_1}^T \frac{dQ_1}{d\tau} \|^2 + \| (I_n - Q_1 {Q_1}^T) \frac{dQ_1}{d\tau} \|^2 . \end{aligned} \tag{29}$$

Define the k-dimensional (time-dependent) linear subspace $\mathcal{S}_1$ by $\mathcal{S}_1 = \mathcal{R}(Q_1)$. Then the term $\| {Q_1}^T \frac{dQ_1}{d\tau} \|$ quantifies the rotational activity of Q_1 within $\mathcal{S}_1$, while $\| (I_n - Q_1 {Q_1}^T) \frac{dQ_1}{d\tau} \|$ quantifies the rotational activity of Q_1 orthogonal to $\mathcal{S}_1$. Now any orthogonal basis $\tilde{Q}_1$ of $\mathcal{S}_1$ can be written as

$$\tilde{Q}_1 = Q_1 Z_{11} , \tag{30}$$

where $Z_{11}: \mathcal{I} \rightarrow \mathbb{R}^{n \times k}$ is an orthogonal matrix function. Observe that

$$\begin{aligned} \| (I_n - \tilde{Q}_1 {\tilde{Q}_1}^T) \frac{d\tilde{Q}_1}{d\tau} \| &= \| (I_n - Q_1 {Q_1}^T) \left(\frac{dQ_1}{d\tau} Z_{11} + Q_1 \frac{dZ_{11}}{d\tau} \right) \| \\ &= \| (I_n - Q_1 {Q_1}^T) \frac{dQ_1}{d\tau} Z_{11} \| \\ &= \| (I_n - Q_1 {Q_1}^T) \frac{dQ_1}{d\tau} \| . \end{aligned} \tag{31}$$

Hence, the rotational activity within ${\mathcal{S}_1}^{\perp}$ is independent of the choice of the (orthogonal) basis. For the rotational activity within $\mathcal{S}_1$ we obtain

$$\begin{aligned} \| {\tilde{Q}_1}^T \frac{d\tilde{Q}_1}{d\tau} \| &= \| {Z_{11}}^T {Q_1}^T (\frac{dQ_1}{d\tau} Z_{11} + Q_1 \frac{dZ_{11}}{d\tau}) \| \\ &= \| {Q_1}^T \frac{dQ_1}{d\tau} Z_{11} + \frac{dZ_{11}}{d\tau} \| . \end{aligned} \tag{32}$$

Therefore, the rotational activity of the column orthogonal basis $\tilde{Q}_1$ of $\mathcal{S}_1$ is minimal if Z_{11} is a solution of the differential equation

$$\frac{dZ_{11}}{d\tau} = -{Q_1}^T \frac{dQ_1}{d\tau} Z_{11} . \tag{33}$$

In that case the rotational activity of $\tilde{Q}_1$ is equal to $\| (I_n - Q_1 {Q_1}^T) \frac{dQ_1}{d\tau} \|$. Remark that (33) implies that

$$\frac{d}{d\tau} ({Z_{11}}^T Z_{11}) = 0 .$$

So, any orthogonal initial value for Z_{11} induces an orthogonal $Z_{11}(\tau)$, for all $\tau \epsilon \mathcal{I}$. This leads to the following (unambiguous) definition.

Definition 1.21

Let X_1 be a given $n \times k$ matrix function, having full rank. The rotational activity *at τ of the k-dimensional, linear subspace $\mathcal{S}_1 = \mathcal{R}(X_1)$ is defined by*

$$\| \left(I_n - Q_1(\tau) {Q_1}^T(\tau) \right) \frac{dQ_1}{d\tau}(\tau) \| ,$$

where Q_1 is a column orthogonal (time-dependent) basis of $\mathcal{S}_1$.

■

Let $X_1 : \mathcal{I} \rightarrow \mathbb{R}^{n \times k}$ $(k < n)$ be a given continuously differentiable matrix function having full rank. Define $\mathcal{S}_1 = \mathcal{R}(X_1)$. One might think that the rotational activity of X_1 is at least as large as the rotational activity of $\mathcal{S}_1$. However, this is not the case, as is illustrated by the next example.

Example 1.22

Let, for $0 < |\varepsilon| \leq 1$, $X_1(\tau) = \begin{bmatrix} 1 & \sqrt{1-\varepsilon^2} \\ 0 & \varepsilon \cos\tau \\ 0 & \varepsilon \sin\tau \end{bmatrix}$, $\tau \in [0, 1]$. Then X_1 is identical to its direction matrix. Therefore, the rotational activity of $X_1(\tau)$ is equal to $\| \frac{dX_1}{d\tau} \| = \varepsilon$. Moreover, the matrix function $Q_1(\tau) = \begin{bmatrix} 1 & 0 \\ 0 & \cos\tau \\ 0 & \sin\tau \end{bmatrix}$ is column orthogonal and satisfies $\mathcal{R}(Q_1) = \mathcal{R}(X_1)$. Hence, the rotational activity of $\mathcal{R}\Big(X_1(\tau)\Big)$ is equal to $\| \Big(I_n - Q_1(\tau)\, Q_1(\tau)^T\Big) \frac{dQ_1}{d\tau}(\tau) \| = 1$.

■

The above example shows that an ill-conditioned matrix function X_1 may have a much lower rotational activity than the corresponding subspace $\mathcal{R}(X_1)$. A relation between the two follows from

Property 1.23

Let $X_1 : \mathcal{I} \rightarrow \mathbb{R}^{n \times k}$ $(k < n)$ be a given continuously differentiable matrix function having full rank. Let T_1 be the direction matrix corresponding to X_1. Decompose T_1 as $T_1 = Q_1 Z_{11}$, where Q_1 is column orthogonal. Then

$$\| \frac{dT_1}{d\tau} \| \geq \| (I_n - Q_1\, Q_1{}^T) \frac{dQ_1}{d\tau} \| \ \mathrm{glb}(T_1) \ . \tag{34}$$

Proof:

From $T_1 = Q_1 Z_{11}$ we obtain

$$\begin{aligned} \| \frac{dT_1}{d\tau} \| &= \| \frac{dQ_1}{d\tau} Z_{11} + Q_1 \frac{dZ_{11}}{d\tau} \| \\ &\geq \| (I_n - Q_1\, Q_1{}^T)\, (\frac{dQ_1}{d\tau} Z_{11} + Q_1 \frac{dZ_{11}}{d\tau}) \| \\ &= \| (I_n - Q_1\, Q_1{}^T) \frac{dQ_1}{d\tau} Z_{11} \| \\ &\geq \| (I_n - Q_1\, Q_1{}^T) \frac{dQ_1}{d\tau} \| \ \mathrm{glb}(Z_{11}) \end{aligned}$$

$$= \| (I_n - Q_1 Q_1{}^T) \frac{dQ_1}{d\tau} \| \text{ glb}(T_1) .$$

■

We shall return to the concept of rotational activities of matrix functions and of time-dependent linear subspaces in Section 3.2. In connection with solutions of linear differential equations we shall consider, for instance, a direct method for the computation of an orthogonal basis with minimal rotational activity.

Chapter 2

Fundamental concepts for BVPs

2.1 Introduction

In this chapter we consider the linear system of (*ordinary*) *differential equations* (DEs)

$$\frac{dx}{dt} = A(t)x + f(t)\ , \qquad t \epsilon \mathcal{I}\ , \tag{1}$$

where x and f are n-vectors and A an $n \times n$ matrix function. The interval of integration $\mathcal{I}$ may be finite or infinite. In the former case we shall always assume that $\mathcal{I} = [0,1]$ and in the latter case $\mathcal{I} = [0,\infty)$. For the moment we let $\mathcal{I}$ be finite, so $\mathcal{I} = [0,1]$, and consider *boundary conditions* (BCs) of the form:

$$B^0\, x(0) + B^1\, x(1) = b \tag{2}$$

($B^0, B^1 \epsilon \mathbb{R}^{n\times n}$ and $b \epsilon \mathbb{R}^n$). In the sequel we shall also consider BCs having a special form, such as separated ones, where $B^0 = \begin{bmatrix} 0 \\ B^{02} \end{bmatrix} \begin{matrix} \updownarrow k \\ \updownarrow n-k \end{matrix}$ and $B^1 = \begin{bmatrix} B^{11} \\ 0 \end{bmatrix} \begin{matrix} \updownarrow k \\ \updownarrow n-k \end{matrix}$, for some integer k.

A *fundamental solution* X of (1) is defined by

$$\frac{d}{dt}X = A(t)\,X\ , \qquad t \epsilon \mathcal{I}\ , \tag{3}$$

where X is an $n \times n$ matrix function with $X(0)$ non-singular (the columns of X form a complete set of solutions of the homogeneous part of (1)). The basic existence theorem for the two-point *boundary value problem* (BVP) (1) and (2) can be stated as follows

Theorem 2.1 ([29])
Let $A, f \in C^p[0,1]$ and X be a fundamental solution. Then the BVP (1) and (2) has a unique solution $x \in C^{p+1}[0,1]$ if and only if $\mathcal{B}(X) = B^0 X(0) + B^1 X(1)$ is non-singular.

■

In the sequel we assume that (1), subject to (2), has a unique solution (is a well-posed problem). For any fundamental solution X the solution x of (1), (2) can be written as

$$x(t) = X(t)\,\mathcal{B}(X)^{-1}\,b + \int_0^1 G(t,s) f(s)\,ds\ , \tag{4}$$

where the *Green's (matrix) function* G is given by

$$G(t,s) = \begin{cases} X(t)\,\mathcal{B}(X)^{-1}\,B^0\,X(0)\,X^{-1}(s) & ,\ t \geq s \\ -X(t)\,\mathcal{B}(X)^{-1}\,B^1\,X(1)\,X^{-1}(s) & ,\ t < s \end{cases}\ . \tag{5}$$

Define

$$\alpha = \max_{0 \leq t,s \leq 1} \| G(t,s) \| \tag{6}$$

and

$$\beta = \max_{0 \leq t \leq 1} \| X(t)\,\mathcal{B}(X)^{-1} \|\ . \tag{7}$$

Then we obtain, for all $t \in [0,1]$, the (*stability*) *inequality*

$$\| x(t) \| \leq \beta \| b \| + \alpha \int_0^1 \| f(s) \|\, ds\ . \tag{8}$$

The choice of the function norm is in some sense arbitrary, but may have a great influence on the magnitude of the stability constants, as is for instance seen in

Example 2.2 (cf. [56])
Consider, for $0 < \varepsilon \ll 1$, the DE

$$\varepsilon \frac{d^2u}{dt^2} + 2t \frac{du}{dt} = 0 , \quad t\epsilon[-1,1] , \tag{9a}$$

subject to

$$u(-1) = 0 \quad \text{and} \quad u(1) = 1 . \tag{9b}$$

Transforming (9a) into a first order system with $x_1 = \frac{du}{dt}$ and $x_2 = u$ yields

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{bmatrix} -2t/\varepsilon & 0 \\ 1 & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} . \tag{10}$$

Define

$$E(t) = e^{-t^2/\varepsilon} \quad \text{and} \quad I(t) = \frac{1}{\sqrt{\pi\varepsilon}} \int_{-\infty}^{t} E(\tau)\, d\tau .$$

The scaling factor $\sqrt{\pi\varepsilon}$ has been chosen such that $I(1) \approx 1$, for ε sufficiently small. A fundamental solution X of (10) is given by

$$X(t) = \begin{bmatrix} \frac{1}{\sqrt{\pi\varepsilon}} E(t) & -\frac{1}{\sqrt{\pi\varepsilon}} E(t) \\ I(t) & 1 - I(t) \end{bmatrix} .$$

Observe that $\mathcal{B}(X) = B^{-1} X(-1) + B^1 X(1) \approx I_2$. Hence,

$$\max_t \| X(t)\, \mathcal{B}(X)^{-1} \| \approx \max_t \| X(t) \| = O(\frac{1}{\sqrt{\varepsilon}}) ,$$

but

$$\int_0^1 \| X(t)\, \mathcal{B}(X)^{-1} \|\, dt \approx \int_0^1 \| X(t) \|\, dt = O(1) .$$

On the other hand, with $x_1 = \sqrt{\pi\varepsilon}\, \frac{du}{dt}$ and $x_2 = u$ we obtain the system

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{bmatrix} -2t/\varepsilon & 0 \\ \frac{1}{\sqrt{\pi\varepsilon}} & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} . \tag{11}$$

A fundamental solution X of (11) is given by

$$X(t) = \begin{bmatrix} E(t) & -E(t) \\ I(t) & 1 - I(t) \end{bmatrix} ,$$

which implies that both $\max_t \| X(t)\, \mathcal{B}(X)^{-1} \|$ and $\int_0^1 \| X(t)\, \mathcal{B}(X)^{-1} \|\, dt$ are of order 1.

If, however, one is interested in the value of $\frac{du}{dt}$ then at $t = 0$ one is confronted again with the factor $1/\sqrt{\varepsilon}$.

■

Observe that both α and β are independent of the choice of the fundamental solution X, that α is invariant for row scaling of the BCs, but β is not. Therefore we first look more precisely at the BCs.

Let $U\,\Sigma\,V^T$ be the SVD (cf. (1.13)) of $B = \left[\, B^0 \mid B^1 \,\right] \epsilon\, \mathbb{R}^{n\times 2n}$. Now the *condition of the BCs* is defined as $\kappa(B)$ (cf. (1.12)). Observe that this condition is not influenced by an orthogonal transformation. Hence, without loss of generality we may assume that $U = I_n$. In order to normalize the BCs we observe that

$$1 = \kappa(V) \leq \kappa(DB) \ ,$$

for any $n \times n$ diagonal matrix D. Therefore, by writing the BCs as

$$V^0\, x(0) + V^1\, x(1) = \Sigma^{-1} b \tag{12}$$

we have, with respect to the 2-norm, optimized the BCs by an appropriate row scaling (Σ^{-1}). In the sequel we shall make the following assumption

Assumption 2.3
The $n \times 2n$ matrix $B = \left[\, B^0 \mid B^1 \,\right]$ is row orthogonal. Moreover, the BVP (1) and (2) is well-conditioned, i.e., the stability constants α and β are of moderate size.

■

A straightforward way of solving the BVP (1), (2) is the so-called *shooting method.* In that case the solution x is constructed by superposition of the columns of a fundamental solution X and a particular solution, as for instance has been done in (4). However, in many applications the solution space of (1), (2) is dichotomic, which, roughly speaking, means that homogeneous modes, i.e., solutions of the homogeneous part of (1), exhibit a widely varying growth behaviour. In general this implies that the columns of the fundamental solution X will become nearly dependent, which, of course, influences the numerical computations where this X is involved. How this numerical instability can be overcome will be discussed in the next chapter.

2.2 Dichotomy

In this section we shall have a closer look at the concept of *dichotomy*. For various reasons we assume that the system of differential equations is homogeneous and defined on $\overline{\mathbb{R}^+}$ ($\mathcal{I} = \overline{\mathbb{R}^+}$). Hence,

$$\frac{dx}{dt} = A(t)\,x\ , \qquad t \geq 0\ . \tag{13}$$

Let X be a fundamental solution of (13), i.e.,

$$\begin{cases} \dfrac{d}{dt}X = A(t)\,X\ , & t \geq 0 \\ X(0) \text{ non-singular} \end{cases} . \tag{14}$$

Denote the *solution space* (or *flow*) $\mathcal{S}$ of this DE by

$$\mathcal{S} = \{\, Xc \,|\, c \in \mathbb{R}^n \,\}\ . \tag{15}$$

Now we shall concentrate on solution subspaces $\mathcal{S}$ that can be split into two subspaces of (exponentially) decreasing solutions; one for decreasing time ($\mathcal{S}_1$) and the other for increasing time ($\mathcal{S}_2$).

Definition 2.4
A solution subspace $\mathcal{S}_1 \subset \mathcal{S}$ is called a dominant subspace *or a* subspace of unstable solutions *if there exist positive constants k_1 and λ_1 such that for any non-trivial $\phi_1 \in \mathcal{S}_1$ we have*

$$\frac{\| \phi_1(t) \|}{\| \phi_1(s) \|} \leq k_1\, e^{-\lambda_1 (s-t)}, \qquad t \leq s\ . \tag{16}$$

A solution subspace $\mathcal{S}_2 \subset \mathcal{S}$ is called a dominated subspace *or a* subspace of stable solutions *if there exist a positive constant k_2 and a non-negative constant λ_2 such that for any non-trivial $\phi_2 \in \mathcal{S}_2$ we have*

$$\frac{\| \phi_2(t) \|}{\| \phi_2(s) \|} \leq k_2\, e^{-\lambda_2 (t-s)}, \qquad t \geq s\ . \tag{17}$$

■

Here k_1 and k_2 are assumed to be of moderate size. The option that λ_2 may be equal to 0 implies that smooth solutions are assumed to be in $\mathcal{S}_2$.

Remark 2.5

(i) On an infinite interval the mere existence of finite k_1 and k_2 is sufficient for analytic applications. For numerical purposes and on finite intervals we have to be more careful. That is why k_1 and k_2 should not be excessively large. The relation between the magnitudes of k_i, $\lambda_i\,(i = 1,2)$ and the length of the interval (when this is not normalized to length 1) becomes important. We shall return to this question at the end of this section.

(ii) If A is a constant matrix, then any non-trivial solution of (13) satisfies

$$\frac{\| x(t+h) \|}{\| x(t) \|} \leq p_n(h)\, e^{h\,\lambda_{\max}(A)}\,,$$

where $p_n(h)$ is an n-th degree polynomial in h with $p_n(0) = 1$ ([59]). Corresponding to this inequality a sharper bound (especially for small values of h) may be obtained if we replace Definition 2.4 by

$$\phi_1 \in \mathcal{S}_1 \;\Rightarrow \| \phi_1(t) \| \leq \| \phi_1(t+h) \|\, (1 + k_1 h)\, e^{-\lambda_1 h} \quad (h > 0)$$

and

$$\phi_2 \in \mathcal{S}_2 \;\Rightarrow \| \phi_2(t) \| \leq \| \phi_2(t+h) \|\, (1 + k_2 h)\, e^{-\lambda_2 h} \quad (h > 0)\,.$$

However, with these definitions the formulations and proofs of theorems like Theorem 2.9 become more complex.

(iii) The subspace $\mathcal{S}_1$ is *not* uniquely defined. Nevertheless, for t going to infinity the subspace $\mathcal{S}_1(t)$ is unique (cf. Theorem 2.19). The subspace $\mathcal{S}_2$, however, is unique, since it can be regarded as the subspace of solutions of (13) that are uniformly bounded. It is not restrictive to assume that $\mathcal{S}_1(0)$ is orthogonal to $\mathcal{S}_2(0)$, which makes both $\mathcal{S}_1$ and $\mathcal{S}_2$ uniquely defined.

■

Definition 2.6
A solution space $\mathcal{S}$ that can be split in $\mathcal{S}_1 \oplus \mathcal{S}_2$ with $\mathcal{S}_i\,(i = 1,2)$ satisfying, respectively, (16) *and* (17) *is said to be* weakly exponentially dichotomic. (In [7] this property is called 'comparative exponential dichotomy'.)

■

In this defintion the adverb 'weakly' is used to distinguish between this property of the solution space and the better known property of exponential dichotomy.

Definition 2.7 ([11])
The solution space $\mathcal{S}$ is called exponentially dichotomic *if for each fundamental*

solution X there exists a projection P ($P \neq 0, I_n$) such that

$$\begin{aligned} \| X(t)\, P\, X^{-1}(s) \| &\leq m_1\, e^{-\lambda_1 (s-t)}, \quad 0 \leq t \leq s\,, \\ \| X(t)\, (I-P)\, X^{-1}(s) \| &\leq m_2\, e^{-\lambda_2 (t-s)}, \quad t \geq s \geq 0\,. \end{aligned}$$

For the same reasons as before the quantities m_1 and m_2 should not be large. Moreover, we assume that $\lambda_1 > 0$ and $\lambda_2 \geq 0$.

■

Remark 2.8
If in Definition 2.7 both λ_1 and λ_2 are zero, then the solution space $\mathcal{S}$ is called *dichotomic*. In [26] it is shown that well-conditioning of the BVP, corresponding to the stability constants given in (6) and (7), implies dichotomy of the solution space $\mathcal{S}$. This concept will be used in Chapter 5, when we are dealing with singular perturbation problems.

■

The following result is well-known (cf. [7]). However, since its proof is quite instructive, we shall recapitulate it here.

Theorem 2.9
If the solution space $\mathcal{S}$ is exponentially dichotomic then it is also weakly exponentially dichotomic.

Proof:
Assume $\mathcal{S}$ is exponentially dichotomic and let X be a fundamental solution of (14). Since P of Definition 2.7 is a projection, there exists a non-singular matrix H such that

$$H\, P\, H^{-1} = \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} \quad (1 \leq k < n)\,.$$

Define $Y = X\, H^{-1}$ and write $Y = \begin{bmatrix} \underset{k}{\longleftrightarrow} & \underset{n-k}{\longleftrightarrow} \\ Y_1 & Y_2 \end{bmatrix}$. Then

$$\begin{aligned} \| X(t)\, P\, X^{-1}(s) \| &= \| Y(t) \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} Y^{-1}(s) \| \\ &= \| \begin{bmatrix} Y_1(t) & 0 \end{bmatrix} Y^{-1}(s) \|\,. \end{aligned} \tag{18}$$

Let

$$\mathcal{S}_1 = \{ Y_1 c_1 \mid c_1 \in \mathbb{R}^k \} . \tag{19}$$

Thus $\phi_1 \in \mathcal{S}_1 \Rightarrow \phi_1(t) = Y_1(t)\, c_1$, $t \geq 0$ $(c_1 \in \mathbb{R}^k)$.
Hence, for any non-trivial $\phi_1 \in \mathcal{S}_1$ we obtain

$$\begin{aligned} \frac{\| \phi_1(t) \|}{\| \phi_1(s) \|} &= \frac{\| Y_1(t)\, c_1 \|}{\| Y_1(s)\, c_1 \|} = \frac{\| \begin{bmatrix} Y_1(t) & 0 \end{bmatrix} \begin{pmatrix} c_1 \\ 0 \end{pmatrix} \|}{\| \begin{bmatrix} Y_1(s) & Y_2(s) \end{bmatrix} \begin{pmatrix} c_1 \\ 0 \end{pmatrix} \|} \\ &\leq \| \begin{bmatrix} Y_1(t) & 0 \end{bmatrix} Y^{-1}(s) \| \\ &\leq m_1 e^{-\lambda_1 (s-t)} , \quad t \leq s , \end{aligned}$$

by (18).
In a similar way we define

$$\mathcal{S}_2 = \{ Y_2 c_2 \mid c_2 \in \mathbb{R}^{n-k} \} \tag{20}$$

and derive the relation $(\phi_2 \neq 0)$

$$\phi_2 \in \mathcal{S}_2 \Rightarrow \frac{\| \phi_2(t) \|}{\| \phi_2(s) \|} \leq m_2 e^{-\lambda_2 (t-s)} , \quad t \geq s .$$

■

Weakly exponential dichotomy is not enough to guarantee exponential dichotomy (cf. [38]) as is illustrated by the next example.

Example 2.10

Let $X(t) = \begin{bmatrix} 1 & \sqrt{1-e^{-2t}} \\ 0 & e^{-t} \end{bmatrix} \begin{bmatrix} e^{\lambda t} & 0 \\ 0 & e^{-\lambda t} \end{bmatrix}$, $t \geq 0$ $(\lambda > 0)$. Clearly, $\mathcal{S}$ is weakly exponentially dichotomic with the dominant subspace $\mathcal{S}_1(t) = \text{span} \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\}$ and the dominated subspace $\mathcal{S}_2(t) = \text{span} \left\{ \begin{pmatrix} \sqrt{1-e^{-2t}} \\ e^{-t} \end{pmatrix} \right\}$, $t > 0$. However, for any projection $P \neq 0, I_n$ the quantity $\| X(t)\, P\, X^{-1}(t) \|$ is not uniformly bounded.

■

As will be demonstrated in the sequel exponential dichotomy implies that the gap between $\mathcal{S}_1$ and $\mathcal{S}_2$ remains bounded away from zero (cf. Definition 1.10).

Assume S is weakly exponentially dichotomic with some given $\mathcal{S}_1$ and $\mathcal{S}_2$. Let $Y = \begin{bmatrix} \underset{k}{\longleftrightarrow} & \underset{n-k}{\longleftrightarrow} \\ Y_1 & Y_2 \end{bmatrix}$ be a fundamental solution of (14) such that the columns of Y_1 span $\mathcal{S}_1$ and the columns of Y_2 span $\mathcal{S}_2$. Make, for $t \geq 0$, the following QR-decompositions:

$$Y_1(t) = \tilde{Q}_1(t)\, R_{11}(t) \tag{21}$$

and

$$Y_2(t) = \tilde{Q}_2(t)\, R_{22}(t) \tag{22}$$

($\tilde{Q}_1$, $\tilde{Q}_2$ column orthogonal matrices and R_{11}, R_{22} upper triangular matrices). Hence,

$$\begin{aligned} Y(t) &= \Big[Y_1(t)\; Y_2(t)\Big] = \Big[\tilde{Q}_1(t)\; \tilde{Q}_2(t)\Big] \begin{bmatrix} R_{11}(t) & 0 \\ 0 & R_{22}(t) \end{bmatrix} \\ &= \Big[Q_1(t)\; Q_2(t)\Big] \begin{bmatrix} I_k & {Q_1}^T(t)\,\tilde{Q}_2(t) \\ 0 & {Q_2}^T(t)\,\tilde{Q}_2(t) \end{bmatrix} \begin{bmatrix} R_{11}(t) & 0 \\ 0 & R_{22}(t) \end{bmatrix} \end{aligned} \tag{23}$$

where $Q(t) = \Big[Q_1(t)\; Q_2(t)\Big]$ is an orthogonal matrix with $Q_1(t) = \tilde{Q}_1(t) (t \geq 0)$. Using (23) and Property 1.14 we obtain

Property 2.11

$$\mathrm{GAP}\Big(\mathcal{S}_1(t), \mathcal{S}_2(t)\Big) = \| \left({Q_2}^T(t)\,\tilde{Q}_2(t)\right)^{-1} \|^{-1}, \quad t \geq 0\,.$$

■

Before the main result of this subsection is formulated we first need

Lemma 2.12

$$\begin{aligned} 1/\mathrm{GAP}\Big(\mathcal{S}_1(t), \mathcal{S}_2(t)\Big) &= \| Y(t) \begin{bmatrix} 0 & 0 \\ 0 & I_{n-k} \end{bmatrix} Y^{-1}(t) \| \\ &= \| Y(t) \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} Y^{-1}(t) \|\,, \quad t \geq 0\,. \end{aligned}$$

Proof:

$$\| Y(t) \begin{bmatrix} 0 & 0 \\ 0 & I_{n-k} \end{bmatrix} Y^{-1}(t) \| =$$

$$= \| \begin{bmatrix} \tilde{Q}_1(t) & \tilde{Q}_2(t) \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & I_{n-k} \end{bmatrix} \begin{bmatrix} I_k & {Q_1}^T(t)\,\tilde{Q}_2(t) \\ 0 & {Q_2}^T(t)\,\tilde{Q}_2(t) \end{bmatrix}^{-1} Q^{-1}(t) \|$$

$$= \| \begin{bmatrix} \tilde{Q}_1(t) & \tilde{Q}_2(t) \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \left({Q_2}^T(t)\,\tilde{Q}_2(t) \right)^{-1} \end{bmatrix} \|$$

$$= \| \tilde{Q}_2(t) \left({Q_2}^T(t)\,\tilde{Q}_2(t) \right)^{-1} \|$$

$$= \| \left({Q_2}^T(t)\,\tilde{Q}_2(t) \right)^{-1} \| = 1/\mathrm{GAP}\left(\mathcal{S}_1(t), \mathcal{S}_2(t) \right)$$

by Property 2.11. This proves the first equality.

To derive the second equality we observe that $Y(t) \begin{bmatrix} 0 & 0 \\ 0 & I_{n-k} \end{bmatrix} Y^{-1}(t)$ is a projection. Now the result follows directly from the observation that $\| P \| = \| I_n - P \|$, for any projection $P \neq 0, I_n$.

■

Theorem 2.13 (cf. [11], p.11)
The solution space $\mathcal{S}$ is exponentially dichotomic if and only if $\mathcal{S}$ is weakly exponentially dichotomic and the gap between the corresponding $\mathcal{S}_1$ and $\mathcal{S}_2$ is uniformly bounded away from zero.

Proof:
if-part: Assume $\mathcal{S}$ is weakly exponentially dichotomic and $\mathrm{GAP}\left(\mathcal{S}_1(t), \mathcal{S}_2(t) \right) \geq d > 0$, for all $t > 0$. Let Y be a fundamental solution as in (21) and (22). Then, for any non-trivial $\phi_2 \in \mathcal{S}_2$ and $t \geq s$, we have

$$\begin{aligned} k_2\, e^{-\lambda_2 (t-s)} &\geq \frac{\| \phi_2(t) \|}{\| \phi_2(s) \|} \\ &= \frac{\| \tilde{Q}_2(t)\, R_{22}(t)\, c_2 \|}{\| \tilde{Q}_2(s)\, R_{22}(s)\, c_2 \|} \qquad (0 \neq c_2 \in \mathbb{R}^n) \\ &= \frac{\| R_{22}(t)\, c_2 \|}{\| R_{22}(s)\, c_2 \|} . \end{aligned}$$

Hence,

$$\begin{aligned} \| R_{22}(t)\, {R_{22}}^{-1}(s) \| &= \max_{d_2 \in \mathbb{R}^{n-k}} \left\{ \frac{\| R_{22}(t)\, {R_{22}}^{-1}(s)\, d_2 \|}{\| d_2 \|} \;\middle|\; d_2 \neq 0 \right\} \\ &\leq k_2\, e^{-\lambda_2 (t-s)} . \end{aligned} \tag{24}$$

Therefore,

$$\| Y(t) \begin{bmatrix} 0 & 0 \\ 0 & I_{n-k} \end{bmatrix} Y^{-1}(s) \| = \| R_{22}(t) {R_{22}}^{-1}(s) \left({Q_2}^T(s) \; \tilde{Q}_2(s) \right)^{-1} \|$$

$$\leq \frac{\| R_{22}(t) {R_{22}}^{-1}(s) \|}{\mathrm{GAP}\left(\mathcal{S}_1(s), \mathcal{S}_2(s) \right)} \leq \frac{k_2}{d} e^{-\lambda_2 (t-s)}, \quad t \geq s .$$

In a similar way we may deduce

$$\| Y(t) \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} Y^{-1}(s) \| \leq \frac{k_1}{d} e^{-\lambda_1 (s-t)}, \quad t \leq s .$$

By the relation $X = YH$ (H non-singular), $P = H^{-1} \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} H$ and the foregoing results we obtain validity for any fundamental solution X, which completes the proof.

The 'only if'-part is a direct consequence of Theorem 2.9 and Lemma 2.12.

■

Without an explicit calculation of a fundamental solution it is hard to check, in general, whether $\mathcal{S}$ is exponentially dichotomic or not. A rather trivial case is obtained if A is a non-singular, constant matrix with eigenvalues in both the left and right (open) halfplane of $\mathbb{C}$ ([59]). Less trivial is the following result, originally derived by A. Lazer ([32]).

Theorem 2.14
Let $A(t) = \left(a_{ij}(t)\right)$ be a continuous $n \times n$ matrix function on the half line $\overline{\mathbb{R}^+}$. Suppose there exist a constant $\delta > 0$ and an integer k, $1 \leq k < n$, such that, for all $t \geq 0$,

(i) $|\, a_{ii}(t) \,| \; \geq \; \delta + \sum_{\substack{j=1 \\ j \neq i}}^{n} |\, a_{ij}(t) \,| \qquad (i = 1, \ldots, n)$

(ii) $a_{ii}(t) > 0 \quad (i = 1, \ldots, k)$ *and* $a_{ii}(t) < 0 \quad (i = k+1, \cdots, n)$.

Then $\mathcal{S}$ is weakly exponentially dichotomic and $\dim\left(\mathcal{S}_1(t)\right) = k$.

■

A generalization of this result will be given in Chapter 4.

A sometimes easily verifiable property is that of (exponentially) bounded growth (cf. [11], p.9).

Definition 2.15
The solution space $\mathcal{S}$ of the homogeneous DE

$$\frac{dx}{dt} = A(t)\,x\ , \qquad t \geq 0\ , \tag{25}$$

has exponentially bounded growth *if there exist an $\alpha \geq 0$ and a $c > 0$ such that every non-trivial solution of* (25) *satisfies*

$$\frac{\| x(t) \|}{\| x(s) \|} \leq c e^{\alpha |t-s|}\ , \quad \textit{for all } t, s \geq 0\ .$$

■

This definition directly leads to the next two results.

Corollary 2.16
$\mathcal{S}$ has exponentially bounded growth if and only if for any fundamental solution X of (14) *we have*

$$\| X(t) X^{-1}(s) \| \leq c e^{\alpha |t-s|}, \quad t, s \geq 0\ .$$

■

Corollary 2.17
If $\int_s^t \| A(\tau) \|\, d\tau \leq \alpha |t-s|$ $(t, s \geq 0)$, then $\mathcal{S}$ has exponentially bounded growth.

■

A relation between exponentially bounded growth and a separation of subspaces of the solution space $\mathcal{S}$ is given by

Theorem 2.18
Assume that the solution space $\mathcal{S}$ has exponentially bounded growth. If, moreover, $\mathcal{S}$ is weakly exponentially dichotomic then the gap between the corresponding $\mathcal{S}_1(t)$ and $\mathcal{S}_2(t)$ is uniformly bounded away from zero.

Proof:
(We shall construct a solution of (25) that, for some $t > s \geq 0$, will grow as fast as $\frac{1}{k_1} e^{\lambda_1 (t-s)} / \mathrm{GAP}\big(\mathcal{S}_1(s), \mathcal{S}_2(s)\big)$. By the 'exponentially bounded growth'-condition this implies that the gap must be bounded away from zero.)

Let $s \geq 0$ be fixed. Choose $u \in \mathcal{S}_1(s)$, $v \in \mathcal{S}_2(s)$ and $\theta \in (0, \pi/2]$ such that $u^T u = v^T v = 1$ and $\mathrm{GAP}\Big(\mathcal{S}_1(s), \mathcal{S}_2(s)\Big) = \mathrm{GAP}\Big(\mathrm{span}\{u\}, \mathrm{span}\{v\}\Big) = \sin\theta$. Define $\phi_1 \in \mathcal{S}_1$ and $\phi_2 \in \mathcal{S}_2$ by $\phi_1(s) = u$ and $\phi_2(s) = v$. Let

$$\phi(t) = \frac{1}{\sin\theta}\,\phi_1(t) \; - \; \frac{\cos\theta}{\sin\theta}\,\phi_2(t)\;, \qquad t \geq 0\;.$$

Then $\phi \in \mathcal{S}$ and $\| \phi(s) \| = 1$. Moreover, for any $h > 0$, we have

$$\begin{aligned}
\| \phi(s+h) \| &= \| \phi_1(s+h) - \cos\theta\, \phi_2(s+h) \| \,/\, \sin\theta \\
&\geq \frac{\| \phi_1(s+h) \| \; - \cos\theta \; \| \phi_2(s+h) \|}{\mathrm{GAP}\Big(\mathcal{S}_1(s), \mathcal{S}_2(s)\Big)} \\
&\geq \frac{\frac{1}{k_1}\, e^{\lambda_1 h} - \cos\theta\, k_2\, e^{-\lambda_2 h}}{\mathrm{GAP}\Big(\mathcal{S}_1(s), \mathcal{S}_2(s)\Big)} \\
&\geq \frac{\frac{1}{k_1}\, e^{\lambda_1 h} - k_2\, e^{-\lambda_2 h}}{\mathrm{GAP}\Big(\mathcal{S}_1(s), \mathcal{S}_2(s)\Big)}
\end{aligned}$$

Since $\| \phi(s+h) \| \leq c\, e^{\alpha h} \| \phi(s) \| = c\, e^{\alpha h}$, we have obtained, for any $h > 0$, the relation

$$\mathrm{GAP}\Big(\mathcal{S}_1(s), \mathcal{S}_2(s)\Big) \; \geq \; \frac{\frac{1}{k_1}\, e^{\lambda_1 h} - k_2\, e^{-\lambda_2 h}}{c\, e^{\alpha h}}\;.$$

Define

$$d = \sup_{h>0} \frac{\frac{1}{k_1}\, e^{\lambda_1 h} - k_2\, e^{-\lambda_2 h}}{c\, e^{\alpha h}}\;,$$

then $d > 0$ and independent of s. So

$$\mathrm{GAP}\Big(\mathcal{S}_1(s), \mathcal{S}_2(s)\Big) \geq d > 0\;, \quad \text{for all } s \geq 0\;.$$

■

From these results we for instance conclude that exponential dichotomy is obtained if the matrix function A is uniformly bounded and satisfies the conditions of Theorem 2.14 (cf. [11], proposition 6.3). It is even sufficient that the system can be transformed (by a sufficiently smooth transformation) into a system that satisfies these properties. We shall return to this aspect in the Chapters 3 and 4.

2.3 Consistency

In section 2.2 we obtained some insight in the meaning of a (weakly) exponentially dichotomic solution space $\mathcal{S}$. Here we want to deduce some convergence and growth properties for solutions in such a solution space. To this end we need Definition 1.12 of distance between subspaces.

2.3.1 Infinite intervals

With Property 1.16 we are able to prove

Theorem 2.19
Let the solution subspace $\mathcal{S}$ be weakly exponentially dichotomic and $Y = \underset{k}{\left[Y_1 \right.} \; \underset{n-k}{\left. Y_2 \right]}$ *a fundamental solution such that* $\mathcal{R}\big(Y_1(t)\big) = \mathcal{S}_1(t)$ *and* $\mathcal{R}\big(Y_2(t)\big) = \mathcal{S}_2(t)$, *for all* $t \geq 0$. *Let* $X(t) = \underset{k}{\left[X_1(t) \right.} \; \underset{n-k}{\left. X_2(t) \right]} = Y(t)\,H$, $t \geq 0$ (H non-singular). *Then*

(i) H_{11} *non-singular* $\Rightarrow$ $\mathrm{DIST}\big(\mathcal{R}\big(X_1(t)\big), \mathcal{S}_1(t)\big) = \mathrm{O}\big(e^{-(\lambda_1 + \lambda_2)\,t}\big)$, $t \to \infty$.

(ii) H_{11} *singular* $\Rightarrow$ $\mathrm{DIST}\big(\mathcal{R}\big(X_1(t)\big), \mathcal{S}_1(t)\big) \geq \mathrm{GAP}\big(\mathcal{S}_2(t), \mathcal{S}_1(t)\big)$.

Proof:

(i) Observe that, using the notation of (21), (22) and the partitioning $H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{matrix} \updownarrow k \\ \updownarrow n-k \end{matrix}$, $X_1(t) = \begin{bmatrix} \tilde{Q}_1 & \tilde{Q}_2 \end{bmatrix} \begin{bmatrix} R_{11}(t)\,H_{11} \\ R_{22}(t)\,H_{21} \end{bmatrix}$. In (24) we have seen that

$$\| R_{22}(t)\,{R_{22}}^{-1}(s) \| \leq k_2\, e^{-\lambda_2 (t-s)} , \quad t \geq s ,$$

Similarly we obtain

$$\| R_{11}(t)\,{R_{11}}^{-1}(s) \| \leq k_1\, e^{-\lambda_1 (s-t)} , \quad t \leq s .$$

Hence, if H_{11} is non-singular, then

$$\begin{aligned} F(t) &:= \| R_{22}(t)\,H_{21}\,{H_{11}}^{-1}\,{R_{11}}^{-1}(t) \| \\ &\leq \| R_{22}(t)\,{R_{22}}^{-1}(0) \|\; F(0) \;\| R_{11}(0)\,{R_{11}}^{-1}(t) \| \end{aligned}$$

$$\leq \; k_1 k_2 e^{-(\lambda_1 + \lambda_2)t} F(0) \; . \tag{26}$$

From this and Property 1.16 result (i) follows.

(ii) If H_{11} is singular then there exists a vector $v_1 \in \mathbb{R}^k$ $(v_1 \neq 0)$ such that $X_1 v_1 = \tilde{Q}_2 R_{22} H_{21} v_1$, which is in $\mathcal{S}_2$.

■

Remark 2.20

(i) Observe that to obtain $\mathrm{DIST}\Big(\mathcal{R}\Big(X_1(t)\Big), \mathcal{S}_1(t)\Big) \to 0$ as $t \to \infty$ it is not strictly necessary that both λ_1 and λ_2 are positive. If H_{11} is non-singular we only have to require that $\lambda_1 + \lambda_2 > 0$.

(ii) The condition 'H_{11} non-singular' is identical to

$$\mathcal{S}_2(0) \cap \mathcal{R}\Big(X_1(0)\Big) = \{0\} \; . \tag{27}$$

Hence, as soon as (27) is satisfied we have

$$\mathrm{DIST}\Big(\mathcal{R}\Big(X_1(t)\Big), \mathcal{S}_1(t)\Big) = \mathrm{O}(e^{-(\lambda_1 + \lambda_2)t}) \; .$$

■

Definition 2.21 (cf. [38])
Let the solution space $\mathcal{S}$ be weakly exponentially dichotomic with corresponding subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$. Then a fundamental solution $X = \begin{bmatrix} \underset{k}{\longleftrightarrow} & \underset{n-k}{\longleftrightarrow} \\ X_1 & X_2 \end{bmatrix}$ *is called*

consistent *if it satisfies* $\mathcal{S}_2(0) \cap \mathcal{R}\Big(X_1(0)\Big) = \{0\}$.

■

What we conclude from (26) is that for large values of t the distance between $\mathcal{R}\Big(X_1(0)\Big)$ and $\mathcal{S}_1(0)$ is of minor influence. This is what we would expect, since $\mathcal{S}_1(0)$ is not uniquely defined by (16). As is illustrated by Theorem 2.19 only the asymptotic behaviour of $\mathcal{S}_1(t)$ is unique.

To measure consistency of a fundamental solution X we may use the quantity $1/\mathrm{GAP}\Big(\mathcal{R}\Big(X_1(0)\Big), \mathcal{S}_2(0)\Big)$. This can be seen from the following observation. Write, using the same notation as in (21) and (22),

$$X_1(0) = Y(0) \begin{bmatrix} H_{11} \\ H_{21} \end{bmatrix} = \begin{bmatrix} \tilde{Q}_1(0) & \tilde{Q}_2(0) \end{bmatrix} \begin{bmatrix} K_{11} \\ K_{21} \end{bmatrix} \; .$$

Then $\| R_{22}(0) H_{21} {H_{11}}^{-1} {R_{11}}^{-1}(0) \| = \| K_{21} {K_{11}}^{-1} \|$. Now construct an orthogonal matrix $U = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \in \mathbb{R}^{n \times n}$ such that $X_1(0) = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} V_{11} \\ 0 \end{bmatrix}$ (V_{11} non-singular). Then

$$\begin{aligned} \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} V_{11} \\ 0 \end{bmatrix} &= X_1(0) = \begin{bmatrix} \tilde{Q}_1(0) & \tilde{Q}_2(0) \end{bmatrix} \begin{bmatrix} K_{11} \\ K_{21} \end{bmatrix} \\ &= \begin{bmatrix} \tilde{Q}_1(0) & \tilde{Q}_2(0) \end{bmatrix} \begin{bmatrix} I_k \\ K_{21} {K_{11}}^{-1} \end{bmatrix} K_{11} \, . \end{aligned}$$

Hence,

$$\begin{bmatrix} V_{11} {K_{11}}^{-1} \\ 0 \end{bmatrix} = U^T \begin{bmatrix} \tilde{Q}_1(0) & \tilde{Q}_2(0) \end{bmatrix} \begin{bmatrix} I_k \\ K_{21} {K_{11}}^{-1} \end{bmatrix} ,$$

from which we obtain the relation

$$K_{21} {K_{11}}^{-1} = -\Big({U_2}^T \tilde{Q}_2(0) \Big)^{-1} {U_2}^T \tilde{Q}_1(0) \, .$$

Using Property 1.14 we have

$$\begin{aligned} \| R_{22}(0) H_{21} {H_{11}}^{-1} {R_{11}}^{-1}(0) \| &= \| K_{21} {K_{11}}^{-1} \| \\ &\leq \| \Big({U_2}^T \tilde{Q}_2(0) \Big)^{-1} \| \, \| {U_2}^T \tilde{Q}_1(0) \| \\ &\leq \frac{\mathrm{DIST}\Big(\mathcal{R}\Big(X_1(0) \Big), \mathcal{S}_1(0) \Big)}{\mathrm{GAP}\Big(\mathcal{R}\Big(X_1(0) \Big), \mathcal{S}_2(0) \Big)} \\ &\leq \frac{1}{\mathrm{GAP}\Big(\mathcal{R}\Big(X_1(0) \Big), \mathcal{S}_2(0) \Big)} \end{aligned} \tag{28}$$

We can always find a dominant subspace $\mathcal{S}_1$ with $\mathrm{DIST}\Big(\mathcal{R}\Big(X_1(0) \Big), \mathcal{S}_1(0) \Big) = 0$. So, the consistency of X is most fairly (and uniquely) quantified by $1/\mathrm{GAP}\Big(\mathcal{R}\Big(X_1(0) \Big), \mathcal{S}_2(0) \Big)$.

We shall finish this section with a statement conserning the growth of solutions that are determined by $X_1(0)$.

Theorem 2.22

Assume $\mathcal{S}$ is weakly exponentially dichotomic and $\mathrm{GAP}\Big(\mathcal{S}_1(t), \mathcal{S}_2(t) \Big) \geq d > 0$, *for all $t \geq 0$. Let $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$ be a consistent fundamental solution. For any non-trivial solution $x_1 = X_1 c_1$ ($0 \neq c_1 \in \mathbb{R}^k$) of (25) we have*

$$\frac{\| x_1(t) \|}{\| x_1(s) \|} \leq K_1(t)\, e^{-\lambda_1 (s-t)}\ , \quad 0 \leq t \leq s\ ,$$

where $K_1(t) = 2m_1 \left(1 + \dfrac{k_1 k_2\, e^{-(\lambda_1+\lambda_2)t}}{\mathrm{GAP}\Big(\mathcal{R}\big(X_1(0)\big), \mathcal{S}_2(0)\Big)}\right)$ $(m_1 = k_1/d)$.

Proof:
Since X is a consistent fundamental solution we have

$$X_1(t) = \begin{bmatrix} Y_1(t) & Y_2(t) \end{bmatrix} \begin{bmatrix} H_{11} \\ H_{21} \end{bmatrix}\ , \quad t \geq 0\ ,$$

with H_{11} non-singular. Hence,

$$\frac{\| x_1(t) \|}{\| x_1(s) \|} = \frac{\| X_1(t)\, c_1 \|}{\| X_1(s)\, c_1 \|}$$

$$= \frac{\| \begin{bmatrix} \tilde{Q}_1(t) & \tilde{Q}_2(t) \end{bmatrix} \begin{bmatrix} R_{11}(t) H_{11} \\ R_{22}(t) H_{21} \end{bmatrix} c_1 \|}{\| \begin{bmatrix} \tilde{Q}_1(s) & \tilde{Q}_2(s) \end{bmatrix} \begin{bmatrix} R_{11}(s) H_{11} \\ R_{22}(s) H_{21} \end{bmatrix} c_1 \|}$$

$$\leq \| \begin{bmatrix} \tilde{Q}_1(s)\ \tilde{Q}_2(s) \end{bmatrix}^{-1} \| \| \begin{bmatrix} \tilde{Q}_1(t) & \tilde{Q}_2(t) \end{bmatrix} \| \frac{\| \begin{bmatrix} R_{11}(t) H_{11} \\ R_{22}(t) H_{21} \end{bmatrix} c_1 \|}{\| \begin{bmatrix} R_{11}(s) H_{11} \\ R_{22}(s) H_{21} \end{bmatrix} c_1 \|}\ .$$

Moreover, for $0 \leq t \leq s$ we have

$$\max_{0 \neq c_1 \in \mathbb{R}^k} \left(\frac{\| \begin{bmatrix} R_{11}(t) H_{11} \\ R_{22}(t) H_{21} \end{bmatrix} c_1 \|}{\| \begin{bmatrix} R_{11}(s) H_{11} \\ R_{22}(s) H_{21} \end{bmatrix} c_1 \|} \right) \leq$$

$$\leq \| R_{11}(t) {R_{11}}^{-1}(s) \| \sqrt{\frac{1 + \| R_{22}(t)\, H_{21} {H_{11}}^{-1} {R_{11}}^{-1}(t) \|^2}{1 + \mathrm{glb}^2\Big(R_{22}(s)\, H_{21} {H_{11}}^{-1} {R_{11}}^{-1}(s)\Big)}}$$

$$\leq \| R_{11}(t) {R_{11}}^{-1}(s) \| \Big(1 + \| R_{22}(t)\, H_{21} {H_{11}}^{-1} {R_{11}}^{-1}(t) \|\Big)$$

$$\leq k_1\, e^{-\lambda_1 (s-t)} \left(1 + \frac{k_1 k_2\, e^{-(\lambda_1+\lambda_2)t}}{\mathrm{GAP}\Big(\mathcal{R}\big(X_1(0)\big), \mathcal{S}_2(0)\Big)}\right)$$

(cf. (26) and (28)).
By Lemma 1.17 the proof is completed.

■

In Chapter 3 we shall see why this result is important in relation with invariant imbedding techniques.

2.3.2 Finite intervals

The generalization of the consistency concept to *finite intervals* is not straightforward. Consider the homogeneous DE

$$\frac{dx}{dt} = A(t)\,x\ , \qquad t \in [0,1]\ . \tag{29}$$

Now we make the following definition.

Definition 2.23
Let Z be the fundamental solution, corresponding to (29)*, with $Z(0) = I_n$ and let $U\,\Sigma\,V^T$, with $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$, be the SVD of $Z(1)$. The solution space $\mathcal{S}$ has a* γ-observable dichotomy $(\gamma > 1)$ *if for some integer k $(1 \leq k < n)$ we have*

$$\sigma_k = \gamma\,\sigma_{k+1}\ .$$

Let $V = \left[\ \underset{k}{\underset{\longleftrightarrow}{V_1}}\ \ \underset{n-k}{\underset{\longleftrightarrow}{V_2}}\ \right]$*. For $t \in [0,1]$, we define the subspaces*

$$\mathcal{S}_1(t) = \mathcal{R}\Big(Z(t)\,V_1\Big) \quad \textit{and} \quad \mathcal{S}_2(t) = \mathcal{R}\Big(Z(t)\,V_2\Big)\ . \tag{30}$$

Then $\mathcal{S}_1$ is called a dominant subspace *and $\mathcal{S}_2$* a dominated subspace.

■

A relation between dichotomy on a finite and on an infinite interval is given in

Property 2.24
Assume that the homogeneous DE

$$\frac{dx}{dt} = A(t)\,x\ , \qquad t \in [0,\infty) \tag{31}$$

has an exponential dichotomic solution space $\mathcal{S}$ with $\dfrac{e\,\lambda_1 + \lambda_2}{m_1\,m_2} > 1$ *(see Definition 2.7). Then the DE* (29) *has a γ-observable dichotomy, for some* $\gamma \geq \dfrac{e\,\lambda_1 + \lambda_2}{m_1\,m_2}$.

Proof:
Let Z be the fundamental solution of (31) with $Z(0) = I_n$ and let $U\,\Sigma\,V^T$ be the SVD of $Z(1)$. From the exponential dichotomy of the solution space $\mathcal{S}$ we obtain that there exists a projection P such that

$$m_1\, e^{-\lambda_1} \geq \| PZ^{-1}(1) \| = \| PV\Sigma^{-1} \| \ .$$

Since $\dim\Big(\mathcal{R}(P)\Big) = k$ one can show that $\| PV\Sigma^{-1} \| \geq 1/\sigma_k$. Hence, $\sigma_k \geq \dfrac{e^{\lambda_1}}{m_1}$. Similarly we obtain that

$$m_2\, e^{-\lambda_2} \geq \| Z(1)(I_n - P) \| = \| \Sigma V^T(I_n - P) \| \geq \sigma_{k+1} \ .$$

Therefore,

$$\gamma = \frac{\sigma_k}{\sigma_{k+1}} \geq \frac{e^{\lambda_1 + \lambda_2}}{m_1\, m_2} \ .$$

■

If in Definition 2.7 we have the inequalities $m_1 < e^{\lambda_1}$ and $m_2 \leq e^{\lambda_2}$ then the above property implies that the exponential dichotomy of (31) can already be observed at $[0, 1]$.

In the Chapters 3 and 4 we discuss so-called *decoupling transformations.* Such a transformation tries to separate the solutions in $\mathcal{S}_1$ from those in $\mathcal{S}_2$. It will turn out that the stability of such a decoupling transformation is strongly influenced by the dichotomy factor γ. Therefore, in our applications the Definition 2.23 is only appropriate if γ is significantly larger than 1. In Chapter 5 we shall consider singularly perturbed problems, which in general have dichotomies with $\gamma \gg 1$.
Although it is not essential we shall make the following restriction.

Assumption 2.25
If the solution space $\mathcal{S}$ has a γ-observable dichotomy, then the integer k is such that $\sigma_{k+1} \leq 1 < \sigma_k$.

■

Remark 2.26

(i) $\mathrm{GAP}\Big(\mathcal{S}_1(0), \mathcal{S}_2(0)\Big) = \mathrm{GAP}\Big(\mathcal{S}_1(1), \mathcal{S}_2(1)\Big) = 1$.

(ii) $\mathcal{R}(V_2)$ is the solution of the minimization problem

$$\min_{V \subset \mathbb{R}^n} \max_{\substack{v \in V \\ v \neq 0}} \left\{ \frac{\| Z(1)\, v \|}{\| v \|} \;\middle|\; \dim(V) = n - k \right\}.$$

Equivalently, $\mathcal{R}(V_1)$ is the solution of the maximization problem

$$\max_{V \subset \mathbb{R}^n} \min_{\substack{v \in V \\ v \neq 0}} \left\{ \frac{\| Z(1)\, v \|}{\| v \|} \;\middle|\; \dim(V) = k \right\}.$$

(iii) Let X be some fundamental solution, then $U \Sigma V^T$ is also the SVD of the *incremental matrix* $X(1)\, X^{-1}(0)$ (cf. [26]).

■

Definition 2.27

With the definitions of (30) *a fundamental solution* $X = \begin{bmatrix} \underset{k}{\longleftrightarrow} & \underset{n-k}{\longleftrightarrow} \\ X_1 & X_2 \end{bmatrix}$ *is called*

consistent at $t = 0$ *if* $\mathrm{GAP}\Big(\mathcal{R}\Big(X_1(0)\Big), \mathcal{S}_2(0)\Big) > 0$.

Equivalently, X *is called* consistent at $t = 1$ *if* $\mathrm{GAP}\Big(\mathcal{R}\Big(X_2(1)\Big), \mathcal{S}_1(1)\Big) > 0$.

■

Similarly to Theorem 2.19 we may now formulate

Theorem 2.28

Assume the solution space $\mathcal{S}$ *has a* γ*-observable dichotomy* $(\gamma > 1)$. *Write* $\mathcal{S} = \mathcal{S}_1 \oplus \mathcal{S}_2$ *as in* (30). *Define a fundamental solution* X *by* $X_1(0) = V \begin{bmatrix} H_{11} \\ H_{21} \end{bmatrix}$ *(with* $\mathrm{rank}\Big(\begin{bmatrix} H_{11} \\ H_{21} \end{bmatrix}\Big) = k$*) and* $X_2(1) = U \begin{bmatrix} H_{12} \\ H_{22} \end{bmatrix}$ *(with* $\mathrm{rank}\Big(\begin{bmatrix} H_{12} \\ H_{22} \end{bmatrix}\Big) = n-k$*).*
Then

(i) *if* H_{11} *is singular* *then* $\mathrm{DIST}\Big(\mathcal{R}\Big(X_1(1)\Big), \mathcal{S}_1(1)\Big) = 1$
else $\mathrm{DIST}\Big(\mathcal{R}\Big(X_1(1)\Big), \mathcal{S}_1(1)\Big) \leq \dfrac{\| H_{21} {H_{11}}^{-1} \|}{\gamma}$.

(ii) *if* H_{22} *is singular* *then* $\mathrm{DIST}\Big(\mathcal{R}\Big(X_2(0)\Big), \mathcal{S}_2(0)\Big) = 1$
else $\mathrm{DIST}\Big(\mathcal{R}\Big(X_2(0)\Big), \mathcal{S}_2(0)\Big) \leq \dfrac{\| H_{12} {H_{22}}^{-1} \|}{\gamma}$.

Proof:

(i) Note that, with $\Sigma = \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}$, we obtain

$$X_1(1) = U\,\Sigma \begin{bmatrix} H_{11} \\ H_{21} \end{bmatrix} = U_1\Sigma_{11}H_{11} + U_2\Sigma_{22}H_{21}\ .$$

If H_{11} is non-singular, then

$$\begin{aligned} \mathrm{DIST}\Big(\mathcal{R}\big(X_1(1)\big), \mathcal{S}_1(1)\Big) &= \mathrm{DIST}\Big(X_1(1), U_1\Big) \\ &= \frac{\| \Sigma_{22}\, H_{21}\, {H_{11}}^{-1}\, {\Sigma_{11}}^{-1} \|}{\sqrt{1 + \| \Sigma_{22}\, H_{21}\, {H_{11}}^{-1}\, {\Sigma_{11}}^{-1} \|^2}} \\ &\leq \frac{\sigma_{k+1}}{\sigma_k} \| H_{21}\, {H_{11}}^{-1} \| = \frac{\| H_{21}\, {H_{11}}^{-1} \|}{\gamma}\ . \end{aligned}$$

If H_{11} is singular the result is obtained by Property 1.15 and the observation that $\mathrm{GAP}\Big(\mathcal{R}\big(X_1(1)\big), \mathcal{S}_2(1)\Big) = 0$.

(ii) Since $X_2(1) = V_1\,{\Sigma_{11}}^{-1}\,H_{12} + V_2\,{\Sigma_{22}}^{-1}\,H_{22}$ the proof goes similarly to (i).

■

Note that $\| H_{21}\, {H_{11}}^{-1} \|$ is directly related to $\mathrm{GAP}\Big(\mathcal{R}\big(X_1(0)\big), \mathcal{S}_2(0)\Big)$: $\mathrm{GAP}\Big(\mathcal{R}\big(X_1(0)\big), \mathcal{S}_2(0)\Big) = 1/\sqrt{1 + \| H_{21}\, {H_{11}}^{-1} \|^2}$ (cf. Property 1.15). Similarly we have that $\mathrm{GAP}\Big(\mathcal{R}\big(X_2(1)\big), \mathcal{S}_1(1)\Big) = 1/\sqrt{1 + \| H_{12}\, {H_{22}}^{-1} \|^2}$.

Theorem 2.28 tells us something about the direction of solutions of the DE. Like in Theorem 2.22 also results about the growth of solutions can be derived. Again let X be a fundamental solution with $X_1(0) = V \begin{bmatrix} H_{11} \\ H_{21} \end{bmatrix}$ and $X_2(1) = U \begin{bmatrix} H_{21} \\ H_{22} \end{bmatrix}$. Assume X is consistent at both $t = 0$ and $t = 1$. Then, for $x_1 = X_1\, c_1$ ($c_1 \neq 0$), we have

$$\frac{\| x_1(1) \|^2}{\| x_1(0) \|^2} = \frac{\| U\,\Sigma \begin{bmatrix} H_{11} \\ H_{21} \end{bmatrix} c_1 \|^2}{\| V \begin{bmatrix} H_{11} \\ H_{21} \end{bmatrix} c_1 \|^2} = \frac{\| \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix} \begin{bmatrix} H_{11} \\ H_{21} \end{bmatrix} c_1 \|^2}{\| \begin{bmatrix} H_{11} \\ H_{21} \end{bmatrix} c_1 \|^2}$$

$$= \frac{\dfrac{\| \Sigma_{11}\, d_1 \|^2}{\| d_1 \|^2} + \dfrac{\| \Sigma_{22}\, H_{21}\, {H_{11}}^{-1}\, {\Sigma_{11}}^{-1}\, d_1 \|^2}{\| d_1 \|^2}}{1 + \dfrac{\| H_{21}\, {H_{11}}^{-1}\, d_1 \|^2}{\| d_1 \|^2}} \qquad (d_1 = H_{11}\, c_1)$$

$$\geq \frac{{\sigma_k}^2}{1 + \| H_{21} {H_{11}}^{-1} \|^2} = {\sigma_k}^2 \, \mathrm{GAP}^2\Big(\mathcal{R}\Big(X_1(0)\Big), \mathcal{S}_2(0)\Big) \, .$$

So,

$$\frac{\| x_1(0) \|}{\| x_1(1) \|} \leq \frac{1}{\sigma_k \, \mathrm{GAP}\Big(\mathcal{R}\Big(X_1(0)\Big), \mathcal{S}_2(0)\Big)} \, . \tag{32}$$

Equivalently, for $x_2 = X_2 \, c_2$ ($c_2 \neq 0$), we may derive

$$\frac{\| x_2(1) \|}{\| x_2(0) \|} \leq \frac{\sigma_{k+1}}{\mathrm{GAP}\Big(\mathcal{R}\Big(X_2(1)\Big), \mathcal{S}_1(1)\Big)} \, . \tag{33}$$

The splitting of the solution space $\mathcal{S}$ is based on the values of the fundamental solution Z at the boundary points. Hence, we can only say something about a dichotomy at these boundary points. The concept is therefore not directly applicable for the whole interval; it does not guarantee a different growth behaviour of solutions all over the interval. However, we do know that the larger the factor γ, the better the situation of infinite intervals is approximated (cf. Property 2.24).

2.4 Conditioning

In this section we consider the finite interval $[0, 1]$. We shall examine the relation between the stability constants (cf.(6) and (7))

$$\begin{aligned} \beta &= \max_{t \in [0,1]} \| X(t) \, B(X)^{-1} \| \\ &= \max_{t \in [0,1]} \| X(t) \Big[B^0 \, X(0) + B^1 \, X(1) \Big]^{-1} \| \end{aligned} \tag{34}$$

and

$$\alpha = \max_{t,s \in [0,1]} \| G(t, s) \| \tag{35}$$

with (exponential) dichotomy of the solution space $\mathcal{S}$. How closely these quantities are related has already been explained by de Hoog and Mattheij ([26]). Here we shall derive similar results using our own normalizations and definitions.

A relation between α and β is given by

Theorem 2.29
With the definitions (34) *and* (35) *we obtain the relation*

$$\beta \leq \sqrt{2}\,\alpha\,.$$

Proof:
For the stability constant α we have

$$\alpha = \max_{t,s\,\epsilon\,[0,1]} \| G(t,s) \| \geq \max_{t\,\epsilon\,[0,1]} \| G(t,0) \| = \max_{t\,\epsilon\,[0,1]} \| X(t)\,\mathcal{B}(X)^{-1}\,B^0 \| \;.$$

Equivalently, $\alpha \geq \max_{t\,\epsilon\,[0,1]} \| X(t)\,\mathcal{B}(X)^{-1}\,B^1 \|$.

Hence, by the row orthogonality of $\left[B^0 \mid B^1 \right]$ (Assumption 2.3) we have, for all $t\,\epsilon\,[0,1]$,

$$\| X(t)\,\mathcal{B}(X)^{-1} \| = \| X(t)\,\mathcal{B}(X)^{-1} \left[B^0 \mid B^1 \right] \|$$

$$= \| \left[X(t)\,\mathcal{B}(X)^{-1}\,B^0 \mid X(t)\,\mathcal{B}(X)^{-1}\,B^1 \right] \| \leq \sqrt{2}\,\alpha$$

by Lemma 1.17.

■

Now we first assume that the BCs are *separated*, i.e.,

$$B^0 = \begin{bmatrix} 0 \\ B^{02} \end{bmatrix} \begin{matrix} \updownarrow m \\ \updownarrow n-m \end{matrix} \quad \text{and} \quad B^1 = \begin{bmatrix} B^{11} \\ 0 \end{bmatrix} \begin{matrix} \updownarrow m \\ \updownarrow n-m \end{matrix} \;. \tag{36}$$
$$\qquad\qquad \underset{n}{\longleftrightarrow} \qquad\qquad\qquad\quad \underset{n}{\longleftrightarrow}$$

The value of the integer m is at this moment unspecified. Define a fundamental solution $X = \underset{m}{\underset{\longleftrightarrow}{X_1}} \; \underset{n-m}{\underset{\longleftrightarrow}{X_2}}$ by

$$\text{(i)} \quad B^{02} X(0) = \left[0 \;\; I_{n-m} \right], \tag{37a}$$

$$\text{(ii)} \quad B^{11} X(1) = \left[I_m \;\; 0 \right]. \tag{37b}$$

Observe that by the well-posedness of the problem and Assumption 2.3 these conditions are not contradicting and the matrix function X is a uniquely defined fundamental solution, with $X_2(0)$ and $X_1(1)$ column orthogonal.

Since $B^0\,X(0) + B^1\,X(1) = I_n$, the Green's function G has the simple expression

$$G(t,s) = \begin{cases} X(t) \begin{bmatrix} 0 & 0 \\ 0 & I_{n-m} \end{bmatrix} X^{-1}(s) & , \; t \geq s \\ -X(t) \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix} X^{-1}(s) & , \; t < s \end{cases} \;. \tag{38}$$

As in Lemma 2.12 we now have

Property 2.30

$$\mathrm{GAP}\Big(X_1(t), X_2(t)\Big) \;\geq\; \alpha^{-1}\,, \quad \textit{for all } t \in [0,1].$$

■

Moreover, for all non-trivial $c_1 \in \mathbb{R}^m$,

$$\frac{\| X_1(t)\, c_1 \|}{\| c_1 \|} \;=\; \frac{\| X_1(t)\, c_1 \|}{\| X_1(1)\, c_1 \|} \;=\; \frac{\| \begin{bmatrix} X_1(t) & 0 \end{bmatrix} \begin{pmatrix} c_1 \\ 0 \end{pmatrix} \|}{\| \begin{bmatrix} X_1(1) & X_2(1) \end{bmatrix} \begin{pmatrix} c_1 \\ 0 \end{pmatrix} \|}$$

$$\leq\; \| X(t) \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix} X^{-1}(1) \| \leq\; \alpha\,. \tag{39}$$

Similarly, for all non-trivial $c_2 \in \mathbb{R}^{n-m}$,

$$\frac{\| X_2(t)\, c_2 \|}{\| c_2 \|} \;\leq\; \alpha\,. \tag{40}$$

To obtain a relation between dichotomy, consistency and conditioning we again look at the SVD of $Z(1)$, where Z is the fundamental solution starting with the identity. So, $Z(1) = U\,\Sigma\,V^T$, where $U = \Big[u_1 | \cdots | u_n\Big]$ and $V = \Big[v_1 | \cdots | v_n\Big]$ are orthogonal matrices and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$. Define l_1 as the largest index such that $\sigma_{l_1} > \alpha \geq 1$ and l_2 the smallest index such that $\sigma_{l_2+1} < 1/\alpha \leq 1$. From (40) we obtain that solutions in X_2 (satisfying the homogeneous BCs at $t = 1$) grow at most a factor α. Hence,

$$\mathrm{GAP}\Big(X_2(1), \Big[u_1 | \cdots | u_{l_1}\Big]\Big) > 0\,. \tag{41}$$

Similar arguments show that for $X_1(0)$ defined by (37a) we have

$$\mathrm{GAP}\Big(X_1(0), \Big[v_{l_2+1} | \cdots | v_n\Big]\Big) > 0\,. \tag{42}$$

Definition 2.31

Consider a BVP with a given stability constant α. Let Z be the corresponding fundamental solution with $Z(0) = I_n$. Let $U\,\Sigma\,V^T$ be the SVD of $Z(1)$. Define l_1 as the largest index such that $\sigma_{l_1} > \alpha \geq 1$ and l_2 the smallest index such that $\sigma_{l_2+1} < 1/\alpha \leq 1$. If $l_1 > 0$, then $\mathcal{R}\Big(Z(t)\Big[v_1 | \cdots | v_{l_1}\Big]\Big)$ is called the subspace of fast increasing modes.

Correspondingly, if $l_2 < n$, then $\mathcal{R}\Big(Z(t)\Big[v_{l_2+1} | \cdots | v_n\Big]\Big)$ is called the sub-

space of fast decaying modes.
Finally, if $l_1 \neq l_2$ *then a solution in* $\mathcal{R}\Big(Z(t)\Big[v_{l_1+1}|\cdots|v_{l_2}\Big]\Big)$ *is called a* smooth solution.

■

From (41) it follows that

$$B^{11}\Big[u_1|\cdots|u_{l_1}\Big]c = 0 \;\Rightarrow\; c = 0\ .$$

Hence, we may say that the fast increasing modes are controled by the BCs at $t = 1$. Similarly, the fast decaying modes are controled by the BCs at $t = 0$. The factor of smooth solutions which is contained in the solution x of (1), subject to (36), may be determined on either side.

Now let the solution space $\mathcal{S}$ have a γ-observable dichotomy, i.e., there exist a constant $\gamma > 1$ and an index k such that $\gamma = \sigma_k/\sigma_{k+1}$ and $\sigma_{k+1} \leq 1 < \sigma_k$. The properties (41) and (42) do imply consistency at one of both ends for the fundamental solution X, defined by (37a,b) (cf. Definition 2.23), as soon as $k = l_1$ or $k = l_2$.

Property 2.32
If $\gamma > \alpha^2$ *then* $l_1 \leq k \leq l_2$ *with at least one equality.*

Proof:
Since $\sigma_{l_2+1} < 1/\alpha \leq 1 \leq \alpha < \sigma_{l_1}$ and $\sigma_{k+1} \leq 1 < \sigma_k$ we know that $l_1 \leq k \leq l_2$. If $l_2 - l_1 \leq 1$ then we are ready. Now assume $l_1 < k < l_2$. Then $\gamma = \sigma_k/\sigma_{k+1} \leq \alpha^2$, which is in contradiction with the assumption.

■

Hence, if $\gamma > \alpha^2$ then the separation of the solution subspace $\mathcal{S}$, induced by the dichotomy, coincides with our definitions of fast and smooth solutions; the smooth solutions are grouped together with the fast decaying solutions ($k = l_1$) or with the fast increasing solutions ($k = l_2$). Since our orientation is from left to right we shall assume, without loss of generality, that in such case $k = l_1$.
If $l_1 = l_2$ (implying $m = k$) then there are no non-trivial smooth solutions of the homogeneous part of (1) and we have consistency of the fundamental solution X at both ends. Hence, no decaying solutions are present in X_1 and no increasing solutions in X_2.

In general, however, $l_1 \neq l_2$ and m may be different from k. In that case smooth solutions will be present in both X_1 and X_2. Although not dramatic (cf. Property 2.30) this is a less desirable situation for continuous decoupling

transformations, as will be discussed in the Chapters 3 and 4, since the stability of such methods depends on the ratio σ_m/σ_{m+1}.

If the BCs are *non-separated*, then they do not induce a consistent fundamental solution X in a straightforward way. However, one can always construct separated BCs which define the same solution x and which cause just a small perturbation of the stability constants ([26]). If the solution space $\mathcal{S}$ has a γ-observable dichotomy, then these separated BCs can be written as

$$V_2{}^T x(0) = \tilde{b}_2 \quad \text{and} \quad U_1{}^T x(1) = \tilde{b}_1 \ , \tag{43}$$

where V_2 and U_1 are given in Definition 2.23 and $\tilde{b} = \begin{pmatrix} \tilde{b}_1 \\ \tilde{b}_2 \end{pmatrix}$ has been chosen properly.

Denoting the stability constants corresponding to the DE (1) and the BCs (43) by $\tilde{\alpha}$ and $\tilde{\beta}$ (see (34) and (35)), then one can, for instance, show that $\tilde{\beta} \leq \sqrt{2}\,\beta$. Now, with (α, β) replaced by $(\tilde{\alpha}, \tilde{\beta})$ we can use the earlier obtained results for separated BCs. We have to realize, however, that the derivation of (43) is not constructive, in the sense that it does not yield a numerically stable algorithm for the computation of a consistent fundamental solution X. For a discussion of how the initial values of X may be chosen in order to obtain consistency we refer to Section 3.4.

Chapter 3

Decoupling methods

3.1 Introduction

In Section 2.1 we have already seen that any solution of (2.1) can be expressed as a linear combination of the columns of a fundamental solution X and a particular solution p. Assume that the solution space $\mathcal{S}$ is exponentially dichotomic. Hence, there exist a k-dimensional dominant subspace $\mathcal{S}_1$ and an $(n-k)$-dimensional dominated subspace $\mathcal{S}_2$. In general, each column of $X(0)$ will contain a non-trivial part of modes from $\mathcal{S}_1$, unless $X(0)$ has been chosen very specifically. This implies that, possibly after some initial effect, all columns of X will grow correspondingly to the fastest modes. The directions of these modes will be in $\mathcal{S}_1$ (cf. Theorem 2.19). Therefore, the numerical independence of the columns of X will be destroyed. In this situation we cannot expect to find by superposition an accurate approximation for the solution of a BVP.

This difficulty can be overcome by the use of a so-called *multiple shooting method* ([14],[29],[42],[50]). To reduce the effect of dominant growth behaviour of solutions the interval $[0,1]$ is divided into m subintervals $[t_i, t_{i+1}]$ $(i = 0,\ldots,m-1)$, where

$$0 = t_0 < t_1 < \cdots < t_{m-1} < t_m = 1\,. \tag{1}$$

For ease of notation we define $t_{m+1} = 1$. On each interval $[t_i, t_{i+1}]$ $(i = 0,\ldots,m)$ a fundamental solution X^i and a particular solution p^i are computed. This implies that for any solution x of

$$\frac{dx}{dt} = A(t)\,x + f(t)\,, \qquad t \in [0,1]\,, \tag{2}$$

there exists a sequence of vectors $\{ c^i \}_{i=0}^m$ such that

$$x(t) = X^i(t)\,c^i + p^i(t)\,, \qquad t \in [t_i, t_{i+1}]\,, \quad (i = 0, \ldots, m)\,.$$

Continuity of the solutions at the nodes requires that

$$X^{i-1}(t_i)\,c^{i-1} + p^{i-1}(t_i) = x(t_i) = X^i(t_i)\,c^i + p^i(t_i)\,, \quad (i = 1, \ldots, m)\,. \tag{3}$$

Together with the transformed BCs

$$B^0\,X^0(t_0)\,c^0 + B^1\,X^m(t_m)\,c^m = b - B^0\,p^0(t_0) - B^1\,p^m(t_m)\,, \tag{4}$$

these relations lead to the multiple shooting system

$$M\,\tilde{c} = \tilde{d}\,, \tag{5}$$

where $\left[M \,|\, \tilde{c} \,|\, \tilde{d} \right] =$

$$\left[\begin{array}{ccccc|c|c}
B^0X^0(t_0) & & & & B^1X^m(t_m) & c^0 & b - B^0p^0(t_0) - B^1p^m(t_m) \\
-X^0(t_1) & X^1(t_1) & & & & c^1 & p^0(t_1) - p^1(t_1) \\
 & -X^1(t_2) & X^2(t_2) & & & \vdots & \vdots \\
 & & \ddots & \ddots & & \vdots & \vdots \\
 & & & -X^{m-1}(t_m) & X^m(t_m) & c^m & p^{m-1}(t_m) - p^m(t_m)
\end{array}\right] .$$

The performance of a multiple shooting method depends on
- the way (and the accuracy) in which these X^i and p^i are computed,
- the choice of the initial values $X^i(t_i)$ and $p^i(t_i)$,
- the strategy that determines the *shooting points* t_i,

and, last but not least,
- the linear solver for (5).

If the solution manifold of (2) contains mildly varying modes only, then an explicit Runge-Kutta method may be used to approximate $X^{i-1}(t_i)$. The matrix $X^i(t_i)$ is often chosen such that it is a well-conditioned matrix, whose first k columns give an accurate approximation of $\mathcal{S}_1$. These properties can be obtained by the construction of the QR-decomposition of $X^{i-1}(t_i)\,(i = 1, \ldots, m)$ ([39],[54]). Let

$$X^{i-1}(t_i) = Q^i R^i = \underbrace{\begin{bmatrix} Q_1{}^i & Q_2{}^i \end{bmatrix}}_{\begin{matrix} \longleftrightarrow \\ k \end{matrix} \quad \begin{matrix} \longleftrightarrow \\ n-k \end{matrix}} \begin{bmatrix} R_{11}{}^i & R_{12}{}^i \\ 0 & R_{22}{}^i \end{bmatrix}, \tag{6}$$

with Q^i orthogonal and R^i (block) upper triangular. Take Q^i as the initial value $X^i(t_i)\,(i = 1,\ldots,m)$. Then, by Theorem 2.19, $\mathcal{R}(Q_1{}^i)$ will approximate $\mathcal{S}_1$ quite accurately as soon as $X^0(t_0)$ has been chosen correctly (see Section 3.4). The growth of solutions within $\mathcal{R}(Q_1{}^i)$ over the interval $[t_{i-1}, t_i]$ is given by $R_{11}{}^i$. Moreover, as is shown by geometrical arguments in [42], $R_{22}{}^i$ indicates the growth of dominated solutions over the interval $[t_{i-1}, t_i]$.

The shooting points are chosen adaptively such that $\| X^i(t) \|$ is bounded by some fixed constant, for all $t \epsilon [t_i, t_{i+1}]$. This criterion implies that the QR-decomposition of (6) yields the information about direction and growth of solutions within a prescribed tolerance. Finally (5) is solved by a forward and a backward sweep (decaying solutions in a forward direction and increasing solutions backward ([39]). During this backward sweep we implicitly find the direction of the dominated subspace. It can be shown ([33],[42]) that by such a multiple shooting method the solution of a well-conditioned BVP with only mildly varying solutions is obtained in a numerically stable way.

The algorithm sketched above can be simplified if the BCs are separated. Assume

$$B^0 = \begin{bmatrix} 0 \\ B^{02} \end{bmatrix} \begin{matrix} \updownarrow k \\ \updownarrow n-k \end{matrix}, \quad B^1 = \begin{bmatrix} B^{11} \\ 0 \end{bmatrix} \begin{matrix} \updownarrow k \\ \updownarrow n-k \end{matrix} \text{ and } b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \begin{matrix} \updownarrow k \\ \updownarrow n-k \end{matrix} \tag{7}$$

and $X(t) = \underbrace{\begin{bmatrix} X_1(t) & X_2(t) \end{bmatrix}}_{\begin{matrix} \longleftrightarrow \\ k \end{matrix} \quad \begin{matrix} \longleftrightarrow \\ n-k \end{matrix}}$. If the initial values $X^0(t_0)$ and $p^0(t_0)$ are chosen such that

$$B^{02} \left[X_1{}^0(t_0) \mid p^0(t_0) \right] = \underbrace{[\, 0 \mid b_2 \,]}_{\begin{matrix} \longleftrightarrow \\ k \end{matrix} \quad \begin{matrix} \leftrightarrow \\ 1 \end{matrix}}, \tag{8}$$

then the k-dimensional solution manifold determined by the BCs at $t = 0$ is spanned by p^0 and the columns of $X_1{}^0$. Similarly to (6) the columns of $X_1{}^{i-1}\,(i = 1,\ldots,m)$ can be orthogonalized at the shooting points t_i. This implies that we only need to compute particular solutions p^i and the first k columns of each fundamental solution X^i, whose span does not contain dominated solutions (cf. (2.42)). This method is called the *Godunov-Conte* or *sta-*

bilized march algorithm and is proven to be numerically stable too ([50]).

Observe that any multiple shooting algorithm, using an explicit integration routine, will be inefficient if there are very rapidly varying solutions, since then the number of integration steps needed to compute $X^{i-1}(t_i)$, will become prohibitively large. Such so-called *stiff problems* (see Chapter 5) are not just artificial ([12],[48]). Therefore we shall look at generalizations of multiple shooting algorithms that take advantage of such rapidly varying solutions. It will appear that these generalizations decompose the fundamental solutions X^i continuously throughout the subinterval $[t_i, t_{i+1}]$ $(i = 0, \ldots, m-1)$. As we shall see this implies that the directions of solutions are computed independently of their growth behaviour.

3.2 Continuous decoupling

3.2.1 General description

In [39] it is shown that by the construction of the QR-decompositions (6) a decoupling is obtained of the dominant and dominated subspaces of the homogeneous part of

$$\frac{dx}{dt} = A(t)\,x + f(t)\,, \qquad t \in [0,1]\,, \tag{9}$$

at the shooting points t_i $(i = 1, \ldots, m)$. This decoupling yields the direction of the dominant subspace and (separately) the growth of solutions in both the dominant and the dominated subspace. In this section we try to perform such a decoupling in a continuous way (cf. [42]).

Let X be a fundamental solution corresponding to (9) which is to be continuously *decomposed* in

$$X = TY\,, \tag{10}$$

where the $n \times n$ matrix function T (and consequently also Y) is non-singular, for all t. The matrix function T has to represent the directions of solutions, while the matrix function Y has to indicate the growth behaviour of the various solutions. In [4] the computation of matrix functions T and Y that satisfy (10) is called a *factorization method.*
From (10) we deduce

$$A\,T\,Y = A\,X = \frac{dX}{dt} = \frac{dT}{dt}\,Y + T\,\frac{dY}{dt}\ . \tag{11}$$

Let $\tilde{A}$ be some $n \times n$ matrix function, generally depending on A and T and to be specified later on, such that

$$\frac{dY}{dt} = \tilde{A}(T,t)\,Y\ , \qquad t \epsilon [0,1]\ . \tag{12}$$

Then (11) yields that T has to satisfy the, generally non-linear, *Lyapunov equation*

$$\frac{dT}{dt} = A(t)\,T - T\,\tilde{A}(T,t)\ , \qquad t \epsilon [0,1]\ . \tag{13}$$

Hence, for any $n \times n$ matrix function $\tilde{A}$ and initial values $T(0)$ and $Y(0)$ such that $X(0) = T(0)\,Y(0)$, a decomposition of the form (10) is obtained as soon as T and Y satisfy the DEs (13) and (12).

The equation (13) can also be viewed as the relation between $\tilde{A}$ and T that results if (9) is *transformed* into

$$\frac{dy}{dt} = \tilde{A}(T,t)\,y + \tilde{f}(t)\ , \qquad t \epsilon [0,1]\ , \tag{14a}$$

by the substitutions

$$x(t) = T(t)\,y(t) \quad \text{and} \quad f(t) = T(t)\,\tilde{f}(t)\ . \tag{14b}$$

Then (9) and (14a,b) imply that

$$\tilde{A}(T,t) = T^{-1}(t)\,A(t)\,T(t) - T^{-1}(t)\,\frac{dT}{dt}(t)\ , \tag{14c}$$

which is equivalent with (13), and the $n \times n$ matrix function Y, defined by (12), is nothing but a fundamental solution of the transformed system (14a).

Observe that in the decomposition formulation we choose $\tilde{A}$, whereafter T and Y follow from (13) and (12). In the transformation formulation we start with some T, from which $\tilde{A}$ and Y are obtained. Practical algorithms are often a combination of the two. To obtain special structures for T and Y (like (block) upper/lower triangular or (column) orthogonal) we put some requirements on both $\tilde{A}$ and T. In principle one can say that (13) consists of n^2 equations for $2n^2$ variables (the elements of $\tilde{A}$ and T), so that n^2 degrees of freedom are left.

For a decoupling of direction and growth the decomposition (10) must be such that Y is (block) upper triangular and T well-conditioned and properly scaled,

uniformly in t (cf. [42]). Hence, $\tilde{A}$ has to be (block) upper triangular, for all t, and (14a) is a partially decoupled system, obtained by transformation of (9). For the sake of convenience we shall call the system (14a) *decoupled* if $\tilde{A}$ is (block) upper triangular.

For the form of the matrix function T in (13) we have two genuine possibilities: orthogonal, leading to a QR-decomposition, or lower triangular, leading to an LU-decomposition. The best known member of the latter kind is the so-called *Riccati transformation*, which will be discussed in the next chapter. In this section we shall concentrate mostly on the former group. A solution method for a linear BVP based on the computation of the decomposition (10) with orthogonal T is called *continuous orthonormalization.*

From now on we shall assume that $\tilde{A}$ depends on T in a continuous way and that $\tilde{A}$ is block upper triangular, i.e.,

$$\tilde{A} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ \underset{k}{0} & \underset{n-k}{\tilde{A}_{22}} \end{bmatrix} \begin{matrix} \updownarrow k \\ \updownarrow n-k \end{matrix} , \tag{15}$$

where the dimension k is not yet specified (actually, if possible, we shall choose k equal to the dimension of the dominant subspace). Observe that (15) implies that Y can be taken block upper triangular (if Y is partitioned like $\tilde{A}$, then $Y_{21}(0) = 0$ implies that $Y_{21} \equiv 0$). From now on we shall assume that Y is block upper triangular with the same partitioning as $\tilde{A}$. Let $T = \begin{bmatrix} \underset{k}{T_1} & \underset{n-k}{T_2} \end{bmatrix}$. Then

from (12), (13) and (15) it follows that one may obtain differential equations for T_1 and Y_{11} that are decoupled from those for T_2, Y_{12} and Y_{22}, namely

$$\frac{d}{dt}T_1 = A(t)\,T_1 - T_1\,\tilde{A}_{11}(T_1, t) , \qquad t \in [0,1] , \tag{16}$$

$$\frac{d}{dt}Y_{11} = \tilde{A}_{11}(T_1, t)\,Y_{11} , \qquad t \in [0,1] . \tag{17}$$

Let $X = \begin{bmatrix} \underset{k}{X_1} & \underset{n-k}{X_2} \end{bmatrix}$ be a given fundamental solution. Define the matrix functions T_1 and Y_{11} by (16) and (17), respectively. If the initial values have been chosen such that $X_1(0) = T_1(0)\,Y_{11}(0)$, then $X_1 = T_1\,Y_{11}$, from which we obtain that $\mathcal{R}\Big(T_1(t)\Big) = \mathcal{R}\Big(X_1(t)\Big)$, for all t, independently of the choice of $\tilde{A}_{11}$. Hence, if $\tilde{A}_{11}$ is chosen such that T_1 is well-conditioned and properly

scaled, then the direction of X_1 is represented by T_1 and its growth by Y_{11}. This result can be formalized as follows:

Property 3.1
Let X_1 and T_1 be $n \times k$ matrix functions satisfying, respectively, the DEs

$$\frac{d}{dt}X_1 = A(t)\,X_1\ , \qquad t\epsilon[0,1]\ , \tag{18a}$$

$$\frac{d}{dt}T_1 = A(t)\,T_1 - T_1\,\tilde{A}_{11}(T_1,t)\ , \qquad t\epsilon[0,1]\ , \tag{18b}$$

where $\tilde{A}_{11}$ is a continuous $k \times k$ matrix function, depending on A and T_1, such that T_1 exists on $[0,1]$. If $\mathcal{R}\Big(T_1(0)\Big) = \mathcal{R}\Big(X_1(0)\Big)$ then $\mathcal{R}\Big(T_1(t)\Big) = \mathcal{R}\Big(X_1(t)\Big)$, for all $t\epsilon[0,1]$. Moreover, if $T_1(0)$ has full rank, then $T_1(t)$ has full rank, for all t.

■

Remark 3.2
If the solution space $\mathcal{S}$ is exponentially dichotomic then we obtain, from Theorem 2.19, that the subspace $\mathcal{R}\Big(T_1(t)\Big)$ will asymptotically be equal to the dominant subspace as soon as $\mathcal{S}_2(0) \cap \mathcal{R}\Big(T_1(0)\Big) = \{\,0\,\}$ (see also Theorem 3.5).

■

3.2.2 Possible choices

Now we shall discuss various possibilities for the choice of $\tilde{A}_{11}$. For notational convenience we suppress the dependence of $\tilde{A}_{11}$ on T_1.

Our main goal is to choose $\tilde{A}_{11}$ such that

(i) $T_1(t)$ is uniformly well-conditioned

(ii) the DE (16) for T_1 is stable

(iii) the rotational activity of T_1 is as small as possible (see Definition 1.19).

The requirements (i) and (ii) are somewhat related, as will be shown in Theorem 3.5. If the DE (16) is solved by an automatic integration routine, then the third requirement hopefully maximizes the stepsizes that are taken.
It can be shown that the choice $\tilde{A}_{11} = \mathrm{diag}({T_1}^T A\,T_1)$ leads to a matrix function T_1 with all columns having unit lengths. In that case $\| \frac{d}{dt}T_1 \|$ is the

rotational activity of T_1 and, at the same time, the rotational activity of X_1. However, this choice will, in general, be in conflict with all three of the above requirements. Therefore we look for better choices of $\tilde{A}_{11}$.

To obtain an idea how the first requirement can be fulfilled, we make the following observation. Let T_1 satisfy the DE (16). Then

$$\begin{aligned}\frac{d}{dt}(I_k - T_1^T T_1) &= -\Big(\frac{dT_1}{dt}\Big)^T T_1 - T_1{}^T\Big(\frac{dT_1}{dt}\Big) \\ &= -T_1{}^T(A + A^T)T_1 + \tilde{A}_{11}^T T_1{}^T T_1 + T_1{}^T T_1 \tilde{A}_{11} \\ &= \tilde{A}_{11} + \tilde{A}_{11}^T - T_1{}^T(A + A^T)T_1 \qquad (19) \\ &\qquad -(I_k - T_1^T T_1)\tilde{A}_{11} - \tilde{A}_{11}^T(I_k - T_1^T T_1)\,.\end{aligned}$$

From this relation we obtain

Property 3.3
The matrix function T_1 is column orthogonal, for all t, if and only if $T_1(0)$ is column orthogonal and

$$\text{symm}(\tilde{A}_{11}) = \text{symm}(T_1{}^T A T_1)\,. \qquad (20)$$

Proof:
If (20) is satisfied and $T_1(0)$ is column orthogonal, then with $Z_{11} = I_k - T_1{}^T T_1$ the DE (19) reduces to the linear DE

$$\frac{d}{dt}Z_{11} = -Z_{11}\tilde{A}_{11}(t) - \tilde{A}_{11}^T(t)\,Z_{11}\,. \qquad (21)$$

Hence, $Z_{11}(0) = 0$ implies $Z_{11} \equiv 0$, which proves the 'if'-part.
On the other hand, if $T_1(t)$ is column orthogonal, for all t, then (19) reduces to

$$0 = (\tilde{A}_{11} + \tilde{A}_{11}^T) - T_1{}^T(A + A^T)T_1\,,$$

which is identical to (20).

■

This property is valid only for exact computations. If we want to have a similar property for the numerical approximation of T_1 we need something like asymptotic stability for (19). This is formalized in

Theorem 3.4
Consider the DE

$$\frac{dT_1}{dt} = A(t)\,T_1 - T_1\,\tilde{A}_{11}(t)\ , \qquad t \geq 0\ , \tag{22}$$

where the $k \times k$ matrix function $\tilde{A}_{11}$ satisfies

- $\text{symm}(\,\tilde{A}_{11}(t)\,) = \text{symm}\Big(\,(\,T_1{}^T A\,T_1\,)(t)\,\Big)$ (23a)
- $\mu\Big(-\tilde{A}_{11}(t)\Big) \leq -\alpha_1 < 0$ (23b)
- $\|\,\tilde{A}_{11}(t) - \text{symm}\Big(\tilde{A}_{11}(t)\Big)\,\|$ *is bounded, uniformly in t .* (23c)

Then T_1 exists on $[\,0,\infty\,)$, for any value of $T_1(0)$. Moreover,

$$\lim_{t\to\infty} T_1{}^T(t)\,T_1(t) = I_k\ ,$$

(i.e., T_1 is column orthogonal at ∞).

Proof:
Since $\text{symm}(\,\tilde{A}_{11}\,) = \text{symm}(\,T_1{}^T A\,T_1\,)$, Property 3.3 asserts that $Z_{11} = I_k - T_1{}^T T_1$ satisfies the DE (21). Define the $k \times k$ matrix function R_{11} by

$$\left\{\begin{array}{rcl} \dfrac{d}{dt}R_{11} & = & -R_{11}\,\tilde{A}_{11}(t)\ , \qquad t \geq 0 \\ R_{11}(0) & = & I_k\ , \end{array}\right. \tag{24}$$

Then one verifies directly that, for all $t \geq 0$,

$$Z_{11}(t) = R_{11}{}^T(t)\,Z_{11}(0)\,R_{11}(t) \tag{25a}$$

From (24) we find, using (23b) and Property 1.8,

$$\|\,R_{11}(t)\,\| \leq\ e^{-\alpha_1 t}\ . \tag{25b}$$

Hence,

$$\|\,Z_{11}(t)\,\| \leq\ e^{-2\alpha_1 t}\,\|\,Z_{11}(0)\,\| \tag{25c}$$

from which we obtain

$$\|\,T_1(t)\,\|^2 \leq\ 1 + e^{-2\alpha_1 t}\,\|\,Z_{11}(0)\,\|\ .$$

So, $T_1(t)$ is uniformly bounded.
Observe that (22) can be written as

$$\frac{dT_1}{dt} = A(t)\,T_1 - T_1\,\text{symm}\Big(\tilde{A}_{11}(t)\Big) - T_1\,S_{11}(t)\ , \qquad t \geq 0\ ,$$

where $S_{11}(t) = \tilde{A}_{11}(t) - \text{symm}\Big(\tilde{A}_{11}(t)\Big)$, the skew-symmetric part of $\tilde{A}_{11}$. By the boundedness of T_1, A and S_{11} we obtain that $\frac{d}{dt}T_1$ is uniformly bounded, from which the existence of T_1 follows by standard reasoning.
Moreover, by (25c) the DE (21) has $Z_{11} \equiv 0$ as asymptotically stable solution, which yields the column orthogonality of T_1 at ∞.

■

If (23a,b,c) are satisfied, then the requirements (i) and (ii) will be fulfilled. The obvious choice

$$\tilde{A}_{11} = {T_1}^T A T_1 \ , \tag{26}$$

satisfies the conditions (23a) and (23c), if T_1 is uniformly bounded. The condition (23b), however, is rather strong and will, in many cases, not be satisfied. In the proof of Theorem 3.4 we actually have used only the property that $\| R_{11}(t) \| \to 0$, as $t \to \infty$. This condition is less restrictive and will be satisfied, in general, as soon as T_1 is uniformly well-conditioned and properly scaled (in the sense that $\Big|\ 1 - \| T_1 \|\ \Big|$ is sufficiently small). This is shown in

Theorem 3.5
Assume that the solution space $\mathcal{S}$ is exponentially dichotomic with dim$(\mathcal{S}_1) = k$ (see Definition 2.7). Let T_1 and R_{11} satisfy, respectively, the DEs (22) and (24). If $T_1(0)$ has full rank and $\mathcal{R}\Big(T_1(0)\Big) \cap \mathcal{S}_2(0) = \{0\}$, then, for all t, we have

$$\| R_{11}(t) \| \leq \Big(\| T_1(t) \| / \text{glb}\Big(T_1(0)\Big) \Big)\ K_1(0)\ e^{-\lambda_1 t} \ , \tag{27}$$

where the function K_1 is given in Theorem 2.22.

Proof:
Let Y_{11} be the solution of (17) with $Y_{11}(0) = I_k$. Define $X_1 = T_1 Y_{11}$, which forms the first k columns of a fundamental solution X. By the condition for $\mathcal{R}\Big(T_1(0)\Big)\Big(= \mathcal{R}\Big(X_1(0)\Big)\Big)$ this X is consistent. Hence, using Theorem 2.22, we have, for any $c_1 \in \mathbb{R}^k$ and for all $t \geq 0$,

$$\begin{aligned} \frac{1}{K_1(0)}\, e^{\lambda_1 t} \ &\leq\ \frac{\| X_1(t) c_1 \|}{\| X_1(0) c_1 \|} \ = \ \frac{\| T_1(t) Y_{11}(t) c_1 \|}{\| T_1(0) c_1 \|} \\ &\leq\ \Big(\| T_1(t) \| / \text{glb}\Big(T_1(0)\Big) \Big)\ \frac{\| Y_{11}(t) c_1 \|}{\| c_1 \|} \ . \end{aligned}$$

Therefore,

$$\frac{1}{\mathrm{glb}\Big(Y_{11}(t)\Big)} \leq \Big(\| T_1(t) \| / \mathrm{glb}\Big(T_1(0)\Big) \Big) K_1(0)\, e^{-\lambda_1 t} .$$

Note that, by definition, $R_{11}(t) = \Big(Y_{11}(t)\Big)^{-1}$. Together with Property 1.1 this proves the theorem.

■

By this theorem we see how closely the requirements (i) and (ii) are connected. If $\tilde{A}_{11}$ has been chosen such that T_1 is uniformly well-conditioned, then the DE (22) will be stable, in general. If, for some reason, $\| T_1(t) \|$ will become large, then at such a point we can always restart the DE (22) with a column orthogonal matrix, spanning the same subspace. Such restart techniques will be discussed thoroughly in Section 4.3.

In order to fulfill the third requirement the following observation will be useful.

Property 3.6
The choice $\tilde{A}_{11} = T_1{}^+ A T_1$ *minimizes the quantity* $\| \frac{dT_1}{dt}(t) \|$, *for each* t.

Proof:
For each $t \in [0,1]$ and any $c_1 \in \mathbb{R}^k$ with $\| c_1 \| = 1$, we have

$$\begin{aligned}
\| \frac{dT_1}{dt}(t)\, c_1 \|^2 &= \| T_1(t)\, T_1{}^+(t) \frac{dT_1}{dt}(t)\, c_1 \|^2 + \\
&\qquad \| \Big(I_n - T_1(t)\, T_1{}^+(t) \Big) \frac{dT_1}{dt}(t)\, c_1 \|^2 \\
&= \| T_1(t) \Big(T_1{}^+(t)\, A(t)\, T_1(t) - \tilde{A}_{11}(t) \Big) c_1 \|^2 + \\
&\qquad \| \Big(I_n - T_1(t)\, T_1{}^+(t) \Big) A(t)\, T_1(t)\, c_1 \|^2 \\
&\geq \| \Big(I_n - T_1(t)\, T_1{}^+(t) \Big) A(t)\, T_1(t)\, c_1 \|^2 .
\end{aligned}$$

Hence, for any $\tilde{A}_{11}$, $\| \frac{d}{dt} T_1 \| \geq \| \Big(I_n - T_1\, T_1{}^+ \Big) A T_1 \|$ and this lower bound can be achieved if $\tilde{A}_{11} = T_1{}^+ A T_1$, since then these two matrix functions are identical.

■

If the choice $\tilde{A}_{11} = T_1{}^+ A T_1$ would lead to a column orthogonal matrix function T_1, then the rotational activity of T_1 is equal to the rotational activity of the subspace $\mathcal{R}(T_1)$, which is in some sense the best we can do.

Observe that with the choice $\tilde{A}_{11} = T_1{}^+ A T_1$ we have (see (19)):

$$\frac{d}{dt}\left(I_k - T_1{}^T T_1 \right) = 0 \ ,$$

independently of the value of $T_1(0)$. Hence, as soon as $T_1(0)$ is column orthogonal, then $T_1(t)$ is column orthogonal, for all t. So we have derived

Property 3.7
Let $\mathcal{S}_1$ be the solution subspace spanned by the columns of X_1, where X_1 is the solution of

$$\frac{d}{dt} X_1 = A(t)\, X_1 \ , \qquad t \epsilon [0,1] \ ,$$

with $X_1(0)$ a given column orthogonal $n \times k$ matrix.
Define the $n \times k$ matrix function $\hat{T}_1$ by

$$\begin{cases} \frac{d}{dt}\hat{T}_1 &= A(t)\hat{T}_1 - \hat{T}_1 \hat{T}_1^+ A(t) \hat{T}_1 \\ \hat{T}_1(0) &= X_1(0) \end{cases} \tag{28}$$

Then the columns of $\hat{T}_1$ form an orthogonal basis of $\mathcal{S}_1$ with at each time a minimal rotational activity.

■

If T_1 is column orthogonal, then $T_1{}^+ A T_1 = T_1{}^T A T_1$. The numerical approximation that is obtaine will be just nearly column orthogonal. However, Theorem 3.4 suggests that, in general, the column orthogonality is a stable property. Therefore, the choice $\tilde{A}_{11} = T_1{}^T A T_1$ will generally suffice too.

Remark 3.8
For BVPs with a singularity of the first kind (see Chapter 6) the choice $\tilde{A}_{11} = T_1{}^+ A T_1$ has already been suggested by Abramov in 1961 ([1]).

■

In order to obtain column orthogonality the symmetric part of $\tilde{A}_{11}$ is prescribed. Hence, we still have $\frac{1}{2}k(k-1)$ degrees of freedom left. This is just sufficient to make $\tilde{A}_{11}$ upper triangular, uniformly in t. This (unique) $\tilde{A}_{11}$ can be constructed using a mapping ψ_1, defined as follows.
Let $M \epsilon \mathbb{R}^{k \times k}$ be decomposed as

$$M = L + D + U \ , \tag{29}$$

where L is strictly lower triangular, U is strictly upper triangular and D diagonal. Define $\psi_1(M)$ by

$$\psi_1(M) = L^T + D + U \ . \tag{30}$$

Note that $\psi_1(M)$ is the unique upper triangular matrix with the same symmetric part as M. We now have

Property 3.9
Let the matrix functions X_1, T_1 and Y_{11} satisfy, respectively, the DEs (18a), (22) *and* (17). *Assume that the initial values have been chosen such that $X_1(0) = T_1(0)\,Y_{11}(0)$ is the QR-decomposition of $X_1(0)$.*
Then $X_1(t) = T_1(t)\,Y_{11}(t)$ is the QR-decomposition of $X_1(t)$, continuously in time, if and only if $\tilde{A}_{11} = \psi_1({T_1}^T A\,T_1)$.

Proof:
The matrix function Y_{11} is upper triangular, for all $t \geq 0$, if and only if $Y_{11}(0)$ and $\tilde{A}_{11}(t)$ are upper triangular. In Property 3.3 it is shown that T_1 is column orthogonal, for all $t \geq 0$, if and only if $T_1(0)$ is column orthogonal and $\mathrm{symm}(\tilde{A}_{11}) = \mathrm{symm}({T_1}^T A\,T_1)$. By the construction and uniqueness of the effect of the operator ψ_1 the result follows.

■

Although triangularity of $\tilde{A}_{11}$ and Y_{11} is a nice property it probably does not lead to the most efficient computation of a well-conditioned basis of $\mathcal{R}\Big(X_1(t)\Big)$, for all t. This is illustrated by the next (perhaps contrived) example. Therefore we stay to our choice $\tilde{A}_{11} = {T_1}^T A\,T_1$.

Example 3.10
Let

$$A(t) = \begin{bmatrix} \lambda_1 \cos^2\omega t + \lambda_2 \sin^2\omega t & (\lambda_2 - \lambda_1)\sin\omega t\cos\omega t + \omega & 0 \\ (\lambda_2 - \lambda_1)\sin\omega t\cos\omega t - \omega & \lambda_1 \sin^2\omega t + \lambda_2 \cos^2\omega t & 0 \\ 0 & 0 & -\lambda_1 \end{bmatrix}.$$

Then a fundamental solution X is given by

$$X(t) = \begin{bmatrix} \cos\omega t & \sin\omega t & 0 \\ -\sin\omega t & \cos\omega t & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} e^{\lambda_1 t} & & \\ & e^{\lambda_2 t} & \\ & & e^{-\lambda_1 t} \end{bmatrix}.$$

This expression directly gives the unique QR-decomposition of X. So, with $k = 2$ we obtain for $\bar{T}_1$, as defined by Property 3.9,

$$\bar{T}_1(t) = \begin{bmatrix} \cos\omega t & \sin\omega t \\ -\sin\omega t & \cos\omega t \\ 0 & 0 \end{bmatrix} .$$

Note that the rotational activity of $\bar{T}_1$ is equal to ω, for all t, which may be very large.
On the other hand, the matrix function $\hat{T}_1$, defined in (28), is given by

$$\hat{T}_1(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} ,$$

which is stationary. Of course, the rotational activity of X_1 (also equal to ω) is now completely moved to Y_{11}, but this is in general less troublesome, since Y_{11} satisfies a linear DE. Moreover, if an absolute accuracy is required, then the impact of this rotational activity on the integration routine may be reduced drastically by the *invariant imbedding* technique of Section 3.3.

■

3.2.3 The complementary subspace

So far we have discussed the role of T_1 only. In some situations, for instance if the BCs are separated (cf. the marching techniques), this will be sufficient. In the general case, however, also T_2 has to be considered.
In Theorem 3.5 we have shown that under certain conditions the matrix function $\tilde{A}_{11}$ governs the growth behaviour of the k most dominant solutions. Now we want to choose the $n \times (n-k)$ matrix function T_2 such that the growth behaviour of the solutions in the $(n-k)$-dimensional dominated solution subspace are reasonably well governed by $\tilde{A}_{22}$. Therefore it is not sufficient for T_2 to be well-conditioned and properly scaled. As has been shown in [42], we also need that GAP(T_1, T_2) is not too small. This implies that the $n \times n$ matrix function $T = \begin{bmatrix} T_1 & T_2 \end{bmatrix}$ has to be well-conditioned and properly scaled.
If we want T to be orthogonal, for all t, then we obtain similarly to Property 3.3, that

$$\text{symm}(\tilde{A}) = \text{symm}(T^T A\, T) .$$

Hence, since $\tilde{A}_{21} \equiv 0$, we get

$$\text{symm}(\tilde{A}_{11}) = \text{symm}({T_1}^T A\, T_1) \qquad (31a)$$

$$\text{symm}(\tilde{A}_{22}) = \text{symm}(T_2{}^T A\, T_2) \tag{31b}$$

and

$$\tilde{A}_{12} = T_1{}^T (A + A^T)\, T_2 \ . \tag{31c}$$

For T_1 we have found already in Section 3.2.2 the DE

$$\frac{d}{dt} T_1 = A(t)\, T_1 - T_1\, \tilde{A}_{11}(t) \ , \qquad t \geq 0 \ . \tag{32}$$

From the Lyapunov equation (13), the conditions (31b,c) and the orthogonality of T we obtain for T_2 the DE

$$\begin{aligned} \frac{d}{dt} T_2 &= A(t)\, T_2 - T_1\, \tilde{A}_{12}(t) - T_2\, \tilde{A}_{22}(t) \\ &= A(t)\, T_2 - T_1\, T_1{}^T \left(A(t) + A^T(t) \right) T_2 - T_2\, \tilde{A}_{22}(t) \\ &= A(t)\, T_2 + (T_2\, T_2{}^T - I_n) \left(A(t) + A^T(t) \right) T_2 - T_2\, \tilde{A}_{22}(t) \\ &= -A^T(t)\, T_2 + T_2 \left(T_2{}^T \left(A(t) + A^T(t) \right) T_2 - \tilde{A}_{22}(t) \right) \\ &= -A^T(t)\, T_2 + T_2\, \tilde{A}_{22}^T(t) \ , \qquad t \geq 0 \ . \end{aligned} \tag{33}$$

Let $Z_{21} = T_2{}^T\, T_1$. Then from (32) and (33) we derive

$$\left\{ \begin{array}{lcl} \dfrac{d}{dt} Z_{21} &=& \tilde{A}_{22}(t)\, Z_{21} - Z_{21}\, \tilde{A}_{11}(t) \ , \qquad t \geq 0 \\ Z_{21}(0) &=& T_2{}^T(0)\, T_1(0) \end{array} \right. \ . \tag{34}$$

Under mild conditions, which certainly will be satisfied if the solution space is exponentially dichotomic, $Z \equiv 0$ is the asymptotically stable solution of the DE of (34). Hence, the mutual orthogonality of T_1 and T_2 is an asymptotically stable property.

Properties, similar to those derived for T_1 (like column orthogonality), can also be obtained for T_2 by replacing A by $-A^T$. This implies, for instance, that for any fundamental solution X there exists a unique orthogonal $n \times n$ matrix function T such that $Y = T^{-1} X$ is upper triangular, for all t (cf. [37]).

In order to make T_1 (nearly) column orthogonal we found in Section 3.2.2 as the most obvious choice $\tilde{A}_{11} = T_1{}^T A\, T_1$. Similarly we obtain for T_2 the choice $\tilde{A}_{22} = T_2{}^T A\, T_2$. Then

$$\tilde{A} = \left[\begin{array}{cc} T_1{}^T A\, T_1 & T_1{}^T (A + A^T)\, T_2 \\ 0 & T_2{}^T A\, T_2 \end{array} \right] \ . \tag{35}$$

Hence, to complete the decomposition of the fundamental solution X we have to compute a non-singular block upper triangular solution Y of the DE

$$\frac{d}{dt}Y = \begin{bmatrix} T_1{}^T(t)\,A(t)\,T_1(t) & T_1{}^T(t)\Big(A(t)+A^T(t)\Big)\,T_2(t) \\ 0 & T_2{}^T(t)\,A(t)\,T_2(t) \end{bmatrix} Y \ . \qquad (36)$$

Remark 3.11
Although the matrix function $\tilde{A}$ is partially decoupled, the columns of the fundamental solution Y will in general become nearly dependent. Therefore the construction by superposition of y from a particular solution and the columns of Y may cause a loss of accuracy. One way to circumvent this instability is the use of a (generalized) multiple shooting method. Another way will be discribed in the next section.

■

3.2.4 Determination of the dominated subspace

To finish Section 3.2 we consider the following question: is it possible to compute a uniformly well-conditioned matrix function T such that Y is block diagonal? In other words: can we find a well-conditioned transformation T such that the transformed system is completely decoupled?
If $\tilde{A}_{12} \equiv 0$, then we obtain from the Lyapunov equation (13) that

$$\frac{d}{dt}T_2 = A(t)\,T_2 - T_2\,\tilde{A}_{22}(t) \ .$$

By Property 3.1 and Theorem 2.19 this implies that the gap between $\mathcal{R}(T_1)$ and $\mathcal{R}(T_2)$ will become small, unless $\mathcal{R}\Big(T_2(0)\Big)$ contains only dominated solutions. This subspace, however, is generally unknown, and even if it was known, then the DE for T_2 would be poorly conditioned. Therefore we may conclude that it is impossible, in general, to determine the dominated solution subspace in forward direction, or, which is actually the same, to decouple completely in just one sweep.

To determine the dominated solution subspace we need a double sweep. Therefore, let T be defined by (32) and (33), where $\tilde{A}_{11}$ and $\tilde{A}_{22}$ are chosen such that T is orthogonal, uniformly in t. Then any regular solution Y of (36) is a fundamental solution of the transformed system. Let $\hat{Y}$ be such a fundamental solution, satisfying

$$\Big[\hat{Y}_{21}(0)\ \ \hat{Y}_{22}(0)\Big] = \Big[0\ \ I_{n-k}\Big] \quad \text{and} \quad \Big[\hat{Y}_{11}(1)\ \ \hat{Y}_{12}(1)\Big] = \Big[I_k\ \ 0\Big] \ . (37)$$

This indeed defines a block upper triangular fundamental solution, since $\hat{Y}_{21} \equiv 0$ and $\hat{Y}_{22}(1)$ is a non-singular matrix. Hence, it can be decomposed as

$$\hat{Y}(t) = \begin{bmatrix} \hat{Y}_{11}(t) & \hat{Y}_{12}(t) \\ 0 & \hat{Y}_{22}(t) \end{bmatrix} = \begin{bmatrix} I_k & \hat{R}_{12}(t) \\ 0 & I_{n-k} \end{bmatrix} \begin{bmatrix} \hat{Y}_{11}(t) & 0 \\ 0 & \hat{Y}_{22}(t) \end{bmatrix}, \quad (38)$$

where $\hat{R}_{12}(t) = \hat{Y}_{12}(t)\hat{Y}_{22}^{-1}(t)$, for all t. One easily verifies that $\hat{R}_{12}$ satisfies the DE

$$\begin{cases} \frac{d}{dt}\hat{R}_{12} = \tilde{A}_{12} + \tilde{A}_{11}\hat{R}_{12} - \hat{R}_{12}\tilde{A}_{22}\,, & t \in [0,1] \\ \hat{R}_{12}(1) = 0 \end{cases} . \quad (39)$$

Now a fundamental solution X of the original system is given by

$$\begin{aligned} X &= T\hat{Y} = \begin{bmatrix} T_1 & T_2 \end{bmatrix} \begin{bmatrix} I_k & \hat{R}_{12} \\ 0 & I_{n-k} \end{bmatrix} \begin{bmatrix} \hat{Y}_{11} & 0 \\ 0 & \hat{Y}_{22} \end{bmatrix} \\ &= \begin{bmatrix} T_1 & T_1\hat{R}_{12} + T_2 \end{bmatrix} \begin{bmatrix} \hat{Y}_{11} & 0 \\ 0 & \hat{Y}_{22} \end{bmatrix} \\ &= \begin{bmatrix} T_1 & \hat{T}_2 \end{bmatrix} \begin{bmatrix} \hat{Y}_{11} & 0 \\ 0 & (\hat{R}_{12}^T\hat{R}_{12} + I_{n-k})^{\frac{1}{2}}\hat{Y}_{22} \end{bmatrix}, \end{aligned}$$

where $\hat{T}_2 = (T_1\hat{R}_{12} + T_2)(\hat{R}_{12}^T\hat{R}_{12} + I_{n-k})^{-\frac{1}{2}}$ is column orthogonal, for all $t \in [0,1]$, and mutually orthogonal to T_1 at $t = 1$.
Define, for $t \in [0,1]$,

$$\tilde{a}_{12}(t) = \| \tilde{A}_{12}(t) \|\,, \quad (40a)$$

$$\tilde{c}(t) = -\mu\Big(-\tilde{A}_{11}(t)\Big) - \mu\Big(\tilde{A}_{22}(t)\Big)\,. \quad (40b)$$

Then we have

Theorem 3.12
Define, corresponding to (40a,b),

$$\alpha_{12} = \max_{0\le t\le 1} \tilde{a}_{12}(t) \quad \textit{and} \quad \gamma = \min_{0\le t\le 1} \tilde{c}(t)\,.$$

If $\gamma > 0$ *then* $\| \hat{R}_{12}(t) \| \le \alpha_{12}/\gamma$, *for all* $t \in [0,1]$.

Proof:
Define the $k \times k$ incremental matrix function $W_{11}(t,\tau)$, for each $\tau \in [0,1]$, by

$$\begin{cases} \frac{d}{dt} W_{11}(t,\tau) = \tilde{A}_{11}(t)\, W_{11}(t,\tau)\,, \quad t \in [0,1] \\ W_{11}(\tau,\tau) = I_k \end{cases}, \tag{41}$$

and the $(n-k)\times(n-k)$ incremental matrix function $W_{22}(t,\tau)$, for each $\tau \in [0,1]$, by

$$\begin{cases} \frac{d}{dt} W_{22}(t,\tau) = -W_{22}(t,\tau)\, \tilde{A}_{22}(t)\,, \quad t \in [0,1] \\ W_{22}(\tau,\tau) = I_{n-k} \end{cases}. \tag{42}$$

Then $\hat{R}_{12}$ of (39) satisfies

$$\hat{R}_{12}(t) = -\int_t^1 W_{11}(t,\tau)\, \tilde{A}_{12}(\tau)\, W_{22}(t,\tau)\, d\tau\,. \tag{43}$$

From this relation it follows that

$$\begin{aligned} \| \hat{R}_{12}(t) \| \;&\leq\; \int_t^1 \| \tilde{A}_{12}(\tau) \| \, \{\exp(\int_t^\tau \mu(-\tilde{A}_{11}(s)) + \mu(\tilde{A}_{22}(s))\, ds)\}\, d\tau \\ &\leq\; \int_t^1 \tilde{a}_{12}(\tau)\, \{\exp(-\int_t^\tau \tilde{c}(s)\, ds)\}\, d\tau \end{aligned}$$

If $\gamma > 0$ then one easily verifies that the right hand side of this relation is smaller then α_{12}/γ, which proves the theorem.

■

If there exist positive constants α_1 and α_2 such that $\mu(-\tilde{A}_{11}) \leq -\alpha_1$ and $\mu(\tilde{A}_{22}) \leq -\alpha_2$ then $\gamma = \alpha_1 + \alpha_2$. Moreover, for any solution $x_1 \in \mathcal{R}(X_1)$ we obtain (cf. (25b)):

$$\| x_1(t) \| \geq e^{\alpha_1 t} \| x_1(0) \|\,, \qquad t \in [0,1]$$

In Definition 2.23 we have defined a dominant solution subspace $\mathcal{S}_1$ and a dominated solution subspace $\mathcal{S}_2$ that are mutually orthogonal at both ends $t = 0$ and $t = 1$. If $T_1(0)$ is chosen such that $\mathcal{R}\left(T_1(0)\right) = \mathcal{S}_1(0)$ then it follows, via $\mathcal{R}\left(T_1(1)\right) = \mathcal{S}_1(1)$ and $\mathcal{R}\left(\hat{T}_2(1)\right) = \mathcal{S}_2(1)$, that $\mathcal{R}\left(\hat{T}_2(0)\right) = \mathcal{S}_2(0)$. Hence, $T_1(0)$ and $\hat{T}_2(0)$ are mutually orthogonal, which implies that by this choice $\hat{R}_{12}(0) = 0$. Under these conditions we obtain, for any solution $x_2 \in \mathcal{R}(X_2)$ that

$$\| x_2(t) \| \;\leq\; e^{-\alpha_2 t} \| x_2(0) \| \sqrt{1 + \| \hat{R}_{12}(t) \|^2}$$

$$= \frac{e^{-\alpha_2 t} \parallel x_2(0) \parallel}{\mathrm{GAP}\left(T_1(t), \hat{T}_2(t)\right)} .$$

From this result we may conclude that, in general, $\mathcal{R}(T_1)$ will be a dominant solution subspace (determined in a forward sweep) and that $\mathcal{R}(\hat{T}_2)$ will be a dominated solution subspace (determined in a backward sweep).

3.3 Invariant imbedding

3.3.1 General description

One of the first inventors of the term '*invariant imbedding*' was Bellman (for instance [6]). Since the early sixties the term has been used quite often, but not always with the same meaning. Denman ([13]) uses the following definition:

> 'Invariant imbedding is a mathematical procedure by which a particular problem is imbedded within a family of related problems. The family of related problems are initial value problems and easily solved by a digital computer.'

However, sometimes one has to guess which imbedding is used and the corresponding IVPs may be very stiff. Often it is suggested that invariant imbedding is necessarily connected with the Riccati transformation ([53]). This misconception can be explained by historical arguments. The particle transport problem ([6]), for instance, is a classical example where the Riccati transformation (transmission) and invariant imbedding (reflection) are connected.

In this section we shall explain what we mean by invariant imbedding. It will turn out that invariant imbedding is just a special technique to perform the necessary backward sweep of a decoupled system. Therefore it can be combined with any decoupling transformation, not necessarily being a Riccati transformation (which will be discussed in Chapter 4). The combination of some decoupling transformation with invariant imbedding results in a method that can be seen as a generalization of a multiple shooting method (see Section 3.1).

Let T be any (time-dependent) transformation, defined by (13) and existing for $t \in [0,1]$, such that with $y = T^{-1}x$ the original system (9) becomes decoupled, i.e.,

$$\frac{dy}{dt} = \begin{pmatrix} \frac{dy_1}{dt} \\ \frac{dy_2}{dt} \end{pmatrix} \begin{matrix} \updownarrow k \\ \updownarrow n-k \end{matrix} = \begin{bmatrix} \tilde{A}_{11}(t) & \tilde{A}_{12}(t) \\ 0 & \tilde{A}_{22}(t) \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} \tilde{f}_1(t) \\ \tilde{f}_2(t) \end{pmatrix} , \quad (44)$$

$t \in [0,1]$. The transformed BCs read

$$B^0\, T(0)\, y(0) + B^1\, T(1)\, y(1) = b\ . \quad (45)$$

A fundamental solution Y corresponding to (44) is defined by

$$\frac{d}{dt} \begin{bmatrix} Y_{11} & Y_{12} \\ 0 & Y_{22} \end{bmatrix} = \begin{bmatrix} \tilde{A}_{11}(t) & \tilde{A}_{12}(t) \\ 0 & \tilde{A}_{22}(t) \end{bmatrix} \begin{bmatrix} Y_{11} & Y_{12} \\ 0 & Y_{22} \end{bmatrix} , \qquad t \in [0,1] . \quad (46)$$

If T is well-conditioned and properly scaled, then we have seen in the foregoing section that Y_{11} represents the growth behaviour of the unstable solutions, whereas the growth behaviour of dominated solutions is represented by Y_{22}. Hence, the $k \times k$ matrix function Y_{22} can be computed in forward direction. However, a forward computation of the $k \times n$ matrix function $\begin{bmatrix} Y_{11} & Y_{12} \end{bmatrix}$ would lead to numerical instability. Therefore, it seems better to compute this part of the fundamental solution from right to left. So we look for the fundamental solution $\hat{Y}$, satisfing the BCs (cf. (37))

$$\begin{bmatrix} \hat{Y}_{21}(0) & \hat{Y}_{22}(0) \end{bmatrix} = \begin{bmatrix} 0 & I_{n-k} \end{bmatrix} \text{ and } \begin{bmatrix} \hat{Y}_{11}(1) & \hat{Y}_{12}(1) \end{bmatrix} = \begin{bmatrix} I_k & 0 \end{bmatrix} . \quad (47)$$

Similarly, a particular solution $\hat{p}$ may be computed by

$$\frac{d}{dt} \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \end{pmatrix} = \begin{bmatrix} \tilde{A}_{11}(t) & \tilde{A}_{12}(t) \\ 0 & \tilde{A}_{22}(t) \end{bmatrix} \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \end{pmatrix} + \begin{pmatrix} \tilde{f}_1(t) \\ \tilde{f}_2(t) \end{pmatrix} , \qquad t \in [0,1] , \quad (48a)$$

subject to

$$\hat{p}_2(0) = 0 \quad \text{and} \quad \hat{p}_1(1) = 0\ . \quad (48b)$$

By the principle of superposition we know that there exists a vector $c \in \mathbb{R}^n$ such that

$$y(t) = \hat{Y}(t)\, c + \hat{p}(t) , \qquad \text{for all } t \in [0,1] .$$

By the choices (47) and (48b) we directly obtain that $c = \begin{pmatrix} y_1(1) \\ y_2(0) \end{pmatrix}$. Hence, for all $t \in [0,1]$, we have the relations

$$y_1(t) = \hat{Y}_{11}(t)\, y_1(1) + \hat{Y}_{12}(t)\, y_2(0) + \hat{p}_1(t) , \quad (49a)$$

$$y_2(t) = \hat{Y}_{22}(t)\, y_2(0) + \hat{p}_2(t)\ . \quad (49b)$$

The values of $y_1(1)$ and $y_2(0)$ are obtained by substituting (49a) with $t = 0$ and (49b) with $t = 1$ into the BCs (45). Hereafter, the value of $y(t)$, for any $t \in [0,1]$, can be computed by (49a,b).

The transformation T (and therefore $\tilde{A}$) has been computed in a forward sweep. Therefore, the computation in a forward sweep of $\hat{Y}_{22}$ and $\hat{p}_2$ will not cause serious difficulties. However, the computation of $\hat{Y}_{11}$, $\hat{Y}_{12}$ and $\hat{p}_1$ in a backward sweep requires the storage and interpolation of intermediate results. For separated BCs Meyer ([43]) suggests to overcome this difficulty by computing the solution of the original problem in a backward sweep, and to project this solution from time to time onto the manifold determined by $\mathcal{R}(T_1)$ and $T_2\, p_2$. Although such a method overcomes the problem of an overwhelming memory access, the instability of the original problem still remains. Moreover, it is a priori unknown which points should be taken as projection points.

The computation of $\left[\hat{Y}_{22} \mid \hat{p}_2\right]$ in a forward sweep and $\left[\hat{Y}_{11}\ \hat{Y}_{12} \mid \hat{p}_1\right]$ in a backward sweep (and also the update of Meyer) is based on the sound idea to express a solution of (44) in terms of $y_2(0)$ and $y_1(1)$ (a classical shooting method expresses the solution in terms of $y(0)$, and multiple shooting in terms of $y^i(t_i)$). What usually is meant by *invariant imbedding* ([6],[53]) *is not to express* $y_1(t)$ *in terms of* $y_1(1)$ *and* $y_2(0)$, *but* $y_1(0)$ *in terms of* $y_1(t)$ *and* $y_2(0)$. In other words: instead of (49a) we look for functions $\left[\underset{\longleftrightarrow}{R_{11}}\ \underset{\longleftrightarrow}{R_{12}} \mid \underset{\leftrightarrow}{g_1}\right] \updownarrow k$ (under the braces: k, $n-k$, 1) such that, for $t \in [0,1]$,

$$y_1(0) = R_{11}(t)\, y_1(t) + R_{12}(t)\, y_2(0) + g_1(t)\ . \tag{50}$$

(This is a generalization of the so-called *recovery transformation*, cf. [53]).

Let Y be the fundamental solution satisfying (46), with $Y(0) = I_n$, and let p be a particular solution satisfying $p(0) = 0$. Then

$$y_1(t) = Y_{11}(t)\, y_1(0) + Y_{12}(t)\, y_2(0) + p_1(t)\ , \tag{51a}$$

$$y_2(t) = Y_{22}(t)\, y_2(0) + p_2(t)\ . \tag{51b}$$

Comparing (50) with (51a) yields the relations

$$R_{11} = {Y_{11}}^{-1}\ , \quad R_{12} = -{Y_{11}}^{-1} Y_{12} \quad \text{and} \quad g_1 = -{Y_{11}}^{-1} p_1 \tag{52}$$

From these relations we derive that (50) holds if and only if $\left[R_{11}\ R_{12} \mid g_1\right]$ satisfies

$$\frac{d}{dt} R_{11} = -R_{11}\, \tilde{A}_{11}(t)\ , \quad t \in [0,1]\ , \quad R_{11}(0) = I_k\ , \tag{53a}$$

$$\frac{d}{dt}R_{12} = -R_{11}(t)\,\tilde{A}_{12}(t)\,Y_{22}(t)\;, \quad t\in[0,1]\,, \quad R_{12}(0)=0\;, \tag{53b}$$

$$\frac{d}{dt}g_1 = -R_{11}(t)\left(\tilde{A}_{12}(t)\,p_2(t)+\tilde{f}_1(t)\right)\;, \quad t\in[0,1]\,, \quad g_1(0)=0\;. \tag{53c}$$

The equation (53a) will generally be numerically stable (cf. Remark 3.2). Moreover, the equations (53b,c) are actually not DEs.

Taken together, the relations (50) and (51b) can be written as

$$\begin{bmatrix} I_k & -R_{12}(t) \\ 0 & -Y_{22}(t) \end{bmatrix}\begin{pmatrix} y_1(0) \\ y_2(0) \end{pmatrix}+\begin{bmatrix} -R_{11}(t) & 0 \\ 0 & I_{n-k} \end{bmatrix}\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} g_1(t) \\ p_2(t) \end{pmatrix}. \tag{54}$$

At $t=1$, this yields n relations between $y(0)$ and $y(1)$. Together with the BCs (45) this is sufficient to determine $y(0)$ and $y(1)$:

$$\begin{bmatrix} B^0\,T(0) & B^1\,T(1) \\ \begin{bmatrix} I_k & -R_{12}(1) \\ 0 & -Y_{22}(1) \end{bmatrix} & \begin{bmatrix} -R_{11}(1) & 0 \\ 0 & I_{n-k} \end{bmatrix} \end{bmatrix}\begin{pmatrix} y(0) \\ y(1) \end{pmatrix} = \begin{pmatrix} b \\ g_1(1) \\ p_2(1) \end{pmatrix}. \tag{55}$$

In combination with (10) we observe that a simple relation with the original problem is given by

Property 3.13
Let X be the fundamental solution corresponding to (9) *with $X(0)=T(0)$. Then we have the relations $T_1 = X_1\,R_{11}$ and $T_2\,Y_{22} = X_1\,R_{12}+X_2$.*
In matrix notation:

$$X\begin{bmatrix} R_{11} & R_{12} \\ 0 & I_{n-k} \end{bmatrix} = T\begin{bmatrix} I_k & 0 \\ 0 & Y_{22} \end{bmatrix}\;. \tag{56}$$

Proof:
The relations in (52) can be written as

$$\begin{bmatrix} Y_{11} & Y_{12} \\ 0 & Y_{22} \end{bmatrix}\begin{bmatrix} R_{11} & R_{12} \\ 0 & I_{n-k} \end{bmatrix} = \begin{bmatrix} I_k & 0 \\ 0 & Y_{22} \end{bmatrix}$$

Now the result follows directly from the relation $X = T\,Y$.

■

Let X be a consistent fundamental solution. Then we know (Property 3.1 and Theorem 2.19) that $\mathcal{R}(T_1) = \mathcal{R}(X_1) \to \mathcal{S}_1$, the dominant subspace. On the other hand, $\mathcal{R}\left(X\begin{bmatrix} R_{12} \\ I_{n-k} \end{bmatrix}\right)$ cannot contain dominant solutions, since Y_{22} rep-

resents the growth of dominated solutions. Actually, the invariant imbedding technique described above will asymptotically deliver the dominated subspace at $t = 0$, as is shown by

Theorem 3.14
Consider the homogeneous DE

$$\frac{dx}{dt} = A(t)\,x\ , \qquad t \in [0, \infty)\ . \tag{57}$$

Assume that the solution space $\mathcal{S}$ has an exponential dichotomy ($\mathcal{S} = \mathcal{S}_1 \oplus \mathcal{S}_2$, with $\dim(\mathcal{S}_1) = k$). Let T be a decoupling transformation as in (44), such that

(i) *T is uniformly well-conditioned and properly scaled*

(ii) *The matrix function R_{11}, as defined in (53a), satisfies $\| R_{11}(t) \| \to 0$ as $t \to \infty$.*

Then

$$\mathrm{DIST}(\,\mathcal{S}_2(0), \mathcal{R}\Big(T(0)\begin{bmatrix} R_{12}(t) \\ I_{n-k} \end{bmatrix}\Big)\,) = \mathrm{O}(\,\| R_{11}(t) \|\,)\,, \ \text{as } t \to \infty\ .$$

Proof:
From (50), where $y = T^{-1}x$, we obtain the relation

$$\begin{bmatrix} I_{n-k} & -R_{12}(t) \end{bmatrix} T^{-1}(0)\,x(0) = \begin{bmatrix} R_{11}(t) & 0 \end{bmatrix} T^{-1}(t)\,x(t)\ .$$

Hence, for any solution x of (57), there exists a vector $c_2 \in \mathbb{R}^{n-k}$ such that, for all $t \geq 0$,

$$x(0) = T(0)\begin{bmatrix} R_{12}(t) \\ I_{n-k} \end{bmatrix} c_2 + T(0)\begin{bmatrix} R_{11}(t) & 0 \\ 0 & 0 \end{bmatrix} T^{-1}(t)\,x(t)\ . \tag{58}$$

By assumption (i) the quantity $\| T(0) \| \| T^{-1}(t) \|$ is uniformly bounded. Moreover, the solution subspace $\mathcal{S}_2$ consists of all uniformly bounded solutions. Hence, all solutions x in $\mathcal{S}_2$ with $\| x(0) \| = 1$ are uniformly bounded by the same constant. So, if $x \in \mathcal{S}_2$ and $\| x(0) \| = 1$, then

$$\| T(0)\begin{bmatrix} R_{11}(t) & 0 \\ 0 & 0 \end{bmatrix} T^{-1}(t)\,x(t) \| = \mathrm{O}(\,\| R_{11}(t) \|\,)\ , \quad t \to \infty\ .$$

This implies, together with (58), that for each $x \in \mathcal{S}_2$ we have

$$\mathrm{GAP}(\,\mathrm{span}\{x(t)\}, \mathcal{R}\Big(T(0)\begin{bmatrix} R_{12}(t) \\ I_{n-k} \end{bmatrix}\Big)\,) \;=\; \mathrm{O}(\|\, R_{11}(t)\,\|)\,.$$

With Definition 1.12 of distance this shows the required result.

■

As a direct consequence of this theorem and the Properties 1.14 and 1.15 we have

Corollary 3.15

(i) *The matrix* $\bar{R}_{12} = \lim_{t\to\infty} R_{12}(t)$ *exists.*

(ii) *If* $T(0)$ *is orthogonal, then* $\mathrm{GAP}\Big(\mathcal{S}_2(0), \mathcal{R}\Big(T_1(0)\Big)\Big) = 1/\sqrt{1+\|\,\bar{R}_{12}\,\|^2}$.

■

Using (2.28) this corollary implies that the measure of consistency of an orthogonal matrix T(0) is quantified by $\sqrt{1+\|\,\bar{R}_{12}\,\|^2}$.

3.3.2 Restarts

The relations (51b) and (50) do contain the proper information to determine the solution somewhere in the interior of the interval, once $y(0)$ is known. Unfortunately, however, the computation of $y_1(t)$ from $y_1(0)$, using (50), would be unstable, since R_{11} will generally be ill-conditioned with respect to inversion. If we know a priori that the solution is wanted at the points $0 = t_0 < t_1 < \cdots < t_m = 1$, then instead of determining $\left[\begin{array}{cc|c} R_{11} & R_{12} & g_1 \\ 0 & Y_{22} & p_2 \end{array}\right]$, for all $t\,\epsilon\,[0,1]$, we could determine $\left[\begin{array}{cc|c} {R_{11}}^i & {R_{12}}^i & {g_1}^i \\ 0 & {Y_{22}}^i & {p_2}^i \end{array}\right]$, for $t\,\epsilon\,[t_i, t_{i+1}]$ $(i = 0, \ldots, m-1)$, which satisfy the DEs (see also Section 4.3):

$$\frac{d}{dt}{Y_{22}}^i = \tilde{A}_{22}(t)\,{Y_{22}}^i\,, \quad {Y_{22}}^i(t_i) = I_{n-k}\,,$$

$$\frac{d}{dt}{R_{11}}^i = -{R_{11}}^i\,\tilde{A}_{11}(t)\,, \quad {R_{11}}^i(t_i) = I_k\,,$$

$$\frac{d}{dt}{R_{12}}^i = -{R_{11}}^i(t)\,\tilde{A}_{12}(t)\,{Y_{22}}^i(t)\,, \quad {R_{12}}^i(t_i) = 0$$

$$\frac{d}{dt}{p_2}^i = \tilde{A}_{22}(t)\,{p_2}^i + \tilde{f}_2(t)\,, \quad {p_2}^i(t_i) = 0\,,$$

$$\frac{d}{dt} g_1{}^i = -R_{11}{}^i(t) \left(\tilde{A}_{12}(t)\, p_2{}^i(t) + \tilde{f}_1(t) \right) \ , \quad g_1{}^i(t_i) = 0 \ .$$

With the solutions of these DEs at $t = t_{i+1}$ we obtain the relation

$$\begin{bmatrix} I_k & -R_{12}{}^i(t_{i+1}) \\ 0 & -Y_{22}{}^i(t_{i+1}) \end{bmatrix} \begin{pmatrix} y_1(t_i) \\ y_2(t_i) \end{pmatrix} + \begin{bmatrix} -R_{11}{}^i(t_{i+1}) & 0 \\ 0 & I_{n-k} \end{bmatrix} \begin{pmatrix} y_1(t_{i+1}) \\ y_2(t_{i+1}) \end{pmatrix} = \begin{pmatrix} g_1(t_{i+1}) \\ p_2(t_{i+1}) \end{pmatrix} .$$

Together with the BCs (45) this yields the exact number of relations to determine $y(t_i)$ $(i = 0, \ldots, m)$. Actually, we have obtained a system which is an extension of (55) and very similar to the multiple shooting system (5), namely

$$\bar{M}\, \bar{y} = \bar{b} \ , \tag{59}$$

where $\bar{M} =$

$$\begin{bmatrix} B^0\, T(0) & & & & B^1\, T(1) \\ \begin{bmatrix} I_k & -R_{12}{}^0(t_1) \\ 0 & -Y_{22}{}^0(t_1) \end{bmatrix} & \begin{bmatrix} -R_{11}{}^0(t_1) & 0 \\ 0 & I_{n-k} \end{bmatrix} & & & \\ & \ddots & & \ddots & \\ & & \begin{bmatrix} I_k & -R_{12}{}^{m-1}(t_m) \\ 0 & -Y_{22}{}^{m-1}(t_m) \end{bmatrix} & \begin{bmatrix} -R_{11}{}^{m-1}(t_m) & 0 \\ 0 & I_{n-k} \end{bmatrix} & \end{bmatrix}$$

and

$$\left[\bar{y} \,|\, \bar{b} \right]^T = \begin{bmatrix} y(t_0)^T & y(t_1)^T & & \cdots & & y(t_m)^T \\ b^T & g_1{}^0(t_1)^T & p_2{}^0(t_1)^T & \cdots & g_1{}^{m-1}(t_m)^T & p_2{}^{m-1}(t_m)^T \end{bmatrix} .$$

In this multiple shooting system the dominant and dominated solutions have been decoupled already. Therefore, a relation between $y(0)$ and $y(1)$ can be otained by a straightforward substitution: define the sequence $\left\{ \left[\begin{array}{cc|c} F_{11}{}^i & F_{12}{}^i & d_1{}^i \\ 0 & F_{22}{}^i & d_2{}^i \end{array} \right] \right\}_{i=0}^{m}$ by the relation

$$\begin{pmatrix} y_1(t_i) \\ y_2(t_i) \end{pmatrix} = \begin{bmatrix} F_{11}{}^i & F_{12}{}^i \\ 0 & F_{22}{}^i \end{bmatrix} \begin{pmatrix} y_1(1) \\ y_2(0) \end{pmatrix} + \begin{pmatrix} d_1{}^i \\ d_2{}^i \end{pmatrix} . \tag{60}$$

Then one obtain the numerically stable recursions (cf. (4.58a-e)):
for $i = 0, \ldots, m-1$ (forward sweep)

$$F_{22}{}^{i+1} = Y_{22}{}^i(t_{i+1})\, F_{22}{}^i \ , \qquad F_{22}{}^0 = I_{n-k} \ ,$$

$$d_2{}^{i+1} = Y_{22}{}^{i}(t_{i+1})\, d_2{}^{i} + p_2{}^{i}(t_{i+1}) \;, \qquad d_2{}^{0} = 0 \;,$$

for $i = m, \ldots, 1$ (backward sweep)

$$F_{11}{}^{i-1} = R_{11}{}^{i-1}(t_i)\, F_{11}{}^{i} \;, \qquad F_{11}{}^{m} = I_k \;,$$

$$F_{12}{}^{i-1} = R_{11}{}^{i-1}(t_i)\, F_{12}{}^{i} + R_{12}{}^{i-1}(t_i)\, F_{22}{}^{i-1} \;, \qquad F_{12}{}^{m} = 0 \;,$$

$$d_1{}^{i-1} = R_{11}{}^{i-1}(t_i)\, d_1{}^{i} + R_{12}{}^{i-1}(t_i)\, d_2{}^{i-1} + g_1{}^{i-1}(t_i) \;, \qquad d_1{}^{m} = 0 \;.$$

The multiple shooting system (59) now reduces to the $n \times n$ system

$$\left(B^0 T(0) \begin{bmatrix} F_{11}{}^{0} & F_{12}{}^{0} \\ 0 & I_{n-k} \end{bmatrix} + B^1 T(1) \begin{bmatrix} I_k & 0 \\ 0 & F_{22}{}^{m} \end{bmatrix} \right) \begin{pmatrix} y_1(1) \\ y_2(0) \end{pmatrix} =$$

$$b - B^0 T(0) \begin{pmatrix} d_1{}^{0} \\ 0 \end{pmatrix} - B^1 T(1) \begin{pmatrix} 0 \\ d_2{}^{m} \end{pmatrix} . \tag{61}$$

With the solution of this system and the relation (60) we can find the solution y (and therefore $x = Ty$) at all points t_i $(i = 0, \ldots, m)$.

Remark 3.16
Comparing the relations (60) and (49a,b) we obtain

$$\left[\begin{array}{cc|c} F_{11}{}^{i} & F_{12}{}^{i} & d_1{}^{i} \\ 0 & F_{22}{}^{i} & d_2{}^{i} \end{array} \right] = \left[\begin{array}{cc|c} \hat{Y}_{11}(t_i) & \hat{Y}_{12}(t_i) & \hat{p}_1(t_i) \\ 0 & \hat{Y}_{22}(t_i) & \hat{p}_2(t_i) \end{array} \right] .$$

This result will be elaborated in Property 4.24.

■

Restarting at any point where output is required will be expensive if there are many. In that case it seems worth while first to construct some major subintervals and to maintain some of the information at the intermediate output points. On such a major subinterval, say $[\xi_i, \xi_{i+1}]$, we have a BVP with separated BCs, namely $y_2(\xi_i)$ and $y_1(\xi_{i+1})$ given. Using the relation

$$y_2(t) = Y_{22}{}^{i}(t)\, y_2(t_i) + p_2{}^{i}(t_i) \;, \qquad t \in [\xi_i, \xi_{i+1}] \;,$$

we obtain $(n-k)$ conditions at any point t in this subinterval. In order to derive at some specified point t a full set of n conditions we have to transform the information that is contained in $y_1(\xi_{i+1})$ backward to this value of t. To this end another (simple) decoupling transformation is needed (see, for instance, Remark 4.19).

Remark 3.17
Under the proper conditions the result of Theorem 3.14 can easily be generalized to

$$\mathrm{DIST}(\mathcal{S}_2(t_i), \mathcal{R}\Big(T(t_i)\begin{bmatrix} R_{12}{}^i(t) \\ I_{n-k} \end{bmatrix}\Big)) \to 0 \quad \text{as } t \to \infty .$$

In general we will have that (cf. Property 3.1 and Theorem 2.19)

$$\mathrm{DIST}\Big(\mathcal{S}_1(t_i), \mathcal{R}\Big(T_1(t_i)\Big)\Big) = \mathrm{O}(e^{-\lambda_1 t_i})$$

and

$$\| R_{11}{}^i(t) \| = \mathrm{O}\Big(e^{-\lambda_1(t-t_i)}\Big) \quad (t \geq t_i) .$$

Hence, if both quantities are small and $T(t_i)$ is orthogonal, then $\| R_{12}{}^i(t) \|$ yields a proper indication of $\mathrm{GAP}\Big(\mathcal{S}_1(t_i), \mathcal{S}_2(t_i)\Big)$ (cf. Corollary 3.15). This implies that under these conditions a large value of $\| R_{12}{}^i(t) \|$ indicates a large stability constant α (cf. Lemma 2.12).

■

Remark 3.18
One can show that the resulting system can also be derived from the classical multiple shooting system (5) with a fundamental solution X, satisfying (56). The necessary decoupling of dominant and dominated solutions in $X(t_i)$ ($i = 1, \ldots, m$) is performed in a numerically stable way by the transformation T.

■

3.4 Initial values

So far we have not discussed the difficulty of choosing proper values for $T(0)$ and k. Even in the description of the (standard) multiple shooting technique in Section 3.1 we have not touched on this subject.

Assume that the solution space $\mathcal{S}$ has an exponential dichotomy. Then the initial values have to be chosen such that they induce a decoupling, which is in correct order, i.e., such that the growth of the increasing modes is governed by $\tilde{A}_{11}$. By the well-conditioning of T the growth of the decaying modes is then governed by $\tilde{A}_{22}$. Hence, if possible, the integer k should be equal to the dimension of the dominant subspace $\mathcal{S}_1$. Unfortunately, this quantity is generally not known beforehand. In case of separated BCs we get an indication.

However, when smooth solutions are present this indication may be wrong (see also Section 2.4).

If $k \neq \dim(\mathcal{S}_1)$ we do not necessarily obtain an inefficient algorithm. A crucial role in the stability analysis is played by the fundamental solution R_{11} in (53a). As long as R_{11} does not contain fast increasing modes, then we still have a method that is stable (but not asymptotically stable). This condition is identical to the requirement that $\mathcal{R}(T_1)$ does not contain fast decaying modes.

Suppose we have separated BCs of the form $B^{02}x(0) = b_2$ and $B^{11}x(1) = b_1$, where $B^{02} \in \mathbb{R}^{(n-k)\times n}$ and $B^{11} \in \mathbb{R}^{k\times n}$. Then this value of k will do. Moreover, if $T(0)$ is chosen such that its first k columns satisfy the homogeneous part of the BCs at $t = 0$, i.e., $B^{02}T_1(0) = 0$, then $\mathcal{R}(T_1)$ does not contain fast decaying modes (see Section 2.4).

With non-separated BCs the situation is more complex. If $T(0)$ has been chosen randomly, then every column will have probability 1 that it contains a non-trivial component of the fastest growing mode. Similarly, with probability 1, the k fastest increasing modes are contained in the span of the first k columns of $T(0)$. This explains why the choice $T(0) = I_n$ will suffice, in general.

If the eigenvalues of A give a proper indication of the growth behaviour of the solutions of

$$\frac{dx}{dt} = A(t)\,x\ , \qquad t \in [0,1]\ ,$$

then a Schur transformation of $A(0)$ seems to generate a safe start: let $U^0 \in \mathbb{R}^{n\times n}$ be such that $A^0(0) = (U^0)^T A(0)\,U^0$ is (quasi) upper triangular ([20], p.192). In [58] it is indicated how the diagonal blocks (1×1 or 2×2) can be ordered along the diagonal. Therefore we may choose U^0 such that the real parts of the eigenvalues of the diagonal blocks are well ordered; the most positive ones to the left and the most negative ones to the right. Then we can take $T(0) = U^0$. In the case of constant coefficients we indeed obtain with this starting value a consistent fundamental solution. However, the eigenvalues of A give only local information of the growth behaviour. For the global behaviour of solutions they may be very misleading.

Example 3.19
Let

$$A(t) = \begin{bmatrix} 10\dfrac{1-20t}{1+20t} & 0 \\ 0 & -10\dfrac{1-20t}{1+20t} \end{bmatrix}\ , \qquad t > 0\ .$$

Then $A(0) = \begin{bmatrix} 10 & 0 \\ 0 & -10 \end{bmatrix}$, which seems to be correctly ordered. However, the corresponding fundamental solution, starting with the identity, is given by

$$X(t) = \begin{bmatrix} (1+20t)\,e^{-10t} & 0 \\ 0 & \dfrac{1}{1+20t}\,e^{10t} \end{bmatrix}.$$

This fundamental solution is not consistent.

■

Fortunately, the possibility remains to check whether the initial ordering has been done correctly or not (cf. [39]). If the gap between T_1 and T_2 stays sufficiently large, then the matrix function Y_{22} of (46) has to govern the growth behaviour of non-increasing solutions. Hence, if some elements of Y_{22} turn out to become large, say at $t = \hat{t}$, then some columns of $T(0)$ have to be permuted. Since all the necessary information is available at $\hat{t}$, this can be performed without an explicit restart of the integration at $t = 0$, although the resulting formulas do not look simple.

In the general case of non-separated BCs the Schur transformation may also indicate which value for k should be chosen. Assume that the rotation of the invariant subspaces of A is moderate (compared to the growth behaviour of solutions of the DE). Then the dichotomy of the solution space will correspond to a separation of the spectrum of A. However, again we have to be cautious, as is illustrated by the next example.

Example 3.20

Take the matrix function A of Example 3.10. Basis solutions of the DE grow, respectively, like $e^{\lambda_1 t}$, $e^{\lambda_2 t}$ and $e^{-\lambda_1 t}$. For $t = 0$ we obtain

$$A(0) = \begin{bmatrix} \lambda_1 & \omega & 0 \\ -\omega & \lambda_2 & 0 \\ 0 & 0 & -\lambda_1 \end{bmatrix}$$

and $\lambda\big(A(0)\big) = \left\{ \dfrac{\lambda_1+\lambda_2}{2} \pm \sqrt{\left(\dfrac{\lambda_1-\lambda_2}{2}\right)^2 - \omega^2}\,,\ -\lambda_1 \right\}$. Hence, if both λ_1 and λ_2 are positive, then we have two eigenvalues in $\mathbb{C}^+$ and one eigenvalue in $\mathbb{C}^-$, independent of the frequency ω. The choice $k = 2$ corresponds to the dimension of any dominant solution subspace.

If, however, λ_1 and λ_2 have opposite signs, then the situation may be quite dif-

ferent. In the general situation that $\lambda_1 > 0$ and $\lambda_2 < 0$ (and therefore $k = 1$) we have two eigenvalues in $\mathbb{C}^-$ and one in $\mathbb{C}^+$ as long as $|\omega| < \sqrt{-\lambda_1\lambda_2}$. For stiff problems (in this case: $\max\{|\lambda_1|, |\lambda_2|\} \gg 1$) this condition will generally be satisfied.
For instance, let $\lambda_2 = -\lambda_1$. Then $\lambda\Big(A(0)\Big) = \left\{ \pm\lambda_1\sqrt{1-(\omega/\lambda_1)^2}\,, -\lambda_1 \right\}$. As long as $|\omega/\lambda_1|$ is sufficiently smaller than 1, this spectrum nicely indicates the growth behaviour of solutions. However, if $|\omega/\lambda_1| \approx 1$ (or even more so $|\omega/\lambda_1| > 1$) then all information about increasing solutions is lost and it is impossible to decide which value of k would be the correct one.

■

This example illustrates once again that at $t = 0$ we do not have, in general, sufficient information to decide how the partitioning should be chosen and which initial value we have to take. Local information may be misleading. The effect of the initial choices can be checked during the integration, but the reparation in the case of a misconstruction may sometimes be time consuming.

3.5 Separated BCs

In Section 3.1 we have already seen that in a multiple shooting method we can reduce the number of equations involved in the computation if the BCs are separated. This is true also for the method sketched above. Assume that the BCs are given by

$$B^0\,x(0) + B^1\,x(1) = \begin{bmatrix} 0 \\ B^{02} \end{bmatrix} x(0) + \begin{bmatrix} B^{11} \\ 0 \end{bmatrix} x(1) = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} .$$

Let $T = \underset{k}{\left[\begin{matrix} T_1 \\ \longleftrightarrow \end{matrix} \right.} \underset{n-k}{\left. \begin{matrix} T_2 \\ \longleftrightarrow \end{matrix} \right]}$ be a decoupling transformation, existing on $[0,1]$, such that

$$B^{02}\,T_1(0) = 0\ . \tag{62}$$

Then

$$b_2 = B^{02}x(0) = B^{02}\,T(0)\,y(0) = B^{02}\,T_2(0)\,y_2(0)\ . \tag{63}$$

By the well-posedness of the BVP and the regularity of $T(0)$ the matrix $B^{02}\,T_2(0)$ will be non-singular. Hence, $y_2(0)$ is known. Moreover, by T the DE for y_2,

$$\frac{d}{dt}y_2 = \tilde{A}_{22}(t)\,y_2 + \tilde{f}_2(t)\ , \qquad t\,\epsilon\,[0,1]\ ,$$

is decoupled from y_1. Therefore, y_2 can be computed at the same time as T.

Since $y_2(0)$ is known the recovery transformation (50) can be reduced to the simpler form

$$y_1(0) = R_{11}(t)\,y_1(t) + g_1(t)\ , \qquad \text{for all } t\,\epsilon\,[0,1]\ , \tag{64}$$

where R_{11} and g_1 have to satisfy the DEs (cf. (53a,c)):

$$\frac{d}{dt}R_{11} = -R_{11}\,\tilde{A}_{11}(t)\ , \quad t\,\epsilon\,[0,1], \qquad R_{11}(0) = I_k\ , \tag{65a}$$

$$\frac{d}{dt}g_1 = -R_{11}(t)\left(\tilde{A}_{12}(t)\,y_2(t) + \tilde{f}_1(t)\right)\ , \qquad g_1(0) = 0\ . \tag{65b}$$

Thus, by the initial choices (62) and (63) the computation of Y_{22} and R_{12} has become superfluous.

If T is orthogonal even a further reduction can be obtained, since we are actually interested in $x = T\,y = T_1\,y_1 + T_2\,y_2$. We can avoid the computation of T_2 (just as in marching algorithms) as follows. Define $z = T_2\,y_2$. Then

$$x = T_1\,y_1 + z\ . \tag{66}$$

Since T is orthogonal, the orthogonality condition ${T_1}^T z = 0$ is satisfied over the entire interval.

Assume T is defined by (32),(33), where $\tilde{A}$ is chosen such that T is orthogonal. Then, using $\tilde{f}_2 = {T_2}^T f$ and $\tilde{A}_{22} + \tilde{A}_{22}^T = {T_2}^T(A + A^T)\,T_2$ (cf. (31b), we have

$$\begin{aligned}
\frac{dz}{dt} &= \frac{dT_2}{dt}y_2 + T_2\frac{dy_2}{dt} \\
&= -A^T(t)\,z + T_2(t)\,\left\{\left(\tilde{A}_{22}(t) + \tilde{A}_{22}^T(t)\right)\,{T_2}^T(t)\,z + \tilde{f}_2(t)\right\} \\
&= -A^T(t)\,z + T_2(t)\,{T_2}^T(t)\,\left\{\left(A(t) + A^T(t)\right)\,T_2(t)\,{T_2}^T(t)\,z + f(t)\right\}\ .
\end{aligned}$$

Since T_2 will not be computed we have to use the relations $T_1\,{T_1}^T + T_2\,{T_2}^T = I_n$ and $T_2\,{T_2}^T z = z$, which results in

$$\begin{aligned}
\frac{dz}{dt} &= -A^T(t)\,z + \left(I_n - T_1(t)\,{T_1}^T(t)\right)\left\{\left(A(t) + A^T(t)\right)z + f(t)\right\} \\
&= \left(I_n - T_1(t)\,{T_1}^T(t)\right)\left(A(t)\,z + f(t)\right) - T_1(t)\,{T_1}^T(t)\,A^T(t)\,z\ .
\end{aligned} \tag{67}$$

The initial value $z(0)$ is determined by

$$\begin{bmatrix} T_1{}^T(0) \\ B^{02} \end{bmatrix} z(0) = \begin{pmatrix} 0 \\ b_2 \end{pmatrix} .$$

Hence, as soon as $T_1(1)$ and $z(1)$ have been computed we are able to determine $y_1(1)$ from the BCs at $t = 1$:

$$b_1 = B^{11}x(1) = B^{11}\Big(T_1(1)\,y_1(1) + z(1)\Big)$$

$$\Rightarrow B^{11}T_1(1)\,y_1(1) = b_1 - B^{11}z(1) .$$

With $y_1(1)$ known we can use the recovery transformation (64) to determine the solution at any desired point. By the orthogonality of T we have (cf. (31c) and (14b)):

$$\tilde{A}_{12} = T_1{}^T(A + A^T)\,T_2 \quad \text{and} \quad \tilde{f}_1 = T_1{}^T f .$$

Since in the expression for $\tilde{A}_{12}$ the matrix function T_2 is involved the backward sweep also needs a slight modification. Using $T_2\,T_2{}^T z = z$, (53c) changes into

$$\frac{dg_1}{dt} = -R_{11}(t)\,T_1{}^T(t)\left\{\Big(A(t) + A^T(t)\Big)\,z(t) + f(t)\right\} , \; t\epsilon[0,1] , \quad g_1(0) = 0 .$$

Hence, for separated BCs we only have to solve DEs for the $(n+k) \times (k+1)$ matrix function $\begin{bmatrix} R_{11} & g_1 \\ T_1 & z \end{bmatrix}$ (in marching algorithms the order of the system is equal to $n \times (k+1)$).

3.6 Conclusions

Let us now recapitulate the foregoing results for continuous orthonormalization and invariant imbedding. In the first place we remark that a continuous determination of directions of solutions implies that the number of DEs that is to be solved is larger than in the current (multiple) shooting methods. Moreover, the complexity of these DEs has been increased (a linear problem is solved by non-linear problems).

On the other hand, the stability properties of the corresponding DEs have been improved, in general. Besides, the transformation T can be chosen as smooth as possible (cf. Property 3.6), having rotational activity only within $\mathcal{R}(T_1)^{\perp}$. This implies that, possibly after some initial layer effect, an automatic integration routine may choose large stepsizes.

Using the invariant imbedding technique we obtain that in the resulting method the IVPs we have to solve are all expected to be numerically stable in forward

direction. A qualified automatic initial value integration routine will therefore choose, in general, its stepsizes quite efficiently. Moreover, we do not have to store and interpolate intermediate results.
These benefits become more evident if the growth behaviour of solutions of the original problem is quite extreme, like in singular perturbation problems (see Chapter 5). For such problems also the backward sweep has a great influence on the total performance of the method. If output is required in just a few points, then the invariant imbedding technique is the most promising, since it circumvents the creation of many extra internal layers and excessive memory access has become superfluous.
Hence, continuous orthonormalization combined with invariant imbedding may be very successful for stiff and homogeneous (eigenvalue) problems, like the Orr-Sommerfeld equation with high Reynolds numbers (cf. [12]).
The disadvantages, like the number and the complexity of the DEs that are to be solved can be reduced if more restrictions are imposed on T. In the case of a Riccati transformation (see Chapter 4) the same number of DEs are to be solved as in the multiple shooting case.

Note that one can say that by invariant imbedding the problem is imbedded in the class of BVPs:

$$\frac{dy}{dt} = \tilde{A}(t)\,y + \tilde{f}(t)\,, \qquad t \in [0,\xi]\,,$$

subject to

$$B^0\,T(0)\,y(0) + B^1\,T(\xi)\,y(\xi) = b\,, \qquad \xi \in [0,1]\,.$$

This illustrates why the invariant imbedding technique can be used very well for free boundary problems ([44]).

Chapter 4

Riccati transformations

4.1 Introduction

In Chapter 3 we have discussed a general framework for a continuous decoupling of the DE

$$\frac{dx}{dt} = A(t)\,x + f(t)\;, \qquad t \in [0,1] \tag{1}$$

into

$$\frac{dy}{dt} = \tilde{A}(t)\,y + \tilde{f}(t)\;, \qquad t \in [0,1]\;, \tag{2}$$

where $y = T^{-1}x$, for some transformation T. The relation between T and $\tilde{A}$ is given by the Lyapunov equation (3.13), a system of n^2 equations for $2n^2$ variables (the elements of T and $\tilde{A}$). Hence, we have n^2 degrees of freedom. If $\tilde{A}$ is block upper triangular, for all t, then the DE (1) is said to be *decoupled* by T. Assume that A (and similarly $\tilde{A}$) is partitioned as

$$A(t) = A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{matrix} \updownarrow k \\ \updownarrow n-k \end{matrix} \quad , \tag{3}$$
$$\qquad\quad \underset{k}{\longleftrightarrow} \quad \underset{n-k}{\longleftrightarrow}$$

where the dimension k is to be determined later. The requirement $\tilde{A}_{21} \equiv 0$ reduces the number of degrees of freedom with $k(n-k)$. In Chapter 3 we constructed an orthogonal transformation T. This extra condition prescribes the symmetric part of $\tilde{A}$ (Property 3.3), which reduces the degrees of freedom with another $\frac{1}{2}n(n+1)$. We did not use the resulting degrees of freedom,

although their number turned out to be just sufficient to make $\tilde{A}$ upper triangular (cf. Property 3.9). The resulting system of DEs for T were at least cubic (cf. (3.31a,b)).

In this chapter we shall use the remaining $n^2 - k(n-k)$ degrees of freedom to simplify the form of T and the corresponding DEs. To this end we consider the well-known *Riccati transformation* (cf. [42],[51])

$$T(t) = \begin{bmatrix} I_k & 0 \\ R_{21}(t) & I_{n-k} \end{bmatrix}, \qquad t \in [0,1], \tag{4}$$

in which exactly $n^2 - k(n-k)$ components of $T(t)$ are prescribed. Observe that, for all $t \in [0,1]$,

$$T^{-1}(t) = \begin{bmatrix} I_k & 0 \\ -R_{21}(t) & I_{n-k} \end{bmatrix}$$

and

$$\frac{d}{dt}T(t) = \begin{bmatrix} 0 & 0 \\ \frac{d}{dt}R_{21}(t) & 0 \end{bmatrix}.$$

From the Lyapunov equation (3.13) and the requirement $\tilde{A}_{21} \equiv 0$ we obtain that the matrix function R_{21} has to satisfy the DE

$$\frac{d}{dt}R_{21} = A_{21}(t) + A_{22}(t)\,R_{21} - R_{21}\,A_{11}(t) - R_{21}\,A_{12}(t)\,R_{21}\,, \tag{5}$$

which is the so-called (*matrix*) *Riccati equation.* As long as a solution R_{21} of (5) exists, we have (cf. (3.14b,c)):

$$\tilde{A} = \begin{bmatrix} A_{11} + A_{12}\,R_{21} & A_{12} \\ 0 & A_{22} - R_{21}\,A_{12} \end{bmatrix} \tag{6a}$$

and

$$\tilde{f} = \begin{pmatrix} f_1 \\ -R_{21}\,f_1 + f_2 \end{pmatrix}. \tag{6b}$$

Hence, a corresponding fundamental solution Y of (2) may have a block upper triangular form as well. Like in (3.10) we have that $X = TY$ is a fundamental solution of the original system (1). Therefore, we actually have constructed a block LU-decomposition of such a fundamental solution X.

From the relation $X = TY$ we conclude that

$$R_{21}(t) = X_{21}(t)\, X_{11}{}^{-1}(t) \ . \tag{7}$$

This yields a first result on the existence of $R_{21}(t)$.

Property 4.1

Let $\mathcal{I}$ be some given interval. Then a solution R_{21} of (5) *exists on $\mathcal{I}$ if and only if there exists a fundamental solution X, corresponding to* (1)*, with X_{11} non-singular on $\mathcal{I}$.*

■

Although the existence of R_{21} is important, we also have to take care of its boundedness. As in Chapter 3 one of our main goals is to obtain a well-conditioned and properly scaled transformation T. For the Riccati transformation this condition is fulfilled as long as $\| R_{21} \|$ stays sufficiently bounded. In the next section we will see by which factors $\| R_{21} \|$ is mainly determined.

We may not be able, however, to find a solution of (5) which is sufficiently bounded over the entire interval $[0,1]$. In that case a transformation of the system (also called a restart) has to be considered. The way such difficulties can be handled will be discussed in Section 4.3.

To finish this section we make the following two remarks.

Remark 4.2

Assume that R_{21}, subject to $R_{21}(0) = 0$, exists on $[0,1]$. Then, after the computation of fundamental and particular solutions R_{11}, R_{12}, Y_{22}, g_1 and p_2 by the invariant imbedding technique of Section 3.3.1, we obtain from (3.50), (3.51b) and using that $x_2(0) = y_2(0)$ the well-known relations (cf. [6],[53]):

$$x_1(0) = R_{11}(t)\, x_1(t) + R_{12}(t)\, x_2(0) + g_1(t) \ , \qquad t \in [0,1] \ , \tag{8a}$$

$$x_2(t) = R_{21}(t)\, x_1(t) + Y_{22}(t)\, x_2(0) + p_2(t) \ , \qquad t \in [0,1] \ . \tag{8b}$$

These relations can also be written as

$$\begin{bmatrix} I_k & -R_{12}(t) \end{bmatrix} x(0) \ = \ \begin{bmatrix} R_{11}(t) & 0 \end{bmatrix} x(t) + g_1(t) \ ,$$

$$\begin{bmatrix} -R_{21}(t) & I_{n-k} \end{bmatrix} x(t) \ = \ \begin{bmatrix} 0 & Y_{22}(t) \end{bmatrix} x(0) + p_2(t) \ ,$$

which express the similarity in the two formulas. Various authors have used these relations as a starting point for the derivation of the so-called *Riccati method*, by which the combination of the Riccati transformation and invariant imbedding is meant ([45],[53]).

■

Remark 4.3
If ξ is such that $R_{21}(t)$ exists for all $t \epsilon [0, \xi)$, but $\lim_{t \to \xi} R_{21}(t)$ does not exist, then the homogeneous BVP

$$\frac{dx}{dt} = A(t)\, x \ , \tag{9}$$

subject to

$$\left[-R_{21}(0) \quad I_{n-k} \right] x(0) = 0 \quad \text{and} \quad \left[I_k \quad 0 \right] x(\xi) = 0$$

has a non-trivial solution. This property can be proved with the help of Property 4.1. The parameter ξ is called an *eigenlength* or *characteristic length* of (9) ([53]).

■

4.2 Existence and boundedness of R_{21}

The fact that $\| R_{21} \|$ may become large (implying an ill-conditioned transformation T), or even unbounded, is one of the main arguments used by critics to reject the Riccati transformation as a general solution method for linear BVPs. In this section we discuss the factors that influence the magnitude of the Riccati matrix R_{21}.

As will be shown in the next section the following assumption does not violate the generality of the transformation.

Assumption 4.4

$$R_{21}(0) = 0 \ .$$

■

Example 4.5

(i) Consider the second order DE

$$\frac{d^2 y}{dt^2} + \omega^2 y = 0 \ , \qquad t \epsilon [0, 1] \ .$$

Transforming it to a first order system by $x_1 = y$ and $x_2 = \frac{d}{dt} y$, we obtain

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega^2 & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad t \epsilon [0,1]\,. \tag{10}$$

The corresponding 1×1 Riccati matrix is given by

$$R_{21}(t) = -\omega \tan(\omega t)\,.$$

Hence, it blows up very fast if ω is large.
There are two main reasons for this phenomenon. In the first place the solution space of (10) is not exponentially dichotomic, and in the second place the basic modes $\begin{pmatrix} \cos(\omega t) \\ -\sin(\omega t) \end{pmatrix}$ and $\begin{pmatrix} \sin(\omega t) \\ \cos(\omega t) \end{pmatrix}$ are fast rotating.

(ii) The second order DE

$$\frac{d^2 y}{dt^2} - \omega^2 y = 0\,, \qquad t \epsilon [0,1]\,,$$

can be written as

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{bmatrix} 0 & 1 \\ \omega^2 & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad t \epsilon [0,1]\,.$$

The corresponding Riccati matrix is now given by

$$R_{21}(t) = \omega \tanh(\omega t)\,,$$

which is bounded by ω on the entire interval.
In this example the solution space $\mathcal{S}$ is exponentially dichotomic and the directions of the corresponding dominant and dominated subspaces are independent of time.

■

By only taking into account the norm of A one can state the following theorem.

Theorem 4.6
Let A be partitioned as in (3) *and define* $M = \max_{0 \le t \le 1} \| A(t) \|$. *Then the solution* R_{21} *of*

$$\begin{cases} \dfrac{d}{dt} R_{21} = A_{21}(t) + A_{22}(t)\, R_{21} - R_{21}\, A_{11}(t) - R_{21}\, A_{12}(t)\, R_{21} \\ R_{21}(0) = 0 \end{cases},$$

exists at least on $[0, \xi)$, *where* $\xi = \max\left(\frac{1}{M}, 1\right)$.

Proof:

Let $r(t) = \| R_{21}(t) \|_F$, the Frobenius norm of $R_{21}(t)$. Then $r(0) = 0$ and

$$\frac{dr}{dt} \leq M(1 + 2r + r^2) .$$

Hence, $\frac{dr}{(1+r)^2} \leq M\,dt$, from which we obtain $\frac{r(t)}{1+r(t)} \leq Mt$. Therefore,

$$r(t) \leq \frac{Mt}{1 - Mt} .$$

Now the theorem can be proved by standard reasoning.

■

This result may be sharpened if more specific information about A is used. To this end define, for any value of τ, the incremental fundamental solutions M_{11} and M_{22} by

$$\begin{cases} \frac{d}{dt} M_{11}(t,\tau) = -M_{11}(t,\tau)\, A_{11}(t) \\ M_{11}(\tau,\tau) = I_k \end{cases} , \tag{11}$$

$$\begin{cases} \frac{d}{dt} M_{22}(t,\tau) = A_{22}(t)\, M_{22}(t,\tau) \\ M_{22}(\tau,\tau) = I_{n-k} \end{cases} . \tag{12}$$

Define, for $R \in \mathbb{R}^{(n-k)\times k}$,

$$f(t,R) = A_{21}(t) - R\,A_{12}(t)\,R . \tag{13}$$

Then, using Assumption 4.4, the DE (5) can be written in integral form as

$$R_{21}(t) = \int_0^t M_{22}(t,\tau)\, f\Big(\tau, R_{21}(\tau)\Big)\, M_{11}(t,\tau)\, d\tau . \tag{14}$$

Therefore,

$$\| R_{21}(t) \| \leq \tag{15}$$

$$\int_0^t \exp\Big(\int_\tau^t \mu\Big(A_{22}(s)\Big)\, ds\Big)\, \| f\Big(\tau, R_{21}(\tau)\Big) \|\, \exp\Big(\int_\tau^t \mu\Big(-A_{11}(s)\Big)\, ds\Big)\, d\tau .$$

Now define

$$a_{21}(t) = \| A_{21}(t) \| , \tag{16a}$$

$$a_{12}(t) = \| A_{12}(t) \| , \tag{16b}$$

$$c(t) = -\frac{1}{2}\Big(\mu(-A_{11}(t)) + \mu(A_{22}(t))\Big) \ . \tag{16c}$$

Then we are able to show

Lemma 4.7
Let ρ be the solution of the IVP

$$\frac{d\rho}{dt} = a_{21}(t) - 2\,c(t)\,\rho + a_{12}(t)\,\rho^2 \ , \qquad \rho(0) = 0 \ . \tag{17}$$

As long as ρ exists, we have the inequality $\| R_{21}(t) \| \leq \rho(t)$.

Proof:
Define $r(t) = \| R_{21}(t) \|$. Then (15) yields

$$r(t) \ \leq \ \int_0^t \Big(a_{21}(\tau) + a_{12}(\tau)\,r^2(\tau)\Big)\ \exp(-2\int_\tau^t c(s)\,ds)\,d\tau \ .$$

Moreover, ρ satisfies the equality

$$\rho(t) = \int_0^t \Big(a_{21}(\tau) + a_{12}(\tau)\,\rho^2(\tau)\Big)\ \exp(-2\int_\tau^t c(s)\,ds)\,d\tau \ .$$

Hence,

$$r(t) - \rho(t) \ \leq \ \int_0^t g(t,\tau)\,\Big(r(\tau) - \rho(\tau)\Big)\,d\tau \ ,$$

where $g(t,\tau) = a_{12}(\tau)\,\Big(r(\tau) + \rho(\tau)\Big)\ \exp(-2\int_\tau^t c(s)\,ds)$, which is a non-negative function. Now, using the Lemma of Grönwall, the lemma is proved.

■

Theorem 4.8
Let ξ be such that, for all $t \in [0,\xi]$, (see (16a,b,c))

(i) $c(t) \geq \gamma > 0$

(ii) $K(t) = \dfrac{a_{21}(t)\,a_{12}(t)}{c^2(t)} < 1$

(iii) $p_-(t) = \dfrac{a_{21}(t)}{c(t)}\,\dfrac{1}{1+\sqrt{1-K(t)}}$, $p_+(t) = \dfrac{c(t)}{a_{12}(t)}\,\Big(1+\sqrt{1-K(t)}\Big)$ *and* $m \in \mathbb{R}$ *satisfy*

$$p_-(t) \leq m < p_+(t) , \qquad 0 \leq t \leq \xi .$$

Then $0 \leq \| R_{21}(t) \| \leq m$, *for all* $t \in [0, \xi]$.

Proof:
Observe that p_- and p_+ are defined such that (17) is identical to

$$\frac{d\rho}{dt} = a_{12}(t) \Big(\rho - p_-(t)\Big) \Big(\rho - p_+(t)\Big) .$$

Therefore,

$$\frac{d\rho}{dt} = \begin{cases} < 0 & \text{if } p_-(t) < \rho(t) < p_+(t) \\ = 0 & \text{if } p_-(t) = \rho(t) \text{ or } p_+(t) = \rho(t) \\ > 0 & \text{elsewhere} \end{cases} .$$

Since $\rho(0) = 0 < p_-(0)$ the result follows from Lemma 4.7 in a straightforward way.

■

Sometimes the upper bound in Lemma 4.7 may be overly pessimistic, as is illustrated by the next contrived example.

Example 4.9
Let the matrix function A be defined by

$$A(t) = \begin{bmatrix} e^{\lambda t} & e^{2\lambda t} \\ 1 & -e^{\lambda t} \end{bmatrix} \quad (\lambda > 0) , \quad t \geq 0 .$$

Then one easily checks that the corresponding Riccati matrix R_{21} satisfies $R_{21} \to 0$ as $t \to \infty$. However, for ρ we obtain the DE

$$\frac{d\rho}{dt} = (1 - e^{\lambda t}\rho)^2 , \qquad t \geq 0 ,$$

whose solution will grow rather fast.

■

Now we shall obtain a result slightly more general than Theorem 4.8. To this end we define the non-decreasing functions s, s_{21} and s_{12} by (see (11) and (12))

$$s(t) = 2 \int_0^t \| M_{22}(t, \tau) \| \| M_{11}(t, \tau) \| \, d\tau , \tag{18a}$$

$$s_{21}(t) = \max_{0\leq\tau\leq t} \| A_{21}(\tau) \| , \tag{18b}$$

and

$$s_{12}(t) = \max_{0\leq\tau\leq t} \| A_{12}(\tau) \| . \tag{18c}$$

Then we have

Theorem 4.10
As long as t is such that, correspondingly to (18a-c),

$$s^2(t)\, s_{21}(t)\, s_{12}(t) < 1 , \tag{19}$$

then $R_{21}(t)$ exists and $\| R_{21}(t) \| \leq s(t)\, s_{21}(t)$.

Proof:
Observe that $s(0) = 0$, which implies that (19) is satisfied for $t = 0$. Now let $t > 0$ be such that (19) is satisfied. Existence of a solution of (5) can be proved by the method of successive substitution, applied to the non-linear integral equation (14). Hence, define the sequence of matrix functions $\{R^i(t)\}_{i=0}^{\infty}$ by

$$\begin{aligned} &R^0(t) \equiv 0 , \\ &R^{i+1}(t) = \int_0^t M_{22}(t,\tau)\, f\Big(\tau, R^i(\tau)\Big)\, M_{11}(t,\tau)\, d\tau . \end{aligned} \tag{20}$$

Let

$$w^i(t) = \max_{0\leq\tau\leq t} \| R^i(\tau) \| , \quad i = 0, 1, \cdots . \tag{21}$$

Then from (20) and (13) it follows that

$$\begin{aligned} w^{i+1}(t) &\leq \frac{1}{2}\, s(t) \max_{0\leq\tau\leq t} \| f\Big(\tau, R^i(\tau)\Big) \| \\ &\leq \frac{1}{2}\, s(t) \Big(s_{21}(t) + s_{12}(t)\, w^i(t)^2 \Big) . \end{aligned}$$

This implies that the requirement (19) guarantees boundedness of the numbers $w^i(t)$ by $s(t)\, s_{21}(t)$ (cf. [57]). In order to show that the $R^i(t)$ converge we observe that

$$\begin{aligned} &\| f(\tau, R^i(\tau)) - f(\tau, R^{i-1}(\tau)) \| \\ &\quad \leq \| A_{12}(\tau) \| \Big(\| R^i(\tau) \| + \| R^{i-1}(\tau) \| \Big) \| R^i(\tau) - R^{i-1}(\tau) \| . \end{aligned}$$

Hence, for all $\xi \in [0, t]$ we have

$$
\begin{aligned}
&\| R^{i+1}(\xi) - R^{i}(\xi) \| \\
&\leq \int_0^{\xi} \| M_{22}(\xi,\tau) \|\| f(\tau, R^{i}(\tau)) - f(\tau, R^{i-1}(\tau)) \|\| M_{11}(\xi,\tau) \| \, d\tau \\
&\leq \frac{1}{2}\, s(\xi)\, s_{12}(\xi)\, 2\, s(\xi)\, s_{21}(\xi) \max_{0\leq\tau\leq t} \| R^{i}(\tau) - R^{i-1}(\tau) \| \\
&\leq s^2(t)\, s_{12}(t)\, s_{21}(t) \max_{0\leq\tau\leq t} \| R^{i}(\tau) - R^{i-1}(\tau) \| \ .
\end{aligned}
$$

Therefore, again using (19), $\lim_{i\to\infty} R^{i}(t)$ exists, is bounded by $s(t)\, s_{21}(t)$ and, moreover, it satisfies the integral equation (14).

■

Remark 4.11
Observe that R_{21} also satisfies the DE

$$
\frac{d}{dt} R_{21} = A_{21}(t) + \tilde{A}_{22}(t)\, R_{21} - R_{21}\, \tilde{A}_{11}(t) + R_{21}\, A_{12}(t)\, R_{21} \ ,
$$

where $\tilde{A}_{11} = A_{11} + A_{12}\, R_{21}$ and $\tilde{A}_{22} = A_{22} - R_{21}\, A_{12}$ (see (6a)). Hence, a similar result as that of Theorem 4.8 can be formulated with c replaced by

$$
\tilde{c}(t) = -\frac{1}{2}\Big(\mu(-\tilde{A}_{11}(t)) + \mu(\tilde{A}_{22}(t)) \Big) \ .
$$

Equivalently, Theorem 4.10 is also valid with M_{11} and M_{22} replaced by R_{11} and Y_{22}, the fundamental solutions corresponding to, respectively, $-\tilde{A}_{11}$ and $\tilde{A}_{22}$. These solutions are actually computed when the invariant imbedding technique of Section 3.3.1 is used (see (3.53a) and (3.46)).

■

From the foregoing we may conclude that it will be hard to give sharp conditions that imply boundedness of solutions of the Riccati DE. Only in special applications, which occur for instance in control theory, we have conditions (of symmetry and definiteness) that guarantee boundedness of the Riccati matrix (cf. [52]).
Remark 4.11 shows already that the magnitude of R_{21} is strongly influenced by the growth behaviour of the fundamental solutions R_{11} and Y_{22} and by $\| A_{21} \|$ and $\| A_{12} \|$. Therefore it is important that the distance between $\mathcal{R}\Big(\begin{bmatrix} I_k \\ R_{21}(t) \end{bmatrix} \Big)$ and $\mathcal{S}_1(t)$, the dominant subspace at time t, is sufficiently small (cf. Theorem 3.5).

Using geometrical arguments similar conclusions can be drawn. To this end we look at the *algebraic Riccati equation*

$$0 = A_{21}(t) + A_{22}(t)\,P_{21}(t) - P_{21}(t)\,A_{11}(t) - P_{21}(t)\,A_{12}(t)\,P_{21}(t)\ . \qquad (22)$$

Observe that if $P_{21}(t)$ satisfies (22), then

$$A(t)\begin{bmatrix} I_k \\ P_{21}(t) \end{bmatrix} = \begin{bmatrix} I_k \\ P_{21}(t) \end{bmatrix}\Big(A_{11}(t) + A_{12}(t)\,P_{21}(t)\Big)\ , \qquad (23)$$

which implies that $\mathcal{R}\Big(\begin{bmatrix} I_k \\ P_{21}(t) \end{bmatrix}\Big)$ is an invariant subspace of $A(t)$. Moreover, one can show that

$$\lambda\Big(A(t)\Big) = \lambda\Big(A_{11}(t) + A_{12}(t)\,P_{21}(t)\Big) \cup \lambda\Big(A_{22}(t) - P_{21}(t)\,A_{12}(t)\Big)$$

(cf. [57]). For the remaining part of this section we make the following assumption.

Assumption 4.12
The matrix $A(0)$ is (block) upper triangular and correctly ordered, i.e.,

$$A(0) = \begin{bmatrix} A_{11}(0) & A_{12}(0) \\ 0 & A_{22}(0) \end{bmatrix}$$

and $\lambda_{\min}\Big(A_{11}(0)\Big) > \lambda_{\max}\Big(A_{22}(0)\Big)$ (which can always be obtained by an (orthogonal) Schur transformation (see Section 3.4)).

■

With this assumption and the definition of separation of matrices (Definition 1.5) one can prove the following generalization of Theorem 3.5 in [57].

Theorem 4.13
Define

$$a_{21}(t) = \|\ A_{21}(t)\ \|\ ,$$

$$a_{12}(t) = \|\ A_{12}(t)\ \|\ ,$$

and

$$d(t) = \frac{1}{2}\,\mathrm{sep}\Big(\,A_{11}(t), A_{22}(t)\,\Big)\ . \qquad (24)$$

As long as $d(t) > 0$ and

$$K(t) = \frac{a_{21}(t)\, a_{12}(t)}{d^2(t)} < 1$$

then there exists a unique continuously differentiable solution P_{21} *of* (22) *with* $P_{21}(0) = 0$ *and satisfying*

$$\| P_{21}(t) \| \leq \frac{a_{21}(t)}{d(t)} \frac{1}{1+\sqrt{1-K(t)}} < \frac{a_{21}(t)}{d(t)} \ . \tag{25}$$

Proof:
Stewart ([57]) shows that, for any fixed t with $K(t) < 1$, there exists a unique solution $P_{21}(t)$, which satisfies (25). From Property 1.6 we obtain

$$\text{sep}(\, A_{11} + A_{12}\, P_{21}, A_{22} - P_{21}\, A_{12}\,) \ \geq \ 2\, d - 2\frac{a_{12}\, a_{21}}{d} > 0 \ .$$

Hence, using Assumption 4.12,

$$\lambda_{\min}(A_{11} + A_{12}\, P_{21}) > \lambda_{\max}(A_{22} - P_{21}\, A_{12}) \ . \tag{26}$$

This implies that $\mathcal{R}\left(\begin{bmatrix} I_k \\ P_{21}(t) \end{bmatrix}\right)$ is the (unique) invariant subspace corresponding to the k eigenvalues with largest real parts $\Big(\lambda(A_{11} + A_{12}\, P_{21})\Big)$. With (26) it follows that this invariant subspace changes in a continuously differentiable way.

■

Observe that the quantities s in (18a) and d in (24) have a relation similar to the one obtained in Theorem 1.9. By this relation the conditions of Theorem 4.8 and Theorem 4.13 are coupled.

Assume the solution P_{21} of (22) defined in Theorem 4.13 exists on some interval $[\,0, \xi\,]$. Define the matrix function E_{21} as the solution of the Riccati DE

$$\begin{aligned} \frac{d}{dt}E_{21} = -\frac{d}{dt}P_{21} \ &+ \ \Big(A_{22}(t) - P_{21}(t)\, A_{12}(t)\Big)\, E_{21} \\ &- \ E_{21}\, \Big(A_{11}(t) + A_{12}(t)\, P_{21}(t)\Big) - E_{21}\, A_{12}(t)\, E_{21} \ , \end{aligned}$$

subject to $E_{21}(0) = 0$. Then E_{21} is nothing but the difference between R_{21} and P_{21}. The magnitude of E_{21} is, cf. Theorem 4.10, in the first place determined by the growth behaviour of the fundamental solutions corresponding to $A_{22} - P_{21}\, A_{12}$ and $-(A_{11} + A_{12}\, P_{21})$ and in the second place by the values $\| \frac{d}{dt}P_{21} \|$ and $\| A_{12} \|$. Since $R_{21} = P_{21} + E_{21}$ we observe that, if both P_{21} and E_{21} remain sufficiently bounded, then also R_{21} remains bounded. These conditions

will be fulfilled if:
- the eigenvalues of $A(t)$ are sufficiently well separated into two clusters; globally speaking: one cluster containing eigenvalues with positive real parts and the other cluster containing eigenvalues with non-positive real parts.
Relative to this separation we need that
- the invariant subspace belonging to the cluster of eigenvalues with positive real parts does not change too rapidly
- the quantity $\| A_{12} \|$ is not too large
- $E_{21}(0)$ is sufficiently small.

Remark 4.14
Differentiating the algebraic Riccati equation (22) yields the (time-dependent) Sylvester equation

$$(A_{22} - P_{21}A_{12})(\frac{d}{dt}P_{21}) - (\frac{d}{dt}P_{21})(A_{11} + A_{12}P_{21}) =$$
$$-\Big(\frac{d}{dt}A_{21} + \Big(\frac{d}{dt}A_{22}\Big)P_{21} - P_{21}\Big(\frac{d}{dt}A_{11}\Big) - P_{21}\Big(\frac{d}{dt}A_{12}\Big)P_{21}\Big) .$$

Hence, the magnitude of $\| \frac{d}{dt}P_{21} \|$ is determined by $\| \frac{d}{dt}A \|$, $\| P_{21} \|$ and $\mathrm{sep}\Big(A_{11} + A_{12}P_{21}, A_{22} - P_{21}A_{12}\Big)$.

■

4.3 Separated BCs

In this section we shall discuss in more detail what the Riccati method looks like in case of a BVP with separated BCs. We have seen already in Section 3.5 that a reduction in the number of DEs can be obtained. By choosing the proper initial values (Section 4.3.1) the Riccati method requires the computation of the $n \times (k+1)$ system $\begin{bmatrix} R_{11} & g_1 \\ R_{21} & y_2 \end{bmatrix}$. Observe that the order of this system is as large as for a stabilized march algorithm (see Section 3.1).
However, the Riccati DE for R_{21} will be non-linear. This non-linearity may cause some additional orthogonalization steps, in order to keep the decoupling transformation well-conditioned. How such a restart can be performed will be treated in Section 4.3.2.
The resulting method is described algorithmically in Section 4.3.3.
For the computational aspects of the Riccati method we refer to Section 4.5.

4.3.1 Initial values

Assume that the solution space $\mathcal{S}$ has an exponential dichotomy ($\mathcal{S} = \mathcal{S}_1 \oplus \mathcal{S}_2$). In Section 3.4 we have already seen that, if possible, k should be chosen equal to $\dim(\mathcal{S}_1)$. The Riccati matrix R_{21} is constructed such that

$$\mathcal{R}\Big(X_1(t)\Big) = \mathcal{R}\Big(\begin{bmatrix} I_k \\ R_{21}(t) \end{bmatrix}\Big) , \tag{27}$$

where X_1 is an $n \times k$ matrix function consisting of the first k columns of a fundamental solution X (cf. (7)). As follows from Remark 4.11, this fundamental solution X should actually be consistent, since then $\mathrm{DIST}\Big(\mathcal{R}\Big(X_1(t)\Big), \mathcal{S}_1(t)\Big) \rightarrow 0$ as $t \rightarrow \infty$ (cf. Theorem 2.19). Hence, we have to satisfy the relation

$$\{0\} = \mathcal{R}\Big(X_1(0)\Big) \cap \mathcal{S}_2(0) = \mathcal{R}\Big(\begin{bmatrix} I_k \\ R_{21}(0) \end{bmatrix}\Big) \cap \mathcal{S}_2(0) . \tag{28}$$

For *separated* BCs, i.e.,

$$B^{02}x(0) = b_2 \quad \text{and} \quad B^{11}x(1) = b_1 ,$$

where $B^{02} \epsilon \mathbb{R}^{(n-k)\times n}$ and $B^{11} \epsilon \mathbb{R}^{k\times n}$, this value of k will do and we may choose a fundamental solution $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$ that satisfies the homogeneous BCs at $t = 0$, i.e., $B^{02} X_1(0) = 0$ (see Section 3.4). Assuming that $X_{11}(0)$ is non-singular, the corresponding initial value for the Riccati transformation, satisfying

$$B^{02} \begin{bmatrix} I_k \\ R_{21}(0) \end{bmatrix} = 0 , \tag{29}$$

would be $R_{21}(0) = X_{21}(0)\, {X_{11}}^{-1}(0)$ (cf. (7)). This quantity, however, may be large, implying, at least for small t, a poorly conditioned Riccati transformation.

There are some ways to decrease the magnitude of this initial value. Allowing that the columns of B^{02} (and correspondingly the elements of x) are permuted one may derive

Theorem 4.15 ([60])
There exists a permutation matrix $\Pi^0 = [\underset{\longleftrightarrow\, k}{\Pi_1{}^0} \; \underset{\longleftrightarrow\, n-k}{\Pi_2{}^0}]$ *such that* $B^{02}\,\Pi_2{}^0$ *is non-singular and all elements of the matrix* ${R_{21}}^0(0)$ *that satisfies*

$$B^{02}\,\Pi^0 \begin{bmatrix} I_k \\ {R_{21}}^0(0) \end{bmatrix} = 0 \tag{30}$$

are in absolute value not greater than 1 $\left(\left|R_{21}{}^0(0)\right| \leq 1\right)$.

■

This implies that for the permuted system

$$\frac{dx^0}{dt} = A^0(t)\,x^0 + f^0(t)\,, \qquad t \in [0,1]\,,$$

where $x^0 = (\Pi^0)^T x$, $f^0 = (\Pi^0)^T f$ and $A^0 = (\Pi^0)^T A\,\Pi^0$, we can find a start for the Riccati transformation that satisfies the homogeneous BCs at $t = 0$ and is well-conditioned, for t sufficiently small.

Even more reduction can be obtained by an orthogonal transformation with a generally less simple form. This is a result of

Property 4.16
By the well-posedness of the BVP there exist an orthogonal matrix $U^0 \in \mathbb{R}^{n\times n}$ *and a non-singular matrix* $V_{22}{}^0 \in \mathbb{R}^{(n-k)\times(n-k)}$ *such that*

$$B^{02}\,U^0 = \left[\,0 \;\; V_{22}{}^0\,\right]\,. \tag{31}$$

Using Assumption 2.2 on the row orthogonality of B^{02} *we can obtain* $V_{22}{}^0 = I_{n-k}$.

■

With this choice for the orthogonal transformation U^0 the system

$$\frac{dx^0}{dt} = A^0(t)\,x^0 + f^0(t)\,, \qquad t \in [0,1]\,, \tag{32}$$

where $x^0 = (U^0)^T x$, $f^0 = (U^0)^t f$ and $A^0 = (U^0)^T A\,U^0$, has at $t = 0$ the BCs $b_2 = B^{02}\,U^0\,x^0(0) = \left[\,0 \;\; V_{22}{}^0\,\right] x^0(0)$. Hence, the Riccati transformation, satisfying the homogeneous BCs at $t = 0$ starts with zero ($R_{21}(0) = 0$).

We have to realize, however, that even this starting value does not guarantee boundedness for R_{21} for a long time, as is seen in the next example.

Example 4.17
Consider the BVP with constant coefficients

$$\frac{dx}{dt} = \begin{bmatrix} -10 & 0 \\ 20 & 10 \end{bmatrix} x\,, \qquad t \in [0,1]\,,$$

subject to $x_2(0) = 1$ and $x_1(1) - x_2(1) = 0$. This is a well-conditioned problem, with an exponentially dichotomic solution space. The increasing solution $\begin{pmatrix} e^{-10t} \\ e^{10t} - e^{-10t} \end{pmatrix}$ satisfies the homogeneous BCs at $t = 0$. However, for small t its rotational acticity is high. This is illustrated by the fact that the corresponding Riccati DE, given by

$$\frac{d}{dt} R_{21} = 20 + 20\, R_{21} \ , \qquad R_{21}(0) = 0 \ ,$$

has the fast increasing solution $R_{21}(t) = e^{20t} - 1$.

■

A Riccati transformation fitted to the BCs at $t = 0$ has an interesting property, which is a direct result of (2), (4) and (6a)

Property 4.18
Assume that we have separated BCs with $B^{02} = \begin{bmatrix} 0 & V_{22}{}^0 \end{bmatrix}$ (cf. Property 4.16). *Let* ξ *be such that the solution of the Riccati DE* (5), *subject to* $R_{21}(0) = 0$, *exists on* $[0, \xi]$. *Then any solution of the (incomplete) IVP*

$$\begin{cases} \dfrac{dx}{dt} = A(t)\, x + f(t) \ , \\ B^{02}\, x(0) = b_2 \end{cases} \tag{33}$$

satisfies the relation

$$\begin{bmatrix} -R_{21}(t) & I_{n-k} \end{bmatrix} x(t) = y_2(t) \ , \qquad t \in [0, \xi] \ , \tag{34}$$

where y_2 *is the solution of the IVP*

$$\frac{dy_2}{dt} = \Big(A_{22}(t) - R_{21}(t)\, A_{12}(t) \Big)\, y_2(t) - R_{21}(t)\, f_1(t) + f_2(t) \ ,$$

subject to $V_{22}{}^0\, y_2(0) = b_2$.

■

Remark 4.19
Property 4.18 with $t = \xi$ can be interpreted as follows: with (34) the $(n - k)$ boundary conditions $B^{02}\, x(0) = b_2$ are transformed to the point ξ in the interval. So, what is left is a BVP on $[\xi, 1]$ with separated BCs. We shall take advantage of this in Chapter 6, when dealing with BVPs having a singularity of the first kind.
Similarly, we can define a Riccati matrix S_{12} and a vector function z_1 such

that, if S_{12} exists on $[\xi, 1]$, any solution x of the DE, fitting the k boundary conditions $B^{11}\, x(1) = b_1$, satisfies

$$\left[I_k \quad -S_{12}(t) \right] x(t) = z_1(t) \,, \qquad t \in [\xi, 1] \,.$$

Using both R_{21} and S_{12} we obtain the proper number of conditions to determine $x(\xi)$:

$$\begin{bmatrix} I_k & -S_{12}(\xi) \\ -R_{21}(\xi) & I_{n-k} \end{bmatrix} x(\xi) = \begin{pmatrix} z_1(\xi) \\ y_2(\xi) \end{pmatrix} .$$

Sometimes S_{12} is called the *inverse Riccati matrix* ([15]).

■

Of course, the final solution x can not be computed by just constructing a continuous description of the solution manifold corresponding to the BCs at $t = 0$. As we have seen in Section 3.2.4, some kind of *backward sweep* is needed too. For such a backward sweep we prefer the invariant imbedding formulation of Section 3.3. For separated BCs (cf. Section 3.5) this implies that (in a *forward sweep*) we have to compute the functions R_{11}, R_{12} and g_1, which are determined by the relation (cf. (3.64))

$$y_1(0) = R_{11}(t)\, y_1(t) + g_1(t) \,. \tag{35}$$

These functions satisfy the DEs (cf. (3.65a,b))

$$\frac{d}{dt} R_{11} = -R_{11} \left(A_{11}(t) + A_{12}(t)\, R_{21}(t) \right) , \quad t \geq 0 \,, \qquad R_{11}(0) = I_k \,, \tag{36a}$$

$$\frac{d}{dt} g_1 = -R_{11}(t) \left(A_{12}(t)\, y_2(t) + f_1(t) \right) , \quad t \geq 0 \,, \qquad g_1(0) = 0 \,. \tag{36b}$$

Remark 4.20

The DEs we have to solve for the Riccati transformation (R_{21} and y_2) and for the invariant imbedding technique (R_{11}, R_{12} and g_1) can be written in one $n \times (k+1)$ system:

$$\frac{d}{dt} \left[\begin{array}{c|c} R_{11} & g_1 \\ R_{21} & y_2 \end{array}\right] = \left[\begin{array}{c|c} 0 & 0 \\ A_{21}(t) & f_2(t) \end{array}\right] +$$

$$\begin{bmatrix} 0 & 0 \\ 0 & A_{22}(t) \end{bmatrix} \left[\begin{array}{c|c} R_{11} & g_1 \\ R_{21} & y_2 \end{array}\right] - \left[\begin{array}{c|c} R_{11} & g_1 \\ R_{21} & y_2 \end{array}\right] \left[\begin{array}{c|c} A_{11}(t) & f_1(t) \\ \hline 0 & 0 \end{array}\right]$$

$$-\left[\begin{array}{c|c} R_{11} & g_1 \\ R_{21} & y_2 \end{array}\right]\left[\begin{array}{cc} 0 & A_{12}(t) \\ \hline 0 & 0 \end{array}\right]\left[\begin{array}{c|c} R_{11} & g_1 \\ R_{21} & y_2 \end{array}\right].$$

When the BCs at $t = 0$ satisfy $B^{02} = \begin{bmatrix} 0 & V_{22}{}^0 \end{bmatrix}$ (cf. Property 4.16), then the initial values are given by

$$\left[\begin{array}{c|c} R_{11}(0) & g_1(0) \\ R_{21}(0) & y_2(0) \end{array}\right] = \left[\begin{array}{c|c} I_k & 0 \\ 0 & (V_{22}{}^0)^{-1} b_2 \end{array}\right].$$

This is actually how they are computed, since then no intermediate results have to be stored and interpolated.

■

4.3.2 Restarting techniques

Even the knowledge that no dominated solutions are contained in $\mathcal{R}\left(\begin{bmatrix} I_k \\ R_{21}(t) \end{bmatrix}\right)$ does not guarantee boundedness of R_{21} over the entire interval (cf. Example 4.17). Suppose that at some point within $(0,1)$, say at $t = t_1$, the Riccati matrix has grown such that we have to decide for a new basis of the subspace $\mathcal{R}\left(\begin{bmatrix} I_k \\ R_{21}(t) \end{bmatrix}\right)$. We indicate some (old and new) possibilities for such a new basis, when the Riccati transformation has been fitted to the BCs at $t = 0$, i.e., when (34) holds.

(i) Often BVPs have an equal number of BCs at both ends (for instance: a second order system that is transformed to a first order one). This implies a square Riccati matrix. If in that case $R_{21}(t_1)$ is non-singular, then the solution manifold, described by (34) (with $t = t_1$) can be written as

$$\begin{bmatrix} I_{n-k} & -S_{12}(t) \end{bmatrix} x(t) = y_2{}^1(t), \qquad t \geq t_1,$$

where the initial value of $y_2{}^1$ is given by $y_2{}^1(t_1) = -\left(R_{21}(t_1)\right)^{-1} y_2(t_1)$. Hence, for $t > t_1$, the matrix function S_{12} (which also satisfies a Riccati DE) and the vector function $y_2{}^1$ are computed instead of R_{21} and y_2. In other words: the transformation T of (4) is, for $t > t_1$, replaced by

$$T^1(t) = \begin{bmatrix} S_{12}(t) & I_k \\ I_{n-k} & 0 \end{bmatrix}.$$

Important drawbacks of this restarting technique are
- it is far from general

- problems may be expected with the numerical stability of the method
- the implementation is not straightforward.

In [9] Breitenecker discusses a generalization of this concept for non-separated BCs.

More success may be expected from one of the following techniques:

(ii) ([60]) For any permutation matrix $\Pi^1 = \left[\underset{k}{\underset{\longleftrightarrow}{\Pi_1{}^1}} \quad \underset{n-k}{\underset{\longleftrightarrow}{\Pi_2{}^1}} \right]$ such that the matrix $\left[-R_{21}(t_1) \quad I_{n-k} \right] \Pi_2{}^1$ is non-singular, we have that at $t = t_1$ the relation (34) is equivalent to

$$\left[-R_{21}{}^1(t_1) \quad I_{n-k} \right] x^1(t_1) = y_2{}^1(t_1) \,, \tag{37}$$

where

$$R_{21}{}^1(t_1) = \left(\left[-R_{21}(t_1) \quad I_{n-k} \right] \Pi_2{}^1 \right)^{-1} \left[-R_{21}(t_1) \quad I_{n-k} \right] \Pi_1{}^1 \,,$$

$$x^1(t_1) = (\Pi^1)^T x(t_1) \,,$$

and

$$y_2{}^1(t_1) = \left(\left[-R_{21}(t_1) \quad I_{n-k} \right] \Pi_2{}^1 \right)^{-1} y_2(t_1) \,.$$

As a result of Theorem 4.15, Π^1 can be chosen (and computed) such that all elements of $R_{21}{}^1(t_1)$ are in absolute value not greater than 1.
Now, for $t \geq t_1$, the original matrix function A is to be replaced by $A^1 = (\Pi^1)^T A \, \Pi^1$ and the inhomogeneous term f by $f^1 = (\Pi^1)^T f$. Hereafter, the integration continuous with R_{21} and y_2 replaced by $R_{21}{}^1$ and $y_2{}^1$, respectively.

This technique is not straightforwardly generalizable to the case of non-separated BCs.

(iii) ([35]) The permutation matrix Π^1 can be replaced by another orthogonal transformation U^1, such that a further reduction can be achieved. Choose U^1 such that

$$\left[-R_{21}(t_1) \quad I_{n-k} \right] U^1 = \left[\underset{k}{\underset{\longleftrightarrow}{0}} \quad \underset{n-k}{\underset{\longleftrightarrow}{V_{22}{}^1}} \right] \,, \tag{38}$$

where $V_{22}{}^1$ is a non-singular matrix (in fact: $V_{22}{}^1 = (U_{22}{}^1)^{-T}$). Then, with $x^1 = (U^1)^T x$ the relation (34) reads

$$\left[0 \quad V_{22}{}^1 \right] x^1(t_1) = y_2(t_1) \,. \tag{39}$$

This implies that a new Riccati matrix, say ${R_{21}}^1$, and a vector function ${y_2}^1$, can be defined with the initial values, respectively, ${R_{21}}^1(t_1) = 0$ and ${y_2}^1(t_1) = ({V_{22}}^1)^{-1} y_2(t_1)$, describing the same solution manifold, but corresponding to a new basis (namely the columns of U^1).

Such orthogonal transformations have the advantage that they can be applied in the general case of non-separated BCs too. However, the complexity of the final solution method is increased, since we have to replace the matrix function A by $A^1 = (U^1)^T A\, U^1$ and the inhomogeneous term f by $f^1 = (U^1)^T f$.

Remark 4.21
The first k columns of the orthogonal transformation U^1 of (38) form an orthogonal basis of $\mathcal{R}(\begin{bmatrix} I_k \\ R_{21}(t_1) \end{bmatrix})$. In fact: $\begin{bmatrix} U_{11} \\ U_{21} \end{bmatrix} U_{11}^{-1}$ is the QR-decomposition of $\begin{bmatrix} I_k \\ R_{21}(t_1) \end{bmatrix}$. Hence, the orthogonal update technique described above has some similarity with the Godunov-Conte algorithm ([29]); first a more or less easily computable basis of the solution subspace corresponding to the left BCs is computed. At points where this basis is found to be poorly conditioned a reorthogonalization of this basis is carried out.

■

The restart strategy for the invariant imbedding technique has been discussed in Section 3.3.2. For separated BCs it implies that, for $t \geq t_1$, we have to solve the DE

$$\frac{d}{dt}\Big[{R_{11}}^1 \,|\, {g_1}^1\Big] = -{R_{11}}^1 \Big\{\Big[{A_{11}}^1(t)\ {A_{12}}^1(t)\Big]\begin{bmatrix} I_k & 0 \\ {R_{21}}^1(t) & {y_2}^1(t) \end{bmatrix} + \Big[0 \,|\, {f_1}^1(t)\Big]\Big\},$$

subject to $\Big[{R_{11}}^1(t_1) \,|\, {g_1}^1(t_1)\Big] = \Big[I_k \,|\, 0\Big]$. This yields the relation (cf. (35))

$$ {y_1}^1(t_1) = {R_{11}}^1(t)\, {y_1}^1(t) + {g_1}^1(t)\ , \qquad t \geq t_1\ . \tag{40}$$

4.3.3 Algorithmic description

Assume a set of restart points $\{t_i\}_{i=0}^m$, with $0 = t_0 < t_1 < \cdots < t_m = 1$, are determined by the output requirements and the boundedness condition for ${R_{21}}^i$. Then, as is illustrated by Theorem 4.6, the number m will be finite. By these restarts a set of subintervals $[t_i, t_{i+1}](i = 0, \ldots, m-1)$ is generated.

For separated BCs we can give now a complete description of the *Riccati*

method, by which we mean the combination of a Riccati transformation, invariant imbedding and the orthogonal restarting technique (iii).

Algorithm 4.22

step 1. Construct an orthogonal $U^0 \epsilon \mathbb{R}^{n\times n}$ such that

$$B^{02}\,U^0 = [\underbrace{0}_{k} \quad \underbrace{V_{22}{}^0}_{n-k}] \quad (V_{22}{}^0 \text{ non-singular})$$

and set $Q^0 := U^0$ and $y_2{}^{-1}(t_0) := b_2$.

step 2. For $i = 0, 1, \ldots, m-1$ do

a. Define, for $t \epsilon [t_i, t_{i+1}]$, the transformed matrix function

$$A^i(t) := (Q^i)^T A(t)\, Q^i \tag{41a}$$

and the vector function

$$f^i(t) := (Q^i)^T f(t)\ . \tag{41b}$$

b. Solve, for $R^i := \left[\begin{array}{c|c} R_{11}{}^i & g_1{}^i \\ R_{21}{}^i & y_2{}^i \end{array}\right]$ and for $t \epsilon [t_i, t_{i+1}]$, the i-th Riccati DE

$$\frac{d}{dt}R^i = \left[\begin{array}{c|c} 0 & 0 \\ A_{21}{}^i(t) & f_2{}^i(t) \end{array}\right] + \left[\begin{array}{cc} 0 & 0 \\ 0 & A_{22}{}^i(t) \end{array}\right] R^i \tag{42}$$

$$-R^i \left[\begin{array}{c|c} A_{11}{}^i(t) & f_1{}^i(t) \\ \hline 0 & 0 \end{array}\right] - R^i \left[\begin{array}{cc} 0 & A_{12}{}^i(t) \\ \hline 0 & 0 \end{array}\right] R^i\ ,$$

subject to $R^i(t_i) = \left[\begin{array}{c|c} I_k & 0 \\ 0 & (V_{22}{}^i)^{-1} y_2{}^{i-1}(t_i) \end{array}\right]$.

c. Construct an orthogonal matrix $U^{i+1} \epsilon \mathbb{R}^{n\times n}$ and a non-singular matrix $V_{22}{}^{i+1} \epsilon \mathbb{R}^{(n-k)\times(n-k)}$ such that

$$\left[-R_{21}{}^i(t_{i+1}) \quad I_{n-k}\right] U^{i+1} = [\underbrace{0}_{k} \quad \underbrace{V_{22}{}^{i+1}}_{n-k}] \tag{43}$$

and define $Q^{i+1} := Q^i\, U^{i+1}$.

■

Defining

$$x^i = (U^i)^T x^{i-1} = (Q^i)^T x\ , \qquad (i = 0, \ldots, m) \tag{44}$$

we obtain, similarly to (39), the relation

$$\begin{bmatrix} 0 & V_{22}{}^i \end{bmatrix} x^i(t_i) = y_2{}^{i-1}(t_i) \ .$$

Observe that the initial values at $t = t_i$ are chosen such that

$$x_2{}^i(t_i) = y_2{}^i(t_i) \ , \qquad (i = 0, \ldots, m) \ . \tag{45}$$

Hence, $x_2{}^i(t_i)$ is directly obtained during the forward sweep.

When the integration step is fulfilled and $t = 1$ has been reached, then $x_1{}^m(t_m)$ can be computed with the BCs at $t = 1$:

$$\begin{aligned} b_1 \ &= \ B^{11} x(1) = B^{11} Q^m \, x^m(t_m) \\ &\Rightarrow B^{11} Q_1{}^m x_1{}^m(t_m) = b_1 - B^{11} Q_2{}^m x_2{}^m(t_m) \ . \end{aligned}$$

The $k \times k$ matrix $B^{11} Q_1{}^m$ is well-conditioned, since the BVP is well-conditioned.

Now $x_1{}^i(t_i)$ $(i = m-1, \ldots, 0)$ is determined (in a backward sweep) from the generalized recovery transformation (cf. (40)):

$$y_1{}^i(t_i) = R_{11}{}^i(t_{i+1}) \, y_1{}^i(t_{i+1}) + g_1{}^i(t_{i+1}) \ , \qquad (i = 0, \ldots, m-1) \ . \tag{46}$$

By the special form of the Riccati transformation we have $x_1{}^i(t) = y_1{}^i(t)$, for all $t \in [t_i, t_{i+1}]$. Hence, (46) can be written as

$$x_1{}^{i-1}(t_{i-1}) = R_{11}{}^{i-1}(t_i) \, x_1{}^{i-1}(t_i) + g_1{}^{i-1}(t_i) \ . \tag{47}$$

Together with

$$x_1{}^{i-1}(t_i) = \begin{bmatrix} U_{11}{}^i & U_{12}{}^i \end{bmatrix} x^i(t_i)$$

we obtain

$$x_1{}^{i-1}(t_{i-1}) = W_{11}{}^i \, x_1{}^i(t_i) + w_1{}^i \ , \qquad (i = 1, \ldots, m) \ , \tag{48}$$

where

$$W_{11}{}^i = R_{11}{}^{i-1}(t_i) \, U_{11}{}^i$$

and

$$\begin{aligned} w_1{}^i \ &= \ R_{11}{}^{i-1}(t_i) \, U_{12}{}^i \, x_2{}^i(t_i) + g_1{}^{i-1}(t_i) \\ &= \ R_{11}{}^{i-1} \, U_{12}{}^i \, y_2{}^i(t_i) + g_1{}^{i-1}(t_i) \ . \end{aligned}$$

Observe that (48) is generally a numerically stable recursion.

4.4 Non-separated BCs

If the BCs are *non-separated*, say $B^0\,x(0) + B^1\,x(1) = b$, then some steps of the algorithm for the Riccati method become slightly more complicated. For instance, the number of DEs that are to be solved is somewhat larger. Also a proper value of the integer k is generally unknown.

4.4.1 Initial values

One might hope that the eigenvalues of $A(0)$ give correct information about the growth behaviour of homogeneous modes. In general this will be the case if the Riccati matrix R_{21} is relatively close to the solution P_{21} of (22), i.e., if the conditions mentioned at the end of Section 4.2 are fulfilled.

A useful tool for the computation of eigenvalues and invariant subspaces is the Schur decomposition of Theorem 1.3. It directly delivers the eigenvalues and some of the invariant subspaces, depending on the order for the eigenvalues that has been chosen. Now construct the Schur transformation U^0 such that

$$A^0(0) = (U^0)^T A(0)\,U^0 = \begin{bmatrix} {A_{11}}^0(0) & {A_{12}}^0(0) \\ 0 & {A_{22}}^0(0) \end{bmatrix}$$

and $\lambda_{\min}\left({A_{11}}^0(0)\right) > \lambda_{\max}\left({A_{22}}^0(0)\right)$. Then the integer k is chosen such that the separation between ${A_{11}}^0(0)$ and ${A_{22}}^0(0)$ is sufficiently large.

In Section 3.4 we took $T(0) = U^0$. If ${U_{11}}^0$ is non-singular, then a similar choice for the Riccati matrix would be $R_{21}(0) = {U_{21}}^0({U_{11}}^0)^{-1}$, since then $\mathcal{R}\left(\begin{bmatrix} I_k \\ R_{21}(0) \end{bmatrix}\right) = \mathcal{R}({U_1}^0)$. This does not guarantee consistency (cf. Example 3.19), but makes it quite likely. However, $R_{21}(0)$ may be large, implying a poorly conditioned transformation T, at least for small t.

This problem can be circumvented by transforming the DE with U^0, before starting a decoupling. Hence, define $x^0 = (U^0)^T x$. Then

$$\frac{dx^0}{dt} = A^0(t)\,x^0 + f^0(t)\,, \qquad t \in [0,1]\,, \tag{49}$$

where $A^0 = (U^0)^T A\,U^0$ and $f^0 = (U^0)^T f$. This system satisfies Assumption 4.12 and therefore we may choose $R_{21}(0) = 0$ (observe that $\mathcal{R}\left(\begin{bmatrix} I_k \\ 0 \end{bmatrix}\right)$ is the invariant subspace of $A^0(0)$ corresponding to the k eigenvalues with largest real parts). This choice defines a consistent fundamental solution as soon as

$$\mathcal{R}(U_1{}^0) \cap \mathcal{S}_2(0) = \{0\} .$$

Now the Riccati transformation

$$y^0 = (T^0)^{-1} x^0 = \begin{bmatrix} I_k & 0 \\ -R_{21}{}^0 & I_{n-k} \end{bmatrix} x^0 ,$$

yields the relation

$$\left[-R_{21}{}^0(t) \;\; I_{n-k} \right] x^0(t) = y_2{}^0(t) . \tag{50}$$

However, contrary to the case of separated BCs, $y_2{}^0(0)$ is unknown, which implies that the function $y_2{}^0$ can not be computed directly. Therefore, step 2b. of Algorithm 4.22 has to be extended with the computation of a fundamental solution $Y_{22}{}^0$ corresponding to $\tilde{A}_{22}^0 = A_{22}{}^0 - R_{21}{}^0 A_{12}{}^0$ (see (6a)) and a particular solution $p_2{}^0$, satisfying the DE

$$\frac{d}{dt} p_2{}^0 = \Big(A_{22}{}^0(t) - R_{21}{}^0(t) A_{12}{}^0(t) \Big) p_2{}^0 - R_{21}{}^0(t) f_1{}^0(t) + f_2{}^0(t) ,$$

(see (6b)). With the initial values

$$\left[R_{21}{}^0(0) \;\; Y_{22}{}^0(0) \,|\, p_2{}^0(0) \right] = \left[0 \;\; I_{n-k} \,|\, 0 \right]$$

the relation (50) transforms into

$$\left[-R_{21}{}^0(t) \;\; I_{n-k} \right] x^0(t) = Y_{22}{}^0(t)\, x_2{}^0(0) + p_2{}^0(t) , \tag{51}$$

which is, apart from the transformation U^0, equivalent to the relation (8b).

4.4.2 Restarting techniques

From (51) we see that a restart can be performed similarly to the case of separated BCs (step 2c.). Assume a restart turns out to be necessary at $t = t_i$ ($i = 1, \ldots, m$). Then an orthogonal matrix $U^i \in \mathbb{R}^{n \times n}$ is constructed such that (cf. (43))

$$\left[-R_{21}{}^{i-1}(t_i) \;\; I_{n-k} \right] U^i = \left[0 \;\; V_{22}{}^i \right] , \tag{52}$$

with $V_{22}{}^i$ non-singular (again: $V_{22}{}^i = (U_{22}{}^i)^{-T}$). Defining

$$x^i = (U^i)^T x^{i-1} , \tag{53}$$

we obtain, on each subinterval $[t_{i-1}, t_i]$, the equivalent of (51), namely

$$\left[-R_{21}{}^{i-1}(t) \;\; I_{n-k}\right] x^{i-1}(t) = Y_{22}{}^{i-1}(t)\, x_2{}^i(t_{i-1}) + p_2{}^{i-1}(t) \; . \tag{54}$$

Combining the relations (52)–(54) yields the recursion

$$x_2{}^i(t_i) = W_{22}{}^{i-1} x_2{}^{i-1}(t_{i-1}) + w_2{}^{i-1} \; , \tag{55}$$

where

$$W_{22}{}^{i-1} = (V_{22}{}^i)^{-1} Y_{22}{}^{i-1}(t_i) = (U_{22}{}^i)^T Y_{22}{}^{i-1}(t_i) \; ,$$

$$w_2{}^{i-1} = (V_{22}{}^i)^{-1} p_2{}^{i-1}(t_i) = (U_{22}{}^i)^T p_2{}^{i-1}(t_i) \; .$$

Since during the forward sweep the value of $x_2{}^i(t_i)$ is unknown, the invariant imbedding relation (47) has to be generalized too. According to (8a) we have to compute, on each subinterval $[t_{i-1}, t_i]$, the matrix functions $R_{11}{}^{i-1}$ and $R_{12}{}^{i-1}$ and the vector function $g_1{}^{i-1}$ such that

$$x_1{}^{i-1}(t_{i-1}) = R_{11}{}^{i-1}(t_i)\, x_1{}^{i-1}(t_i) + R_{12}{}^{i-1}(t_i)\, x_2{}^{i-1}(t_{i-1}) + g_1{}^{i-1}(t_i) \; .$$

Hereafter we obtain the recursion

$$\begin{aligned} x_1{}^{i-1}(t_{i-1}) &= R_{11}{}^{i-1}(t_i) \left[U_{11}{}^i \;\; U_{12}{}^i\right] x^i(t_i) + \\ &\qquad R_{12}{}^{i-1}(t_i)\, x_2{}^{i-1}(t_{i-1}) + g_1{}^{i-1}(t_i) \\ &= W_{11}{}^i\, x_1{}^i(t_i) + W_{12}{}^i\, x_2{}^{i-1}(t_{i-1}) + w_1{}^i \; , \end{aligned} \tag{56}$$

where

$$W_{11}{}^i = R_{11}{}^{i-1}(t_i)\, U_{11}{}^i \; ,$$

$$W_{12}{}^i = R_{11}{}^{i-1}(t_i)\, U_{12}{}^i\, W_{22}{}^{i-1} + R_{12}{}^{i-1}(t_i)$$

and

$$w_1{}^i = R_{11}{}^{i-1}(t_i)\, U_{12}{}^i\, w_2{}^{i-1} + g_1{}^{i-1}(t_i) \; .$$

The recursions (55) and (56) contain relations between the solution x at consecutive points. This implies a relation between $x^0(0)$ and $x^m(1)$. Such a relation can, for instance, be obtained by the construction of a sequence of matrices $\{F^i\}_{i=0}^m$ and vectors $\{d^i\}_{i=0}^m$ such that (cf. (3.60))

$$x_2{}^i(t_i) = F_{22}{}^i\, x_2{}^0(t_0) + d_2{}^i \; , \tag{57a}$$

$$x_1{}^i(t_i) = F_{11}{}^i\, x_1{}^m(t_m) + F_{12}{}^i\, x_2{}^0(t_0) + d_1{}^i \; . \tag{57b}$$

One immediately derives that the F^i's and d^i's have to satisfy the recursions:

for $i = 0, \ldots, m-1$ (forward sweep)

$$F_{22}{}^{i+1} = W_{22}{}^{i} F_{22}{}^{i} , \quad F_{22}{}^{0} = I_{n-k} , \tag{58a}$$

$$d_2{}^{i+1} = W_{22}{}^{i} d_2{}^{i} + w_2{}^{i} , \quad d_2{}^{0} = 0 , \tag{58b}$$

for $i = m, \ldots, 1$ (backward sweep)

$$F_{11}{}^{i-1} = W_{11}{}^{i} F_{11}{}^{i} , \quad F_{11}{}^{m} = I_k , \tag{58c}$$

$$F_{12}{}^{i-1} = W_{11}{}^{i} F_{12}{}^{i} + W_{12}{}^{i} F_{22}{}^{i-1} , \quad F_{12}{}^{m} = 0 , \tag{58d}$$

$$d_1{}^{i-1} = W_{11}{}^{i} d_1{}^{i} + W_{12}{}^{i} d_2{}^{i-1} + w_1{}^{i} , \quad d_1{}^{m} = 0 . \tag{58e}$$

One easily verifies with an induction argument the following

Property 4.23
For $i = 0, \ldots, m$ we have the relations

$$F_{22}{}^{i} = \prod_{k=0}^{i-1} W_{22}{}^{k} ,$$

$$d_2{}^{i} = \sum_{j=0}^{i-1} \Big(\prod_{k=j+1}^{i-1} W_{22}{}^{k} \Big) w_2{}^{j} ,$$

$$F_{11}{}^{i} = \prod_{k=0}^{m-i-1} W_{11}{}^{m-k} ,$$

$$F_{12}{}^{i} = \sum_{j=0}^{m-i-1} \Big(\prod_{k=j+1}^{m-i-1} W_{11}{}^{m-k} \Big) W_{12}{}^{m-j} F_{22}{}^{m-j-1} ,$$

$$d_1{}^{i} = \sum_{j=0}^{m-i-1} \Big(\prod_{k=j+1}^{m-i-1} W_{11}{}^{m-k} \Big) \Big(W_{12}{}^{m-j} d_2{}^{m-j-1} + w_2{}^{m-j-1} \Big) .$$

■

The values of $x_2{}^{0}(t_0)$ and $x_1{}^{m}(t_m)$ can be determined from the BCs:

$$B^0 x(0) + B^1 x(1) = b \tag{59}$$

$$\Rightarrow \quad B^0 Q^0 x^0(t_0) + B^1 Q^m x^m(t_m) = b$$

$$\Rightarrow \quad \left(B^0 Q^0 \begin{bmatrix} F_{11}{}^{0} & F_{12}{}^{0} \\ 0 & I_{n-k} \end{bmatrix} + B^1 Q^m \begin{bmatrix} I_k & 0 \\ 0 & F_{22}{}^{m} \end{bmatrix} \right) \begin{pmatrix} x_1{}^{m}(t_m) \\ x_2{}^{0}(t_0) \end{pmatrix}$$

$$= \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} - B^0 Q^0 \begin{pmatrix} {d_1}^0 \\ 0 \end{pmatrix} - B^1 Q^m \begin{pmatrix} 0 \\ {d_2}^m \end{pmatrix} . \tag{60}$$

Define, for $i = 0, \ldots, m$,

$$F^i = \begin{bmatrix} {F_{11}}^i & {F_{12}}^i \\ 0 & {F_{22}}^i \end{bmatrix} \quad \text{and} \quad d^i = \begin{pmatrix} {d_1}^i \\ {d_2}^i \end{pmatrix} . \tag{61}$$

Observe that $F^0 = \begin{bmatrix} {F_{11}}^0 & {F_{12}}^0 \\ 0 & I_{n-k} \end{bmatrix}$ and $F^m = \begin{bmatrix} I_k & 0 \\ 0 & {F_{22}}^m \end{bmatrix}$.
Then the solution x of the original problem is, at $t = t_i$ ($i = 0, \ldots, m$), given by

$$x(t_i) = Q^i x^i(t_i) = Q^i F^i \begin{pmatrix} {x_1}^m(t_m) \\ {x_2}^0(t_0) \end{pmatrix} + Q^i d^i . \tag{62}$$

Inspired by this expression we make the following observation.

Property 4.24
Let $Z = \underset{\substack{\longleftrightarrow \\ k}}{Z_1} \; \underset{\substack{\longleftrightarrow \\ n-k}}{Z_2}$ *be a fundamental solution, satisfying*

$$\frac{d}{dt} Z = A(t)\, Z , \qquad t \in [0,1] ,$$

subject to $Z(0) = Q^0 F^0$.
Then, for $i = 0, \ldots, m$, *we have*

$$Q^i F^i = Z(t_i) . \tag{63}$$

Equivalently, define the particular solution z *by*

$$\frac{dz}{dt} = A(t)\, z + f(t) , \qquad t \in [0,1] ,$$

with $z(0) = Q^0 d^0$.
Then, for $i = 0, \ldots, m$, *we have*

$$Q^i d^i = z(t_i) . \tag{64}$$

Proof:
Observe that the initial value $Z(0)$ is such that (63) holds for $i = 0$. Suppose it is also valid for some $i = l - 1$, i.e.,

$$Z(t_{l-1}) = Q^{l-1} F^{l-1} .$$

Define, for $t \in [t_{l-1}, t_l]$, the matrix functions

$$Y^{l-1}(t) = (Q^{l-1})^T Z(t) ,$$
$$A^{l-1}(t) = (Q^{l-1})^T A(t)\, Q^{l-1} .$$

Then,

$$\frac{d}{dt}Y^{l-1} = A^{l-1}(t)\, Y^{l-1} , \qquad t \epsilon [t_{l-1}, t_l] ,$$

and $Y^{l-1}(t_{l-1}) = F^{l-1}$.
From Property 3.19 we conclude that

$$\begin{bmatrix} I_k & 0 \\ {R_{21}}^{l-1}(t) & {Y_{22}}^{l-1}(t) \end{bmatrix} \begin{bmatrix} {R_{11}}^{l-1}(t) & {R_{12}}^{l-1}(t) \\ 0 & I_{n-k} \end{bmatrix}^{-1}$$

is the fundamental solution corresponding to A^{l-1}, which is the identity at $t = t_{l-1}$. Hence, for all $t \epsilon [t_{l-1}, t_l]$,

$$Y^{l-1}(t) = \begin{bmatrix} I_k & 0 \\ {R_{21}}^{l-1}(t) & {Y_{22}}^{l-1}(t) \end{bmatrix} \begin{bmatrix} {R_{11}}^{l-1}(t) & {R_{12}}^{l-1}(t) \\ 0 & I_{n-k} \end{bmatrix}^{-1} F^{l-1} .$$

The definition of U^l is such that, cf. (52),

$$\begin{bmatrix} I_k & 0 \\ {R_{21}}^{l-1}(t) & {Y_{22}}^{l-1}(t) \end{bmatrix} = U^l \begin{bmatrix} ({U_{11}}^l)^{-1} & ({U_{21}}^l)^T {Y_{22}}^{l-1}(t_l) \\ 0 & ({U_{22}}^l)^T {Y_{22}}^{l-1}(t_l) \end{bmatrix} .$$

So we obtain,

$$Y^{l-1}(t_l) = U^l \begin{bmatrix} (W_{11}^l)^{-1} & -(W_{11}^l)^{-1} {R_{12}}^{l-1}(t_l) + (U_{21}^l)^T {Y_{22}}^{l-1}(t_l) \\ 0 & W_{22}^{l-1} \end{bmatrix} F^{l-1} .$$

Using the definition of ${W_{12}}^l$ and the orthogonality of U^l this reduces to

$$Y^{l-1}(t_l) = U^l \begin{bmatrix} ({W_{11}}^l)^{-1} & -({W_{11}}^l)^{-1} {W_{12}}^l \\ 0 & {W_{22}}^{l-1} \end{bmatrix} F^{l-1} = U^l F^l .$$

Therefore,

$$Z(t_l) = Q^{l-1} Y^{l-1}(t_l) = Q^l F^l .$$

By induction result (63) is proved.
The correctness of (64) is seen by the following observation. For all $t \epsilon [0, 1]$ we have (cf. (62))

$$x(t) = Z(t) \begin{pmatrix} {x_1}^m(t_m) \\ {x_2}^0(t_0) \end{pmatrix} + z(t) .$$

Hence,

$$z(t_i) = x(t_i) - Z(t_i)\begin{pmatrix} x_1{}^m(t_m) \\ x_2{}^0(t_0) \end{pmatrix} = Q^i\, d^i\ .$$

■

As a result of this property we conclude that the Riccati method, as it is sketched above, finally delivers, at a discrete set of points $\{t_i\}$, the values of a fundamental and a particular solution of the original problem. At all points t_i a block QR-decomposition of that fundamental solution is available. In often prevailing circumstances the columns of the fundamental solution will be reasonably scaled. We shall return to this aspect of the Riccati method in Section 4.5 .

4.4.3 Algorithmic description

In this subsection we algorithmically describe the Riccati method for general BCs, as it was discussed in the foregoing subsections. So, consider the $n \times n$ BVP

$$\frac{dx}{dt} = A(t)\,x + f(t)\ , \qquad t \epsilon [0,1]\ , \tag{65}$$

subject to the (non-separated) BCs

$$B^0\,x(0) + B^1\,x(1) = b\ . \tag{66}$$

The $n \times 2n$ matrix $\left[\, B^0 \mid B^1 \,\right]$ is assumed to have orthogonal rows. The solution x is required at the $q+1$ output points

$$0 = \xi_0 < \xi_1 < \cdots < \xi_q = 1\ .$$

Algorithm 4.25

step 1. Initialization part

a. Compute an orthogonal matrix $U^0 \epsilon \mathbb{R}^{n \times n}$ such that $(U^0)^T A(0)\, U^0$ is (block) upper triangular and correctly ordered (see Assumption 4.12).

b. Determine a partitioning integer k, $1 \leq k < n$. (If no other information is available one may use as guideline the eigenvalues of $A(0)$).

c. Set $i := 0$, $j := 0$, $t := t_0 := \xi_0$ and $Q^0 := U^0$.

step 2. Integration part

While $j < q$ do

a. Set $A^i := (Q^i)^T A\, Q^i$, $f^i := (Q^i)^T f$ and $x^i := (Q^i)^T x$.

(Then (65) changes into $\frac{d}{dt}x^i = A^i(t)\, x^i + f^i(t)$).

b. Solve, for $R^i = \left[\begin{array}{cc|c} R_{11}{}^i & R_{12}{}^i & g_1{}^i \\ R_{21}{}^i & Y_{22}{}^i & p_2{}^i \end{array}\right]$ and $t \geq t_i$, the Riccati DE

$$\frac{d}{dt}R^i = \left[\begin{array}{cc|c} 0 & 0 & 0 \\ A_{21}{}^i(t) & 0 & f_2{}^i(t) \end{array}\right] + \left[\begin{array}{cc} 0 & 0 \\ 0 & A_{22}{}^i(t) \end{array}\right] R^i \qquad (67)$$

$$-R^i \left[\begin{array}{cc|c} A_{11}{}^i(t) & 0 & f_1{}^i(t) \\ 0 & 0 & 0 \\ \hline 0 & 0 & 0 \end{array}\right] - R^i \left[\begin{array}{cc|c} 0 & A_{12}{}^i(t) & \\ 0 & 0 & \\ \hline 0 & 0 & \end{array}\right] R^i ,$$

subject to

$$R^i(t_i) = \left[\begin{array}{cc|c} I_k & 0 & 0 \\ 0 & I_{n-k} & 0 \end{array}\right] ,$$

until $t = \xi_{j+1}$ or $\left|R^i(t)\right| = a$ (for some given constant $a > 1$).
Set $t_{i+1} := t$.

c. Construct an orthogonal matrix $U^{i+1} \in \mathbb{R}^{n\times n}$ and a non-singular $V_{22}{}^{i+1} \in \mathbb{R}^{(n-k)\times(n-k)}$ such that

$$\left[-R_{21}{}^i(t_{i+1}) \;\; I_{n-k}\right] U^{i+1} = \left[0 \;\; V_{22}{}^{i+1}\right] .$$

d. Generate $\left[W_{22}{}^i \mid w_2{}^i\right]$ by

$$W_{22}{}^i := (V_{22}{}^{i+1})^{-1} Y_{22}{}^i(t_{i+1}) = (U_{22}{}^{i+1})^T\, Y_{22}{}^i(t_{i+1})$$

$$w_2{}^i := (V_{22}{}^{i+1})^{-1} p_2{}^i(t_{i+1}) = (U_{22}{}^{i+1})^T p_2{}^i(t_{i+1})$$

and $\left[W_{11}{}^{i+1} \;\; W_{12}{}^{i+1} \mid w_1{}^{i+1}\right]$ by

$$W_{11}{}^{i+1} := R_{11}{}^i(t_{i+1})\, U_{11}{}^{i+1}$$

$$W_{12}{}^{i+1} := R_{11}{}^i(t_{i+1})\, U_{12}{}^{i+1}\, W_{22}{}^i + R_{12}{}^i(t_{i+1})$$

$$w_1{}^{i+1} := R_{11}{}^i(t_{i+1})\, U_{12}{}^{i+1}\, w_2{}^i + g_1{}^i(t_{i+1})$$

e. Set $i := i + 1$; if $t_i = \xi_{j+1}$ then $j := j + 1$; $Q^i := Q^{i-1}U^i$.

step 3. Completion part

a. Set $m = i$ (and observe that $t_m = 1$).

b. Generate, for $i = 0, \ldots, m$, the matrices ${F_{22}}^i$ and the vectors ${d_2}^i$ by the forward sweep

$$
\begin{aligned}
&{F_{22}}^{i+1} = {W_{22}}^i \, {F_{22}}^i \;, \quad {F_{22}}^0 = I_{n-k} \\
&{d_2}^{i+1} = {W_{22}}^i \, {d_2}^i + {w_2}^i \;, \quad {d_2}^0 = 0
\end{aligned}
$$

and, for $i = m, \ldots, 0$, the matrices $\left[{F_{11}}^i \;\; {F_{12}}^i \right]$ and the vectors ${d_1}^i$ by the backward sweep

$$
\begin{aligned}
&{F_{11}}^{i-1} = {W_{11}}^i \, {F_{11}}^i \;, \quad {F_{11}}^m = I_k \\
&{F_{12}}^{i-1} = {W_{11}}^i \, {F_{12}}^i + {W_{12}}^i \, {F_{22}}^{i-1} \;, \quad {F_{12}}^m = 0 \\
&{d_1}^{i-1} = {W_{11}}^i \, {d_1}^i + {W_{12}}^i \, {d_2}^{i-1} + {w_1}^i \;, \quad {d_1}^m = 0 \;.
\end{aligned}
$$

c. Compute the values of ${x_2}^0(t_0)$ and ${x_1}^m(t_m)$ from the BCs:

$$
\left(B^0 \, Q^0 \begin{bmatrix} {F_{11}}^0 & {F_{12}}^0 \\ 0 & I_{n-k} \end{bmatrix} + B^1 \, Q^m \begin{bmatrix} I_k & 0 \\ 0 & {F_{22}}^m \end{bmatrix} \right) \begin{pmatrix} {x_1}^m(t_m) \\ {x_2}^0(t_0) \end{pmatrix}
$$

$$
= \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} - B^0 \, Q^0 \begin{pmatrix} {d_1}^0 \\ 0 \end{pmatrix} - B^1 \, Q^m \begin{pmatrix} 0 \\ {d_2}^0 \end{pmatrix} .
$$

d. Compute $x^i(t_i)$ $(i = 1, \ldots, m)$ by

$$
\begin{pmatrix} {x_1}^i \\ {x_2}^i \end{pmatrix} = \begin{bmatrix} {F_{11}}^i & {F_{12}}^i \\ 0 & {F_{22}}^i \end{bmatrix} \begin{pmatrix} {x_1}^m(t_m) \\ {x_2}^0(t_0) \end{pmatrix} + \begin{pmatrix} {d_1}^i \\ {d_2}^i \end{pmatrix} .
$$

e. Set $x(t_i) = Q^i \, x^i(t_i)$ $(i = 0, \ldots, m)$.

■

By the integration of the $n \times n$ system of DEs (67) the original continuous problem has been replaced by a discrete problem. This discrete problem finally reduces to solving the $n \times n$ linear system (60) and the linear recursions (57a,b). The condition of the system (60) and the stability of these recursions will be discussed in the next section.

4.5 Computational aspects

The Algorithms 4.22 and 4.25 have been implemented in the package RICCATI. In this section we shall discuss some computational aspects of the method and details of the implementation of Algorithm 4.25.

4.5.1 Initialization

The first step in Algorithm 4.25 is just a minor one, although the effect of a bad choice for the partitioning integer k may be dramatic. In this subsection we only remark that in RICCATI U^0 is computed by an adapted version of the routine HQR3 of Stewart ([58]).

4.5.2 Integration

The most expensive part of the algorithm is step 2b. On $[t_i, t_{i+1}]$ we have to solve the i-th Riccati DE:

$$\begin{cases} \frac{d}{dt}R_{21}{}^i = A_{21}{}^i(t) + A_{22}{}^i(t)\,R_{21}{}^i - R_{21}{}^i\,A_{11}{}^i(t) - R_{21}{}^i\,A_{12}{}^i(t)\,R_{21}{}^i \\ R_{21}{}^i(t_i) = 0 \end{cases} . \tag{68}$$

Furthermore, with $\tilde{A}^i_{22} = A_{22}{}^i - R_{21}{}^i\,A_{12}{}^i$ and $\tilde{A}^i_{11} = A_{11}{}^i + A_{12}{}^i\,R_{21}{}^i$, we must solve the linear DEs:

$$\frac{d}{dt}Y_{22}{}^i = \tilde{A}^i_{22}(t)\,Y_{22}{}^i\ , \quad Y_{22}{}^i(t_i) = I_{n-k} \tag{69a}$$

$$\frac{d}{dt}p_2{}^i = \tilde{A}^i_{22}(t)\,p_2{}^i - R_{21}{}^i(t)\,f_1{}^i(t) + f_2{}^i(t)\ , \quad p_2{}^i(t_i) = 0 \tag{69b}$$

$$\frac{d}{dt}R_{11}{}^i = -R_{11}{}^i\,\tilde{A}^i_{11}(t)\ , \quad R_{11}{}^i(t_i) = I_k \tag{69c}$$

$$\frac{d}{dt}R_{12}{}^i = -R_{11}{}^i(t)\,A_{12}{}^i(t)\,Y_{22}{}^i(t)\ , \quad R_{12}{}^i(t_i) = 0 \tag{69d}$$

$$\frac{d}{dt}g_1{}^i = -R_{11}{}^i(t)\Big(A_{12}{}^i(t)\,p_2{}^i(t) + f_1{}^i(t)\Big)\ , \quad g_1{}^i(t_i) = 0\ . \tag{69e}$$

These DEs may be stiff or non-stiff, where a stiff problem is one in which the solution components of interest are slowly varying, but solutions with very rapidly changing components are possible ([55], p.127). One of the better implementations of a solver that can handle both types of problems is the routine LSODA from ODEPACK ([25]). For non-stiff problems it uses an implicit Adams-method, whereas for stiff problems the so-called Backward Differentiation Formulas (BDF) are used. The routine LSODA has the nice feature that it automatically switches from one method to the other and vica versa. In many applications the integrator works satisfactory, but there are situations (see Example 5.26) in which it fails completely. We shall return to this aspect in Section 5.3.

The accuracy of the result can be controled by the $n \times (n+1)$ matrices $RTOL$ and $ATOL$, containing relative and absolute error tolerance parameters, respectively. The integration routine will choose its stepsizes such that all components of the $n \times (n+1)$ matrix EST of approximated local errors in R, satisfy the inequality

$$EST_{ij} \leq RTOL_{ij} * |R_{ij}| + ATOL_{ij} \qquad (1 \leq i \leq n,\ 1 \leq j \leq n+1)\ . \quad (70)$$

Since we want to consider the integration routine as a black box most parts of the code have been left unaltered. Just some minor updates, increasing its efficiency by using matrix arithmetic, have been performed. These updates are mainly found in the Newton-like step, which computes the correction term in a BDF-method. The only non-linear DE to be solved is the Riccati DE (68). The Jacobian J^i of (68) is given by

$$J^i(.) = \tilde{A}^i_{22} \,.\, - \,.\, \tilde{A}^i_{11} \quad . \qquad (71)$$

Observe that $\tilde{A}^i_{11}$ and $\tilde{A}^i_{22}$ have been computed already in order to solve (69a-e). For computing the correction term in a BDF-method we have to solve a Sylvester equation of the form

$$(I_{n-k} - \mu\, h\, \tilde{A}^i_{22})\, \triangle X_{21} + \mu\, h\, \triangle X_{21}\, \tilde{A}^i_{11} = h\, D_{21}\ , \qquad (72)$$

where h is the suggested stepsize, μ some constant, depending on the order of the method, $\triangle X_{21}$ the correction term in the approximation of the solution R_{21} and the matrix D_{21} is computed from earlier approximations and corresponding function values.

If the Riccati transformation decouples correctly, then $\lambda(\tilde{A}^i_{22})$ will be in $\overline{\mathbb{C}^-}$ and $\lambda(\tilde{A}^i_{11})$ in $\mathbb{C}^+$. This implies that (72) will have a unique (and small) solution (cf. Theorem 1.9).
To solve (72) we transform (using orthogonal transformations) $I_{n-k} - \mu h\, \tilde{A}^i_{22}$ to Hessenberg form and $\tilde{A}^i_{11}$ to (2×2 block) upper triangular form (cf. [21]). Hereafter the solution $\triangle X_{21}$ of (72) is found columnwise, where for each column a (2×2 block) upper triangular system has to be solved. These nice (almost) upper triangular forms are also useful when (69a-e) are solved. Moreover, it gives us the opportunity to compute the eigenvalues of $\tilde{A}^i_{22}$ in a simple way. These eigenvalues give, in general, a nice indication of the (local) growth behaviour of the fundamental solution Y_{22}.

Which value should be given to the upper bound constant a is hard to say. However, it seems better to choose a not too large, for instance $a = 3$ or $a = 5$.

Large values of a may force the integrator to reduce the stepsize, without preventing an additional restart.

4.5.3 Orthogonalization

In analyzing step 2c we make the following observation:

Property 4.26
Let $R_{21} \in \mathbb{R}^{(n-k)\times k}$. *Let* $U = \begin{bmatrix} \underbrace{U_1}_{k} & \underbrace{U_2}_{n-k} \end{bmatrix}$ *be orthogonal and such that*

$$\begin{bmatrix} -R_{21} & I_{n-k} \end{bmatrix} U = \begin{bmatrix} 0 & V_{22} \end{bmatrix} . \tag{73}$$

Then

(i) *the columns of* U_1 *form an orthogonal basis of* $\mathcal{R}\left(\begin{bmatrix} I_k \\ R_{21} \end{bmatrix}\right)$

(ii) $V_{22} = {U_{22}}^{-T}$.

Proof:

(i) From $\begin{bmatrix} -R_{21} & I_{n-k} \end{bmatrix} U_1 = 0$ we obtain that $\begin{bmatrix} I_k \\ R_{21} \end{bmatrix} = U_1 V_{11}$, for some non-singular $V_{11} \in \mathbb{R}^{k\times k}$.

(ii) Extension of (73) gives

$$\begin{bmatrix} I_k & 0 \\ -R_{21} & I_{n-k} \end{bmatrix} U = \begin{bmatrix} U_{11} & U_{12} \\ 0 & V_{22} \end{bmatrix} ,$$

from which we obtain

$$U^T = U^{-1} = \begin{bmatrix} U_{11} & U_{12} \\ 0 & V_{22} \end{bmatrix}^{-1} \begin{bmatrix} I_k & 0 \\ R_{21} & I_{n-k} \end{bmatrix} .$$

This directly yields the required result.

■

The relation (73) only prescribes the subspaces $\mathcal{R}(U_1)$ and $\mathcal{R}(U_2)$. However, we are still free in choosing an orthogonal basis for these subspaces. Hence, each U satisfying (73) may be post-multiplied by a block-diagonal orthogonal matrix, say $Z = \begin{bmatrix} Z_{11} & 0 \\ 0 & Z_{22} \end{bmatrix}$. Then UZ still satisfies (73). In principle, Z_{11}

may be chosen such that $W_{11} = R_{11}\,U_{11}\,Z_{11}$ has a nice form, for instance upper triangular. In the same way Z_{22} may be chosen such that $W_{22} = {Z_{22}}^T {U_{22}}^T Y_{22}$ has an upper triangular form. Although these combinations can be combined partly with the determination of U, it does not seem worth-while the extra operations, since the algorithm is constructed such that none of the matrices W_{11} and W_{22} has to be inverted. However, for the stiff BVPs that will be discussed in Chapter 5, these upper triangular forms may be useful. Like in other multiple shooting techniques the diagonal elements of W_{22} indicate, in general, which solutions have been damped out and which have not. If it turns out that parts of W_{22} are negligible small, then the corresponding columns of Y_{22} (and R_{12}) can be skipped from further computations. The same is true for W_{11}.

Another important aspect in the suggested construction of the matrices $W^i = \begin{bmatrix} {W_{11}}^i & {W_{12}}^i \\ 0 & {W_{22}}^i \end{bmatrix}$ is the memory access of the computer. Observe that after the computation of W^i the matrices ${R_{11}}^i$, ${R_{12}}^i$ and ${Y_{22}}^i$ are not needed anymore. Hence, the memory locations of these matrices can be used to store the matrices W^i. This implies that no extra positions in the memory of the computer are occupied.

Probably this is the best place to remark that in the completion part of the algorithm the matrices W^i are overwritten by the matrices F^i, which are determined in step 3b.

Of course, similar remarks can be made for, respectively, ${w_1}^i$, ${w_2}^i$ and ${d_1}^i$, ${d_2}^i$. These are stored in the memory locations of ${g_1}^i$ and ${p_2}^i$.

4.5.4 The computation of ${x_1}^m(t_m)$ and ${x_2}^0(t_0)$

In this subsection we investigate the condition of the system (60):

$$\left[B^0\,Z(0) + B^1\,Z(1)\right]\begin{pmatrix} {x_1}^m(t_m) \\ {x_2}^0(t_0) \end{pmatrix} = b - B^0\,z(0) - B^1\,z(1)\ . \tag{74}$$

We first remark that, since Z is a fundamental solution and the original problem is well-posed, the matrix $B^0\,Z(0) + B^1\,Z(1)$ is non-singular (cf. Theorem 2.1). In Section 2.1 we have seen already that the sensitivity of the solution x for changes in the BCs is given by the stability constant

$$\beta = \max_{0\le t\le 1} \| \ Z(t)\left(B^0\,Z(0) + B^1\,Z(1)\right)^{-1} \| \ . \tag{75}$$

Now we have

Theorem 4.27

$$\| \left(B^0 Z(0) + B^1 Z(1)\right)^{-1} \| \leq \sqrt{2}\beta \ .$$

Proof:
Let x be such that $\| \left(B^0 Z(0) + B^1 Z(1)\right) x \| = \text{glb}\left(B^0 Z(0) + B^1 Z(1)\right)$ and $\| x \| = 1$. Then

$$\begin{aligned}
\beta &\geq \max_i \{\| Z(t_i) \left(B^0 Z(0) + B^1 Z(1)\right)^{-1} \|\} \\
&\geq \max_i \left\{ \frac{\| Z(t_i)\, x \|}{\| \left(B^0 Z(0) + B^1 Z(1)\right) x \|} \right\} \\
&= \max_i \left\{ \frac{\| Z(t_i)\, x \|}{\text{glb}\left(B^0 Z(0) + B^1 Z(1)\right)} \right\} \\
&= \| \left(B^0 Z(0) + B^1 Z(1)\right)^{-1} \| \ \max_i \{\| Z(t_i)\, x \|\} \\
&\geq \| \left(B^0 Z(0) + B^1 Z(1)\right)^{-1} \| \ \max \{\| F^0 x \|, \| F^m x \|\} \\
&= \| \left(B^0 Z(0) + B^1 Z(1)\right)^{-1} \| \\
&\qquad \max \left\{ \| \begin{bmatrix} {F_{11}}^0 & {F_{12}}^0 \\ 0 & I_{n-k} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \|, \| \begin{bmatrix} I_k & 0 \\ 0 & {F_{22}}^m \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \| \right\} \\
&\geq \| \left(B^0 Z(0) + B^1 Z(1)\right)^{-1} \| \ \max \{\| x_2 \|, \| x_1 \|\} \\
&\geq \frac{1}{2}\sqrt{2} \| \left(B^0 Z(0) + B^1 Z(1)\right)^{-1} \| \ .
\end{aligned}$$

■

Well-conditioning of $B^0 Z(0) + B^1 Z(1)$ is obtained as soon as the quantities $\| F^0 \|$ and $\| F^m \|$ are sufficiently bounded. This follows from Theorem 4.27 together with

Theorem 4.28

$$\| B^0 Z(0) + B^1 Z(1) \| \leq$$
$$(1 + \| {F_{12}}^0 \|) \max \{1, \| {F_{11}}^0 \|\} + \max \{1, \| {F_{22}}^m \|\} \ .$$

Proof:
Using the row orthogonality of $\left[B^0 \mid B^1 \right]$ and relation (63) we obtain

$$\begin{aligned} \| B^0 Z(0) + B^1 Z(1) \| &= \| \left[B^0 \mid B^1 \right] \begin{bmatrix} Z(0) \\ Z(1) \end{bmatrix} \| \leq \| \begin{bmatrix} Z(0) \\ Z(1) \end{bmatrix} \| \\ &= \| \begin{bmatrix} F^0 \\ F^m \end{bmatrix} \| \leq \| F^0 \| + \| F^m \| \ . \end{aligned}$$

Write

$$F^0 = \begin{bmatrix} {F_{11}}^0 & {F_{12}}^0 \\ 0 & I_{n-k} \end{bmatrix} = \begin{bmatrix} I_k & {F_{12}}^0 \\ 0 & I_{n-k} \end{bmatrix} \begin{bmatrix} {F_{11}}^0 & 0 \\ 0 & I_{n-k} \end{bmatrix} .$$

Observe that

$$\| \begin{bmatrix} I_k & {F_{12}}^0 \\ 0 & I_{n-k} \end{bmatrix} \| \leq \ (1 + \| {F_{12}}^0 \|) \ .$$

Hence,

$$\begin{aligned} & \| B^0 Z(0) + B^1 Z(1) \| \\ & \leq \ (1 + \| {F_{12}}^0 \|) \ \| \begin{bmatrix} {F_{11}}^0 & 0 \\ 0 & I_{n-k} \end{bmatrix} \| + \| \begin{bmatrix} I_k & 0 \\ 0 & {F_{22}}^m \end{bmatrix} \| \\ & \leq \ (1 + \| {F_{12}}^0 \|) \ \max \{ 1, \| {F_{11}}^0 \| \} + \max \{ 1, \| {F_{22}}^m \| \} \ . \end{aligned}$$

■

From the foregoing results we see why it is important to choose U^0 such that $\| {F_{ij}}^l \|$ $(i, j = 1, 2;\ l = 0, m)$ are sufficiently bounded. Since $Z(0) = U^0 \begin{bmatrix} {F_{11}}^0 & {F_{12}}^0 \\ 0 & I_{n-k} \end{bmatrix}$ we have $Z_2(0) = U_1^0 {F_{12}}^0 + U_2^0$. Using the Properties 1.13 and 1.14 we directly obtain

Lemma 4.29

$$\mathrm{GAP}\left({U_1}^0, Z_2(0) \right) = \frac{1}{\sqrt{1 + \| {F_{12}}^0 \|^2}} \ .$$

■

How far these results are really useful depends on the character of the original problem. Suppose that the solution space $\mathcal{S}$ has an exponential dichotomy to which the integer k corresponds, i.e., the dominant solution subspace $\mathcal{S}_1$

is k-dimensional. Then we may expect, as a result of Theorem 2.22, that $\| F_{11}{}^0 \| \approx k_1 e^{-\lambda_1}$ and $\| F_{22}{}^m \| \approx k_2 e^{-\lambda_2}$ (see Definition 2.4). As has been indicated already by Lemma 4.29 the magnitude of $\| F_{12}{}^0 \|$ is determined by the initial choice U^0. Let $Q^i = \begin{bmatrix} Q_1{}^i & Q_2{}^i \end{bmatrix}$ $(i = 0, \ldots, m)$. If U^0 is consistent at $t = 0$ then $\mathcal{R}\Big(Z_1(1)\Big)$ $\Big(= \mathcal{R}(Q_1{}^m)\Big)$ will be a dominant subspace (cf. Theorem 2.19). Since $\mathcal{R}\Big(Z_2(1)\Big)$ $\Big(= \mathcal{R}(Q_2{}^m)\Big)$ $\perp \mathcal{R}(Q_1{}^m)$ we conclude that $\mathcal{R}\Big(Z_2(0)\Big)$ almost describes the dominated solution subspace $\mathcal{S}_2$. Hence, as a generalization of Corollary 3.15 we obtain that the consistency of U^0 is in Lemma 4.29 measured by $\sqrt{1 + \| F_{12}{}^0 \|^2}$. Gathering these results yields the estimate

$$\kappa\Big(B^0 Z(0) + B^1 Z(1)\Big) \lessapprox \left(\frac{\sqrt{2}}{\mathrm{GAP}\Big(U_1^0, \mathcal{S}_2(0)\Big)} + 1 \right) \sqrt{2}\beta \,, \tag{76}$$

where $\mathcal{S}_2(0)$ is the dominated solution subspace at $t = 0$.

The general conclusion we can draw from the foregoing analysis is: if $\| F^0 \|$ and $\| F^m \|$ are reasonably bounded, then the system (74) is well-conditioned.

4.5.5 The computation of $x(t_i)$ $(i = 0, \ldots, m)$

First we look at the stability properties of the linear recursions

$$x_2{}^{i+1}(t_{i+1}) = W_{22}{}^i \, x_2{}^i(t_i) + w_2{}^i \,, \qquad i = 0, \ldots, m-1 \,, \tag{77}$$

and

$$x_1{}^{i-1}(t_{i-1}) = W_{11}{}^i \, x_1{}^i(t_i) + W_{12}{}^i \, x_2{}^{i-1}(t_{i-1}) + w_1{}^i \,, \qquad i = m, \ldots, 1 \,. \tag{78}$$

With $c = \begin{pmatrix} x_1{}^m(t_m) \\ x_2{}^0(t_0) \end{pmatrix}$ and using the relations (57a,b) these recursions simplify to

$$x^i(t_i) = F^i c + d^i = \begin{bmatrix} F_{11}{}^i & F_{12}{}^i \\ 0 & F_{22}{}^i \end{bmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} d_1{}^i \\ d_2{}^i \end{pmatrix} \,, \qquad i = 0, \ldots, m \,. \tag{79}$$

The kind of perturbations we are going to consider is affecting the matrices $\{ F^i \}_{i=0}^m$ and the vectors $\{ d^i \}_{i=0}^m$. For deriving stability results we need

Property 4.30
Let

$$x^i = A^i c + d^i \,, \qquad i = 0, \ldots, m \,,$$

where $c, \{ d^i \}_{i=0}^m$ *are given vectors and* $\{ A^i \}_{i=0}^m$ *are given matrices. Consider the perturbed expression*

$$\bar{x}^i = (A^i + \triangle A^i) c + d^i + \triangle d^i \ , \qquad i = 0, \ldots, m \ , \tag{80}$$

where

$$\| \triangle A^i \| \leq \ ae + \| A^i \| re \tag{81a}$$

$$\| \triangle d^i \| \leq \ ae + \| d^i \| re \ , \tag{81b}$$

for some given values of ae and re. Then

$$\| \bar{x}^i - x^i \| \leq \ ae \, (1 + \| c \|) + re \, (\| A^i \| \| c \| + \| d^i \|) \ .$$

Proof:
From (80) we obtain

$$\bar{x}^i - x^i = \triangle A^i c + \triangle d^i \ ,$$

which yields the required result.

■

Applying Property 4.30 to the (decoupled) system (79) with possible perturbations like (81a,b) yields

$$\| \bar{x}_2^i - {x_2}^i \| \leq \ (1 + \| c_2 \|) ae + (\| {F_{22}}^i \| \| c_2 \| + \| {d_2}^i \|) re$$

and

$$\begin{aligned} \| \bar{x}_1^i - {x_1}^i \| \ \leq \ & (1 + \| c_1 \| + \| c_2 \|) ae + \\ & (\| {F_{11}}^i \| \| c_1 \| + \| {F_{12}}^i \| \| c_2 \| + \| {d_1}^i \|) re \ . \end{aligned}$$

Hence, the obtained accuracy of $x^i(t_i)$ is similar to the one of F^i and d^i if the problem has been scaled such that $\| c_1 \|$ and $\| c_2 \|$ are just moderate and

$$\begin{pmatrix} \| {x_1}^i \| \\ \| {x_2}^i \| \end{pmatrix} \approx \begin{bmatrix} \| {F_{11}}^i \| & \| {F_{12}}^i \| \\ 0 & \| {F_{22}}^i \| \end{bmatrix} \begin{pmatrix} \| c_1 \| \\ \| c_2 \| \end{pmatrix} + \begin{pmatrix} \| {d_1}^i \| \\ \| {d_2}^i \| \end{pmatrix} \ .$$

This last condition prevents that, at least for the most significant components of ${x_1}^i(t_i)$ and ${x_2}^i(t_i)$, a subtraction of two almost identical numbers has to be performed, which would lead to a loss of accuracy.
The way the perturbation matrices and vectors have been built up follows from the relations of Property 4.23. These relations show that the amplification factors of errors made during the integration steps are given by $\prod \| {W_{22}}^i \|$,

$\prod \| W_{11}{}^i \|$ or a combination of two such factors. This implies that the partitioning has to be chosen such that both $\| W_{11}{}^i \|$ and $\| W_{22}{}^i \|$ are at least moderately sized, for all $i = 0, \ldots, m$. This condition is for instance satisfied if the solution space $\mathcal{S}$ is exponentially dichotomic and the transformation U^0 happens to be such that $\| F_{12}{}^0 \|$ is not too large.

We finish this subsection with a remark on scaling. As is expressed in (78) the value of $x_1{}^i(t_i)$ ($i = 1, \ldots, m$) is actually computed by

$$x_1{}^{i-1}(t_{i-1}) = W_{11}{}^i \, x_1{}^i(t_i) + W_{12}{}^i \, x_2{}^{i-1}(t_{i-1}) + w_1{}^i \ ,$$

where $x_2{}^{i-1}(t_{i-1})$ is given by $F_{22}{}^{i-1} c_2 + d_2{}^{i-1}$. This implies that the errors in $W_{12}{}^i$ and $x_2{}^{i-1}(t_{i-1})$ are multiplied. Therefore, $\|x_2{}^i(t_i)\|/\|x_1{}^i(t_i)\|$ has to be reasonably bounded, for all i. Observe that this quantity is equal to $\|(Q_2{}^i)^T x(t_i)\|/\|(Q_1{}^i)^T x(t_i)\|$, which is large only if DIST($x(t_i), Q_1{}^i$) ≈ 1. However, $Q_1{}^i$ will, in general, represent a dominant solution subspace, which implies that, at least for $i \neq 0$, this distance will be sufficiently bounded away from 1. Thus, for $i = 1, \ldots, m$, the correct scaling is a result of the dichotomy of the solution space of the DE and the decoupling property of the Riccati transformation.
Remains the possibility that $\| (Q_2{}^0)^T x(t_0) \| / \| (Q_1{}^0)^T x(t_0) \|$ is large, implying that c_2 is (relatively) large. This illustrates again the importance of the initial transformation U^0.

4.6 Examples

In this section we want to show the performance of the implementation of the Algorithms 4.22 and 4.25 in the RICCATI package.

4.6.1 Constant coefficients

The first example we consider has constant coefficients and separated BCs. The DE is given by (cf. Problem 3 in [33]):

$$\frac{d^3 u}{dt^3} = \omega \frac{d^2 u}{dt^2} + \frac{d u}{dt} - \omega \, u \ , \qquad t \epsilon [0, T] \ . \tag{82}$$

The separated BCs are:

$$\begin{aligned} u(0) &= 1 + e^{-\omega T} + e^{-T} \\ u(T) &= 2 + e^{-T} \quad \text{and} \quad \frac{du}{dt}(T) = 1 + \omega - e^{-T} \ . \end{aligned}$$

Exact solution: $u(t) = e^{-t} + e^{\omega(t-T)} + e^{t-T}$.
A first order system was made by $x_1 = \dfrac{d^2u}{dt^2}$, $x_2 = \dfrac{du}{dt}$ and $x_3 = u$.
Most of the tested codes in [33] were already in trouble for relatively small values of ω and T ($\omega = 20$ and T of order 1).
To indicate the performance of the integration routine LSODA ([25]) we have solved (82) with $\omega = 20$ and $T = 1, 10, 100$, respectively, and without any intermediate output points. The required accuracy in all components was 10^{-6}, absolute or relative, depending on the size of the solution. The obtained accuracy for the solution $x = (x_1, x_2, x_3)^T$ was of the same order. Other results are shown in Table 4.1, where the second column contains the number of integration steps, the third column the number of function calls and the last column the stepsize of the last accepted integration step.

T	# steps	# func. calls	stepsize at end
1	63	138	$6.63\ 10^{-2}$
10	171	363	$3.44\ 10^{-1}$
100	192	389	81.4 (!)

Table 4.1

The solution space of a DE with constant coefficients has no rotational activity. Therefore, in case of an exponential dichotomy, the Riccati matrix will converge to a constant matrix. In our case: $[1/20, -21/20]$. This is the main reason why solving the problem with $T = 100$ is almost as expensive as the problem with $T = 10$. Observe the amazing final stepsize.

This example illustrates that invariant imbedding is indeed a working technique. The solution has for increasing ω a boundary layer at $t = T$. However, an accurate solution is obtained by taking small stepsizes at $t = 0$ and (very) large stepsizes at $t = T$.

Similar results have been obtained for larger values of ω. In Table 4.2 one finds the results for $T = 10$ and $\omega = 20, 2000$, respectively. Required intermediate output points where $t = 2.5, 5.0$ and 7.5. Again we observe from the last column that the Riccati matrix converges, which implies that on the last two subintervals the results are nearly the same. Moreover, we see that the increase of ω does not really affect the performance. This is explained by the fact that ω does not influence the convergence of the Riccati matrix, since the number of BCs at $t = 0$ is such that the separation is between the solutions that grow like e^t and e^{-t}, respectively.

t	# steps		# func. calls		abs. error in u		$\lvert R_{21}\rvert$	
	ω:20	2000	20	2000	20	2000	20	2000
2.5	89	110	190	193	$6.9\,10^{-8}$	$8.8\,10^{-7}$	$1.04\,10^{-0}$	$9.87\,10^{-1}$
5.0	87	100	180	188	$2.9\,10^{-8}$	$4.1\,10^{-7}$	$7.04\,10^{-3}$	$6.70\,10^{-3}$
7.5	87	99	181	196	$2.7\,10^{-7}$	$4.8\,10^{-6}$	$4.74\,10^{-5}$	$4.52\,10^{-5}$
10.0	87	99	181	196	0	0	$3.19\,10^{-7}$	$3.05\,10^{-7}$

Table 4.2

4.6.2 Rotational activity

The second example we shall consider is based on Example 9.1 of [42]:

$$\frac{dx}{dt} = \begin{bmatrix} 1+19\cos(2\omega t) & 0 & -\omega+19\sin(2\omega t) \\ 0 & 19 & 0 \\ \omega+19\sin(2\omega t) & 0 & 1-19\cos(2\omega t) \end{bmatrix} \tag{83}$$

subject to the non-separated BCs

$$x(0) + x(\pi) = b\ .$$

The inhomogeneous term f and the vector b are chosen such that the solution becomes $x(t) = (\,e^t, \omega\, e^{-t}, e^t\,)^T$.
A fundamental solution X corresponding to (83) is given by

$$X(t) = \begin{bmatrix} \cos(\omega t) & 0 & -\sin(\omega t) \\ 0 & 1 & 0 \\ \sin(\omega t) & 0 & \cos(\omega t) \end{bmatrix} \begin{bmatrix} e^{20t} & & \\ & e^{19t} & \\ & & e^{-18t} \end{bmatrix}\ . \tag{84}$$

For ω not too large the growth behaviour of solutions is nicely indicated by the eigenvalues of A. For instance, for $\omega = 4$ we obtain $\lambda\Big(A(0)\Big) = \{\,1 + \sqrt{345}, 19, 1 - \sqrt{345}\,\}$. Therefore we choose $k = 2$.
One of the components of the Riccati matrix is expected to behave like $\tan(\omega t)$. Hence, some restarts will be unavoidable. The number of such restarts depends on the value of the boundedness constant a, i.e. $\lvert R_{21}(t)\rvert \le a$ (see step 2.b of Algorithm 4.25). In Table 4.3 the results are shown for $\omega = 4$, the accuracy of integration $= 10^{-6}$ and for different values of a. A small value of a causes a relatively large number of restarts. On the other hand, a large value of a does generally not affect the number of restarts, but the algorithm becomes less efficient, since steprefinement will take place at the end of each subinterval.

a	# restarts	# steps	# func. calls
1	15	641	1333
3	9	526	1137
50	8	720	1724

Table 4.3

The obtained relative accuracy at the boundary and restart points when $a = 3$ is shown in Table 4.4 ($\omega = 4$ and the accuracy of integration $= 10^{-6}$).

t	|rel. error in x_1|	|rel. error in x_2|	|rel. error in x_3|
0	$9.31\ 10^{-8}$	$7.71\ 10^{-8}$	$3.75\ 10^{-6}$
0.342	$3.37\ 10^{-7}$	$7.31\ 10^{-8}$	$1.83\ 10^{-7}$
0.659	$2.37\ 10^{-7}$	$2.11\ 10^{-8}$	$4.12\ 10^{-7}$
0.972	$1.50\ 10^{-8}$	$8.17\ 10^{-8}$	$2.33\ 10^{-7}$
1.286	$3.61\ 10^{-7}$	$2.66\ 10^{-8}$	$1.88\ 10^{-7}$
1.603	$1.20\ 10^{-7}$	$4.93\ 10^{-8}$	$3.50\ 10^{-7}$
1.918	$3.30\ 10^{-7}$	$7.88\ 10^{-8}$	$1.74\ 10^{-7}$
2.234	$1.78\ 10^{-7}$	$7.15\ 10^{-8}$	$4.09\ 10^{-7}$
2.550	$7.47\ 10^{-8}$	$1.12\ 10^{-7}$	$1.90\ 10^{-7}$
2.865	$4.60\ 10^{-7}$	$5.22\ 10^{-8}$	$1.96\ 10^{-7}$
π	$4.02\ 10^{-9}$	$1.78\ 10^{-6}$	$1.62\ 10^{-7}$

Table 4.4

Finally we remark that the Schur transformation induces a consistent fundamental solution (cf. Lemma 4.29), since

$$|{F_{12}}^0| = 0.106\ , \quad |{F_{11}}^0| = 1.19\ 10^{-26} \quad \text{and} \quad |{F_{22}}^m| = 2.78\ 10^{-25}\ .$$

Chapter 5

Stiff Boundary Value Problems

5.1 Introduction

In this chapter we shall investigate various aspects of the Riccati method, derived in the previous chapter, when applied to *stiff* BVPs. A BVP is called stiff if the homogeneous part has solutions with rapidly changing (non-oscillating) components, to be called the *fast modes*. Typically, fast modes have their impact on the final solution of a BVP only within small regions, so-called *layers*. Outside these layers a solution x is composed of *slow* (or 'smooth') *modes* (in that $\| x \|$, $\| \frac{dx}{dt} \|$, etc. are bounded by moderate constants). However, the potentially rapid growth of fast modes on larger intervals makes the BVP numerically hard to solve. This problem is reminiscent of what is well-known for IVPs. There the main question is to let the stepsize of the integration routine be dictated by the activity of the smooth components only, which is called the *stiffness* problem. We shall adopt this terminology as well in the context of BVPs.

The notion of stiffness can be made more precise for a *singular perturbation problem*. There the quotient of the time scales of the smooth and the fast modes is governed by one or more (small) parameters. A most convenient situation is obtained when the system is in so-called *bordered form*:

$$\begin{bmatrix} I_k & 0 \\ 0 & E \end{bmatrix} \frac{d}{dt} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{bmatrix} A_{11}(t) & A_{12}(t) \\ A_{21}(t) & A_{22}(t) \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} f_1(t) \\ f_2(t) \end{pmatrix}, \; t \in [0,1], \quad (1)$$

where $E = \text{diag}\,(\varepsilon_{k+1}, \ldots, \varepsilon_n)$, $0 < \varepsilon_i \ll 1$ $(i = k+1, \ldots, n)$. The integer k may be zero, i.e., the homogeneous problem may have fast modes only.
Generally, the derivative of x_2 grows unboundedly when $E \to 0$. Moreover, when $E \to 0$ the DE reduces to the differential-algebraic system

$$\begin{aligned} \frac{d}{dt}x_1 &= A_{11}(t)\,x_1 + A_{12}(t)\,x_2 + f_1(t) \\ 0 &= A_{21}(t)\,x_1 + A_{22}(t)\,x_2 + f_2(t)\ . \end{aligned}$$

In general, the solution of this system can not satisfy all the BCs simultaneously. This singular behaviour accounts for the name *singular perturbation.*

A numerical method for solving a general (stiff) BVP is based on discretization ([3],[31]). Recalling that the fast modes have a possibly significant contribution to the solution x only inside the layers, we realize that the mesh used for discretization should be chosen quite differently inside and outside these layers. Such a grid should be relatively fine inside (commensurating with the activity of the fast modes) and fairly coarse where the particular solution x is smooth.

Besides purely numerical techniques there are also a number of mixed analytic and numerical methods. Two familiar techniques are (analytic) decoupling of the system when it is in bordered form (possibly combined with regular expansions) ([41],[49]) or matched asymptotic expansions. The latter method, realizing the singular behaviour of the fast mode part when $E \to 0$, uses different power series for the smooth part of the solution (*outer solution*) and the layer part of the solution (*inner solution*). By requiring that the final solution is in $C^p[0,1]$, for some integer p, those two solutions are matched at the layer ends ([17]).

Stiff BVPs have already received much attention in the literature, both analytically and numerically ([3],[19],[23],[31],[61]). Numerically, problems where the layers appear at the boundary only can generally be solved quite satisfactory nowadays. Problems with *internal layers*, often related to so-called *turning points* are less well solved, although there is a growing literature, see e.g. ([10],[30],[62]). Part of the problem for internal layers is that the location of the turning point is often not known beforehand and has to be determined during the solution process. We shall return to this class of difficult problems in Section 5.3.

In this chapter we shall discuss the Riccati method of Chapter 4 for two kind of well-conditioned BVPs:
problems having no internal layer (Section 5.2) and problems that do have

such a layer (Section 5.3). For the first class of problems we shall assume that the solution space $\mathcal{S}$ is *exponentially trichotomic*, which is a refinement of the concept of exponential dichotomy (Definition 2.7). It will turn out that for large problems, with a relatively small number of slowly varying modes, a substantial reduction can be obtained. The quintessence of the Riccati method will not be violated, since it will work too. The proposed adaptation is for reasons of efficiency only.

When internal layers are present it is not directly clear whether the Riccati method will work or not. As we have seen in Section 3.2 the Riccati transformation determines the direction of a dominant subspace. However, in an internal layer this direction may change drastically. The influence of this rotational activity on the Riccati method is investigated in Section 5.3 . To simplify the discussion we shall mainly consider 2-dimensional singular perturbation problems, involving a parameter ε.

5.2 Large systems

In this section we shall present a reduction technique for the Riccati method, discussed in Chapter 4, when applied to large systems of DEs, with no internal layers. In the discussion the role of the Riccati transformation is not essential; any decoupling transformation will do.

5.2.1 Exponential trichotomy

In Chapter 2 we have introduced the concept of exponential dichotomy. Here we want to generalize this concept in such a way that slow modes are isolated from the rapidly varying ones. Let $\mathcal{S}$ be the solution space corresponding to the homogeneous DE

$$\frac{dx}{dt} = A(t)\,x\ , \qquad t \in [0,\infty)\ . \tag{2}$$

Definition 5.1
The solution space $\mathcal{S}$ of (2) *is called* exponentially trichotomic *if for every fundamental solution X there exist projections P_1, P_2 and P_3, with $P_1 + P_2 + P_3 = I_n$, such that*

$$\| X(t)\,P_1\,X^{-1}(s) \| \ \leq m_1\, e^{-\lambda_1(s-t)}\ , \quad 0 \leq t \leq s\ ,$$

$$1/m_2 \leq \ \| X(t)\,P_2\,X^{-1}(s) \| \ \leq m_2\ , \qquad\qquad 0 \leq t, s\ ,$$

$$\| X(t) P_3 X^{-1}(s) \| \leq m_3 e^{-\lambda_3 (t-s)} , \quad 0 \leq s \leq t ,$$

where the constants m_1, m_2 *and* m_3 *are of moderate size and* $\lambda_1, \lambda_3 > 0$.

■

Remark 5.2

(i) An exponential trichotomic solution space is also exponentially dichotomic.

(ii) As in the case of exponential dichotomy, Definition 5.1 is equivalent to the following formulation (cf. the proof of Theorem 2.9):
the solution space $\mathcal{S}$ of (2) is exponentially trichotomic if it can be split into three parts, $\mathcal{S} = \mathcal{S}_1 \oplus \mathcal{S}_2 \oplus \mathcal{S}_3$, with

$$\phi_1 \in \mathcal{S}_1 \Rightarrow \| \phi_1(t) \| \leq m_1 e^{-\lambda_1 (s-t)} \| \phi_1(s) \| , \quad 0 \leq t \leq s ,$$

$$\phi_2 \in \mathcal{S}_2 \Rightarrow \frac{1}{m_2} \| \phi_2(s) \| \leq \| \phi_2(t) \| \leq m_2 \| \phi_2(s) \| , \quad 0 \leq t, s ,$$

$$\phi_3 \in \mathcal{S}_3 \Rightarrow \| \phi_3(t) \| \leq m_3 e^{-\lambda_3 (t-s)} \| \phi_3(s) \| , \quad 0 \leq s \leq t ,$$

where m_1, m_2 and m_3 are positive (moderate) constants and $\lambda_1, \lambda_3 > 0$. Moreover, there exist positive constants q_1, q_2 such that, uniformly in t,

$$\mathrm{GAP}\Big(\mathcal{S}_1(t), \mathcal{S}_2(t) \oplus \mathcal{S}_3(t) \Big) \geq q_1 \quad \text{and} \quad \mathrm{GAP}\Big(\mathcal{S}_1(t) \oplus \mathcal{S}_2(t), \mathcal{S}_3(t) \Big) \geq q_2 .$$

Here the subspaces $\mathcal{S}_i$ $(i = 1, 2, 3)$ are defined as $\mathcal{S}_i = \Big\{ X(t) P_i c \,|\, c \in \mathbb{R}^n \Big\}$.

■

Property 5.3
The exponential trichotomy of $\mathcal{S}$ *is observable on a sufficiently large but finite interval.*

Proof:
Consider the DE (2), having an exponentially trichotomic solution space $\mathcal{S}$, on the finite interval $[0, \xi]$ $(\xi > 0)$. Let the solution subspaces $\mathcal{S}_i$ $(i = 1, 2, 3)$ be defined as in Remark 5.2 (ii). Then we have

$$\phi_1 \in \mathcal{S}_1 \Rightarrow \| \phi_1(\xi) \| \geq \frac{1}{m_1} e^{\lambda_1 \xi} \| \phi_1(0) \|$$

$$\phi_2 \in \mathcal{S}_2 \Rightarrow \frac{1}{m_2} \| \phi_2(\xi) \| \leq \| \phi_2(0) \| \leq m_2 \| \phi_2(\xi) \|$$

$$\phi_3 \epsilon \mathcal{S}_3 \Rightarrow \| \phi_3(\xi) \| \leq m_3 e^{-\lambda_3 \xi} \| \phi_3(0) \| .$$

Hence, if ξ is such that both

$$\frac{e^{\lambda_1 \xi}}{m_1 m_2} > 1 \quad \text{and} \quad \frac{e^{\lambda_3 \xi}}{m_2 m_3} > 1$$

then the exponential trichotomy is observable on $[0, \xi]$.

■

Since our main interest are DEs on a finite interval, say $[0, 1]$, we have to assume that the exponentially trichotomic behaviour of the solution space $\mathcal{S}$ can already be observed on $[0, 1]$. As we have seen in Property 5.3 this will certainly be the case under the following

Assumption 5.4

$$\frac{e^{\lambda_1}}{m_1 m_2} \gg 1 \quad \text{and} \quad \frac{e^{\lambda_3}}{m_2 m_3} \gg 1 .$$

■

Remark 5.5

With the above assumption a mode $\phi_1 \epsilon \mathcal{S}_1$ with $\| \phi_1(1) \| = 1$ is only significantly different from 0 in a (small) $O(1/\lambda_1)$ neighbourhood of 1. Similarly, a mode $\phi_3 \epsilon \mathcal{S}_3$ with $\| \phi_3(0) \| = 1$ is only significantly different from 0 in a (small) $O(1/\lambda_3)$ neighbourhood of 0. Therefore these fast solutions can only play a role in the boundary layers. This implies that by Definition 5.1 internal layers are excluded, since modes in $\mathcal{S}_2$ have no layer behaviour at all.

■

So far we have not specified what the dimensions are of the solution subspaces $\mathcal{S}_i (i = 1, 2, 3)$; anyone of these subspaces may even be empty. In the sequel we partition matrices and vectors accordingly to the dimensions k, l and m of the solution subspaces $\mathcal{S}_i (i = 1, 2, 3)$, respectively. So,

$$A(t) = \begin{bmatrix} A_{11}(t) & A_{12}(t) & A_{13}(t) \\ A_{21}(t) & A_{22}(t) & A_{23}(t) \\ A_{31}(t) & A_{32}(t) & A_{33}(t) \end{bmatrix} \begin{matrix} \updownarrow k \\ \updownarrow l \\ \updownarrow m \end{matrix} \quad \text{and} \quad x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix} \begin{matrix} \updownarrow k \\ \updownarrow l \\ \updownarrow m \end{matrix} \quad (3)$$

$\underset{k}{\longleftrightarrow}$ $\underset{l}{\longleftrightarrow}$ $\underset{m}{\longleftrightarrow}$ $\underset{1}{\longleftrightarrow}$

5.2.2 Single shooting

The reduction technique for the Riccati method that is presented in this section is based on the following observation. Consider a DE

$$\frac{dx}{dt} = A(t)\,x + f(t)\ , \qquad t \in [0,1]\ , \tag{4}$$

with slow and fast decaying modes only. So, $k = 0$ and $\mathcal{S} = \mathcal{S}_2 \oplus \mathcal{S}_3$. In order to compute a fundamental solution X, satisfying

$$\frac{d}{dt}X = A(t)\,X\ , \qquad t \in [0,1]\ , \tag{5}$$

with $X(0)$ non-singular and consistent (see Section 2.3), we need an integration routine by which
– the slow modes are computed accurately
– the influence of the fast decaying modes on the stepsize is restricted to the initial layer at $t = 0$.
This implies that we have to use a stiffly stable integrator ([46]), for instance a BDF-method (such as implemented in the routine LSODA in ODEPACK ([25])). In general, such an integration routine automatically generates the expected grid: small stepsizes in the boundary layer at $t = 0$, say on $[0, \delta]$, and a coarse grid hereafter. This implies that also in the boundary layer accurate solutions are obtained. At the end of the layer, say at $t = \delta$, the fast decaying solutions have been damped out. Hence, $\mathrm{rank}\big(X(\delta)\big)$ is effectively reduced to l, the dimension of $\mathcal{S}_2$. Let the QR-decomposition of $X(\delta)$ (possibly after some column permutation) be given by

$$X(\delta) = \underset{\substack{\longleftrightarrow \\ l}}{\begin{bmatrix} Q_2 \end{bmatrix}} \underset{\substack{\longleftrightarrow \\ m}}{\begin{bmatrix} Q_3 \end{bmatrix}} \begin{bmatrix} \underset{\substack{\longleftrightarrow \\ l}}{R_{22}} & \underset{\substack{\longleftrightarrow \\ m}}{R_{23}} \\ 0 & R_{33} \end{bmatrix} \begin{matrix} \updownarrow l \\ \updownarrow m \end{matrix} \ . \tag{6}$$

The matrix R_{33} will be $O(e^{-\lambda_3 \delta})$ ([42]), which is negligible, for δ large enough (i.e., smaller than the error due to the numerical integration routine (cf. (4.70)).

The inhomogeneous term can be treated in a similar way. Compute the particular solution p, satisfying

$$\frac{dp}{dt} = A(t)\,p + f(t)\ , \quad t \in [0,\delta]\ , \qquad p(0) = 0\ .$$

By superposition we know that there exists a vector $c = \begin{pmatrix} c_2 \\ c_3 \end{pmatrix} \in \mathbb{R}^{l+m}$ such that

$$x(t) = X(t)\,c + p(t)\ , \qquad \text{for all } t \in [0,\delta]\ . \tag{7}$$

Hence,

$$\begin{aligned} x(\delta) &= X(\delta)\,c + p(\delta) \\ &= \begin{bmatrix} Q_2 & Q_3 \end{bmatrix} \begin{bmatrix} R_{22} & R_{23} \\ 0 & R_{33} \end{bmatrix} c + p(\delta) \\ &\approx Q_2 \begin{bmatrix} R_{22} & R_{23} \end{bmatrix} c + p(\delta) \\ &= Q_2 \left(\begin{bmatrix} R_{22} & R_{23} \end{bmatrix} c + {Q_2}^T p(\delta) \right) + (I_{l+m} - Q_2\,{Q_2}^T)\,p(\delta)\ . \end{aligned}$$

Let

$${c_2}^0 = \begin{bmatrix} R_{22} & R_{23} \end{bmatrix} c + {Q_2}^T p(\delta)\ . \tag{8a}$$

Then we approximately have

$$x(\delta) = Q_2\,{c_2}^0 + Q_3\,{Q_3}^T p(\delta)\ . \tag{8b}$$

Remark 5.6
From (8b) we obtain that at $t = \delta$ the manifold of slow modes is approximately described by the relation

$${Q_3}^T x(\delta) = {Q_3}^T p(\delta)\ . \tag{9}$$

The accuracy of this approximation depends on the magnitude of $\| R_{33}\,c_3 \|$. Hence, it is important to scale the DE and to choose the initial value $X(0)$ such that c_3 is not extremely large.

■

For $t \geq \delta$ we want to find a continuous extension of (8b). To this end we write any solution x, satisfying (8b), as

$$x(t) = {X_2}^0(t)\,{c_2}^0 + p^0(t)\ , \qquad t \in [\delta,1]\ . \tag{10}$$

Hence, ${X_2}^0$ is a part of a fundamental solution, satisfying the IVP

$$\frac{d}{dt}{X_2}^0 = A(t)\,{X_2}^0\ , \qquad t \in [\delta,1]\ , \qquad \text{with } {X_2}^0(\delta) = Q_2\ .$$

Similarly, p^0 is a particular solution of (4), satisfying

$$\frac{d}{dt}p^0 = A(t)\,p^0 + f(t)\ , \qquad t \in [\delta,1]\ , \qquad \text{with } p^0(\delta) = Q_3\,{Q_3}^T p(\delta)\ .$$

Observe that ${X_2}^0$ contains slowly varying modes only. Since $\left({X_2}^0(\delta)\right)^T p^0(\delta) = 0$ this implies that, at least for t near δ, a well-conditioned representation of ${X_2}^0 \oplus p^0$ is given by ${X_2}^0$ and p^0.

Substitution of (7) with $t = 0$ and of (10) with $t = 1$ into the (general) BCs

$$B^0\, x(0) + B^1\, x(1) = b$$

yields ($p^0(0) = 0$):

$$B^0\, X(0)\, c + B^1\, {X_2}^0(1)\, {c_2}^0 = b - B^1\, p^0(1)\ .$$

Together with the continuity of x at $t = \delta$ (relation (8a)) this results in the shooting system

$$\begin{bmatrix} B^0\, X(0) & B^1\, {X_2}^0(1) \\ -\begin{bmatrix} R_{22} & R_{23} \end{bmatrix} & I_l \end{bmatrix} \begin{pmatrix} c \\ {c_2}^0 \end{pmatrix} = \begin{pmatrix} b - B^1\, p^0(1) \\ {Q_2}^T p(\delta) \end{pmatrix}\ . \tag{11}$$

From $\begin{pmatrix} c \\ {c_2}^0 \end{pmatrix}$ the solution x can be computed in any desired point.

The above sketched method has some interesting benefits.

(i) The grid is automatically generated by the integration routine.

(ii) The influence of the fast decaying modes is noticeable only in the boundary layer.

(iii) The method is not restricted to singular perturbation problems (where an explicit small parameter ϵ is available).

(iv) For non-stiff problems, having slow modes only, it simplifies to a single shooting method.

(v) A generalization to a solution method for multi time-scale problems is straightforward, without an explicit knowledge of the various time-scales. This makes the method also suitable for mildly stiff problems.

(vi) If necessary, more subintervals can simply be generated, resulting in a multiple shooting system.

(vii) the total number of integration steps does not really depend on the stiffness of the system.

This last observation is an important property. To prove the result we consider the scalar model problem

$$\frac{dx}{dt} = \lambda x \ , \qquad t \geq 0 \ , \tag{12}$$

subject to $x(0) = 1$. This has the exact solution $x(t) = e^{\lambda t}$. Let h_i be the stepsize taken in the $(i+1)^{\text{th}}$ integration step $(i = 0, 1, \cdots)$. Let TOL be the required absolute accuracy per step and EST_i an approximation of the local discretization error (cf. [55], p.115). For a p-th order method a standard technique to choose the stepsizes is given by

$$h_{i+1} = c\, h_i \left(\frac{TOL}{|EST_i|}\right)^{\frac{1}{p+1}} \ , \tag{13}$$

where the constant c is a safety factor, smaller than 1.

Let $t_0 = 0$ and define the nodes t_i by $t_i = \sum_{j=0}^{i-1} h_j$ $(i = 1, 2, \cdots)$. The local discretization error at t_{i+1} satisfies ideally

$$EST_i \approx \xi\, {h_i}^{p+1} x^{(p+1)}(t_{i+1}) = \xi\, {h_i}^{p+1} \lambda^{p+1} e^{\lambda t_{i+1}} \ , \tag{14}$$

where ξ is a positive constant of moderate size. Now we have

Property 5.7

Suppose we want to have the solution of (12) *in some point* T. *Let* p *be the order of the method used and let the initial stepsize* h_0 *be given by* $h_0 = \frac{c}{|\lambda|}\left(\frac{TOL}{\xi}\right)^{\frac{1}{p+1}}$. *Define* $\nu = \frac{\lambda}{p+1}$. *Using the stepsize strategy* (13) *we need approximately* $\frac{1 - e^{\nu T}}{-\nu\, h_0}$ *steps to obtain* $x(T)$.

Proof:

Combining (14) with (13) yields

$$h_i = h_0\, e^{-\nu t_i} \ , \qquad (i = 0, 1, \cdots) \ . \tag{15}$$

Now consider a continuously differentiable function h such that, for all $i = 0, 1, \cdots$,

$$t_i = \sum_{j=0}^{i-1} h_j \approx \int_0^i h(\tau)\, d\tau \ .$$

Then (15) is approximated by its continuous form

$$h(t) = h_0 \exp(-\nu \int_0^t h(\tau)\, d\tau) , \qquad t \geq 0 . \tag{16}$$

Differentiating (16) yields the Riccati DE

$$\begin{cases} \dfrac{dh}{dt} &= -\nu h^2 , \qquad t \geq 0 \\ h(0) &= h_0 \end{cases} ,$$

which has the exact solution $h(t) = \dfrac{1}{\nu}(t + \dfrac{1}{\nu h_0})^{-1}$.

Let N be such that $T = t_N = \sum_{j=0}^{N-1} h_j$. Then

$$h_0 e^{-\nu T} = h_0 e^{-\nu t_N} = h_N \approx h(N) = \frac{1}{\nu}(N + \frac{1}{\nu h_0})^{-1}$$

$$\Rightarrow N \approx \frac{e^{\nu T} - 1}{\nu h_0} .$$

■

Corollary 5.8

Consider the DE (12) *with* $\lambda \ll -1$. *Then the number of integration steps needed to reach a point* T *outside the initial layer is independent of the stiffness parameter* λ.

Proof:

Note that $\nu h_0 = -\dfrac{c}{p+1}\Big(\dfrac{TOL}{\xi}\Big)^{\frac{1}{p+1}}$, which is independent of λ. If $\lambda \ll -1$, then $\nu = \dfrac{\lambda}{p+1} \ll -1$ and $e^{\nu T} \approx 0$. Hence, with Property 5.7, the number of integration steps is approximately given by $\dfrac{p+1}{c}\Big(\dfrac{TOL}{\xi}\Big)^{p+1}$, which is depending on the required accuracy, but not on the stiffness of the DE.

■

In the next section we transform a general stiff BVP, with an exponentially trichotomic solution space $\mathcal{S}$, into a sequence of problems that can be solved by this shooting method. To this end we shall use an adapted version of the Riccati method of Chapter 4. With the above shooting technique (and with invariant imbedding) we are able to reduce the number of DEs that has to be solved. This is especially of interest when we are dealing with a large system. For small systems the Riccati method of Chapter 4 will work too, and will

almost be as efficient as the adapted version discussed in the next section.

5.2.3 Riccati transformation

We now want to treat the general case, where also fast increasing modes are present. For reasons, discussed in the Sections 3.4, 4.4.1 and 5.2.1, we shall make the following assumptions:

- $A(0)$ is in quasi-triangular form and correctly ordered (cf. Assumption 4.12).
- the BVP has been scaled such that $\| x(t) \|$ has a reasonable upper bound, for all t.
- the solution space $\mathcal{S}$ has an exponential trichotomy (see Definition 5.1), with $\dim(\mathcal{S}_1) = k$, $\dim(\mathcal{S}_2) = l$ and $\dim(\mathcal{S}_3) = m$.

Consider the DE (cf. (2), (3)):

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{bmatrix} A_{11}(t) & A_{12}(t) & A_{13}(t) \\ A_{21}(t) & A_{22}(t) & A_{23}(t) \\ A_{31}(t) & A_{32}(t) & A_{33}(t) \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} f_1(t) \\ f_2(t) \\ f_3(t) \end{pmatrix}, \quad (17)$$

$t \epsilon [0,1]$, subject to the BCs

$$B^0 x(0) + B^1 x(1) = b \,. \quad (18)$$

Often the Riccati transformation is used to decouple the fast modes (both increasing and decaying) and the slow modes ([41],[63]). If (17) is a singular perturbation problem in bordered form, like (1), then it can be shown that there exists such a Riccati transformation, having an asymptotic power series expansion in ε ([61]). Unfortunately, this technique can not be generalized to the case of a general stiff BVP, since the DE for this Riccati matrix will be unstable, unless all fast solutions are decaying. In the latter situation, however, it is probably more efficient to decouple only once, outside the initial layer, as is discussed in Section 5.2.2.

The Riccati transformation we propose for a general stiff BVP decouples the fast increasing modes, $\mathcal{S}_1$, from the other ones (both slow and fast decaying modes). For this latter solution manifold we can use the same reduction as obtained by the shooting technique described in Section 5.2.2. Observe that for obtaining accuracy this reduction technique is not necessary; the Riccati method of Chapter 4 will deliver accurate results too. For large systems, however, these reductions will make the algorithm more efficient.

Define the Riccati transformation

$$T(t) = \begin{bmatrix} I_k & 0 & 0 \\ R_{21}(t) & I_l & 0 \\ R_{31}(t) & 0 & I_m \end{bmatrix} . \tag{19}$$

In order to decouple we obtain, as is shown in Chapter 4, the Riccati DE

$$\frac{d}{dt}\begin{bmatrix} R_{21} \\ R_{31} \end{bmatrix} = \begin{bmatrix} A_{21}(t) \\ A_{31}(t) \end{bmatrix} + \begin{bmatrix} A_{22}(t) & A_{23}(t) \\ A_{32}(t) & A_{33}(t) \end{bmatrix}\begin{bmatrix} R_{21} \\ R_{31} \end{bmatrix} - \tag{20a}$$

$$\begin{bmatrix} R_{21} \\ R_{31} \end{bmatrix} A_{11}(t) - \begin{bmatrix} R_{21} \\ R_{31} \end{bmatrix}\begin{bmatrix} A_{12}(t) & A_{13}(t) \end{bmatrix}\begin{bmatrix} R_{21} \\ R_{31} \end{bmatrix} ,$$

subject to

$$\begin{bmatrix} R_{21}(0) \\ R_{31}(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{20b}$$

($A(0)$ is assumed to be in quasi-triangular form and correctly ordered).

Let $T_1 = \begin{bmatrix} I_k \\ R_{21} \\ R_{31} \end{bmatrix}$. If $\mathrm{DIST}\Big(\mathcal{R}\big(T_1(t)\big), \mathcal{S}_1(t)\Big)$ behaves like $(e^{-\lambda_1 t})$, then the decoupling induced by (19) has been done correctly (cf. Theorem 2.19). In that case the DE (20a) will be asymptotically stable, since the corresponding Jacobian has eigenvalues all far in the left halfplane of $\mathbb{C}$ (cf. (4.71)). This implies that, possibly after some initial layer (depending on the distance between $\mathcal{R}\big(T_1(0)\big)$ and $\mathcal{S}_1(0)$), the Riccati matrix will be slowly varying. More precisely, outside the initial layer the variation of the Riccati matrix is governed by the rotational activity of $\mathcal{S}_1$.

Define $y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = T^{-1}x$. Then, as long as T exists, we obtain the transformed system (cf. (4.6a))

$$\frac{dy}{dt} = \begin{bmatrix} \tilde{A}_{11}(t) & A_{12}(t) & A_{13}(t) \\ 0 & \tilde{A}_{22}(t) & \tilde{A}_{23}(t) \\ 0 & \tilde{A}_{32}(t) & \tilde{A}_{33}(t) \end{bmatrix} y + \begin{pmatrix} f_1(t) \\ \tilde{f}_2(t) \\ \tilde{f}_3(t) \end{pmatrix} , \tag{21a}$$

where

$$\tilde{A}_{11} = A_{11} + \begin{bmatrix} A_{12} & A_{13} \end{bmatrix}\begin{bmatrix} R_{21} \\ R_{31} \end{bmatrix} , \tag{21b}$$

and

$$\left[\begin{array}{cc|c} \tilde{A}_{22} & \tilde{A}_{23} & \tilde{f}_2 \\ \tilde{A}_{32} & \tilde{A}_{33} & \tilde{f}_3 \end{array}\right] = \left[\begin{array}{cc|c} A_{22} & A_{23} & f_2 \\ A_{32} & A_{33} & f_3 \end{array}\right] - \left[\begin{array}{c} R_{21} \\ R_{31} \end{array}\right] \left[\begin{array}{cc|c} A_{12} & A_{13} & f_1 \end{array}\right] \quad (21c)$$

Since $\mathcal{R}(T_1)$ describes, in general, a dominant solution subspace, containing fast increasing modes only, the growth of these modes is governed by $\tilde{A}_{11}$ (where we assume that the Riccati matrix stays sufficiently bounded). Moreover, the growth of the slow and fast decaying modes is governed by $\left[\begin{array}{cc} \tilde{A}_{22} & \tilde{A}_{23} \\ \tilde{A}_{32} & \tilde{A}_{33} \end{array}\right]$ and this growth has been decoupled from the fast increasing modes. This implies that for the decoupled part of (21a) (involving y_2 and y_3) the technique described in Section 5.2.2 can be used. Hence, we have to solve the IVP

$$\frac{d}{dt}\left[\begin{array}{cc|c} Y_{22} & Y_{23} & p_2 \\ Y_{32} & Y_{33} & p_3 \end{array}\right] = \quad (22a)$$

$$\left[\begin{array}{cc} \tilde{A}_{22}(t) & \tilde{A}_{23}(t) \\ \tilde{A}_{32}(t) & \tilde{A}_{33}(t) \end{array}\right] \left[\begin{array}{cc|c} Y_{22} & Y_{23} & p_2 \\ Y_{32} & Y_{33} & p_3 \end{array}\right] + \left[\begin{array}{cc|c} 0 & 0 & \tilde{f}_2(t) \\ 0 & 0 & \tilde{f}_3(t) \end{array}\right], \quad t \geq 0,$$

subject to the initial values

$$\left[\begin{array}{cc|c} Y_{22}(0) & Y_{23}(0) & p_2(0) \\ Y_{32}(0) & Y_{33}(0) & p_3(0) \end{array}\right] = \left[\begin{array}{cc|c} I_l & 0 & 0 \\ 0 & I_m & 0 \end{array}\right]. \quad (22b)$$

By the principle of superposition there exists a vector $\begin{pmatrix} c_2 \\ c_3 \end{pmatrix} \epsilon\, \mathbb{R}^{l+m}$ such that, as long as the Riccati matrix exists, (cf. (4.8b))

$$\left[\begin{array}{ccc} -R_{21}(t) & I_l & 0 \\ -R_{31}(t) & 0 & I_m \end{array}\right] x(t) = \begin{pmatrix} y_2(t) \\ y_3(t) \end{pmatrix}$$

$$= \left[\begin{array}{cc} Y_{22}(t) & Y_{23}(t) \\ Y_{32}(t) & Y_{33}(t) \end{array}\right] \begin{pmatrix} c_2 \\ c_3 \end{pmatrix} + \begin{pmatrix} p_2(t) \\ p_3(t) \end{pmatrix} \quad (23)$$

Observe that by (22b) and (20b) we have

$$\begin{pmatrix} c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} y_2(0) \\ y_3(0) \end{pmatrix} = \begin{pmatrix} x_2(0) \\ x_3(0) \end{pmatrix}. \quad (24)$$

When the fast decaying modes have become smaller than the required (absolute) accuracy for integration, say at $t = \delta$, we make the QR-decomposition (cf. (6))

$$\begin{bmatrix} Y_{22}(\delta) & Y_{23}(\delta) \\ Y_{32}(\delta) & Y_{33}(\delta) \end{bmatrix} = \begin{bmatrix} \bar{Q}_2^0 & \bar{Q}_3^0 \end{bmatrix} \begin{bmatrix} {R_{22}}^0 & {R_{23}}^0 \\ 0 & {R_{33}}^0 \end{bmatrix} \tag{25}$$

($\bar{Q}^0 = \begin{bmatrix} \bar{Q}_2^0 & \bar{Q}_3^0 \end{bmatrix} \epsilon \, \mathbb{R}^{(l+m)\times(l+m)}$ orthogonal and $R^0 = \begin{bmatrix} {R_{22}}^0 & {R_{23}}^0 \\ 0 & {R_{33}}^0 \end{bmatrix}$ $\epsilon \, \mathbb{R}^{(l+m)\times(l+m)}$ upper triangular).

Since δ has been chosen such that $\| \, {R_{33}}^0 \, \|$ is below a prescribed tolerance (generally $\| \, {R_{33}}^0 \, \| = \mathrm{O}(e^{-\lambda_3 \delta})$), we obtain that

$$\begin{bmatrix} Y_{22}(\delta) & Y_{23}(\delta) \\ Y_{32}(\delta) & Y_{33}(\delta) \end{bmatrix} \approx \bar{Q}_2^0 \begin{bmatrix} {R_{22}}^0 & {R_{23}}^0 \end{bmatrix} .$$

Therefore, using the boundedness of $\| \, x(t) \, \|$, (23) reduces at $t = \delta$ effectively to (cf. 8a,b)):

$$\begin{aligned} \begin{bmatrix} -R_{21}(\delta) & I_l & 0 \\ -R_{31}(\delta) & 0 & I_m \end{bmatrix} x(\delta) &= \bar{Q}_2^0 \begin{bmatrix} {R_{22}}^0 & {R_{23}}^0 \end{bmatrix} \begin{pmatrix} c_2 \\ c_3 \end{pmatrix} + \begin{pmatrix} p_2(\delta) \\ p_3(\delta) \end{pmatrix} \\ &= \bar{Q}_2^0 {c_2}^0 + \bar{Q}_3^0 {\bar{Q}_3^0}^T \begin{pmatrix} p_2(\delta) \\ p_3(\delta) \end{pmatrix} , \end{aligned} \tag{26}$$

where (cf. (24))

$${c_2}^0 = \begin{bmatrix} {R_{22}}^0 & {R_{23}}^0 \end{bmatrix} \begin{pmatrix} x_2(0) \\ x_3(0) \end{pmatrix} + {\bar{Q}_2^0}^T \begin{pmatrix} p_2(\delta) \\ p_3(\delta) \end{pmatrix} . \tag{27}$$

Similarly as has been done in (10) we seek for a continuous extension of (26), for $t \geq \delta$, of the form

$$\begin{bmatrix} -R_{21}(t) & I_l & 0 \\ -R_{31}(t) & 0 & I_m \end{bmatrix} x(t) = \begin{pmatrix} y_2(t) \\ y_3(t) \end{pmatrix} = \begin{bmatrix} {Y_{22}}^0(t) \\ {Y_{32}}^0(t) \end{bmatrix} {c_2}^0 + \begin{pmatrix} {p_2}^0(t) \\ {p_3}^0(t) \end{pmatrix} . \tag{28}$$

One verifies directly that $\begin{bmatrix} R_{21} \\ R_{31} \end{bmatrix}$ still has to satisfy the Riccati DE (20a). As has been shown in Section 5.2.2, for satisfying (28) we moreover need the solution of the IVP

$$\frac{d}{dt} \left[\begin{array}{c|c} {Y_{22}}^0 & {p_2}^0 \\ {Y_{32}}^0 & {p_3}^0 \end{array}\right] = \left[\begin{array}{cc|c} \tilde{A}_{22}(t) & \tilde{A}_{23}(t) & \\ \tilde{A}_{32}(t) & \tilde{A}_{33}(t) & \end{array}\right] \left[\begin{array}{c|c} {Y_{22}}^0 & {p_2}^0 \\ {Y_{32}}^0 & {p_3}^0 \end{array}\right] + \left[\begin{array}{c|c} 0 & \tilde{f}_2(t) \\ 0 & \tilde{f}_3(t) \end{array}\right] , \tag{29a}$$

$t \geq \delta$, subject to

$$\left[\begin{array}{c|c} {Y_{22}}^0(\delta) & {p_2}^0(\delta) \\ {Y_{32}}^0(\delta) & {p_3}^0(\delta) \end{array}\right] = \left[\begin{array}{c|c} \bar{Q}_2^0 & \bar{Q}_3^0 {\bar{Q}_3^0}^T \begin{pmatrix} p_2(\delta) \\ p_3(\delta) \end{pmatrix} \end{array}\right] . \tag{29b}$$

Remark 5.9
At $t = \delta$, the solution manifold which does not contain fast decaying modes $\Big(\mathcal{S}_1(\delta) \oplus \mathcal{S}_2(\delta)\Big)$ is approximately described by (cf. (9))

$$\bar{Q}_3^{0^T} \begin{bmatrix} -R_{21}(\delta) & I_l & 0 \\ -R_{31}(\delta) & 0 & I_m \end{bmatrix} x(\delta) \;=\; \bar{Q}_3^{0^T} \begin{pmatrix} p_2(\delta) \\ p_3(\delta) \end{pmatrix} . \tag{30}$$

Note that this relation has been obtained without using any information concerning the BCs.

■

5.2.4 Invariant imbedding

In Section 5.2.3 we have used forward integration in order to obtain the solution manifold at $t = \delta$ that does not contain fast decaying modes. Similarly, one might integrate backward from $t = 1$ for obtaining at, say, $t = 1-\delta$ a description of the solution manifold that does not contain fast increasing modes. However, on the remaining interval $[\delta, 1-\delta]$ all the fast solutions are still potentially present.

In order to find the solution manifold which does not contain fast increasing modes we shall use the invariant imbedding technique discussed in Section 3.4. If the decoupling has been done correctly, then the growth of the fast increasing modes is governed by $\tilde{A}_{11}$. So fundamental solutions corresponding to the adjoint equation

$$\frac{d}{dt} R_{11} = -R_{11}\, \tilde{A}_{11}(t) , \qquad t \geq 0 ,$$

will damp out fast (generally $\| R_{11}(t) \| = O(e^{-\lambda_1 t})$). To obtain the recovery transformation (3.50) we moreover have to solve the DE

$$\frac{d}{dt} \Big[R_{12} \;\; R_{13} \mid g_1 \Big] =$$

$$-R_{11}(t) \Big(\Big[A_{12}(t) \;\; A_{13}(t) \Big] \left[\begin{array}{cc|c} Y_{22}(t) & Y_{23}(t) & p_2(t) \\ Y_{32}(t) & Y_{33}(t) & p_3(t) \end{array}\right] + \Big[0 \;\; 0 \mid f_1(t) \Big] \Big) ,$$

subject to $\Big[R_{12}(0) \;\; R_{13}(0) \mid g_1(0) \Big] = \Big[0 \;\; 0 \mid 0 \Big]$.
The right hand side of this last equation strongly depends on the magnitude of R_{11}. As soon as $\| R_{11}(t) \|$ has become negligible (and $\left[\begin{array}{cc|c} Y_{22}(t) & Y_{23}(t) & p_2(t) \\ Y_{32}(t) & Y_{33}(t) & p_3(t) \end{array}\right]$

stays sufficiently bounded) then $\left[R_{12} \quad R_{13} \mid g_1 \right]$ attains a constant value (cf. Corollary 3.15).

Assume δ has been chosen such that both $\| R_{33}{}^0 \|$ and $\| R_{11}(\delta) \|$ can be neglected ($\delta \approx \max(1/\lambda_1, 1/\lambda_3)$). From the relation (cf. (4.8a)):

$$\left[I_k \quad -R_{12}(t) \quad -R_{13}(t) \right] x(0) = R_{11}(t)\, x_1(t) + g_1(t) \ ,$$

and using the boundedness assumption on $\| x(t) \|$, we obtain that the solution manifold at $t = 0$ which does not contain fast increasing modes $\left(\mathcal{S}_2(0) \oplus \mathcal{S}_3(0) \right)$ is approximately described by (cf. Theorem 3.14)

$$\left[I_k \quad -R_{12}(\delta) \quad -R_{13}(\delta) \right] x(0) = g_1(\delta) \ . \tag{31}$$

5.2.5 Computing the solution

Assume that the Riccati matrix stays sufficiently bounded over the entire interval (and that no intermediate output points are required). Then, using (28) with $t = 1$ and (27) we obtain the relation:

$$\begin{bmatrix} -R_{21}(1) & I_l & 0 \\ -R_{31}(1) & 0 & I_m \end{bmatrix} x(1) = \tag{32}$$

$$\begin{bmatrix} Y_{22}{}^0(1) \\ Y_{32}{}^0(1) \end{bmatrix} \left(\left[R_{22}{}^0 \ R_{23}{}^0 \right] \begin{pmatrix} x_2(0) \\ x_3(0) \end{pmatrix} + \bar{Q}_2{}^{0T} \begin{pmatrix} p_2(\delta) \\ p_3(\delta) \end{pmatrix} \right) + \begin{pmatrix} p_2{}^0(1) \\ p_3{}^0(1) \end{pmatrix} .$$

Together with (31) and the BCs (18) this yields sufficient information to determine the solution at the boundary points:

$$\left[\begin{array}{c|c} B^0 & B^1 \\ \hline \left[\begin{array}{c|cc} I_k & -R_{12}(\delta) & -R_{13}(\delta) \\ \hline 0 & \multicolumn{2}{c}{\left[\begin{array}{c} -Y_{22}{}^0(1) \\ -Y_{32}{}^0(1) \end{array} \right] \left[R_{22}{}^0 \ R_{23}{}^0 \right]} \end{array} \right] & \left[\begin{array}{ccc} 0 & 0 & 0 \\ \hline -R_{21}(1) & I_l & 0 \\ -R_{31}(1) & 0 & I_m \end{array} \right] \end{array} \right] \begin{pmatrix} x(0) \\ x(1) \end{pmatrix}$$

$$= \begin{pmatrix} b \\ g_1(\delta) \\ \begin{pmatrix} p_2{}^0(1) \\ p_3{}^0(1) \end{pmatrix} + \begin{bmatrix} Y_{22}{}^0(1) \\ Y_{32}{}^0(1) \end{bmatrix} \bar{Q}_2{}^{0T} \begin{pmatrix} p_2(\delta) \\ p_3(\delta) \end{pmatrix} \end{pmatrix} . \tag{33}$$

By its simple structure this system can be reduced straightforwardly to an $n \times n$ system, involving $x_2(0)$, $x_3(0)$ and $x_1(1)$ only.

This system has been obtained by integrating on $[0, \delta]$ a complete $n \times (n+1)$ set of DEs, involving $\left[\begin{array}{ccc|c} R_{11} & R_{12} & R_{13} & g_1 \\ R_{21} & Y_{22} & Y_{23} & p_2 \\ R_{31} & Y_{32} & Y_{33} & p_3 \end{array}\right]$, and on $[\delta, 1]$ an $(l+m) \times (k+l+1)$ set of DEs, involving $\left[\begin{array}{cc|c} R_{21} & {Y_{22}}^0 & {p_2}^0 \\ R_{31} & {Y_{32}}^0 & {p_3}^0 \end{array}\right]$. Parts of these DEs are quadratic and parts are linear.

Remark 5.10
In the general case the length of the interval may be such that also slow modes increase too much for obtaining an accurate result. In that case the general decoupling technique as described in [42] can still be used to distinguish between the slowly increasing and slowly decreasing modes.

■

Remark 5.11
The influence of the boundary layer at $t = 1$ on the solution x at $t = 0$ is effectively approximated by the computation of a boundary layer at $t = 0$ for the corresponding adjoint equation. This implies that the integration routine will generally use large stepsizes at the end of the interval, passing the boundary layer at $t = 1$ in just one step. This is allowed, since we have not asked for the solution somewhere near this boundary, but at the endpoints only.
If we are interested in the solution x at $t = 1 - \delta$, for some small δ, then a restart at $1 - \delta$ has to be performed (see Section 5.2.6), which induces small stepsizes.

■

5.2.6 Restarts

In general one also wants to know the solution at intermediate points, which requires a restarting procedure at such points (cf. Section 4.3.2). Moreover, some restarts in order to control the magnitude of the Riccati matrix may be necessary. At each restart a new layer has to be resolved accurately. Although this layer resolution is not important for the solution x as such, it is unavoidable when the invariant imbedding technique is used. However, the intervals where these extra steps have to be taken are again relatively short $\left(O(1/\lambda_1)\right)$ and, as we have seen in Property 5.7, the number of integration steps is independent of λ_1. So, the overhead will be moderate if the number of such restarts is fairly small.

Assume a restart has to be performed at $t = t_1$ ($t_1 > \delta$). So, at $t = t_1$ we possess the values of

$$\left[\begin{array}{cc|c} R_{21}(t_1) & {Y_{22}}^0(t_1) & {p_2}^0(t_1) \\ R_{31}(t_1) & {Y_{32}}^0(t_1) & {p_3}^0(t_1) \end{array}\right].$$

In Chapter 4 a restart at $t = t_1$ was made by the construction of an orthogonal basis for the dominant subspace $\mathcal{S}_1(t_1)$, described by $\mathcal{R}\left(\left[\begin{array}{c} I_k \\ R_{21}(t_1) \end{array}\right]\right)$. This Riccati transformation decoupled between the dominant and dominated solution subspaces. In case of an exponential trichotomy we moreover want to distinguish between the subspaces $\mathcal{S}_1(t_1) \oplus \mathcal{S}_2(t_1)$ and $\mathcal{S}_3(t_1)$. Therefore we continue in a fashion slightly different from the Riccati method of Chapter 4.

We start with the computation of an orthogonal matrix $\bar{Q}^1 = \left[\bar{Q}_2^1 \;\; \bar{Q}_3^1\right] = \left[\begin{array}{cc} \bar{Q}_{22}^1 & \bar{Q}_{23}^1 \\ \bar{Q}_{32}^1 & \bar{Q}_{33}^1 \end{array}\right] \epsilon\, \mathbb{R}^{(l+m)\times(l+m)}$, being a product of elementary Householder transformations and satisfying

$$\left[\begin{array}{cc} \bar{Q}_{22}^1 & \bar{Q}_{23}^1 \\ \bar{Q}_{32}^1 & \bar{Q}_{33}^1 \end{array}\right]^T \left[\begin{array}{c} {Y_{22}}^0(t_1) \\ {Y_{32}}^0(t_1) \end{array}\right] = \left[\begin{array}{c} {R_{22}}^1 \\ 0 \end{array}\right], \tag{34}$$

where ${R_{22}}^1 \epsilon\, \mathbb{R}^{l\times l}$ is upper triangular. The matrix $\bar{Q}^1$ has been constructed such that, at $t = t_1$, $\left[\begin{array}{c} 0 \\ \bar{Q}_2^1 \end{array}\right] \epsilon\, \mathbb{R}^{n\times l}$ represents an orthogonal basis for the subspace of slow modes within the complementary subspace, which is in the Riccati formulation spanned by $\left[\begin{array}{cc} 0 & 0 \\ I_l & 0 \\ 0 & I_m \end{array}\right]$. This construction is possible, since the slow modes dominate the fast decaying modes.

Now we want to construct an orthogonal matrix $U^1 \epsilon\, \mathbb{R}^{n\times n}$ such that

(i) $\mathcal{R}\left(\left[\begin{array}{c} I_k \\ R_{21}(t_1) \\ R_{31}(t_1) \end{array}\right]\right) = \mathcal{R}\left(U^1 \left[\begin{array}{c} I_k \\ 0 \\ 0 \end{array}\right]\right)$, implying that the first k columns of U^1 form an orthogonal basis of $\mathcal{S}_1(t_1)$. (This is similar to the Riccati method of Chapter 4.)

(ii) $\mathcal{R}\left(\left[\begin{array}{cc} I_k & 0 \\ R_{21}(t_1) & \bar{Q}_{22}^1 \\ R_{31}(t_1) & \bar{Q}_{32}^1 \end{array}\right]\right) = \mathcal{R}\left(U^1 \left[\begin{array}{cc} I_k & 0 \\ 0 & I_l \\ 0 & 0 \end{array}\right]\right)$, implying that the first $(k+l)$ columns of U^1 form an orthogonal basis of $\mathcal{S}_1(t_1) \oplus \mathcal{S}_2(t_1)$.

Such a U^1 can be obtained by the QR-decomposition

$$(\bar{Q}^{\,1})^T \begin{bmatrix} -R_{21}(t_1) & I_l & 0 \\ -R_{31}(t_1) & 0 & I_m \end{bmatrix} U^1 = \begin{bmatrix} 0 & V_{22}{}^1 & V_{23}{}^1 \\ 0 & 0 & V_{33}{}^1 \end{bmatrix}, \tag{35a}$$

where $V_{22}{}^1$ and $V_{33}{}^1$ are non-singular and upper triangular. This implies that

$$\begin{bmatrix} I_k & 0 & 0 \\ R_{21}(t_1) & \bar{Q}_{22}^1 & \bar{Q}_{23}^1 \\ R_{31}(t_1) & \bar{Q}_{32}^1 & \bar{Q}_{33}^1 \end{bmatrix} = U^1 \begin{bmatrix} U_{11}{}^1 & U_{12}{}^1 & U_{13}{}^1 \\ 0 & V_{22}{}^1 & V_{23}{}^1 \\ 0 & 0 & V_{33}{}^1 \end{bmatrix}^{-1}. \tag{35b}$$

Now define $x^1 = (U^1)^T x$. Then from the relation (cf. (28))

$$\begin{bmatrix} -R_{21}(t_1) & I_l & 0 \\ -R_{31}(t_1) & 0 & I_m \end{bmatrix} x(t_1) = \begin{pmatrix} y_2(t_1) \\ y_3(t_1) \end{pmatrix} = \begin{bmatrix} Y_{22}{}^0(t_1) \\ Y_{32}{}^0(t_1) \end{bmatrix} c_2{}^0 + \begin{pmatrix} p_2{}^0(t_1) \\ p_3{}^0(t_1) \end{pmatrix},$$

we obtain, using (34) and (35a),

$$\begin{bmatrix} V_{22}{}^1 & V_{23}{}^1 \\ 0 & V_{33}{}^1 \end{bmatrix} \begin{pmatrix} x_2{}^1(t_1) \\ x_3{}^1(t_1) \end{pmatrix} = \begin{bmatrix} R_{22}{}^1 \\ 0 \end{bmatrix} c_2{}^0 + (\bar{Q}^{\,1})^T \begin{pmatrix} p_2{}^0(t_1) \\ p_3{}^0(t_1) \end{pmatrix}. \tag{36}$$

Hence,

$$x_3{}^1(t_1) = (V_{33}{}^1)^{-1} (\bar{Q}_3^1)^T \begin{pmatrix} p_2{}^0(t_1) \\ p_3{}^0(t_1) \end{pmatrix}. \tag{37}$$

The computation of $x_3{}^1(t_1)$ will not introduce large errors, as is seen by

Property 5.12

$$\kappa(V_{jj}{}^1) \;\leq\; 1 + \| \begin{bmatrix} R_{21}(t_1) \\ R_{31}(t_1) \end{bmatrix} \|, \qquad (j = 2 \text{ or } 3).$$

Proof:
From (35a) we obtain the relation ($U^1 = \begin{bmatrix} U_1{}^1 \; U_2{}^1 \; U_3{}^1 \end{bmatrix}$):

$$V_{jj}{}^1 = (\bar{Q}_j^1)^T \begin{bmatrix} -R_{21}(t_1) & I_l & 0 \\ -R_{31}(t_1) & 0 & I_m \end{bmatrix} U_j{}^1.$$

Hence, $\| V_{jj}{}^1 \| \leq \| \begin{bmatrix} -R_{21}(t_1) & I_l & 0 \\ -R_{31}(t_1) & 0 & I_m \end{bmatrix} \| \leq 1 + \| \begin{bmatrix} R_{21}(t_1) \\ R_{31}(t_1) \end{bmatrix} \|$.

Moreover, from (35b) we get $(V_{jj}{}^1)^{-1} = (U_j{}^1)^T \begin{bmatrix} 0 \\ \bar{Q}_j^1 \end{bmatrix}$. So, $\| (V_{jj}{}^1)^{-1} \| \leq 1$.

■

With (37) m components of $x^1(t_1)$ are explicitly known, without using the BCs; the relation holds for *all* solutions x, with $x(0)$ reasonably bounded.

On the interval $[t_1, t_2]$, where t_2 is the next restart point, we compute, corresponding to $A^1(t) = (U^1)^T A(t)\, U^1$ and $f^1(t) = (U^1)^T f(t)$,

(i) a Riccati matrix $\begin{bmatrix} {R_{21}}^1 \\ {R_{31}}^1 \end{bmatrix}$ with $\begin{bmatrix} {R_{21}}^1(t_1) \\ {R_{31}}^1(t_1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$,

(ii) a fundamental solution $\begin{bmatrix} {Y_{22}}^1 \\ {Y_{32}}^1 \end{bmatrix}$ with $\begin{bmatrix} {Y_{22}}^1(t_1) \\ {Y_{32}}^1(t_1) \end{bmatrix} = \begin{bmatrix} I_l \\ 0 \end{bmatrix}$ and

(iii) a particular solution $\begin{pmatrix} {p_2}^1 \\ {p_3}^1 \end{pmatrix}$ with $\begin{pmatrix} {p_2}^1(t_1) \\ {p_3}^1(t_1) \end{pmatrix} = \begin{pmatrix} 0 \\ {x_3}^1(t_1) \end{pmatrix}$, (see (37)).

Then there exists a vector ${c_2}^1 \in \mathbb{R}^l$ such that, for all $t \in [t_1, t_2]$, we have

$$\begin{bmatrix} -{R_{21}}^1(t) & I_l & 0 \\ -{R_{31}}^1(t) & 0 & I_m \end{bmatrix} x^1(t) = \begin{bmatrix} {Y_{22}}^1(t) \\ {Y_{32}}^1(t) \end{bmatrix} {c_2}^1 + \begin{pmatrix} {p_2}^1(t) \\ {p_3}^1(t) \end{pmatrix} . \tag{38}$$

From the initial values at $t = t_1$ we obtain that ${c_2}^1 = {x_2}^1(t_1)$.
A relation between ${c_2}^0$ of (27) and ${c_2}^1$ is found by matching (36)–(38) at $t = t_1$ and requiring just continuity (cf. multiple shooting):

$$\begin{aligned} {V_{22}}^1 {c_2}^1 &= {V_{22}}^1 {x_2}^1(t_1) \\ &= {R_{22}}^1 {c_2}^0 + (\bar{Q}_2^1)^T \begin{pmatrix} {p_2}^0(t_1) \\ {p_3}^0(t_1) \end{pmatrix} - {V_{23}}^1 ({V_{33}}^1)^{-1} (\bar{Q}_3^1)^T \begin{pmatrix} {p_2}^0(t_1) \\ {p_3}^0(t_1) \end{pmatrix} \\ &= {R_{22}}^1 {c_2}^0 + \begin{bmatrix} I_l & -{V_{23}}^1 ({V_{33}}^1)^{-1} \end{bmatrix} (\bar{Q}^1)^T \begin{pmatrix} {p_2}^0(t_1) \\ {p_3}^0(t_1) \end{pmatrix} . \end{aligned} \tag{39}$$

In order to obtain a relation between ${x_2}^i(t_i)$ and ${x_2}^{i+1}(t_{i+1})$ $(i = 1, \ldots, q-1)$ similar steps can be taken at the points t_{i+1} (see Algorithm 5.13).

Similarly to the strategy at $t = 0$ we can use the invariant imbedding technique to obtain at $t = t_1$ a description of the solution manifold which does not contain fast increasing modes. However, since ${x_3}^1(t_1)$ is known explicitly, we can simplify the recovery transformation (3.50) to a relation of the form

$${x_1}^1(t_1) = {R_{11}}^1(t)\, {x_1}^1(t) + {R_{12}}^1(t)\, {x_2}^1(t) + {g_1}^1(t) . \tag{40}$$

Again we obtain that ${R_{11}}^1(t) \approx e^{-\lambda_1 (t - t_1)}$ and therefore ${R_{12}}^1$ and ${g_1}^1$ will converge rapidly to constant values. If $\delta_1 > 0$ is such that ${R_{11}}^1(t_1 + \delta_1)$ has become negligible, then (40) reduces approximately to the relation

$${x_1}^1(t_1) = {R_{12}}^1(t_1 + \delta_1)\, {x_2}^1(t_1) + {g_1}^1(t_1 + \delta_1) . \tag{41}$$

Hence, the k components of ${x_1}^1(t_1)$ can be expressed explicitly in terms of ${x_2}^1(t_1)$, without using the BCs. Again this relation is true for *all* solutions x, with $x(t)$ reasonably bounded.

A similar relation can be obtained at the other restart points. The details will be given in Algorithm 5.13.

5.2.7 The algorithm

Assume a set of restart points $\{t_i\}_{i=0}^{q}$, with $0 < t_1 < t_2 < \cdots < t_q = 1$, is determined by the output requirements and the boundedness condition for ${R_{21}}^i$. We summarize the description of the *Riccati method for a stiff BVP* (with an exponentially trichotomic solution space) in the following

Algorithm 5.13

Step 1. The first subinterval

a. Integrating through the initial layer.

On $[0,\delta]$ solve, for $R = \left[\begin{array}{ccc|c} R_{11} & R_{12} & R_{13} & g_1 \\ R_{21} & Y_{22} & Y_{23} & p_2 \\ R_{31} & Y_{32} & Y_{33} & p_3 \end{array}\right]$ the Riccati DE

$$\frac{d}{dt}R = \left[\begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ A_{21}(t) & 0 & 0 & f_2(t) \\ A_{31}(t) & 0 & 0 & f_3(t) \end{array}\right] + \left[\begin{array}{ccc} 0 & 0 & 0 \\ 0 & A_{22}(t) & A_{23}(t) \\ 0 & A_{32}(t) & A_{33}(t) \end{array}\right] R$$

$$-R\left[\begin{array}{ccc|c} A_{11}(t) & 0 & 0 & f_1(t) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \end{array}\right] - R\left[\begin{array}{ccc} 0 & A_{12}(t) & A_{13}(t) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \hline 0 & 0 & 0 \end{array}\right] R,$$

$$\text{subject to } R(0) = \left[\begin{array}{ccc|c} I_k & 0 & 0 & 0 \\ 0 & I_l & 0 & 0 \\ 0 & 0 & I_m & 0 \end{array}\right].$$

The value of δ is chosen such that both $|R_{11}(\delta)|$ and $|{R_{33}}^0|$ are below a prescribed tolerance, where ${R_{33}}^0$ is obtained from the QR-decomposition (cf. (25)):

$$\left[\begin{array}{cc} Y_{22}(\delta) & Y_{23}(\delta) \\ Y_{32}(\delta) & Y_{33}(\delta) \end{array}\right] = \left[\bar{Q}_2^0 \;\; \bar{Q}_3^0\right] \left[\begin{array}{cc} {R_{22}}^0 & {R_{23}}^0 \\ 0 & {R_{33}}^0 \end{array}\right]. \qquad (42)$$

b. Integration over $[\delta, t_1]$.

On $t \in [\delta, t_1]$ solve, for $R^0 = \left[\begin{array}{cc|c} R_{21} & {Y_{22}}^0 & {p_2}^0 \\ R_{31} & {Y_{32}}^0 & {p_3}^0 \end{array}\right]$ the Riccati DE

$$\frac{d}{dt}R^0 = \left[\begin{array}{cc|c} A_{21}(t) & 0 & f_2(t) \\ A_{31}(t) & 0 & f_3(t) \end{array}\right] + \left[\begin{array}{cc} A_{22}(t) & A_{23}(t) \\ A_{32}(t) & A_{33}(t) \end{array}\right] R^0$$

$$-R^0 \left[\begin{array}{cc|c} A_{11}(t) & 0 & f_1(t) \\ 0 & 0 & 0 \\ \hline 0 & 0 & 0 \end{array}\right] - R^0 \left[\begin{array}{cc} A_{12}(t) & A_{13}(t) \\ 0 & 0 \\ \hline 0 & 0 \end{array}\right] R^0,$$

subject to $R^0(\delta) = \left[\begin{array}{cc|c} R_{21}(\delta) & \bar{Q}_{22}^0 & \bar{Q}_3^0 \bar{Q}_3^{0T} \begin{pmatrix} p_2(\delta) \\ p_3(\delta) \end{pmatrix} \\ R_{31}(\delta) & \bar{Q}_{32}^0 & \end{array}\right]$ (cf. (29b)).

step 2. Restarts and further integration

For $i = 1, \ldots, q$ do

a. Construct the orthogonal matrix $\bar{Q}^i = \left[\begin{array}{cc} \bar{Q}_{22}^i & \bar{Q}_{23}^i \\ \bar{Q}_{32}^i & \bar{Q}_{33}^i \end{array}\right] \in \mathbb{R}^{(l+m)\times(l+m)}$,

satisfying

$$\left[\begin{array}{cc} \bar{Q}_{22}^i & \bar{Q}_{23}^i \\ \bar{Q}_{32}^i & \bar{Q}_{33}^i \end{array}\right]^T \left[\begin{array}{c} {Y_{22}}^{i-1}(t_i) \\ {Y_{32}}^{i-1}(t_i) \end{array}\right] = \left[\begin{array}{c} {R_{22}}^i \\ 0 \end{array}\right].$$

b. Compute an orthogonal matrix $U^i \in \mathbb{R}^{n\times n}$ such that

$$\left[\begin{array}{cc} \bar{Q}_{22}^i & \bar{Q}_{23}^i \\ \bar{Q}_{32}^i & \bar{Q}_{33}^i \end{array}\right]^T \left[\begin{array}{ccc} -{R_{21}}^{i-1}(t_i) & I_l & 0 \\ -{R_{31}}^{i-1}(t_i) & 0 & I_m \end{array}\right] U^i = \left[\begin{array}{ccc} 0 & {V_{22}}^i & {V_{23}}^i \\ 0 & 0 & {V_{33}}^i \end{array}\right].$$

c. Define $Q^i := \prod_{k=1}^{i} U^k$ and, for $t \in [t_i, t_{i+1}]$,

$$\begin{aligned} A^i(t) &:= (Q^i)^T A(t)\, Q^i \\ f^i(t) &:= (Q^i)^T f(t) \\ x^i(t) &:= (Q^i)^T x(t)\,. \end{aligned}$$

d. If $i < q$, then on $t \in [t_i, t_{i+1}]$ solve, for $R^i = \left[\begin{array}{cc|c} {R_{11}}^i & {R_{12}}^i & {g_1}^i \\ {R_{21}}^i & {Y_{22}}^i & {p_2}^i \\ {R_{31}}^i & {Y_{32}}^i & {p_3}^i \end{array}\right]$ the Riccati DE

$$\frac{d}{dt}R^i = \left[\begin{array}{cc|c} 0 & 0 & 0 \\ A_{21}{}^i(t) & 0 & f_2{}^i(t) \\ A_{31}{}^i(t) & 0 & f_3{}^i(t) \end{array}\right] + \left[\begin{array}{ccc} 0 & 0 & 0 \\ 0 & A_{22}{}^i(t) & A_{23}{}^i(t) \\ 0 & A_{32}{}^i(t) & A_{33}{}^i(t) \end{array}\right] R^i$$

$$-R^i \left[\begin{array}{cc|c} A_{11}{}^i(t) & 0 & f_1{}^i(t) \\ 0 & 0 & 0 \\ \hline 0 & 0 & 0 \end{array}\right] - R^i \left[\begin{array}{ccc} 0 & A_{12}{}^i(t) & A_{13}{}^i(t) \\ 0 & 0 & 0 \\ \hline 0 & 0 & 0 \end{array}\right] R^i,$$

subject to $R^i(t_i) = \left[\begin{array}{cc|c} I_k & 0 & 0 \\ 0 & I_l & 0 \\ 0 & 0 & x_3{}^i(t_i) \end{array}\right]$,

where $x_3{}^i(t_i) = (V_{33}{}^i)^{-1} \begin{bmatrix} \bar{Q}_{23}^i \\ \bar{Q}_{33}^i \end{bmatrix}^T \begin{pmatrix} p_2{}^{i-1}(t_i) \\ p_3{}^{i-1}(t_i) \end{pmatrix}$ (cf. (37)).

The first k rows of R^i, i.e., $\left[R_{11}{}^i \quad R_{12}{}^i \mid g_1{}^i\right]$, however, are to be computed only until $\left|R_{11}{}^i\right|$ has become negligible (and, as a consequence, $\left[R_{12}{}^i \mid g_1{}^i\right]$ has reached a constant value).
Assume this is the case at $t = t_i + \delta_i$.

step 3. Computation of the solution at the nodes t_i

a. Solving the multiple shooting system.

Solve the $(n+(q+1)l) \times (n+(q+1)l)$ (multiple shooting) system

$$\tilde{M}\,\tilde{c} = \tilde{d}\,, \tag{43}$$

where $\tilde{M} =$

$$\begin{bmatrix} B^0 \begin{bmatrix} R_{12}(\delta) & R_{13}(\delta) \\ I_l & 0 \\ 0 & I_m \end{bmatrix} & & & & B^1 Q^q \begin{bmatrix} 0 & I_k \\ I_l & 0 \\ 0 & 0 \end{bmatrix} \\ -\begin{bmatrix} R_{22}{}^0 & R_{23}{}^0 \end{bmatrix} & I_l & & & 0 \\ & -R_{22}{}^1 & V_{22}{}^1 & & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & -R_{22}{}^q & V_{22}{}^q & 0 \end{bmatrix}$$

and

$$(\tilde{c}\,|\,\tilde{d}) = \left(\begin{array}{c|l} x_2(0) & b - B^0 \begin{pmatrix} g_1(\delta) \\ 0 \\ 0 \end{pmatrix} - B^1 \begin{pmatrix} 0 \\ 0 \\ {x_3}^q(t_q) \end{pmatrix} \\ x_3(0) & \\ {c_2}^0 & \\ {c_2}^1 & (\bar{Q}_2^{\,0})^T \begin{pmatrix} p_2(\delta) \\ p_3(\delta) \end{pmatrix} \\ \vdots & \left[I_l \quad - {V_{23}}^1 ({V_{33}}^1)^{-1} \right] (\bar{Q}^{\,1})^T \begin{pmatrix} {p_2}^0(t_1) \\ {p_3}^0(t_1) \end{pmatrix} \\ \vdots & \vdots \\ {c_2}^q & \\ {x_1}^q(t_q) & \left[I_l \quad - {V_{23}}^q ({V_{33}}^q)^{-1} \right] (\bar{Q}^{\,q})^T \begin{pmatrix} {p_2}^{q-1}(t_q) \\ {p_3}^{q-1}(t_q) \end{pmatrix} \end{array}\right) .$$

Comment: by the choice of $R^i(t_i)$, we have ${c_2}^i = {x_2}^i(t_i)$ $(i = 1, \ldots, q)$.

b. Computing the remaining parts of $x^i(t_i)$.

Set (cf. (31), (41) and (37), respectively):

$$x_1(0) = R_{12}(\delta)\, x_2(0) + R_{13}(\delta)\, x_3(0) + g_1(\delta) ,$$

$${x_1}^i(t_i) = {R_{12}}^i(t_i + \delta_i)\, {x_2}^i(t_i) + {g_1}^i(t_i + \delta_i) \quad (i = 1, \ldots, q-1)$$

and

$${x_3}^i(t_i) = ({V_{33}}^i)^{-1} (\bar{Q}_3^{\,i})^T \begin{pmatrix} {p_2}^{i-1}(t_i) \\ {p_3}^{i-1}(t_i) \end{pmatrix} \quad (i = 1, \ldots, q) .$$

c. Backtransformation.

With $x(t_i) = Q^i x^i(t_i)$ $(i = 1, \ldots, q)$ the solution x is found at all the restart and boundary points.

■

Remark 5.14

The special structure of the matrix $\tilde{M}$ in (43) can be used to solve the system efficiently. Observe that both ${V_{22}}^i$ and ${R_{22}}^i$ $(i = 1, \ldots, q)$ are upper triangular and of moderate size. Let ${F_{22}}^0 = I_l$ and ${d_2}^0 = 0$. Solve, for $i = 1, \ldots, q$, the upper triangular system

$${V_{22}}^i \left[{F_{22}}^i \mid {d_2}^i \right] = {R_{22}}^i \left[{F_{22}}^{i-1} \mid {d_2}^{i-1} \right] + \left[0 \mid {g_2}^i \right] , \tag{44}$$

where ${g_2}^i = \left[I_l \quad - {V_{23}}^i ({V_{33}}^i)^{-1} \right] (\bar{Q}^{\,i})^T \begin{pmatrix} {p_2}^{i-1}(t_i) \\ {p_3}^{i-1}(t_i) \end{pmatrix}$ (cf. Property 5.12).

Then, for $i = 0, \ldots, q$, we have

(i) ${F_{22}}^i$ is upper triangular

(ii) ${c_2}^i = {F_{22}}^i {c_2}^0 + {d_2}^i$.

Using these matrices and vectors and the relation (27) the multiple shooting system can be reduced to the $n \times n$ system

$$E\,x = e\,, \tag{45}$$

where $E =$

$$\left[B^0 \begin{bmatrix} R_{12}(\delta) & R_{13}(\delta) \\ I_l & 0 \\ 0 & I_m \end{bmatrix} + B^1 Q^q \begin{bmatrix} 0 \\ F_{22}{}^q \begin{bmatrix} R_{22}{}^0 & R_{23}{}^0 \end{bmatrix} \\ 0 \end{bmatrix} \quad B^1 Q^q \begin{bmatrix} I_k \\ 0 \\ 0 \end{bmatrix} \right]$$

and $(\,x \,|\, e\,) =$

$$\left(\begin{pmatrix} x_2(0) \\ x_3(0) \\ x_1{}^q(1) \end{pmatrix} \,\middle|\, b - B^0 \begin{pmatrix} g_1(\delta) \\ 0 \\ 0 \end{pmatrix} - B^1 Q^q \begin{pmatrix} 0 \\ d_2{}^q + F_{22}{}^q (\bar{Q}_2^0)^T \begin{pmatrix} p_2(\delta) \\ p_3(\delta) \end{pmatrix} \\ x_3{}^q(1) \end{pmatrix} \right).$$

The stability of the recursion (44) depends on the values of $\| (V_{22}{}^i)^{-1} R_{22}{}^i \|$. Similar to Property 5.12 we have that $\| (V_{22}{}^i)^{-1} \| \leq 1$, for all i. The values of $\| R_{22}{}^i \|$ can not be large, since they indicate the measure of growth of slow modes on the interval $[t_i, t_{i+1}]$. However, if some of the elements of $R_{22}{}^i$ become large, then a more advanced double sweep technique like in [39] can be used to decouple between the slowly increasing and slowly decaying modes. With 'large' we mean here that, for some i, $\eta / \| F_{22}{}^i \|$ is of the same order as the required accuracy, where η is the machine precision.

■

Remark 5.15
Suppose $l = 0$. Then (41) reduces to $x_1{}^i(t_i) = g_1{}^i(t_i + \delta_i)$. Hence, together with (37), $x^i(t_i)$ ($i = 1, \ldots, q-1$) (i.e., the solution at the intermediate output points) is directly obtained. The solution at the boundary points is determined by (31), (37) with $i = q$, and the BCs (18).

■

Remark 5.16
The efficiency of the proposed method highly depends on a correct choice of the dimensions of the Riccati matrix. In Section 4.4 we have already suggested to use the Schur-transformation in order to find out which dimensions have to be chosen. However, this strategy is, like all possible other ones, not waterproof. In Section 3.4 we saw that even separated BCs do not necessarily give us the

correct dimensions. This is the main reason why the BCs are not explicitly used in step 1 of the algorithm.
If, however, the number of zero rows in B^0 turns out to be equal to k (and $l \neq 0$), then a substantial reduction in the number of operations can be achieved. In that case the value of $\begin{bmatrix} R_{21}(0) \\ R_{31}(0) \end{bmatrix}$ can be chosen such that the computation of the fundamental solutions $\begin{bmatrix} {Y_{22}}^i \\ {Y_{32}}^i \end{bmatrix}$ $(i = 0, \ldots, q-1)$ (and at $[0, \delta]$ the computation of $\begin{bmatrix} Y_{22} & Y_{23} \\ Y_{32} & Y_{33} \end{bmatrix}$) has become superfluous. Moreover, the final multiple shooting system reduces to a $k \times k$ system to determine ${x_1}^q(1)$. Of course, the same is true if the number of zero rows in B^0 is equal to $k+l$, since then all the integrations can be performed from right to left.

■

5.2.8 Example

To demonstrate that the reduction technique of the foregoing sections works indeed we use an adapted version of the second example in [41]. Since the system is only 3×3 we shall not look at the efficiency of the reduction.

Let, for $t \epsilon [0, 10]$ and ε_1, ε_2 some given positive (small) parameters, the matrix function A be defined by $A(t) = \begin{bmatrix} a_{11}(t) & a_{12}(t) & a_{13}(t) \\ a_{21}(t) & a_{22}(t) & a_{23}(t) \\ a_{31}(t) & a_{32}(t) & a_{33}(t) \end{bmatrix}$, with

$$a_{11}(t) = \frac{\sin^2(t) - 3\cos^2(t)}{\varepsilon_1}, \qquad a_{12}(t) = \frac{4\sin(t)\cos(t)}{\varepsilon_1} + 1,$$

$$a_{13}(t) = \frac{\cos(t)\left(3\cos^2(t) - \sin^2(t) - \varepsilon_1/\varepsilon_2\right)}{\varepsilon_1} - \sin(t),$$

$$a_{21}(t) = \frac{4\sin(t)\cos(t)}{\varepsilon_1} - 1, \qquad a_{22}(t) = \frac{\cos^2(t) - 3\sin^2(t)}{\varepsilon_1},$$

$$a_{23}(t) = \cos(t) - \frac{4\sin(t)\cos^2(t)}{\varepsilon_1},$$

$$a_{31}(t) = a_{32}(t) = 0, \qquad a_{33}(t) = -1/\varepsilon_2 .$$

Now consider the linear DE

$$\frac{dx}{dt} = A(t)\,x + f(t), \qquad t \epsilon [0, 1], \tag{46}$$

where the term f has been chosen such that $\bar{x}(t) = (e^{-t}, e^{-t}, e^{-t})^T$ is a particular solution. A fundamental solution X of (46) is given by

$$X(t) = \begin{bmatrix} \cos(t) & \sin(t) & \cos(t) \\ -\sin(t) & \cos(t) & 0 \\ 0 & 0 & 1 \end{bmatrix} \operatorname{diag}\left(e^{-3t/\varepsilon_1}, e^{(t-10)/\varepsilon_1}, e^{-t/\varepsilon_2} \right).$$

The (non-separated) BCs are $x(0) + x(10) = b$, where $b \in \mathbb{R}^3$ is such that the solution we look for is equal to $\bar{x}(t) + X(t) \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$.

If both ε_1 and ε_2 are very small, then we have one fast increasing mode, producing a layer of thickness $O(\varepsilon_1)$ at $t = 10$, and two fast decaying modes, with layers of thickness, respectively, $O(\varepsilon_1)$ and $O(\varepsilon_2)$ at $t = 0$. Hence, $k = 1$, $l = 0$ and $m = 2$. Moreover,

$$\mathcal{S}_1(t) = \operatorname{span}\left\{ \begin{pmatrix} \sin(t) \\ \cos(t) \\ 0 \end{pmatrix} \right\}, \; \mathcal{S}_2 = \{\emptyset\}, \; \mathcal{S}_3(t) = \operatorname{span}\left\{ \begin{pmatrix} \cos(t) \\ -\sin(t) \\ 0 \end{pmatrix}, \begin{pmatrix} \cos(t) \\ 0 \\ 1 \end{pmatrix} \right\}.$$

The precise bound for $\operatorname{GAP}\left(\mathcal{S}_1(t), \mathcal{S}_3(t) \right)$ is given by a complicated expression, but it is amply bounded away from zero. Hence, the solution space $\mathcal{S}$ is exponentially dichotomic.
Since the solution subspace $\mathcal{S}_1$ is rotating, the Riccati transformation will need some restarts.

Firstly we solved the problem with $\varepsilon_2 = 10^{-6}$ and $\varepsilon_1 = 10^{-6}, 10^{-9}$, respectively, without prescribing any internal output point. A restart was made as soon as one of the elements of the Riccati matrix $\begin{bmatrix} R_{21} \\ R_{31} \end{bmatrix}$ became in absolute value larger than 3. Since this Riccati matrix will behave like a tangent and $\arctan(3) \approx 1.25$ we were not surprised to find 7 internal shooting points (almost equally spaced). In Table 5.1 the results are shown. It again shows the nice feature of the adaptive stepsize strategy: in the layer the initial stepsize was, respectively, $1.7\ 10^{-7}$ and $1.7\ 10^{-10}$, and at the end of a shooting interval it was increased to, respectively, $2.3\ 10^{-2}$ and $2.1\ 10^{-2}$. As has been shown in Corollary 5.8 the number of integration steps does actually not depend on ε_1.

Secondly we solved this problem with $\varepsilon_1 = 10^{-6}$ and $\varepsilon_2 = 1$. Now $k = l = m = 1$ and

t	error in x_1		error in x_2		error in x_3	
	$\varepsilon_1 = 10^{-6}$	$\varepsilon_1 = 10^{-9}$	$\varepsilon_1 = 10^{-6}$	$\varepsilon_1 = 10^{-9}$	$\varepsilon_1 = 10^{-6}$	$\varepsilon_1 = 10^{-9}$
0	$1.1\,10^{-05}$	$1.2\,10^{-06}$	$-5.7\,10^{-10}$	$2.9\,10^{-11}$	$-1.5\,10^{-11}$	$1.5\,10^{-11}$
1.35	$3.8\,10^{-07}$	$3.6\,10^{-07}$	$-1.7\,10^{-06}$	$-1.6\,10^{-06}$	$1.8\,10^{-12}$	$-2.8\,10^{-08}$
2.60	$1.5\,10^{-10}$	$1.7\,10^{-10}$	$6.5\,10^{-11}$	$8.2\,10^{-11}$	$2.3\,10^{-13}$	$-3.3\,10^{-11}$
3.87	$-2.8\,10^{-11}$	$-2.9\,10^{-11}$	$-1.3\,10^{-11}$	$8.0\,10^{-12}$	0	$2.8\,10^{-11}$
5.13	$3.4\,10^{-12}$	$4.7\,10^{-12}$	$1.5\,10^{-11}$	$1.3\,10^{-11}$	$-2.8\,10^{-14}$	$-2.9\,10^{-12}$
6.39	$-7.7\,10^{-09}$	$-7.7\,10^{-09}$	$7.9\,10^{-10}$	$8.1\,10^{-10}$	0	$9.0\,10^{-12}$
7.65	$7.1\,10^{-10}$	$6.9\,10^{-10}$	$-3.4\,10^{-09}$	$-3.4\,10^{-09}$	$3.6\,10^{-15}$	$2.5\,10^{-11}$
8.92	$-9.3\,10^{-10}$	$-2.0\,10^{-09}$	$-5.2\,10^{-10}$	$-9.4\,10^{-10}$	0	$-2.7\,10^{-11}$
10	$-1.1\,10^{-05}$	$-1.2\,10^{-06}$	$5.7\,10^{-10}$	$-2.7\,10^{-11}$	$3.2\,10^{-12}$	$-8.4\,10^{-12}$

$\varepsilon_2 = 10^{-6}$.

accuracy of integration: 10^{-4}.

total number of steps: 586 ($\varepsilon_1 = 10^{-6}$), 674 ($\varepsilon_1 = 10^{-9}$).

total number of function evaluations: 1038 ($\varepsilon_1 = 10^{-6}$), 1162 ($\varepsilon_1 = 10^{-9}$).

Table 5.1: Absolute accuracy of the computed solutions on the Burroughs B7900 of the Eindhoven University of Technology (machine accuracy: $\frac{1}{2}8^{-13} \approx 7\ 10^{-12}$).

* *

$$\mathcal{S}_1(t) = \text{span}\left\{\begin{pmatrix} \sin(t) \\ \cos(t) \\ 0 \end{pmatrix}\right\},\ \mathcal{S}_2(t) = \text{span}\left\{\begin{pmatrix} \cos(t) \\ 0 \\ 1 \end{pmatrix}\right\} \quad \text{and}$$

$$\mathcal{S}_3(t) = \text{span}\left\{\begin{pmatrix} \cos(t) \\ -\sin(t) \\ 0 \end{pmatrix}\right\}.$$

Again the gaps between these subspaces are bounded away from zero and therefore the solution space $\mathcal{S}$ is exponentially trichotomic.

In Table 5.2 the errors are shown for different required accuracies for integration. We see that the obtained accuracy in the solution of the BVP is proportional to the accuracy of integration.

5.2.9 Conclusion

The Riccati method for stiff BVPs, having an exponential trichotomic solution space can be summarized as follows.

A Riccati transformation is used to determine the dominant solution subspace $\mathcal{S}_1$. By the decoupling property of this transformation we obtain a decoupled part that contains slow and fast decaying modes only. For this part the shoot-

t	error in x_1		error in x_2		error in x_3	
	10^{-4}	10^{-6}	10^{-4}	10^{-6}	10^{-4}	10^{-6}
0	$1.6\,10^{-05}$	$4.2\,10^{-09}$	$-7.3\,10^{-10}$	$2.9\,10^{-11}$	$-9.1\,10^{-08}$	$-6.1\,10^{-10}$
1.34	$1.5\,10^{-05}$	$1.1\,10^{-07}$	$-1.7\,10^{-06}$	$5.5\,10^{-12}$	$6.2\,10^{-05}$	$4.7\,10^{-07}$
2.59	$-2.7\,10^{-05}$	$-1.4\,10^{-07}$	$-3.6\,10^{-07}$	$-5.8\,10^{-09}$	$3.1\,10^{-05}$	$1.6\,10^{-07}$
3.91	$-9.3\,10^{-06}$	$-1.7\,10^{-07}$	$-8.2\,10^{-08}$	$4.4\,10^{-11}$	$1.3\,10^{-05}$	$2.3\,10^{-07}$
5.19	$2.3\,10^{-06}$	$3.1\,10^{-08}$	$-5.5\,10^{-12}$	$-2.6\,10^{-10}$	$4.8\,10^{-06}$	$6.7\,10^{-08}$
6.45	$2.2\,10^{-06}$	$2.7\,10^{-08}$	$3.0\,10^{-11}$	$2.3\,10^{-10}$	$2.3\,10^{-06}$	$2.9\,10^{-08}$
7.80	$2.9\,10^{-08}$	$5.1\,10^{-10}$	$-6.3\,10^{-13}$	$2.5\,10^{-11}$	$6.7\,10^{-07}$	$7.9\,10^{-09}$
9.05	$-2.0\,10^{-07}$	$-2.1\,10^{-09}$	$-6.2\,10^{-15}$	$-2.9\,10^{-12}$	$2.2\,10^{-07}$	$2.2\,10^{-09}$
10	$-1.6\,10^{-05}$	$-4.2\,10^{-09}$	$7.2\,10^{-10}$	$-3.1\,10^{-11}$	$9.1\,10^{-08}$	$6.2\,10^{-10}$

$\varepsilon_1 = 10^{-6}$, $\varepsilon_2 = 1$.

accuracy of integration: 10^{-4} and 10^{-6}.

total number of steps: 586 (accuracy = 10^{-4}), 1132 (accuracy = 10^{-6}).

total number of function evaluations: 1032 (accuracy = 10^{-4}), 1914 (accuracy = 10^{-6}).

Table 5.2: Absolute accuracy of the computed solutions.

* *

ing technique described in Section 5.2.2 may be used, since outside the initial layer the influence of the fast decaying modes is negligible.
To deal with the influence of the fast increasing modes we use the invariant imbedding technique of Section 3.3. The boundary layer at $t = 1$ is effectively approximated by a boundary layer at $t = 0$ for the adjoint equation (cf. Remark 5.11), or at $t = t_{q-1}$ if intermediate output is required or when the Riccati matrix becomes too large.

This combination implies that on the largest part of the interval $[0, 1]$ we have to solve an $(l + m) \times (k + l + 1)$ system of DEs, which is partially quadratic.

The main advantage of the proposed method is that the involved DEs can all be solved efficiently by a stiffly stable integrator, like LSODA from ODEPACK ([25]), since the solution we are interested in is smooth almost everywhere. This implies that outside some layer the stepsize is controled by the variation of the slow modes and does not depend on the growth behaviour of rapidly changing modes. Moreover, the number of integration steps does not depend on the stiffness of the system.
Also the number of shooting points t_i does *not* depend on the growth behaviour of the fast solutions, but is governed by the rotational activity of $\mathcal{S}_1$ and the output requirements of the user.

Of course, the method also has some disadvantages. In order to solve a linear system of DEs one has to solve a quadratic (matrix) DE. Moreover, at the shooting points t_i $(i = 1, \ldots, q-1)$ steprefinement is done, which does not

correspond to a rapid change of any basis solution. Also the disadvantages of the original Riccati method are still present. This implies that for obtaining efficiency the rotational activity of the dominant subspace $\mathcal{S}_1$ has to be moderate.
Finally, we have to be cautious, since BDF-methods, which are used in LSODA, have the so-called property of *super-stability* (see also Section 5.3). By this we mean that exponentially increasing solutions are numerically damped out when the stepsize h is chosen too large (cf. [16]). Therefore, the stepsize has to be bounded corresponding to the growth behaviour of the most rapidly increasing mode within the subspace of slow modes.

5.3 Some turning point problems

In Section 5.2 we assumed that the solution space $\mathcal{S}$ had an exponential trichotomy. In Remark 5.5 we observed that this assumption excludes the existence of internal layers. Here we shall investigate to some extent how the Riccati method of Chapter 4 (and its implementation) performs when such layers are present. To simplify the discussion we concentrate on singular perturbation problems of the form

$$\varepsilon \frac{dx}{dt} = A(t,\varepsilon)\, x + f(t,\varepsilon)\ , \qquad t \in [-1,1]\ , \tag{47}$$

where ε is a small parameter, $0 < \varepsilon \leq \varepsilon_0$ (ε_0 fixed).

First we shall indicate (and later define) what is meant by a *turning point*. It will turn out that essentially we have two kinds of turning points: in growth or in direction (which may appear both at the same time). These types of turning points have a different effect on the Riccati algorithms of Chapter 4. It will turn out that just some minor implementation modifications will make the algorithms accurate also for two-dimensional turning point problems. Moreover, we conjecture that this result generally will be valid for larger problems too.

5.3.1 Uniform dichotomy

In Remark 2.8 we have noticed already that well-conditioning of a BVP (with the ∞-norm as function norm) implies that the solution space $\mathcal{S}$ is *dichotomic* (but not necessarily exponentially dichotomic). This means that there exists a constant $K > 0$ of moderate size such that $\mathcal{S}$ can be split in $\mathcal{S}_1 \oplus \mathcal{S}_2$, satisfying

$$\phi_1 \epsilon \mathcal{S}_1 \;\Rightarrow\; \| \phi_1(t) \| \leq K \| \phi_1(s) \| , \quad t \leq s ,$$

$$\phi_2 \epsilon \mathcal{S}_2 \;\Rightarrow\; \| \phi_2(t) \| \leq K \| \phi_2(s) \| , \quad t \geq s .$$

Moreover, the gap between $\mathcal{S}_1(t)$ and $\mathcal{S}_2(t)$ is uniformly bounded away from zero.

For singular perturbation problems we shall make an extension of this dichotomy concept.

Definition 5.17
The family of solution space $\mathcal{S}^\epsilon$ of the homogeneous DEs

$$\varepsilon \frac{dx}{dt} = A(t,\varepsilon)\, x , \qquad t \epsilon [-1,1] , \tag{48}$$

is uniformly dichotomic *for $\varepsilon \epsilon (0, \varepsilon_0]$, if each $\mathcal{S}^\epsilon$ can be split in $\mathcal{S}_1^\varepsilon \oplus \mathcal{S}_2^\varepsilon$ such that*

$$\phi_1{}^\epsilon \epsilon \mathcal{S}_1^\varepsilon \;\Rightarrow \| \phi_1{}^\epsilon(t) \| \leq K \| \phi_1{}^\epsilon(s) \| , \quad t \leq s , \tag{49a}$$

$$\phi_2{}^\epsilon \epsilon \mathcal{S}_2^\varepsilon \;\Rightarrow \| \phi_2{}^\epsilon(t) \| \leq K \| \phi_2{}^\epsilon(s) \| , \quad t \geq s , \tag{49b}$$

where the constant $K > 0$ is of moderate size and independent of ε. Moreover, the gap between $\mathcal{S}_1^\varepsilon(t)$ and $\mathcal{S}_2^\varepsilon(t)$ (cf. Definition 1.10) has to be bounded away from zero, uniformly in both t and ε.

■

For the singular perturbation problems we have in mind we shall make the following assumptions:

Assumption 5.18

(i) Both $A(t,\varepsilon)$ and $f(t,\varepsilon)$ are continuous on $[-1,1] \times (0,\varepsilon_0]$. Moreover, if $A(t,0)$ exists, then it is required to be non-zero.

(ii) The family of solution space $\mathcal{S}^\epsilon$ is uniformly dichotomic.

(iii) Solutions of (48) that are slowly varying on the entire interval are grouped together in $\mathcal{S}_2^\varepsilon$.

(iv) The quantity $k = \dim(\mathcal{S}_1^\varepsilon)$ is independent of ε and $0 < k < n$.

(v) Let Z^ϵ be the fundamental solution corresponding to (48) with $Z^\epsilon(-1) = I_n$. Let $U^\epsilon \Sigma^\epsilon V^{\epsilon T}$ be the SVD of $Z^\epsilon(1)$ and define $\gamma^\epsilon = \sigma_k^\varepsilon / \sigma_{k+1}^\varepsilon$, $\varepsilon \epsilon (0,\varepsilon_0]$. Then $\gamma^\epsilon \to \infty$ as $\varepsilon \to 0$.

■

Definition 5.19
According to the SVDs $Z^\varepsilon(1) = U^\varepsilon\,\Sigma^\varepsilon\,V^{\varepsilon T}$ (cf. Assumption 5.18) *the* k*-dimensional subspaces* $\mathcal{S}_1^\varepsilon(t) = \mathcal{R}\Big(Z^\varepsilon(t)\,V_1^\varepsilon\Big)$ *are called the* dominant subspaces. *Similarly the* $(n-k)$*-dimensional subspaces* $\mathcal{S}_2^\varepsilon(t) = \mathcal{R}\Big(Z^\varepsilon(t)\,V_2^\varepsilon\Big)$ *are called the* dominated subspaces.

■

Remark 5.20
Suppose we have a family of BVPs, consisting of the DEs (47) and separated BCs of the form

$$B_\varepsilon^{02}\,x(0) = b_2 \quad \text{and} \quad B_\varepsilon^{11}\,x(1) = b_1\ ,$$

where $B_\varepsilon^{02} \in \mathbb{R}^{(n-k)\times n}$ and $B_\varepsilon^{11} \in \mathbb{R}^{k\times n}$ have full row rank. If there exists a constant $q < 1$ such that both

$$\mathrm{DIST}\Big(\mathcal{R}\Big((B_\varepsilon^{11})^T\Big), \mathcal{S}_1^\varepsilon(1)\Big) \le q \quad \text{and} \quad \mathrm{DIST}\Big(\mathcal{R}\Big((B_\varepsilon^{02})^T\Big), \mathcal{S}_2^\varepsilon(0)\Big) \le q\ ,$$

uniformly in ε, then together with Assumption 5.18 (ii) the BVPs are uniformly well-conditioned (with the ∞-norm as function norm) (cf. [26]).

■

Assumption 5.18 (ii) implies that solutions in $\mathcal{S}_1^\varepsilon$ are nowhere fast decaying and those in $\mathcal{S}_2^\varepsilon$ are nowhere fast increasing. Moreover, in this section we are interested in solution spaces that are not exponentially dichotomic. Therefore we shall assume that the solution subspaces $\mathcal{S}_1^\varepsilon$ contain solutions which are somewhere in the interval slowly varying and somewhere fast increasing. If the DE has solutions with an internal layer, then it is no restriction to assume that this condition is satisfied. The position where, roughly speaking, a solution in $\mathcal{S}_1^\varepsilon$ changes its nature is called a *turning point.* In the sequel we shall give a more refined definition, but first we illustrate by some examples that the above assumptions make sense and allow for a fairly wide class of problems.

Example 5.21
Consider the scalar homogeneous DE

$$\varepsilon\,\frac{dx}{dt} = (1 + e^{t/\varepsilon})^{-1}x\ , \qquad t \in [-1, 1]\ . \tag{50}$$

Solutions of (50) are, for all ε, given by $x_\varepsilon(t) = c\,(1 + e^{-t/\varepsilon})^{-1}$ $(c \in \mathbb{R})$,

where $c_\varepsilon = \| x_\varepsilon(-1) \|$.
Now we may call $t = \xi$ a turning point for (48) if, in a neighourhood of ξ, λ_ε and/or the direction of q_ε changes an order of magnitude. This is formalized in

Definition 5.23
Assume that the DE

$$\varepsilon \frac{dx}{dt} = A(t,\varepsilon)\, x \ , \qquad (t,\varepsilon) \in [-1,1] \times (0,\varepsilon_0] \ , \tag{56}$$

satisfies Assumption 5.18. *Write* $\mathcal{S}^\varepsilon = \mathcal{S}_1^\varepsilon \oplus \mathcal{S}_2^\varepsilon$. *Let the solutions* x_ε *of* (56) *be decomposed in direction and growth as* (cf. (55))

$$x_\varepsilon(t) = c_\varepsilon \ \exp\Big(\int_{-1}^{t} \lambda_\varepsilon(\tau)\, d\tau\Big)\, q_\varepsilon(t) \ .$$

Then (56) *has a* turning point *at* $t = \xi$ *if for* $i = 1$ *and/or* $i = 2$

$$\exists_{L>0} \forall_{h>0} \exists_{\varepsilon_h>0} \forall_{0<\varepsilon<\varepsilon_h} \exists_{x_\varepsilon \in \overline{\mathcal{S}_i^\varepsilon}} \Big[\Big| \ \lambda_\varepsilon(\xi+h) - \lambda_\varepsilon(\xi-h) \ \Big| > L \Big]$$

and/or

$$\exists_{L>0} \forall_{h>0} \exists_{\varepsilon_h>0} \forall_{0<\varepsilon<\varepsilon_h} \exists_{x_\varepsilon \in \overline{\mathcal{S}_i^\varepsilon}} \Big[\| \ q_\varepsilon(\xi+h) - q_\varepsilon(\xi-h) \ \| > L \Big] .$$

In the first case we have a turning point in growth, *whereas in the latter case we have a* directional turning point.

■

Example 5.24

(i) If (56) is a scalar equation, then $\lambda_\varepsilon(t) = \dfrac{A(t,\varepsilon)}{\varepsilon}$. In that case any point where $A(t,\varepsilon)$ changes in size and/or sign is a turning point. For instance: $A(t,\varepsilon) = (1 + e^{t/\varepsilon})^{-1}$ (cf. Example 5.21).

(ii) Consider the DE (51). Then there exists a constant $c \approx 1$ such that $x_\varepsilon(t) = c \begin{pmatrix} E(t,\varepsilon) \\ I(t,\varepsilon) \end{pmatrix} \in \overline{\mathcal{S}_1^\varepsilon}$. Clearly the DE (51) has a directional turning point at $t = 0$, since $I(-h,\varepsilon)/E(-h,\varepsilon) \to 0$, when $\varepsilon \to 0$, and $E(h,\varepsilon)/I(h,\varepsilon) \to 0$, when $\varepsilon \to 0$ ($h > 0$ fixed). Moreover, one can show that

$$\lambda_\varepsilon = \Big(\frac{I\,E}{\sqrt{\pi\,\varepsilon}} - \frac{2t\,E^2}{\varepsilon} \Big) (I^2 + E^2)^{-1} \ ,$$

If the system (47) is given in bordered form, like in (1), then a turning point is sometimes interpreted as a position where an eigenvalue of A_{22} changes sign ([31]). Although intuitively clear, this 'definition' is only appropriate if the eigenvalues are sufficiently indicative for the growth behaviour of solutions. This is true if away from the turning point the rotational activity of the corresponding invariant subspaces is moderate (for instance: independent of ε), as will be the case when A is diagonally dominant and sufficiently smooth ([10]). To obtain a well-conditioned BVP, for all ε, also the eigenvectors of A_{22} have to change drastically at the turning point.

Actually we should consider the *kinematic eigenvalues* (for a definition see [56]). Roughly speaking, these quantities indicate the growth of (pure) fundamental solutions, i.e., of homogeneous modes not containing parts of faster increasing modes.

The definition we shall use is based on quantities like these kinematic eigenvalues and on the Assumption 5.18. To facilitate the notation we introduce *normalized* solution subsets $\overline{\mathcal{S}_1^\varepsilon}$ and $\overline{\mathcal{S}_2^\varepsilon}$, defined by

$$\phi_1{}^\varepsilon \in \overline{\mathcal{S}_1^\varepsilon} \;\Rightarrow\; \phi_1{}^\varepsilon \in \mathcal{S}_1^\varepsilon \wedge \| \phi_1{}^\varepsilon(1) \| = 1 \,, \tag{52a}$$

$$\phi_2{}^\varepsilon \in \overline{\mathcal{S}_2^\varepsilon} \;\Rightarrow\; \phi_2{}^\varepsilon \in \mathcal{S}_2^\varepsilon \wedge \| \phi_2{}^\varepsilon(-1) \| = 1 \,. \tag{52b}$$

Let x_ε be a solution of the homogeneous DE (48). Then x_ε can be decomposed in

$$x_\varepsilon = w_\varepsilon \, q_\varepsilon \,, \tag{53}$$

where $w_\varepsilon = \| x_\varepsilon \|$ is the *size* of x_ε and q_ε indicates the *direction* of x_ε on the unit sphere. One verifies directly that (cf. Section 3.2):

$$\frac{dq_\varepsilon}{dt} = \frac{A(t,\varepsilon)}{\varepsilon} q_\varepsilon - q_\varepsilon \, \lambda_\varepsilon(t) \,, \qquad t \in [-1,1] \,, \tag{54a}$$

$$\frac{dw_\varepsilon}{dt} = \lambda_\varepsilon(t) \, w_\varepsilon \,, \qquad t \in [-1,1] \,, \tag{54b}$$

where

$$\lambda_\varepsilon(t) = q_\varepsilon^T(t) \left(\frac{A(t,\varepsilon)}{\varepsilon} \right) q_\varepsilon(t) \,. \tag{54c}$$

For some modes x_ε this quantity λ_ε is a kinematic eigenvalue corresponding to (48), since (53) can be written as

$$x_\varepsilon(t) = c_\varepsilon \, \exp\Big(\int_{-1}^{t} \lambda_\varepsilon(\tau) \, d\tau \Big) \, q_\varepsilon(t) \,, \tag{55}$$

which are fast increasing for $t \leq 0$ and smooth for $t > \varepsilon$. In the neighbourhood of 0 we have a *layer* or *transient*. Hence, we have a turning point at $t = 0$. If the value of x is given at a non-negative point, then we have a uniformly well-conditioned class of problems.

■

Example 5.22 (cf. Example 2.2)
Consider the BVP

$$\varepsilon \frac{d^2u}{dt^2} + 2t \frac{du}{dt} = 0 , \qquad t \epsilon [-1,1] , \tag{51}$$

subject to $u(-1) = b_2$ and $u(1) = b_1$.

Writing $x_1 = \sqrt{\pi\varepsilon}\, \frac{du}{dt}$ and $x_2 = u$ this DE is transformed into the system

$$\varepsilon \frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{bmatrix} -2t & 0 \\ \sqrt{\varepsilon/\pi} & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} , \qquad t \epsilon [-1,1] ,$$

subject to $x_2(-1) = b_2$ and $x_2(1) = b_1$. Define

$$E(t,\varepsilon) = e^{-t^2/\varepsilon} \quad \text{and} \quad I(t,\varepsilon) = \frac{1}{\sqrt{\pi\varepsilon}} \int_{-\infty}^{t} E(\tau,\varepsilon)\, d\tau .$$

With

$$\mathcal{S}_1^\varepsilon = \text{span}\left\{ \begin{pmatrix} E(t,\varepsilon) \\ I(t,\varepsilon) \end{pmatrix} \right\} \quad \text{and} \quad \mathcal{S}_2^\varepsilon = \text{span}\left\{ \begin{pmatrix} -E(t,\varepsilon) \\ 1 - I(t,\varepsilon) \end{pmatrix} \right\}$$

Assumption 5.18 is satisfied. For all ε sufficiently small we have that the solution

$$x_\varepsilon(t) = \begin{pmatrix} E(t,\varepsilon) \\ I(t,\varepsilon) \end{pmatrix} \epsilon\, \mathcal{S}_1^\varepsilon$$

is fast increasing for $t \leq 0$ and smooth for $t > \sqrt{\varepsilon}$. Again we have a turning point at $t = 0$.

■

5.3.2 The definition of a turning point

In [62] Wasow gives a definition of a turning point by using the (absence of an) asymptotic representation of a formal fundamental solution of (47). He implicitly indicates that a turning point can be caused by a change in the direction matrix and/or in the size matrix (see (1.28)).

which is large for $t < 0$, while $\lambda_\epsilon(h) \to 0$, when $\epsilon \to 0$ ($h > 0$ fixed). Hence, we also have a turning point in growth at $t = 0$.

■

The reason why we explicitly distinguish between the two forms of turning points, direction and/or growth, lies in the fact that they may show up separately in our algorithms. A solution method for a BVP based on a continuous decoupling transformation T (cf. Chapter 3) effectively tries to determine the direction of $\mathcal{S}_1^\epsilon$, whatever the growth activity may be of modes within $\mathcal{S}_1^\epsilon$. Hereafter, the growth of modes is determined by some kind of forward and backward sweep.
The first part of the algorithm, the determination of T, is generally affected only by sudden changes in the direction of $\mathcal{S}_1^\epsilon$, whereas the second part, the computation of a fundamental solution of the transformed system, might be affected by sudden changes in growth. The latter kind of problems we shall consider first.

5.3.3 Turning points in growth

Assume that the BVP has a turning point at $t = 0$. If the rotational activity of $\mathcal{S}_1^\epsilon(t)$ is moderate, even in the neighbourhood of $t = 0$, then the decoupling transformation T will be slowly varying on the entire interval. The dichotomy of $\mathcal{S}^\epsilon$ prohibits solutions in $\mathcal{S}_i{}^\epsilon$ ($i = 1, 2$) to change from being fast decaying into fast increasing or vica versa. Hence, we only have four possible situations for the non-smooth behaviour of solutions at a turning point, which are shown in Figure 5.3

If the invariant imbedding technique of Section 3.3 (based on integration of the adjoint equation) is used to determine the growth behaviour of solutions in $\mathcal{S}_1^\epsilon$, then the situations C and D (see Figure 5.3) change into A and B, respectively. Hence, we only need to consider the first two possibilities:
A: a slow mode changing into a fast decaying mode
B: a fast decaying mode changing into a slow mode.

We shall show by an example that the integration routine LSODA from ODEPACK ([25]) very likely yields the correct grid and the required accuracy. To this end it is sufficient to consider scalar problems.

Example 5.25
Consider the scalar DE

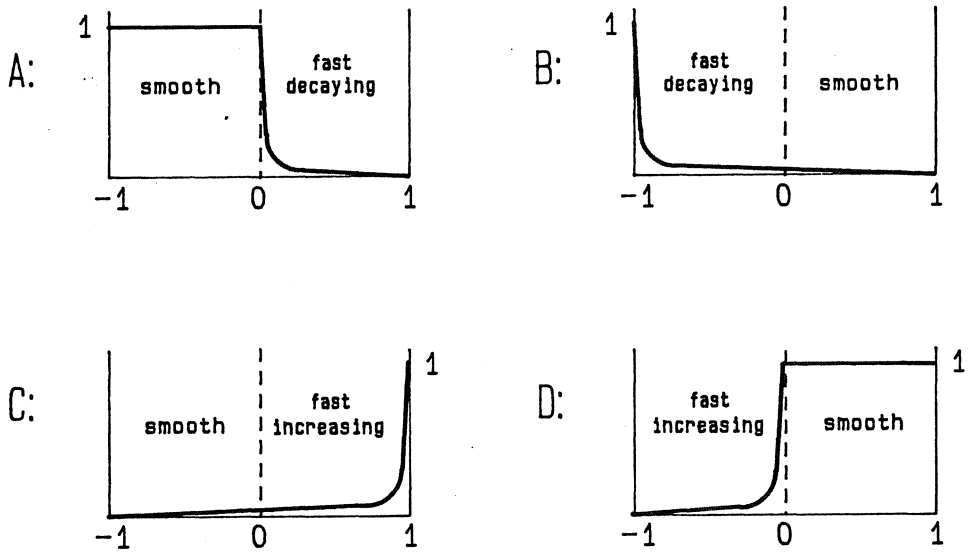

Figure 5.3: Possible turning points in growth

$$\varepsilon \frac{dx}{dt} = a(t)\,x + f(t)\,, \qquad t \in [-1,1]\,, \tag{57}$$

subject to $x(-1) = 1$.

Let

$$a(t) = \begin{cases} 0 & , \ t \le 0 \\ -2t & , \ t > 0 \end{cases} \quad \text{and} \quad f(t) = \begin{cases} -2t\varepsilon & , \ t \le 0 \\ -2t & , \ t > 0 \end{cases} . \tag{58}$$

Then the exact solution $x(t) = \begin{cases} 2 - t^2\,, & t \le 0 \\ 3\,e^{-t^2/\varepsilon} - 1\,, & t > 0 \end{cases}$ is a slow mode that changes into a fast decaying mode.

With

$$a(t) = \begin{cases} 2t & , \ t \le 0 \\ 0 & , \ t > 0 \end{cases} \quad \text{and} \quad f(t) = \begin{cases} 2t & , \ t \le 0 \\ 2t\varepsilon & , \ t > 0 \end{cases} . \tag{59}$$

we obtain the exact solution $x(t) = \begin{cases} 2\,e^{(t^2-1)/\varepsilon} - 1\,, & t \le 0 \\ t^2 + 2\,e^{-1/\varepsilon} - 1\,, & t > 0 \end{cases}$, which is a fast decaying mode that changes into a slow mode.

With $\varepsilon = 10^{-5}$ and a required tolerance (absolute or relative, depending on the size of the solution (cf. (4.70)) of 10^{-6} both problems were solved similarly

by LSODA. In Figure 5.4 the results are shown for situation A. The stepsizes have been indicated on a logarithmic scale and the accuracy on a 1/logarithmic scale.
Since the routine is based on implicit methods, the sudden change in growth is detected and the stepsize is reduced before $t = 0$ is reached. As soon as the layer has been passed, the stepsize is increased significantly. Therefore, the total number of steps taken is reasonable: 88 in situation A and 115 in situation B.

When $\varepsilon = 10^{-8}$ (and $ATOL = RTOL = 10^{-6}$) we get a similar result in situation A: on the entire interval the required accuracy is obtained and the total number of steps taken is equal to 100 (which is in agreement with the results for $\varepsilon = 10^{-5}$ (cf. Corollary 5.8)).
In situation B, however, the computed solution for $t > 0$ turned out to be completely wrong; steprefinement was performed after first accepting one relatively large step passing $t = 0$. The error introduced in this step does not damp out. With a small modification, forcing the integration routine to evaluate a corresponding new Jacobian at each BDF-step, this difficulty could be repaired. In our case this modification is not really expensive ($\approx +10\%$), since in the layers an Adams method with function iteration is used. Moreover, an explicit form of the Jacobian is available (cf. (4.71)), whereas the matrices involved ($\tilde{A}_{11}$ and $\tilde{A}_{22}$) are computed anyway. So it is only the extra decomposition that counts. After this modification the results were similar to the case $\varepsilon = 10^{-5}$.

■

It is to be expected that these observations hold more generally, and so a turning point in growth will be noticed by the integration routine and handled correctly. However, in many 2-dimensional applications the spectrum of $A(t, \varepsilon)$ contains one eigenvalue which is $O(\varepsilon)$ and one eigenvalue of moderate size that changes its sign, say at $t = 0$ (cf. Examples 5.29 and 5.30). If away from $t = 0$ the rotational activities of the corresponding eigenvectors are moderate, then we have at the same time a turning point in growth (both of type A and B) and a directional turning point.

5.3.4 Directional turning points

To investigate the effect of a directional turning point on the Riccati method we concentrate on 2-dimensional homogeneous problems ($n = 2$), of the form

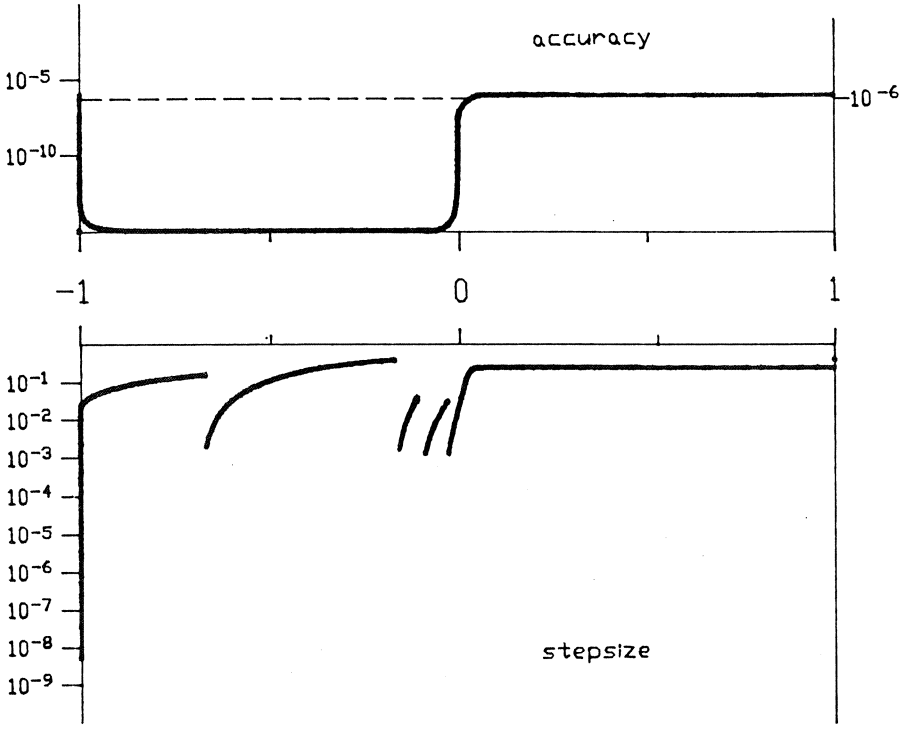

Figure 5.4: Accuracy and stepsizes taken by LSODA

$$\varepsilon \frac{dx}{dt} = A(t,\varepsilon)\, x \ , \qquad t \in [-1,1] \ , \tag{60}$$

since they may give us sufficient insight in possible pitfalls. From now on we shall skip the dependency on ε in the notation of the solution (sub)spaces.

Assume that the solution space $\mathcal{S}$ is uniformly dichotomic and that there exist continuously differentiable functions s_1 (bounded by a constant of moderate size) and s_2 such that, for all $t \in [-1,1]$,

$$\mathcal{S}_1(t) = \mathcal{R}\Big(\begin{bmatrix} 1 \\ s_1(t) \end{bmatrix}\Big) \quad \text{and} \quad \mathcal{S}_2(t) = \mathcal{R}\Big(\begin{bmatrix} 1 \\ s_2(t) \end{bmatrix}\Big) \ .$$

Note that by the restarting technique discussed in Section 4.3.2 the interval $[-1,1]$ is divided in subintervals, such that on each subinterval this condition will be satisfied. Hereby we assume that s_2 may become infinite, implying a direction of $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Now one verifies directly that

(i) $$\mathrm{GAP}\Big(\mathcal{S}_1(t), \mathcal{S}_2(t)\Big) = \frac{|\, s_1(t) - s_2(t)\,|}{\sqrt{1+{s_1}^2(t)}\sqrt{1+{s_2}^2(t)}}$$

(ii) the rotational activity of $\mathcal{S}_i$ ($i = 1,2$) at time t is given by

$$|\frac{ds_i}{dt}(t)\,| / \Big(1+{s_i}^2(t)\Big) \ .$$

Let r be the Riccati matrix for (60), which in this case is a scalar Riccati function. If we have (exponential) difference in growth between solutions in $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively, then $\mathrm{DIST}\Big(\mathcal{R}\Big(\begin{bmatrix} 1 \\ r(t) \end{bmatrix}\Big), \mathcal{S}_1(t)\Big)$ will become small, for almost any initial value $r(-1)$ (cf. Theorem 2.19). This implies that $|\, r(t) - s_1(t)\,|$ will become small. However, in the experiments we sometimes obtain for the numerical approximation $\bar{r}$ that $|\, \bar{r}(t) - s_2(t)\,|$ is small, for $t > 0$. This is explained by means of an example.

Example 5.26

Consider the second order scalar problem

$$\varepsilon \frac{d^2u}{dt^2} = a(t)\frac{du}{dt} \ , \qquad t \in [-1,1] \ , \tag{61}$$

with $a(t) < 0$, for all $t \in [-1,1]$. Suppose we have separated BCs. Let at $t = -1$ the BC be given by $u(-1) + \varepsilon \dfrac{du}{dt}(-1) = 0$. Writing $x_1 = u$ and $x_2 = \varepsilon \dfrac{du}{dt}$ we obtain

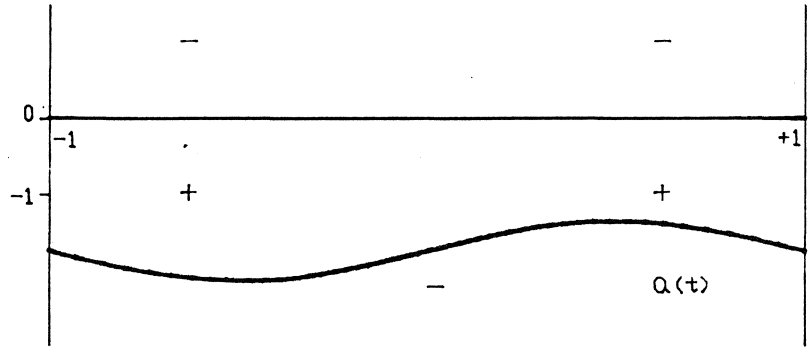

Figure 5.5: Direction field of $\varepsilon \dfrac{dr}{dt} = \Big(a(t) - r\Big)\, r$.

$$\varepsilon \frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & a(t) \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad t \in [-1,1]\,,$$

subject to $x_1(-1) + x_2(-1) = 0$ and a BC at $t = 1$. This system is in correct order (cf. Assumption 4.12), for all $t \in [-1,1]$.

A Riccati transformation of the form $T(t) = \begin{bmatrix} 1 & 0 \\ r(t) & 1 \end{bmatrix}$ that fits to the BC at $t = -1$ is obtained if r satisfies the Riccati DE

$$\varepsilon \frac{dr}{dt} = \Big(a(t) - r\Big)\, r\,, \qquad t \in [-1,1]\,, \tag{62}$$

subject to $r(-1) = -1$. (Observe that we have omitted the orthogonal transformation mentioned in Property 4.16, which would have resulted in a Riccati function r^0, starting with $r^0(-1) = 0$. This is done for reasons of clarity). A sketch of the corresponding direction field is given in Figure 5.5. There the areas where the direction points upward (downward) is indicated by a plus (minus) sign.

Hence, for $0 < \varepsilon \ll |a(t)|$, $r = 0\ (= s_1)$ is a stable fixed point of (62), whereas there is an unstable asymptotic solution $r = s_2(t)$ near the curve $r = a(t)$. Therefore the solution of (62) that fits to the BC at $t = -1$ rapidly tends to $-\infty$ if $s_2(-1) > -1$ and to 0 if $s_2(-1) < -1$.

Assume $s_2(-1) < -1$ and let r_i be a given approximation of $r(t_i)$, for some $t_i \in [-1,1]$. The set of DEs to be solved contains the (stiff) Riccati DE (62), which has to be solved by an implicit method. Typically we shall consider Euler Backward (the simplest BDF-method). Then r_{i+1}, the approximation of

$r(t_{i+1})\Big(= r(t_i + h_i)\Big)$, is a solution of the quadratic equation

$$z^2 + \Big(\frac{\varepsilon}{h_i} - a(t_{i+1})\Big)\, z - \frac{\varepsilon}{h_i} r_i = 0 \ ,$$

which will generally have two solutions in IR. Hence,

$$r_{i+1} = -\frac{1}{2}\Big(\frac{\varepsilon}{h_i} - a(t_{i+1})\Big) \pm \sqrt{\frac{\varepsilon r_i}{h_i} + \frac{1}{4}\Big(\frac{\varepsilon}{h_i} - a(t_{i+1})\Big)^2} \ . \tag{63}$$

It is very natural (and likely) that the integration routine will choose that solution which is the nearest to r_i. Then the wrong sign (minus) will be chosen if $a(t_{i+1}) - 2r_i > \varepsilon/h_i$. If, for some reason, r_i is close to $a(t_{i+1})$ $\Big(\approx s_2(t_{i+1})\Big)$ and h_i is large, then $|\, r_{i+1} - a(t_{i+1})\,|$ will be small too. This does not fit to the prediction that r rapidly tends to zero.

■

The reason for the phenomenon that the integration routine may step on a different solution curve lies in the fact that by BDF-methods fast increasing modes are numerically damped out too. This property is called *super-stability* ([16]). Although almost any exact solution r will move away from s_2, we find numerically that there exists a neighbourhood of s_2 in which the discretized solution $\bar{r}$ will stay. The size of this neighbourhood depends on

- the difference in growth between solutions in $\mathcal{S}_1$ and those in $\mathcal{S}_2$ $(\sim \varepsilon)$
- the stepsize h
- the gap between $\mathcal{S}_1$ and $\mathcal{S}_2$
- the rotational activity of $\mathcal{S}_2$.

There are various reasons why the numerical solution will arrive in this numerically stable, but analytically unstable neighbourhood.
In the first place the initial value of the Riccati function r may be such that it is already close to $s_2(-1)$. This would effectively imply a non-consistent fundamental solution. When the Riccati transformation has been fitted to separated BCs this is only possible for ill-conditioned problems.
The second possibility is that the discretized solution $\{r_i\}$ switches from s_1 to s_2, a situation sketched in Figure 5.6. We see that if this happens the local discretization error is of the order $|\, s_1 - s_2\,|$. This error depends on the stepsize h and the smoothness of s_1, while $|\, s_1 - s_2\,|$ is closely related to the gap between $\mathcal{S}_1$ and $\mathcal{S}_2$. Hence, this situation will not occur if s_1 is smooth on the entire interval and the gap between $\mathcal{S}_1$ and $\mathcal{S}_2$ is sufficiently large. As is shown in Property 2.30 this last requirement is fulfilled as soon as the stability constant

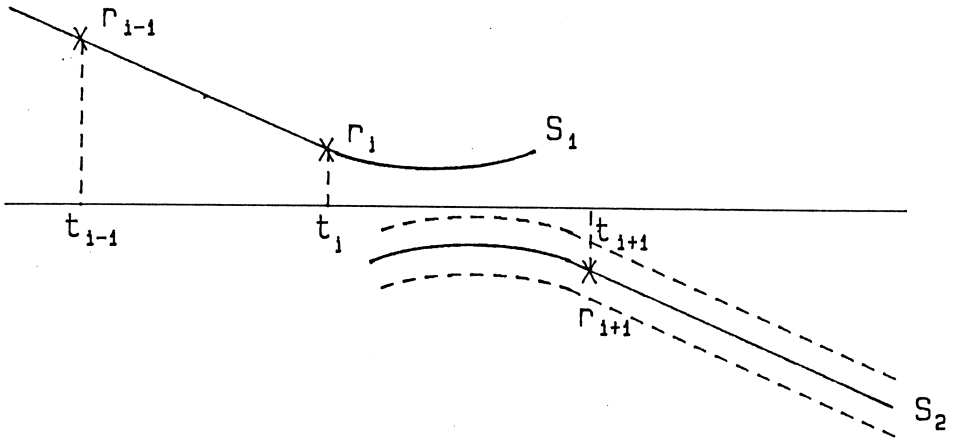

Figure 5.6: A switch to the wrong subspace

α is sufficiently bounded (in ∞-norm).

For some (directional) turning point problems we may expect potential trouble. For such problems the function s_1 is smooth, in general, except for some small region around the turning point, say at $t = 0$. The integration routine may arrive with a relatively large stepsize at $t = 0$, where s_1 changes drastically. Difficulties may then be expected as soon as the first 'guess' for r_{i+1} is in the numerically stable neighbourhood of s_2. This may occur if there exists a function z: $[-1,1] \rightarrow \mathbb{R}^2$ and a positive exponent p (depending on the layer behaviour at $t = 0$) such that

- $z(t) \in \mathcal{S}_1(t)$, for all $t \in [-1, -\varepsilon^p)$
- $z(t) \in \mathcal{S}_2(t)$, for all $t \in (\varepsilon^p, 1]$
- the rotational activity of z at time t is moderate, for all $t \in [-1,1]$.

For a well-conditioned BVP this can happen only at a directional turning point. Therefore a turning point with this property will be called *switchable* and otherwise *non-switchable*.

From the above we conclude that the super-stability phenomenon can only disturb the detection of a switchable turning point. This is illustrated by some examples.

Example 5.27
Let, for some functions p_1 and p_2 (generally depending on ε), the matrix function A be given by

$$A(t) = \begin{bmatrix} -p_2(t) & 1 \\ -p_1(t)\,p_2(t) & p_1(t) \end{bmatrix} .$$

Then

$$A(t) \begin{bmatrix} 1 & 1 \\ p_1(t) & p_2(t) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ p_1(t) & p_2(t) \end{bmatrix} \begin{bmatrix} p_1(t) - p_2(t) & 0 \\ 0 & 0 \end{bmatrix} .$$

Consider the BVP

$$\varepsilon \frac{dx}{dt} = A(t)\,x\ , \qquad t \in [-1, 1]\ ,$$

with the boundary values $x_2(-1)$ and $x_2(1)$ given.

The corresponding Riccati DE, fitting to the BC at $t = -1$, is given by

$$\begin{cases} \varepsilon \dfrac{dr}{dt} = \Big(p_1(t) - r\Big)\Big(r - p_2(t)\Big)\ , & t \geq -1 \\ r(-1) = 0 \end{cases} . \tag{64}$$

Assume that $p_1(t) > p_2(t)$, for $t \in [-1, 0)$, and $p_1(t) < p_2(t)$, for $t \in (0, 1]$. At $t = 0$ we have that $\lambda = 0$ is a defective eigenvalue of $A(t)$ ([20], p.196). This is, however, not essential. For ε sufficiently small, we shall have

$$\mathcal{S}_1(t) = \mathcal{R}\Big(\begin{bmatrix} 1 \\ s_1(t) \end{bmatrix}\Big)\ , \quad \text{with } s_1(t) \approx \begin{cases} p_1(t)\ , & t < 0 \\ p_2(t)\ , & t > 0 \end{cases} ,$$

and

$$\mathcal{S}_2(t) = \mathcal{R}\Big(\begin{bmatrix} 1 \\ s_2(t) \end{bmatrix}\Big)\ , \quad \text{with } s_2(t) \approx \begin{cases} p_2(t)\ , & t < 0 \\ p_1(t)\ , & t > 0 \end{cases} .$$

Solutions in $\mathcal{S}_1$ are fast increasing for $t < 0$ and slowly varying for $t > 0$. On the other hand, solutions in $\mathcal{S}_2$ are slowly varying for $t < 0$ and fast decaying for $t > 0$. Hence, we have a turning point at $t = 0$. Moreover, turning points exist at those points where $p_1 - p_2$ changes drastically.

Both types of turning points are examplified by the following choices for p_1 and p_2:

$$p_1(t) = \Big(1 + e^{\,(t + \frac{1}{2})/\varepsilon}\Big)^{-1} \tag{65a}$$

$$p_2(t) = (t+1)^2 - 1\ . \tag{65b}$$

In Figure 5.7 we have sketched p_1 and p_2 and have indicated upward (downward) directions of the trajectories of the corresponding Riccati DE (64) by

plus (minus) signs as before.

At $t = -\frac{1}{2}$ we have a turning point both in direction and in growth. However, there does not exist a smooth function switching from s_1 to s_2. Therefore, the integration routine will detect the activity of s_1 at $t = -\frac{1}{2}$ and reduce its stepsize before $t = -\frac{1}{2}$ has been reached. Hence, the computed Riccati function stays close to p_1 (and s_1). This kind of turning points can therefore be handled adequately in general.

At $t = 0$, however, a smooth function switching from s_1 to s_2 does exist, namely p_1. By the smoothness of s_1 on $(-\frac{1}{2} + \varepsilon, 0)$ the integration routine will use large stepsizes near $t = 0$. Consequently it will step over easily from s_1 to s_2, without noticing the instability.
This effect is illustrated in Figure 5.8, where for $\varepsilon = 10^{-5}$ and a required tolerance of 10^{-6} the stepsizes found by LSODA and the difference $|r - p_1|$ are shown. We see that $|r - p_1|$ is small, except for the initial layer at $t = -1$ and the 'high activity'-region of p_1 around $t = -\frac{1}{2}$.

■

Remark 5.28
With the functions p_1 and p_2 given in (65a,b) we cannot obtain (in ∞-norm) a uniformly well-conditioned BVP. This results from the observation that both $s_1(0)$ and $s_2(0)$ go to 0, when $\varepsilon \to 0$. Hence, $\mathrm{GAP}\Big(\mathcal{S}_1^\varepsilon(0), \mathcal{S}_2^\varepsilon(0)\Big) \to 0$ and therefore the stability constant $\alpha \to \infty$, when $\varepsilon \to 0$ (cf. Lemma 2.12).
This is not the only reason for a switch to the dominated subspace. In Example 5.22, for instance, the gap between $\mathcal{S}_1^\varepsilon(t)$ and $\mathcal{S}_2^\varepsilon(t)$ is bounded away from zero, uniformly in t and ε. However, as soon as $\sqrt{\varepsilon}$ becomes smaller than the permitted tolerance, then we again obtain the switch from $\mathcal{S}_1^\varepsilon$ to $\mathcal{S}_2^\varepsilon$ ($\mathcal{S}_1^\varepsilon(-t) = \mathcal{S}_2^\varepsilon(t) \approx 0$, for t not too small).

■

In Example 5.27 not only one eigenvalue changes sign at $t = 0$, but at the same time one of the invariant subspaces degenerates. This is a quite common situation in 2-dimensional problems, as is shown in the next, less contrived, example.

Example 5.29
Consider a well-conditioned BVP of the form

$$\varepsilon^2 \frac{d^2u}{dt^2} = a(t)\frac{du}{dt} + u\,, \qquad t \in [-1, 1]\,, \tag{66}$$

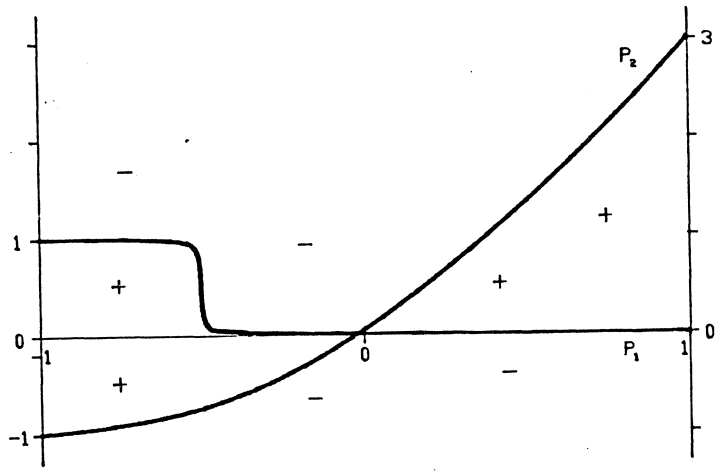

Figure 5.7: Direction field of DE (64).

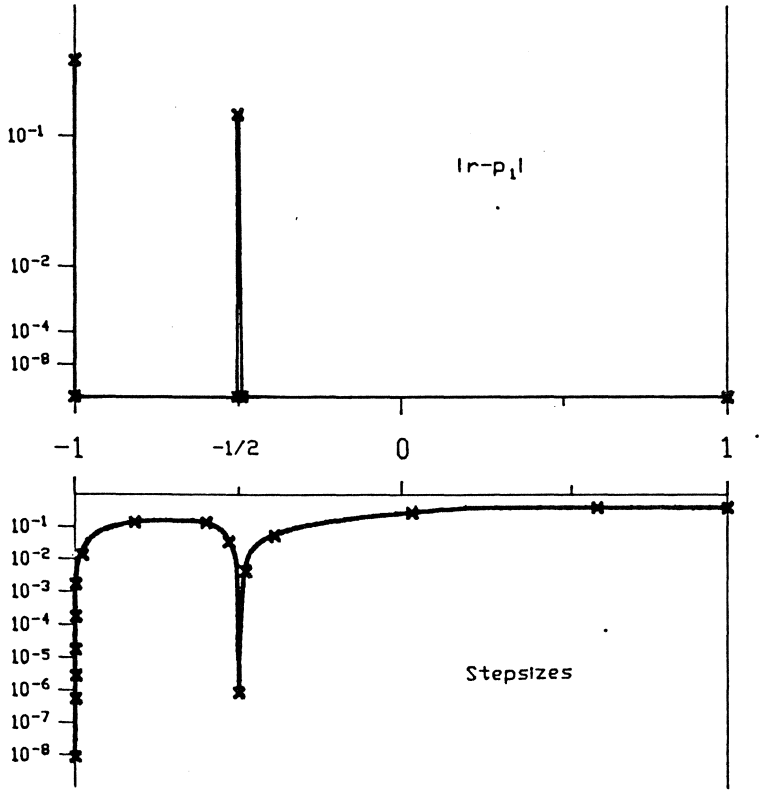

Figure 5.8: Stepsizes taken by LSODA and $|r - p_1|$.

subject to $u(-1) = b_2$ and $u(1) = b_1$.

Let $x_1 = \varepsilon \dfrac{du}{dt}$ and $x_2 = u$. Then

$$\varepsilon \frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{bmatrix} a(t)/\varepsilon & 1 \\ 1 & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = A(t)\, x \ , \qquad t \in [-1, 1] \ , \tag{67}$$

subject to $x_2(-1) = b_2$ and $x_2(1) = b_1$.
Well-conditioning of (67) (but not necessarily of (66)!) implies that $a(-1) \geq 0$ and $a(1) \leq 0$.

The corresponding Riccati DE is given by

$$\begin{cases} \varepsilon \dfrac{dr}{dt} = 1 - \dfrac{a(t)}{\varepsilon} r - r^2 \ , & t \geq -1 \\ r(-1) = 0 \end{cases} . \tag{68}$$

Let the function p be such that $0 = 1 - \dfrac{a(t)}{\varepsilon} p - p^2$, $t \geq -1$. Then $\mathcal{R}(\begin{pmatrix} 1 \\ p \end{pmatrix})$ is an invariant subspace of A, corresponding to the eigenvalue $1/p$. One easily verifies that $p = p_+$ or $p = p_-$, where $p_{\pm} = \dfrac{1}{2\varepsilon}(-a \pm \sqrt{a^2 + 4\varepsilon^2})$. Hence, if $\left| \dfrac{\varepsilon}{a(t)} \right| \ll 1$, then

$$p_+(t) \approx \begin{cases} \dfrac{\varepsilon}{a(t)} & , \text{ if } a(t) > 0 \\ -\dfrac{a(t)}{\varepsilon} - \dfrac{\varepsilon}{a(t)} & , \text{ if } a(t) < 0 \end{cases}$$

and

$$p_-(t) \approx \begin{cases} -\dfrac{a(t)}{\varepsilon} - \dfrac{\varepsilon}{a(t)} & , \text{ if } a(t) > 0 \\ \dfrac{\varepsilon}{a(t)} & , \text{ if } a(t) < 0 \end{cases} .$$

Now the DE (68) can be written as $\varepsilon \dfrac{dr}{dt} = \Big(p_+(t) - r\Big)\Big(r - p_-(t)\Big)$. Therefore, $\mathcal{S}_1(t) = \mathcal{R}(\begin{pmatrix} 1 \\ s_1(t) \end{pmatrix})$, with $s_1(t) \approx p_+(t)$, and $\mathcal{S}_2(t) = \mathcal{R}(\begin{pmatrix} 1 \\ s_2(t) \end{pmatrix})$, with $s_2(t) \approx p_-(t)$. Hence, turning points are found at those places where a changes sign.

If we take $a(t) = t/2 - t^3$ ([30]), then we have turning points at $t = 0$ and $t = \pm \frac{1}{2}\sqrt{2}$. The direction field for the trajectories of the corresponding Riccati DE (68) is sketched in Figure 5.9.

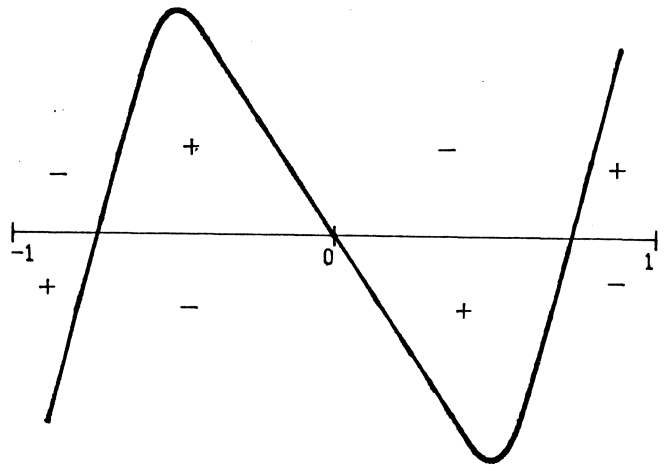

Figure 5.9: Direction field for (68), with $a(t) = t/2 - t^3$.

For $\varepsilon = 10^{-3}$ and $ATOL = RTOL = 10^{-6}$ the turning points are all noticed by LSODA and handled correctly (i.e., steprefinement). However, for $\varepsilon = 10^{-6}$ (and the same tolerances) only the first turning point is noticed, whereafter the numerical solution stays in the neighbourhood of $-a(t)/\varepsilon$. Hence, the last two turning points are missed.

■

From the foregoing we may conclude that an integration routine based on implicit methods will detect a non-switchable turning point. The detection of a switchable turning point, however, depends on the stepsize h. Using large stepsizes the routine may pass the turning point without noticing the rotational activity.

If the position of a turning point is known beforehand, then one can force the integration routine to reduce its stepsize, for instance by requiring an output point just before the turning point is reached. We shall return to this and other aspects in the next section.

5.3.5 The Riccati method

So far we have tested the integration routine LSODA separately for the Riccati DE and for the linear part. However, in the Riccati algorithm all DEs are written in one system, as is described in the Algorithms 4.22 and 4.25. This

means that a turning point undetected by the integration of the quadratic part of this system (direction), may be detected by the integration of the linear part (growth).

As we have seen in Section 5.3.4 a switch into the wrong subspace can occur only when the integration routine is using (relatively) large stepsizes. In some applications the stepsize is not restricted by the quadratic part of the system (the Riccati DE), but by the linear part. This is shown in the next example.

Example 5.30
Consider a general BVP of the form

$$\varepsilon^2 \frac{d^2u}{dt^2} = a(t)\frac{du}{dt} + b(t)\,u\,, \qquad t\,\epsilon\,[-1,1]\,, \tag{69}$$

subject to $u(-1) = 1$ and $u(1) = 2$.

With $x_1 = \varepsilon \dfrac{du}{dt}$ and $x_2 = u$ the DE (69) is transformed into the first order system

$$\varepsilon \frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{bmatrix} a(t)/\varepsilon & b(t) \\ 1 & 0 \end{bmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad t\,\epsilon\,[-1,1]\,, \tag{70}$$

subject to $x_2(-1) = 1$ and $x_2(1) = 2$.

(i) Let $a(t) = t/2 - t^3$ and $b(t) = 1$ (see Example 5.29).
The corresponding direction field can again be found in Figure 5.9.

The dominant subspace $\mathcal{S}_1$ contains modes that are fast increasing when $a(t) > 0$ and smooth when $a(t) < 0$. By the BC at $t = 1$ this implies that these solutions can influence the solution of the BVP on the interval $[\frac{1}{2}\sqrt{2}, 1]$ only. On the other hand, the dominated subspace $\mathcal{S}_2$ contains modes that are smooth when $a(t) > 0$ and fast decaying when $a(t) < 0$. As a consequence, these modes can influence the solution on $[-1, \frac{1}{2}\sqrt{2}]$ only.
Hence, for ε small, the solution x_2 $(= u)$ must have the shape as sketched in Figure 5.10. Since $x_1 = \varepsilon \dfrac{du}{dt}$, it will be small almost everywhere.

We have solved the problem with the code from the RICCATI-package, that is suited for separated BCs (Algorithm 4.22). The tolerance for integration (absolute or relative, depending on the size of the solution) was 10^{-6} in all cases and for all components. A restart was made as soon as the Riccati function became larger than 1 in absolute value.

For $\varepsilon = 10^{-4}$ (so $\varepsilon^2 = 10^{-8}$) and without asking for intermediate output points,

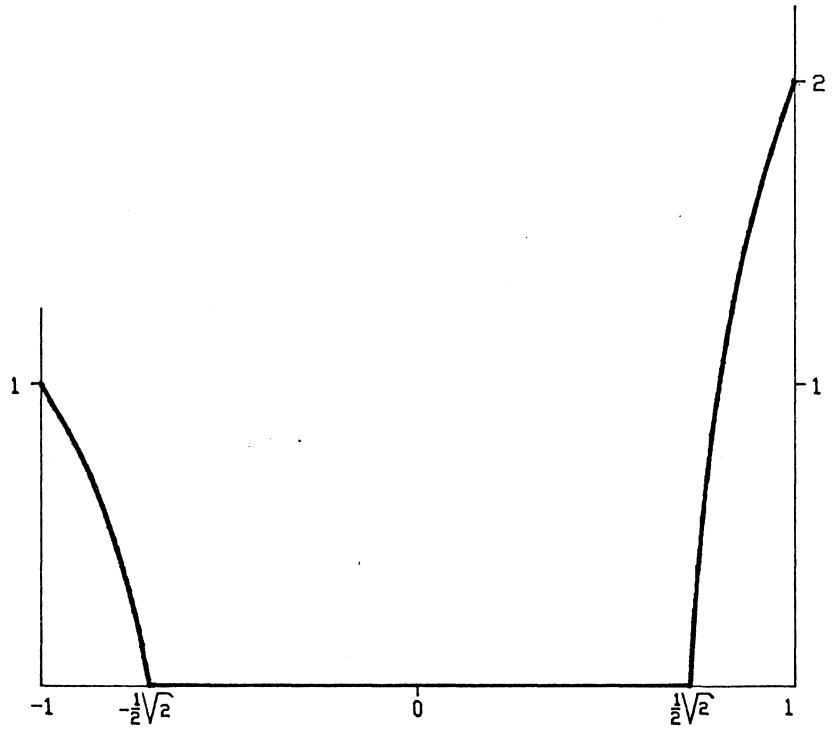

Figure 5.10: Solution of (69), with $a(t) = t/2 - t^3$ and $b(t) = 1$.

the integration restarted at -0.70705, 0.0127 and 0.70716, which is indeed quite close to the turning points $\{0, \pm\frac{1}{2}\sqrt{2}\}$. For $\varepsilon = 10^{-6}$ these restarting points were, respectively, -0.707106, $7.49\ 10^{-5}$ and 0.707107.

That in these cases the super-stability phenomenon does not prevent the detection of a turning point can (partly) be explained as follows.
On the first subinterval $[-1, -\frac{1}{2}\sqrt{2} - \varepsilon)$ we have that both the Riccati function $r(t)$ and $x_1(t)$ are $O(\varepsilon)$. So, $y_2(t) = -r(t)\,x_1(t) + x_2(t) \approx x_2(t)$, which is a slowly varying function. Hence, the stepsize on this subinterval will be controled by the y_2-part of the system.
On the second subinterval $(-\frac{1}{2}\sqrt{2}, 0)$ the dominant modes are smooth. Hence, in contrast to the other (vanishing) parts of the system of DEs, the corresponding fundamental solution and its inverse (R_{11}) are moderately varying. Hence, on this subinterval the stepsize will not be increased significantly. Moreover, as soon as t becomes positive, R_{11} will change drastically, whether the direction of the Riccati transformation is correct or not. Therefore the stepsize will be decreased around $t = 0$.
On the third subinterval $(0, \frac{1}{2}\sqrt{2})$, it might be the activity of the Riccati function which provokes the stepsize to go down at the turning point, since all the other parts of the system are extremely small. This is similar to the situation in Example 5.29, where the activity of the Riccati function was able to detect one turning point too. There it was the first one; here it is the last one.

We also ran the program to compute the solution with specific output at the points -1.0, -0.9, -0.8, -0.7, 0.7, 0.8, 0.9 and 1.0. The results for $\varepsilon = 10^{-6}$ are summarized in Table 5.11. Observe that $\varepsilon^2 = 10^{-12}$, which is smaller than the machine-precision of the Burroughs B7900 ($= \frac{1}{2}\,8^{-13}$), on which the computations were performed.

The last column of Table 5.11 contains the absolute value of the upper left element of the orthogonal transformation Q^i $(i = 0, \ldots, 10)$. It illustrates nicely the direction of the dominant subspace at $t = t_i$ $(i = 1, \ldots, 10)$, which indeed corresponds to the expected directions from Figure 5.9: $Q^i \approx I_2$ when $t_i \in (-1, -\frac{1}{2}\sqrt{2})$ or $t_i \in (0, \frac{1}{2}\sqrt{2})$ and $Q^i \approx \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, when $t_i \in (-\frac{1}{2}\sqrt{2}, 0)$ or $t_i \in (\frac{1}{2}\sqrt{2}, 1]$.

Similar results were obtained for other values of ε.

(ii) Let $a(t) = -\sin^2(\pi t)$ and $b(t) = 1 - t$ ([30]).

Similarly to what has been done in Example 5.29 one can derive that, for

t_i	$u(t_i)$	$\frac{du}{dt}(t_i)$	$r^{i-1}(t_i)$	${R_{11}}^{i-1}(t_i)$	# steps	# funcs.	$\lvert {Q_{11}}^i \rvert$
-1.0	1.00	-2.00	—	—	—	—	1
-0.9	$7.65\ 10^{-1}$	-2.74	$3.58\ 10^{-6}$	$6.23\ 10^{-39}$	94	174	1
-0.8	$4.38\ 10^{-1}$	-3.91	$5.34\ 10^{-6}$	$3.77\ 10^{-36}$	95	164	1
$\approx -\frac{1}{2}\sqrt{2}$	$2.03\ 10^{-9}$	-1.64	1.04	0	219	514	0.69
-0.7	0	0	0.965	$1.45\ 10^{-04}$	320	665	$2\ 10^{-4}$
≈ 0	0	0	1.04	$3.19\ 10^{-14}$	363	705	0.72
0.7	0	0	0.965	0	379	764	1
$\approx \frac{1}{2}\sqrt{2}$	$3.25\ 10^{-5}$	25.3	1.01	0	212	480	0.70
0.8	$8.75\ 10^{-1}$	7.81	0.990	$4.12\ 10^{-05}$	354	741	$9\ 10^{-6}$
0.9	1.53	5.49	$5.34\ 10^{-6}$	0.572	71	101	$4\ 10^{-6}$
1.0	2.00	4.00	$1.59\ 10^{-6}$	0.765	63	89	$2\ 10^{-6}$

Table 5.11: Solution of (70) with $a(t) = t/2 - t^3$, $b(t) = 1$ and $\varepsilon = 10^{-6}$.

* *

$t \neq 0, \pm 1$,

$$\mathcal{S}_1(t) = \mathcal{R}\left(\begin{bmatrix} 1 \\ s_1(t) \end{bmatrix}\right), \quad \text{with } s_1(t) \approx -\frac{a(t)}{\varepsilon\, b(t)} = \frac{\sin^2(\pi t)}{\varepsilon\,(1-t)}$$

and

$$\mathcal{S}_2(t) = \mathcal{R}\left(\begin{bmatrix} 1 \\ s_2(t) \end{bmatrix}\right), \quad \text{with } s_2(t) \approx \frac{\varepsilon}{a(t)} = \frac{\varepsilon}{\sin^2(\pi t)}\,.$$

This situation is sketched in Figure 5.12.

Clearly, the gap between $\mathcal{S}_1(t)$ and $\mathcal{S}_2(t)$ is large, for all t. Moreover, the rotational activity of $\mathcal{S}_1$ at time t is approximated by $\lvert \frac{ds_1}{dt}(t) \rvert / \left(1 + {s_1}^2(t)\right) \approx \frac{\varepsilon\pi \sin(2\pi t)}{\varepsilon^2 + \sin^4(\pi t)}$, which is $O(\varepsilon)$ almost everywhere. For values of t such that $\sin(\pi t) \approx \sqrt{\varepsilon}$, the above approximations are not valid anymore, but we may expect (from the differentiability of $\mathcal{S}_1$) a large rotational activity $\left(O(1/\sqrt{\varepsilon})\right)$.

Running our code with $\varepsilon = 10^{-5}$, an integration tolerance of 10^{-6} and without requiring intermediate output points we found one immediate restart at $t = -0.99896$, whereafter $t = 1$ could be reached with a Riccati function in absolute value not greater than 1.

Observe that in this example we do not have to be afraid for the super-stability

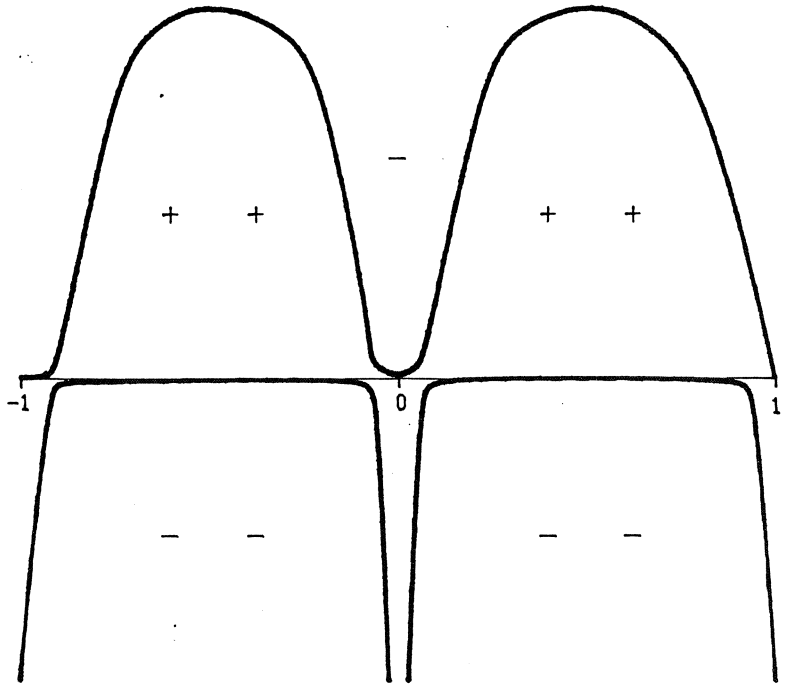

Figure 5.12: Direction field for $a(t) = -\sin^2(\pi t)$ and $b(t) = 1 - t$.

phenomenon, since after the initial rotation ($\approx \begin{bmatrix} \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{bmatrix}$) there is no smooth function switching from the corresponding $\mathcal{S}_1$ to $\mathcal{S}_2$.

As in Example 5.29 we expect solutions in $\mathcal{S}_1$ to grow like $e^{-\int_{-1}^{t} 1/s_1(\tau)\,d\tau}$. Hence, these solutions are fast increasing near $t = -1, 0, 1$ and smooth elsewhere. Similarly, solutions in $\mathcal{S}_2$ are smooth near $t = -1, 0, 1$ and fast decaying elsewhere. Hence, the solution of the BVP is extremely small, at least on $(-1+\varepsilon, 0)$. On $(0, 1)$ we have to be cautious, since $b(1) = 0$, which makes that the boundary layer at $t = 1$ has less influence on the final solution. This is seen also in Table 5.13. There the results are shown for $\varepsilon = 10^{-5}$, intermediate output points -1, 0, 0.1, 0.2, ..., 0.9, 1 and additional output points as soon as $|r(t)| > 1$. If we had chosen $b \equiv 1$, then for instance the last two rows of Table 5.13 would have read:

t_i	$u(t_i)$	$\frac{du}{dt}(t_i)$	$r^{i-1}(t_i)$	${R_{11}}^{i-1}(t_i)$	# steps	# funcs.	$\lvert {Q_{11}}^i \rvert$
-1.0	1.000	$-1.41\ 10^{+06}$	—	—	—	—	1
-0.999	$8.7\ 10^{-10}$	$-1.79\ 10^{-04}$	1.02	$3.87\ 10^{-11}$	97	182	0.699
0.0	$1.5\ 10^{-26}$	$1.50\ 10^{-26}$	$1.05\ 10^{-2}$	$4.32\ 10^{-40}$	435	910	0.707
0.1	0.378	3.56	$1.00\ 10^{-0}$	$5.60\ 10^{-26}$	240	480	$9\ 10^{-5}$
0.2	0.602	1.39	$7.11\ 10^{-5}$	0.627	82	128	$2\ 10^{-5}$
0.3	0.704	0.753	$1.25\ 10^{-5}$	0.855	57	88	$1\ 10^{-5}$
0.4	0.765	0.508	$4.06\ 10^{-6}$	0.920	54	80	$7\ 10^{-6}$
0.5	0.810	0.405	$1.63\ 10^{-6}$	0.945	52	69	$5\ 10^{-6}$
0.6	0.849	0.375	$5.78\ 10^{-7}$	0.955	51	73	$4\ 10^{-6}$
0.7	0.887	0.407	$1.61\ 10^{-7}$	0.957	50	74	$5\ 10^{-6}$
0.8	0.933	0.540	$1.21\ 10^{-6}$	0.951	50	72	$6\ 10^{-6}$
0.9	1.007	1.05	$4.68\ 10^{-6}$	0.927	53	74	$1\ 10^{-5}$
1.0	2.000	$8.10\ 10^{+02}$	$4.04\ 10^{-3}$	0.503	134	223	$4\ 10^{-3}$

Table 5.13: Results for $a(t) = -\sin^2(\pi t)$, $b(t) = 1 - t$ and $\varepsilon = 10^{-5}$.

* *

t_i	$u(t_i)$	$\frac{du}{dt}(t_i)$	$r^{i-1}(t_i)$	${R_{11}}^{i-1}(t_i)$	# steps	# funcs.	$\lvert {Q_{11}}^i \rvert$
0.9	$1.85\ 10^{-25}$	$1.93\ 10^{-24}$	$7.58\ 10^{-5}$	0.582	61	91	$1\ 10^{-4}$
1.0	2.00	$2.00\ 10^{+05}$	1.00	$6.54\ 10^{-26}$	234	399	0.707

The number of steps in the subinterval $[-0.999, 0]$ is twice as large as for the subinterval $[0, 0.1]$, since we have to pass two layers. First the Riccati function r^1 will at $t = -0.999$ move rapidly from 0 to 1. Hereafter its value will stay just below 1, until $t = 0$ is reached. There it changes rapidly back to 0. On $[0, 0.1]$ we have a similar intial layer, but the end layer is missing.

■

5.3.6 Conclusion

In the foregoing sections we have investigated the effect of turning points on the accuracy of the Riccati method of Chapter 4. To this end we mainly have restricted the discussion to two-dimensional problems. We conjecture that the conclusions we draw from them are valid for larger systems also.

We have categorized four kind of turning points, namely

I: turning points in growth
  A: a slow mode changing into a fast mode
  B: a fast mode changing into a slow mode
II: directional turning points
  A: non-switchable
  B: switchable.

A turning point of the form I:A or II:A will be detected by the integration routine and the stepsize will generally have been decreased before the turning point is reached.

For a turning point of the form I:B it turns out that without any modification steprefinement may be performed too late. This can be circumvented by evaluating a new Jacobian at each BDF-step. After this, such turning points will generally be handled correctly.

The remaining category, the switchable turning points, is the most troublesome. The super-stability property of BDF-methods may cause an incorrect approximation of the dominant subspace. This can happen only if (relatively) large stepsizes are used. By writing all the IVPs in one system the stepsize strategy will be rather conservative. Moreover, as soon as one of the components of this system detects the turning point, then it will be handled correctly, in general. The examples indicate that for two-dimensional problems switchable turning points will generally be noticed too.
If not, however, we have the possibility to check whether the direction of the Riccati transformation is correct or not. As we have seen in Chapter 4 the matrix $\tilde{A}_{11} = A_{11} + A_{12} R_{21}$ governs the growth of the dominant solutions. Similarly the growth of the dominated solutions is governed by $\tilde{A}_{22} = A_{22} - R_{21} A_{12}$. For two-dimensional problems both matrix functions are scalar, say $\tilde{a}_{11}$ and $\tilde{a}_{22}$, respectively. Since we have a dichotomic solution space we may expect that $\begin{pmatrix} 1 \\ r(t) \end{pmatrix}$ will represent the direction of a dominant mode and therefore $\tilde{a}_{11}(t)$ will not be much smaller than $\tilde{a}_{22}(t)$, for all $t \in [-1, 1]$. If, however, $\begin{pmatrix} 1 \\ r(t) \end{pmatrix}$ happens to be the direction of the dominated subspace, then generally $\tilde{a}_{11}(t) \ll \tilde{a}_{22}(t)$. Hence, by checking the values of $\tilde{a}_{11}(t)$ and $\tilde{a}_{22}(t)$ an incorrect switch may be noticed.

With this safety checking, the extra Jacobian evaluations and the conservative stepsize strategy the Riccati method is able to detect and to handle correctly all kind of turning points in case of a two-dimensional problem.

Remark 5.31
Checking the values of $\tilde{a}_{11}(t)$ and $\tilde{a}_{22}(t)$ can be generalized to a checking condition for larger problems. In that case the best we can do is comparing $\lambda\Big(\tilde{A}_{11}(t)\Big)$ with $\lambda\Big(\tilde{A}_{22}(t)\Big)$. Observe that for such a check the main part of the determination has already been done. In (4.71) it is indicated that for the determination of the correction term in a BDF-step we have to solve a Sylvester equation of the form

$$\Big(I_{n-k} - \mu\, h\, \tilde{A}_{22}\Big)\, \triangle X_{21} + \triangle X_{21}\, \mu\, h\, \tilde{A}_{11} = h\, D_{21}\ .$$

Therefore $\tilde{A}_{11}$ has already been transformed to quasi-triangular form and $\tilde{A}_{22}$ to Hessenberg form.

■

Chapter 6

Singular Boundary Value Problems

6.1 Introduction

The last kind of problems for which we shall investigate the properties of a continuous decoupling transformation are BVPs with a *singularity of the first kind* ([24], p.114). In this chapter we shall often use the concept of analytic functions, which is defined in

Definition 6.1
A real (matrix/vector) function ϕ, defined on $[0,1]$, is called analytic at $t = 0$ *if there exists a power series $\sum_{k=0}^{\infty} t^k \phi^k$ with positive radius of convergence δ, such that $\phi(t) = \sum_{k=0}^{\infty} t^k \phi^k$, for $0 \leq t < \min(1, \delta)$.*

■

Consider the DE

$$t \frac{dx}{dt} = A(t)\, x + f(t)\,, \qquad t \epsilon (0,1]\,, \tag{1}$$

subject to the BCs

$$x(0) = \lim_{t \downarrow 0} x(t) \ \text{ exists (and finite)} \tag{2a}$$

and

$$B^0\,x(0) + B^1\,x(1) = b\ , \tag{2b}$$

where $B^0, B^1 \in \mathbb{R}^{s\times n}$ are such that $\text{rank}(\left[\,B^0 \mid B^1\,\right]) = s$ and $b \in \mathbb{R}^s$ (s will be specified later such that the BVP has a unique solution).

Assumption 6.2
The matrix function A and the vector function f are continuous on $[\,0,1\,]$ and analytic at $t = 0$.
For t sufficiently small they have the power series expansions

$$A(t) = \sum_{k=0}^{\infty} t^k\,A^k \quad \textit{and} \quad f(t) = \sum_{k=0}^{\infty} t^k\,f^k\ . \tag{3}$$

(Remark: A^k is not the k-th power of A.)

■

In Section 6.2 we shall show that the solution space $\mathcal{S}$ of the homogeneous DE corresponding to (1) generally is the sum of a subspace $\mathcal{S}_1$ of solutions x having a finite limit at $t = 0$ (the *dominant subspace*) and a subspace $\mathcal{S}_2$ of solutions x for which $\lim_{t\downarrow 0} x(t)$ does not exist. The solutions in $\mathcal{S}_2$ are fast decaying for ascending t. Moreover, if $A(0)$ has purely imaginary eigenvalues then $\mathcal{S}_2$ may contain solutions that behave like $\cos(\ln t)$ and $\sin(\ln t)$. Together these solutions constitute the *dominated subspace.*

As has been done in [34] we shall use a decoupling transformation to decouple $\mathcal{S}_1$ and $\mathcal{S}_2$. As a result the existence condition (2a) is replaced by a linear boundary condition at $t = \delta$, for some $\delta > 0$ (cf. Remark 4.19). Next, the invariant imbedding technique of Section 3.3 will be used to find a relation between the undetermined parts of $x(0)$ and $x(\delta)$. By a combination of these two techniques the singular BVP (1), (2a,b) is replaced by a regular BVP on $[\,\delta, 1\,]$. This regular problem can be solved by any of the methods discussed in the foregoing chapters.

The resulting IVPs on $(\,0,\delta\,]$ for the decoupling transformation T and parts of the fundamental solution Y (cf. (3.12) and (3.13)) have nice stability properties: at $t = 0$ the spectra of the corresponding Jacobians lie in the closed left halfplane $\overline{\mathbb{C}^-}$. As we shall see in Theorem 6.17 this implies that all these DEs have solutions that are analytic at $t = 0$. This is surprising, since we do not assume that the solution x is itself analytic; a fundamental solution corresponding to (1) and (2a) generally has singularities at $t = 0$, involving non-integral powers of t and integral powers of $\log(t)$ ([24], §9.5).

By the computation of a number of terms of the corresponding power series expansions we can obtain an accurate approximation at $t = \delta$ of the decoupling transformation T and parts of the fundamental solution Y. It is this property that gives us the oppurtunity to move away from the singularity at $t = 0$.

Remark 6.3
An approximation at $t = \delta$ of the solution subspace $\mathcal{S}_1$ could also be obtained in a more straightforward way. Integrating away from the singular point rapidly gives a good approximation, for almost any initial condition. However, the performance of most numerical integration methods is strongly influenced by the regularity of the solution. Hence, such an integration might be costly. Theorem 6.17 implies that the span of the solutions that satisfy (2a) ($\mathcal{S}_1$) can be described by analytic functions that may be obtained more easily than the solutions itself. Moreover, if $x(0)$ is not completely determined by the condition (2a) (cf. Theorem 6.5), then, in some way or another, we have to take care for the influence of some solutions that start at $t = 0$.

■

In the case of a homogeneous singular BVP with separated BCs, i.e., $B^0 \equiv 0$, these nice properties of a continuous decoupling transformation have been noticed before. In the sixties continuous orthonormalization has been quite popular in the Russian literature ([1]), whereas benefits of the Riccati transformation are, for instance, discussed in [47] and [5].

Of course, a singular BVP can also be solved by other techniques, like finite differences ([27]). However, the performance of such a method depends on the smoothness of the solution x (which in some, but by no means all, interesting applications is analytic too), whereas the efficiency of continuous decoupling and invariant imbedding is mainly influenced by the rotational activity of $\mathcal{S}_1$ and the output requirements.

6.2 Preliminaries

In this section we shall derive and state some elementary results for DEs with a singularity of the first kind. To this end we use the power series expansions given in (3) and, without loss of generality,

Assumption 6.4
The $n \times n$ matrix A^0 has the block-diagonal structure

$$A^0 = \begin{bmatrix} A^0_{kk} & 0 & 0 & 0 \\ 0 & 0 & A^0_{pq} & 0 \\ 0 & 0 & A^0_{qq} & 0 \\ 0 & 0 & 0 & A^0_{mm} \end{bmatrix} \begin{matrix} \updownarrow k \\ \updownarrow p \\ \updownarrow q \\ \updownarrow m \end{matrix} \quad , \tag{4a}$$

$$\underset{k}{\longleftrightarrow} \quad \underset{p}{\longleftrightarrow} \quad \underset{q}{\longleftrightarrow} \quad \underset{m}{\longleftrightarrow}$$

where the dimensions k, p, q and m are determined by the conditions

$$\lambda(A^0_{kk}) \subset \mathbb{C}^+,\ \lambda(A^0_{mm}) \subset \mathbb{C}^-,\ \lambda(A^0_{qq}) \subset i\mathbb{R} \ \textit{and}\ p = \dim\Big(\ker(A^0)\Big)\ . \tag{4b}$$

■

Observe that 0 may or may not belong to $\lambda(A^0_{qq})$. However, the conditions (4b) imply that

$$\operatorname{rank}\Big(\begin{bmatrix} A^0_{pq} \\ A^0_{qq} \end{bmatrix}\Big) = q\ . \tag{5}$$

For $x \in \mathbb{R}^n$ we shall use two kind of partitions:

$$x = \begin{pmatrix} x_k \\ x_p \\ x_q \\ x_m \end{pmatrix} \begin{matrix} \updownarrow k \\ \updownarrow p \\ \updownarrow q \\ \updownarrow m \end{matrix} \quad \text{and} \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \begin{matrix} \updownarrow k+p \\ \updownarrow q+m \end{matrix}\ . \tag{6}$$

Correspondingly to this last partition the matrix function A is partitioned in

$$A(t) = \begin{bmatrix} A_{11}(t) & A_{12}(t) \\ A_{21}(t) & A_{22}(t) \end{bmatrix} \begin{matrix} \updownarrow k+p \\ \updownarrow q+m \end{matrix}\ . \tag{7}$$

$$\underset{k+p}{\longleftrightarrow} \quad \underset{q+m}{\longleftrightarrow}$$

With Assumption 6.4 we obtain that $\lambda(A^0_{11}) \subset \Big(\mathbb{C}^+ \cup \{0\}\Big)$, $\lambda\Big(A^0_{22}\Big) \subset \overline{\mathbb{C}^-}$ and $A^0_{21} = 0$.

By the condition of existence of $x(0)$, parts of $x(0)$ are determined directly, as is shown in

Theorem 6.5
Let x satisfy (1) *and* (2a). *Then*

$$A^0\, x(0) + f^0 = 0\ , \tag{8}$$

where A^0 and f^0 are defined in (3).

Proof:
Assume $A^0\, x(0) + f^0 \neq 0$. From the DE (1) and the continuity of the solution

we obtain, for t sufficiently small,

$$t\frac{dx}{dt} = A^0 x(0) + f^0 + \varepsilon(t) \ , \qquad \text{where } \varepsilon(t) \to 0 \text{ as } t \to 0. \tag{9}$$

Choose $t_0 > 0$. Then integration of (9) leads to

$$x(t) - x(t_0) = \log(\frac{t}{t_0}) \left(A^0 x(0) + f^0 \right) + \int_{t_0}^{t} \frac{\varepsilon(\tau)}{\tau} d\tau \ .$$

For t and t_0 sufficiently small we have, since $\varepsilon(t) \to 0$ as $t \to 0$,

$$\| x(t) - x(t_0) \| \geq \frac{1}{2} \| A^0 x(0) + f^0 \| \left| \log(\frac{t}{t_0}) \right| \ .$$

Hence, $\lim_{t \to 0} \| x(t) - x(t_0) \|$ does not exist, which contradicts (2a).

■

Assumption 6.6

The matrix A^0 and the vector f^0 are such that

$$\begin{pmatrix} f_p{}^0 \\ f_q{}^0 \end{pmatrix} \epsilon \mathcal{R}\left(\begin{bmatrix} A^0_{pq} \\ A^0_{qq} \end{bmatrix} \right) \ . \tag{10}$$

■

Corollary 6.7

Assumption 6.6 *is a necessary condition for the existence of solutions of* (1) *subject to* (2a).

■

Observe that (8) can be written as

$$\begin{cases} A^0_{kk} x_k(0) + f_k{}^0 = 0 \\ \begin{bmatrix} A^0_{pq} \\ A^0_{qq} \end{bmatrix} x_q(0) + \begin{pmatrix} f_p{}^0 \\ f_q{}^0 \end{pmatrix} = 0 \\ A^0_{mm} x_m(0) + f_m{}^0 = 0 \end{cases} \ . \tag{11}$$

Hence, using (5) and Corollary 6.7, $x_k(0)$, $x_q(0)$ and $x_m(0)$ are uniquely defined by the existence condition (2a). This implies, for instance, that if the DE (1) is homogeneous, then only $x_p(0)$ may be non-zero. However, this does not mean that we have p degrees of freedom left, as is seen in

Theorem 6.8
With the Assumptions 6.2 *and* 6.4 *the DE* (1) *has solutions* x, *subject to* (2a), *if and only if* (10) *is satisfied. In that case these solutions form a* $(k+p)$*-dimensional linear manifold.*

Proof: (sketch)
The necessity of (10) for the existence of a solution x has already been given in Corollary 6.7.
Assume (10) is satisfied. Then $x_k(0)$ and $x_2(0)$ are uniquely defined by (11). Define $\bar{x}_0 = \begin{pmatrix} x_k(0) \\ 0 \\ x_2(0) \end{pmatrix} \in \mathbb{R}^n$ and $g(t) = A(t)\,\bar{x}_0 + f(t)$. Then g is analytic at $t = 0$ and $g(0) = 0$. Now consider the DE

$$t\frac{d\phi}{dt} = A(t)\,\phi + g(t)\,, \qquad t \in (0,1]\,. \tag{12}$$

If (12) has a solution ϕ such that $\phi(0)$ exists, then $x(t) = \phi(t) + \bar{x}_0$ is a solution of (1), subject to (2a).

By Theorem 6.5 we know that $\phi_2(0) = 0$. Let $t_0 > 0$ be given. With the variation of constants formula ([24], p.99) we find that ϕ has to satisfy

$$\phi_1(t) = \left(\frac{t}{t_0}\right)^{A_{11}^0}\xi_1 + \int_{t_0}^{t}\left(\frac{t}{\tau}\right)^{A_{11}^0}\left(\frac{A_{12}^0\,\phi_2(\tau) + \left[A_{11}^*(\tau)\;A_{12}^*(\tau)\right]\phi(\tau) + g_1(\tau)}{\tau}\right)d\tau \tag{13a}$$

$$\phi_2(t) = \int_{0}^{t}\left(\frac{t}{\tau}\right)^{A_{22}^0}\left(\frac{\left[A_{21}^*(\tau)\;A_{22}^*(\tau)\right]\phi(\tau) + g_2(\tau)}{\tau}\right)d\tau\,, \tag{13a}$$

where $A^*(t) = A(t) - A^0$ and $\xi_1 \in \mathbb{R}^{k+p}$.
Existence of a solution ϕ of (13a,b) for $t \in (0, t_0]$ and any $\xi_1 \in \mathbb{R}^{k+p}$ can, for sufficiently small t_0, be proved by the method of successive approximations, starting with $\phi^{(0)} \equiv 0$. At the same time, using that $\lambda({A_{11}}^0) \subset \left(\mathbb{C}^+ \cup \{0\}\right)$ (0 being a non-defective eigenvalue ([20], p.196)) and $\lambda({A_{22}}^0) \subset \overline{\mathbb{C}^-}$, we obtain that there exist constants $c_1, c_2 > 0$ such that, for t sufficiently small,

$$\| \phi_1(t) \| \leq c_1 \quad \text{and} \quad \| \phi_2(t) \| \leq c_2\, t\,.$$

For other values of t the boundedness of ϕ is obvious, which implies that ϕ satisfies (12) and $\phi(0)$ exists.

Moreover, ϕ depends linearly on ξ_1. This proves that the solution manifold of (13a,b) is $(k+p)$-dimensional.

■

Corollary 6.9 (cf. [27])
In order to obtain a unique solution of (1), *subject to* (2a,b), *the number s of independent BCs* (2b) *must be equal to* $k+p$.

Proof:
By Theorem 6.8 we see that the existence condition (2a) imposes $q+m$ linear restrictions on the solution. Hence, the solution is uniquely determined by another $k+p$ independent linear restrictions.

■

The solution method we propose is based on continuous decoupling transformations. Then a non-linear IVP has to be solved (cf. Section 3.2). In the next section we shall prove that this non-linear IVP has a solution which is analytic at $t = 0$. To this end we investigate the behaviour of a solution of a special class of non-linear singular IVPs in the complex plane. This complexification is needed, since the space of analytic functions at $t = 0$, defined on $(0,1]$, is not complete.

Theorem 6.10
Let $\delta_1, \delta_2 > 0$ *be given. Define the set* $\mathcal{A}$ *by*

$$\mathcal{A} = \{ (z,x) \in \mathbb{C} \times \mathbb{C}^n \mid |z| \leq \delta_1 \wedge \| x \| \leq \delta_2 \} .$$

Let $F: \mathbb{C} \times \mathbb{C}^n \to \mathbb{C}^n$ *be a given function, being analytic in both arguments on* $\mathcal{A}$, *i.e., the formal power series*

$$F(z,x) = \sum_{i,j_1,\ldots,j_n=0}^{\infty} f_{ij_1\cdots j_n} z^i {x_1}^{j_1} {x_2}^{j_2} \cdots {x_n}^{j_n} ,$$

where $f_{ij_1\cdots j_n} \in \mathbb{C}^n$, *converges for all* $(z,x) \in \mathcal{A}$. *Define*

$$F_1 = \frac{\partial F}{\partial x}(0,0) .$$

Suppose that

(i) $F(0,0) = 0$

(ii) $\lambda(F_1) \subset \overline{\mathbb{C}^-}$.

Then there exists a $\delta_0 > 0$ *such that the IVP*

$$z\frac{dx}{dz} = F(z,x)\ , \tag{14a}$$

subject to

$$\lim_{\substack{t\downarrow 0 \\ t\,\epsilon\,\mathbb{R}^+}} x(t) = 0\ , \tag{14b}$$

has exactly one analytic solution for $|z| \leq \delta_0$.

Proof: (sketch)
By the variation of constants formula we obtain that for analytic functions x (14a,b) is equivalent to the integral equation

$$x(z) = \int_0^1 \tau^{-F_1}\,\bar{F}\Big(\tau z, x(\tau z)\Big)\,\frac{d\tau}{\tau}\ , \tag{15}$$

where $\bar{F}(z,x) = F(z,x) - F_1\,x$.

On $\Big\{\, z\,\epsilon\,\mathbb{C} \;\Big|\; |z| \leq \delta \,\Big\}$ define the sequence of functions $\{\, x^i \,\}$ by

$$\begin{cases} x^0 \equiv 0 \\ x^{i+1}(z) = \displaystyle\int_0^1 \tau^{-F_1}\,\bar{F}\Big(\tau z, x^i(\tau z)\Big)\,\frac{d\tau}{\tau} \end{cases} .$$

In [34] it is shown that there exists a δ, with $0 < \delta \leq \delta_1$, such that for all $|z| \leq \delta$ and $i = 0, 1, 2, \cdots$

(i) $x^i(z)$ is well-defined and analytic at $z = 0$

(ii) $\exists_{c\,\epsilon\,\mathbb{R}^+}$ such that $\| x^i(z) \| \leq c\,|z|$

(iii) on $\Big\{ z\,\epsilon\,\mathbb{C} \;\Big|\; |z| \leq \delta \Big\}$ the sequence $\{\, x^i \,\}$ converges uniformly.

This proves the existence of an analytic solution of (14a,b).

Let y be another analytic solution of (14a,b) and define $e(z) = x(z) - y(z)$. Then e is an analytic function at $z = 0$, satisfying the integral equation

$$e(z) = \int_0^1 \tau^{-F_1}\,\Big(\,\bar{F}\Big(\tau z, x(\tau z)\Big) - \bar{F}\Big(\tau z, y(\tau z)\Big)\,\Big)\,\frac{d\tau}{\tau}\ . \tag{16}$$

It can be shown ([61], p.93) that there exists a constant c_1, such that

$$\| \tau^{-F_1} \| \;\leq\; c_1\,\Big(\,|\ln(\tau)|^{\nu-1} + 1\,\Big)\ , \qquad 0 < \tau \leq 1\ ,$$

where ν is the size of the largest Jordan-block corresponding to an eigenvalue λ of F_1 with $\mathrm{Re}(\lambda) = 0$. Moreover, since $\dfrac{\partial \bar{F}}{\partial x}(0,0) = 0$, there exists a constant

c_2 such that

$$\| \frac{\partial \bar{F}}{\partial x}(z,x) \| \leq c_2 \left(|z| + \| x \| \right) .$$

Therefore, using (16) and the analyticity of x and y, we obtain that

$$\| e(z) \| \leq c_3 |z| \max_{0 \leq \tau \leq 1} \| e(\tau z) \| ,$$

for some positive constant c_3. This implies that $e(z) = 0$, for $|z|$ sufficiently small. By uniqueness of the power series expansion of an analytic function we obtain that $x(z) = y(z)$, for $|z| \leq \delta$.

■

Remark 6.11

Under the assumptions of Theorem 6.10 (14a,b) does not necessarily have a unique solution. For example, let $F(z,x) = x^2$. Then (14a,b) has infinitely many solutions, namely $x(z) = 0 \vee x(z) = (c - \log z)^{-1}$ $(c \in \mathbb{C})$. Of course, only the first one of these solutions is analytic at $z = 0$.
However, one can show that each of the following restrictions makes the solution unique:

- restrictions to the DE:
  - (a) $\lambda(F_1) \subset \mathbb{C}^-$ or
  - (b) F linear in x,
- restrictions to the solution:
  - (c) $\| x(z) \| = O(|z|^\varepsilon)$ $(\varepsilon > 0)$.

■

Clearly we have the following result:

Property 6.12

If all coefficients of F in Theorem 6.10 *are real, i.e., $f_{ij_1 \cdots j_n} \in \mathbb{R}^n$, then the analytic solution of* (14a,b), *restricted to the interval* $[0, \delta_0]$, *is real.*

■

Example 6.13

Consider the linear IVP

$$\begin{cases} t \dfrac{dx}{dt} = A(t)\, x(t) + f(t) , & t \in (0,1] , \\ x(0) = x_0 \end{cases} , \qquad (17a)$$

where A and f are analytic at $t = 0$ and $\lambda\Big(A(0)\Big) \subset \overline{\mathbb{C}^-}$. By Theorem 6.5 we have to assume that $A(0)\,x_0 + f(0) = 0$.
Define, for $t \epsilon [0,1]$, $\phi(t) = x(t) - x_0$. Then ϕ is a solution of

$$\begin{cases} t\dfrac{d\phi}{dt} = A(t)\,\phi + g(t)\ , & t \epsilon (0,1]\ , \\ \phi(0) = 0 \end{cases} \quad , \tag{17b}$$

where $g(t) = A(t)\,x_0 + f(t)$. Hence, $g(0) = 0$.
The linearity of (17b) implies that ϕ is unique and analytic (cf. Remark 6.11) and as a consequence this holds for the solution x of (17a) as well.

■

6.3 The Riccati method

In Section 6.2 we have seen that the existence condition of $x(0)$ implies that $x_k(0)$ and $x_2(0)$ can be computed directly (cf. (11)). Write

$$B^0 = \begin{bmatrix} \underset{\longleftarrow}{B^{0k}} & \underset{\longleftarrow}{B^{0p}} & \underset{\longleftarrow}{B^{02}} \\ k & p & q+m \end{bmatrix} .$$

Then the BC (2b) reduces to the $k + p$ conditions

$$B^{0p}\,x_p(0) + B^1\,x(1) = b - B^0 \begin{pmatrix} x_k(0) \\ 0 \\ x_2(0) \end{pmatrix} , \tag{18}$$

where $x_k(0)$ and $x_2(0)$ are given.
This seems to be too many restrictions, but is not; the fact that $x_k(0)$ is known does not impose extra restrictions to the solution, as may be derived from (13a).

6.3.1 The Riccati transformation

Now we want to use a decoupling transformation to separate between the solution subspace $\mathcal{S}_1$, consisting of all homogeneous modes of (1) existing at $t = 0$, and $\mathcal{S}_2$, the subspace of fast decaying solutions. To this end we shall use a Riccati transformation, although similar results can be derived for continuous orthonormalization methods.

From Theorem 6.8 we observe that the size of the Riccati matrix function R_{21}

must be $(q+m) \times (k+p)$. From the Lyapunov equation (3.13) (and (4.5)) we obtain that R_{21} has to satisfy the DE

$$t \frac{d}{dt} R_{21} = A_{21}(t) + A_{22}(t) R_{21} - R_{21} A_{11}(t) - R_{21} A_{12}(t) R_{21} . \qquad (19)$$

Observe that by Assumption 6.4 the matrix $A(0) = A^0$ has already a nice form: block upper triangular and correctly ordered (cf. Assumption 4.12). By the definition of $\mathcal{S}_1$ the choice $R_{21}(0) = 0$ leads to a consistent fundamental solution (cf. Section 4.4.1). Moreover, the boundedness of R_{21} is mainly determined by the rotational activity of $\mathcal{S}_1$ and the difference in growth between solutions in $\mathcal{S}_1$ and $\mathcal{S}_2$ (cf. Section 4.2). This difference grows unboundedly when $t \downarrow 0$. Therefore we may expect the Riccati matrix to be a rather smooth function, at least for t sufficiently small. Indeed, we have

Theorem 6.14
There exists a $\xi > 0$ such that (19), subject to $R_{21}(0) = 0$, has exactly one solution on $[0, \xi)$ that is analytic at $t = 0$.

Proof:
Consider the entries of a matrix function $U(t) \in \mathbb{R}^{(q+m)\times(k+p)}$ as entries of a vector function $u(t) \in \mathbb{R}^{(q+m)(k+p)}$. Let $F(t, U) = A_{21}(t) + A_{22}(t) U - U A_{11}(t) - U A_{12}(t) U$ and let $\tilde{F}(t, u)$ be the associated vector function. Now it is evident that $\tilde{F}(t, u)$ satisfies the conditions (i) and (ii) of Theorem 6.10 and the condition of Property 6.12. Moreover, for $F_1 = \dfrac{\partial \tilde{F}}{\partial u}(0, 0)$ we have

$$F_1 u = \left({A_{22}}^0 \oplus (-{A_{11}}^0) \right) u ,$$

where $\oplus$ denotes the Kronecker sum. So

$$\lambda(F_1) = \lambda({A_{22}}^0) + \lambda(-{A_{11}}^0) = \left\{ \gamma \in \mathbb{C} \;\middle|\; \gamma = \beta - \alpha , \beta \in \lambda({A_{22}}^0) , \alpha \in \lambda({A_{11}}^0) \right\} .$$

The partition has been chosen such that $\lambda(F_1) \subset \overline{\mathbb{C}^-}$. So, with $R_{21}(0) = 0$, all conditions are satisfied to guarantee on $[0, \xi)$ $(\xi > 0)$ the existence of exactly one solution of (19) that is analytic at $t = 0$.

■

From now on we shall mean by R_{21} this unique analytic solution of (19), starting with $R_{21}(0) = 0$.

Remark 6.15

Write, for t sufficiently small, $R_{21}(t) = \sum_{k=1}^{\infty} t^k C^k$. Using the power series expansion of A we obtain by formal differentiation and multiplication the following relation for C^k ($k = 1, 2, \cdots$):

$$(kI_{q+m} - A_{22}{}^0)\,C^k + C^k\,A_{11}{}^0 = \tag{20}$$

$$A_{21}{}^k + \sum_{m=1}^{k-1}\Big(A_{22}{}^m C^{k-m} - C^{k-m}A_{11}{}^m - C^m \sum_{n=0}^{k-m-1} A_{12}{}^n C^{k-m-n}\Big).$$

By Lemma 1.4 this Sylvester equation has a unique solution, for all $k \geq 1$. Hence, $C^1, C^2, \cdots$ can be calculated consecutively.

■

To show the close relationship with the Riccati matrix in the regular case we formulate the following result (cf. Property 4.1):

Theorem 6.16

Let $\xi > 0$ be such that the Riccati matrix R_{21} exists on $[0, \xi)$. Then there exists a fundamental solution $X = \underset{\substack{\longleftrightarrow \\ k+p}}{\left[X_1\right.} \quad \underset{\substack{\longleftrightarrow \\ q+m}}{\left.X_2\right]}$ *, satisfying the homogeneous DE*

$$t\frac{d}{dt}X = A(t)\,X\ , \qquad t \epsilon (0, 1]\ ,$$

such that, for all $t \epsilon (0, \xi)$, we have the relation

$$R_{21}(t) = X_{21}(t)\,X_{11}{}^{-1}(t)\ .$$

Proof:

If X is a fundamental solution with $X_{11}(t)$ non-singular, for all $t \epsilon (0, \xi)$, then it follows by simple manipulation that $X_{21}\,X_{11}{}^{-1}$ satisfies the same DE as R_{21}. In that case, using Theorem 6.10, it suffices to show that $X_{21}\,X_{11}{}^{-1}$ is analytic at $t = 0$ and $\lim_{t \to 0} X_{21}(t)\,X_{11}{}^{-1}(t) = 0$.

If A^0 has no eigenvalues that differ by a positive integer, then there exists a fundamental solution X of the form

$$X(t) = P(t)\,t^{A^0}\ , \qquad t \epsilon (0, 1]\ ,$$

where P is analytic at $t = 0$ and $P(0) = I_n$ ([24], Theorem 9.5.c). This matrix X already has the desired property, since $X_{21}(t)\,X_{11}{}^{-1}(t) = P_{21}(t)\,P_{11}{}^{-1}(t)$.

If A^0 does have eigenvalues that differ by a positive integer, then the proof

becomes rather technical and therefore we refer to ([34]).

■

Note that the initial value $R_{21}(0) = 0$ exactly corresponds to the BC '$x_2(0)$ given' (see (4.29)). This implies that we may use the technique for separated BCs, as described in Section 4.3 and summarized in Property 4.18. This results in

Theorem 6.17
Let $\xi > 0$ be such that the analytic solution R_{21} of (19) exists on $[0,\xi)$. Then any solution of the incomplete singular IVP

$$t\frac{dx}{dt} = A(t)\,x + f(t)\;, \tag{21a}$$

subject to (cf. (11))

$$\begin{bmatrix} A_{pq}^{0} & 0 \\ A_{qq}^{0} & 0 \\ 0 & A_{mm}^{0} \end{bmatrix} x_2(0) + \begin{pmatrix} f_p{}^{0} \\ f_q{}^{0} \\ f_m{}^{0} \end{pmatrix} = 0\;, \tag{21b}$$

satisfies the relation

$$\begin{bmatrix} -R_{21}(t) & I_{q+m} \end{bmatrix} x(t) = y_2(t)\;, \qquad t\in[0,\xi)\;, \tag{22}$$

where y_2 is the unique and analytic solution of the singular IVP

$$t\frac{dy_2}{dt} = \Big(A_{22}(t) - R_{21}(t)\,A_{12}(t)\Big)\,y_2 - R_{21}(t)\,f_1(t) + f_2(t)\;, \quad t\in(0,\xi)\;, \tag{23}$$

subject to $y_2(0) = x_2(0)$.

Proof:
Observe that the DE (23) can be obtained from the relation (22) by formal differentiation. By the definition of q and Assumption 6.6 the system (21b) has a unique solution. So, (22) is satisfied for $t = 0$. This implies that (22) is valid, for all t.

Define $\tilde{A}_{22}(t) = A_{22}(t) - R_{21}(t)\,A_{12}(t)$ and $\tilde{f}_2(t) = -R_{21}(t)\,f_1(t) + f_2(t)$. Then (23) can be written as

$$t\frac{dy_2}{dt} = \tilde{A}_{22}(t)\,y_2 + \tilde{f}_2(t)\;, \qquad t\in(0,\xi)\;.$$

Note that $\tilde{A}_{22}$ and $\tilde{f}_2$ are analytic functions at $t = 0$. Moreover, $\lambda\Big(\tilde{A}_{22}(0)\Big) = \lambda(A_{22}{}^{0}) \subset \overline{\mathbb{C}^-}$ and, by Theorem 6.5, $\tilde{A}_{22}(0)\,y_2(0) + \tilde{f}_2(0) = 0$. Hence, by

Theorem 6.10, (23) has exactly one solution which is analytic at $t = 0$. Since the DE is linear this solution is the unique solution of (23) (cf. Remark 6.11).

■

So, we have derived that the BC '$x_2(0)$ given' can be transferred to $q + m$ independent linear BCs at a position, say δ, inside the interval $(0, 1]$, away from the singularity. To this end we need to compute the Riccati matrix R_{21}, satisfying (19) with $R_{21}(0) = 0$, and a vector function y_2, satisfying (23) with $y_2(0) = x_2(0)$. Both these functions are analytic at $t = 0$. Hence, we may use the first terms of their power series expansions to move away from the singularity.

6.3.2 Invariant imbedding

By the Riccati transformation of Section 6.3.1 the BC '$x_2(0)$ given' has been transferred to a BC at $t = \delta$. In this section we shall use the invariant imbedding technique of Section 3.3 to transfer the remaining $k + p$ BCs (cf. (18))

$$B^{0p} x_p(0) + B^1 x(1) = b - B^0 \begin{pmatrix} x_k(0) \\ 0 \\ x_2(0) \end{pmatrix} , \tag{24}$$

into BCs of the form

$$B^\delta x(\delta) + B^1 x(1) = \tilde{b} , \tag{25}$$

where $B^\delta \in \mathbb{R}^{(k+p)\times n}$ and $\tilde{b} \in \mathbb{R}^{k+p}$. To this end we have to express $x_p(0)$ in terms of $x_1(\delta)$.

Since $y_2(0)$ is known and, by the special form of the Riccati transformation, $y_1 \equiv x_1$, the recovery transformation (3.50) can be reduced to the simpler form (cf. (4.35))

$$x_1(0) = R_{11}(t)\, x_1(t) + g_1(t) , \qquad t \geq 0 .$$

However, in our situation $x_k(0)$ is known also. Hence, we can restrict ourselves to the relation

$$x_p(0) = R_{p1}(t)\, x_1(t) + g_p(t) , \qquad t \geq 0 , \tag{26}$$

where $R_{p1}(t) \in \mathbb{R}^{p\times(k+p)}$ and $g_p(t) \in \mathbb{R}^p$.
As in Sections 3.5 and 4.3 we obtain that R_{p1} and g_p have to satisfy the IVPs

$$t\,\frac{d}{dt}R_{p1} = -R_{p1}\left(A_{11}(t) + A_{12}(t)\,R_{21}(t)\right),\ t>0,\quad R_{p1}(0) = \begin{bmatrix} 0 & I_p \end{bmatrix} \tag{27a}$$

$$t\,\frac{d}{dt}g_p = -R_{p1}(t)\left(A_{12}(t)\,y_2(t) + f_1(t)\right),\ t>0,\quad g_p(0) = 0\,. \tag{27b}$$

Define $\tilde{A}_{11}(t) = A_{11}(t) + A_{12}(t)\,R_{21}(t)$. Observe that $\tilde{A}_{11}$ is analytic at $t=0$. Moreover, since $R_{21}(0) = 0$, $\lambda\left(-\tilde{A}_{11}(0)\right) = \lambda(-{A_{11}}^0) \subset \overline{\mathbb{C}^-}$. Hence, by Theorem 6.10, R_{p1} is analytic at $t=0$. Since A_{12}, y_2 and f_1 are analytic at $t=0$ and $A_{12}(0)\,y_2(0) + f_1(0) = A_{12}(0)\,x_2(0) + f_1(0) = 0$, the vector function g_p is analytic at $t=0$ as well.

If R_{p1} and g_p have been computed until $t=\delta$, then the BCs (24) are transferred into the BCs

$$B^{0p}\,R_{p1}(\delta)\,x_1(\delta) + B^1 x(1) = b - B^0 \begin{pmatrix} x_k(0) \\ g_p(\delta) \\ x_2(0) \end{pmatrix} \tag{28}$$

Remark 6.18
In [34] it is shown that R_{p1} is the matrix consisting of the last p rows of ${X_{11}}^{-1}$, where X_{11} is the $k \times k$ upper diagonal block of the fundamental solution X, defined in Theorem 6.16.

■

6.3.3 The regular BVP on $[\delta, 1]$

Together with the condition (22) the BCs (28) give a complete set of BCs on $[\delta, 1]$:

$$\begin{bmatrix} B^{0p}\,R_{p1}(\delta) & 0 \\ -R_{21}(\delta) & I_{q+m} \end{bmatrix} x(\delta) + \begin{bmatrix} B^1 \\ 0 \end{bmatrix} x(1) = \begin{pmatrix} b - B^0 \begin{pmatrix} x_k(0) \\ g_p(\delta) \\ x_2(0) \end{pmatrix} \\ y_2(\delta) \end{pmatrix}, \tag{29}$$

where $x_k(0)$ and $x_2(0)$ are given vectors, determined by (11).

Hence, a solution of the singular BVP (1), subject to (2a,b), is on $[\delta, 1]$ also a solution of the regular BVP (1), subject to (29). Finally we show that this solution is unique.

Theorem 6.19
The regular BVP (1), (29) *has a unique solution if and only if the singular BVP* (1), (2a,b) *has a unique solution.*

Proof:
Let $\bar{x}$ be a particular solution of (1), with

$$\bar{x}(0) = \begin{pmatrix} x_k(0) \\ 0 \\ x_2(0) \end{pmatrix} ,$$

where $x_k(0)$ and $x_2(0)$ being determined by (11).

Let $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$ be a fundamental solution such that $\mathcal{R}(X_1)$ contains all homogeneous modes of (1), satisfying (2a). Hence, $R_{21} = X_{21} {X_{11}}^{-1}$ (cf. Theorem 6.16) and for any solution of (1), subject to (2a), there exists a vector $c_1 \in \mathbb{R}^{k+p}$ such that

$$x(t) = X_1(t)\, c_1 + \bar{x}(t) .$$

By the special choice of X_1 and $\bar{x}(0)$ we have that (11) is indeed satisfied. Moreover, from the BCs (2b) we obtain that c_1 has to satisfy

$$\left(\begin{bmatrix} 0 & B^{0p} & 0 \end{bmatrix} X_1(0) + B^1 X_1(1) \right) c_1 = b - B^0 \bar{x}(0) - B^1 \bar{x}(1) .$$

So the singular BVP has a unique solution if and only if the $(k+p) \times (k+p)$ matrix

$$\begin{bmatrix} 0 & B^{0p} & 0 \end{bmatrix} X_1(0) + B^1 X_1(1) = \begin{bmatrix} 0 & B^{0p} \end{bmatrix} + B^1 X_1(1) \qquad (30)$$

is non-singular (cf. Theorem 2.1).

Now define (cf. (29))

$$B^\delta = \begin{bmatrix} B^{0p} R_{p1}(\delta) & 0 \\ -R_{21}(\delta) & I_{q+m} \end{bmatrix} .$$

Then the regular BVP has a unique solution if and only if the $n \times n$ matrix

$$\mathcal{B}(X) = B^\delta X(\delta) + \begin{bmatrix} B^1 \\ 0 \end{bmatrix} X(1)$$

is non-singular. Since $R_{21} = X_{21} {X_{11}}^{-1}$ we observe that $\mathcal{B}(X)_{21} = 0$. Moreover, (see (30))

$$\mathcal{B}(X)_{11} = \begin{bmatrix} 0 & B^{0p} \end{bmatrix} + B^1 X_1(1) ,$$

which is non-singular if and only if the singular BVP has a unique solution. Hence, it suffices to show that $\mathcal{B}(X)_{22}$ is non-singular.
Note that, since $X(\delta)$ and $X_{11}(\delta)$ are non-singular,

$$\begin{aligned} \mathcal{B}(X)_{22} &= R_{21}(\delta)\,X_{12}(\delta) + X_{22}(\delta) \\ &= -X_{21}(\delta)\,{X_{11}}^{-1}(\delta)\,X_{12}(\delta) + X_{22}(\delta) \\ &= \left(\left(X^{-1}(\delta)\right)_{22}\right)^{-1}. \end{aligned}$$

Therefore $\mathcal{B}(X)$ is non-singular.

■

To solve the regular BVP any of the techniques discussed in the foregoing chapters may be used. However, for a continuous decoupling method, like the Riccati transformation, one probably has to choose another partition. If $q \neq 0$ then the original partition used on $(0, \delta]$ will generally not correspond to an (exponential) dichotomy on $[\delta, 1]$. Therefore, on $[\delta, 1]$ it might be better to use a Riccati matrix which is of order $(n-k) \times k$ or $m \times (n-m)$. For example: let $A(t) \approx \operatorname{diag}\left(\lambda(t), \begin{bmatrix} t & 1 \\ 0 & t \end{bmatrix}, -\lambda(t)\right)$, where $\lambda(t) \gg 1$, for all t. Then we shall have one fast increasing mode, one fast decaying mode and two slowly varying modes. Hence, on $[\delta, 1]$ difference in growth behaviour is more pronounced if we take $\dim(\mathcal{S}_1) = 1$ or 3.

Remark 6.20
As in the regular case it is not necessary to store and interpolate intermediate results. All the required functions can be computed by solving just one $(n-k) \times (k+p+1)$ Riccati DE. Write

$$R = \left[\begin{array}{c|c} R_{p1} & g_p \\ R_{21} & y_2 \end{array}\right].$$

Then, for $t \geq 0$, we have

$$t\frac{dR}{dt} = \left[\begin{array}{c|c} 0 & 0 \\ A_{21}(t) & f_2(t) \end{array}\right] + \begin{bmatrix} 0 & 0 \\ 0 & A_{22}(t) \end{bmatrix} R - R\left[\begin{array}{c|c} A_{11}(t) & f_1(t) \\ \hline 0 & 0 \end{array}\right] - R\begin{bmatrix} 0 & A_{12}(t) \\ 0 & 0 \end{bmatrix} R,$$

subject to

$$R(0) = \left[\begin{array}{cc|c} 0 & I_p & 0 \\ 0 & 0 & x_2(0) \end{array}\right].$$

By Theorem 6.10 this IVP has exactly one solution which is analytic at $t = 0$.

■

Remark 6.21

If the existence condition (2a) is replaced by the boundedness condition

$$\sup_{t \in (0,1]} \| x(t) \| < \infty , \tag{31}$$

then still a reduction to a regular BVP on $[\delta, 1]$ can be obtained. In that case we have to transform A^0 such that the purely imaginary eigenvalues and the corresponding invariant subspaces of A^0 have been isolated too. Hereafter, a technique similar to the one derived above may be used. For details we refer to [34].

■

Bibliography

[1] Abramov, A.A., *On the transfer of the condition of boundedness for some systems of ordinary differential equations*, USSR Comp. Math. and Math. Phys., 1 (4), 1961, pp.875-881.

[2] Ascher, U.M., R.M.M. Mattheij and R.D. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Prentice-Hall, 1988.

[3] Ascher, U.M. and R. Weiss, *Collocation for singular perturbation problems II: linear first order systems without turning points*, Math. Comp., 43, 1984, pp.157-187.

[4] Babuška, I., *The connection between the finite difference like methods and the methods based on initial value problems for ODE*, in 'Numerical Solutions of Boundary Value Problems for Ordinary Differential Equations', ed. A.K. Aziz, Academic Press, New York, 1975.

[5] Balla, K., *On the replacement of the condition on boundedness for certain systems of linear ODEs with regular singularity*, in Colloquia Mathematica Societatis Janos Bolyai 22, 'Numerical Methods', ed. P. Rósza, North-Holland, Amsterdam, 1980.

[6] Bellman, R. and G.M. Wing, *An Introduction to Invariant Imbedding*, Pure and Applied Mathematics, John Wiley & Sons, New York, 1975.

[7] Berkey D.D., *Comparative exponential dichotomies and column diagonal dominance*, J. Math. Anal. Appl., 55, 1976, pp.140-149.

[8] Brabston, D.C. and H.B. Keller, *Numerical Methods for Singular Two-point BVPs*, SIAM J. Numer. Anal., 14, 1977, pp.779-791.

[9] Breitenecker, F., *On the Solution of Linear Boundary Value Problems by Extended Invariant Imbedding*, Computing, 28, 1982, pp.333-343.

[10] Brown, D.L. and J. Lorenz, *A High-Order Method for Stiff BVPs with Turning Points*, submitted to SIAM J. Sci. Statist. Comput.

[11] Coppel, W.A., *Dichotomies in Stability Theory*, Lecture Notes in Mathematics 629, Springer Verlag, 1978.

[12] Davey, A., *An automatic orthonormalization method for solving stiff BVPs*, J. Comput. Phys., 51, 1983, pp.343-356.

[13] Denman, E.D., *An Overview of Invariant Imbedding Algorithms and Two-Point Boundary Value Problems*, in 'Codes for Boundary Value Problems in Ordinary Differential Equations', eds. B. Childs, a.o., Lecture Notes in Computer Science 76, Springer-Verlag, Berlin, 1979.

[14] Deuflhard, P. and G. Bader, *Multiple-shooting techniques revisited*, in 'Numerical Treatment of Inverse Problems in Differential and Integral Equations', eds. P. Deuflhard and E. Hairer, Birkhäuser, Boston, 1983.

[15] Dieci, L., M.R. Osborne and R.D. Russell, *A Riccati Transformation Method for Solving BVPs I: Theoretical Aspects*, report 1986, Simon Fraser University, Vancouver.

[16] Dieci, L., M.R. Osborne and R.D. Russell, *A Riccati Transformation Method for Solving BVPs II: Computational Aspects*, report 1986, Simon Fraser University, Vancouver.

[17] Eckhaus, W., *Matched Asymptotic Expansions and Singular Perturbations*, North Holland, Amsterdam, 1973.

[18] Eirola, T., *A Study of the Back-and Forth Shooting Method*, Ph.D. Thesis, Helsinki University of Technology, Finland, 1985.

[19] Flaherty, J.E. and R.E. O'Malley, Jr., *Numerical Methods for Stiff Systems of Two-Point Boundary Value Problems*, SIAM J. Sci. Stat. Comput., 5, 1984, pp.865-886.

[20] Golub, G.H. and C.F. van Loan, *Matrix Computations*, John Hopkins University Press, 1983.

[21] Golub, G.H., Nash, S. and C.F. van Loan, *A Hessenberg-Schur Method for the Problem* $AX + XB = C$, IEEE Trans. on Math. Softw., 24, 1979, pp.903-913.

[22] Hautus, M.L.J., *Formal and Convergent Solutions of Ordinary Differential Equations*, Proc. Kon. Ned. Akad. Wetenschappen, Amsterdam, Serie A, 81, 1978, pp.216-229.

[23] Hemker, P., *A Numerical Study of Stiff BVPs*, Ph.D. Thesis, Math. Centrum, Amsterdam, 1977.

[24] Henrici, P., *Applied and Computational Complex Analysis*, Vol. 1 & 2, Pure and Applied Mathematics, Wiley, London, 1977.

[25] Hindmarsch, A., *ODEPACK, A Systematized Collection of ODE solvers*, in 'Scientific Computing', eds. R.S. Stepleman a.o., Amsterdam, 1983, pp.55-64.

[26] de Hoog, F. and R.M.M. Mattheij, *On Dichotomy and Well-conditioning in Boundary Value Problems*, SIAM J. Numer. Anal., 24, 1987, pp.89-105.

[27] de Hoog, F. and R. Weiss, *Difference methods for BVPs with a singularity of the first kind*, SIAM J. Numer. Anal., 13, 1976, pp.775-813.

[28] Keller, H.B. and M. Lentini, *Invariant Imbedding, the Box-scheme and an Equivalence between them*, SIAM J. Numer. Anal., 19, 1982, pp.942-962.

[29] Keller, H.B., *Numerical Solution of Two-point Boundary Value Problems*, SIAM Regional Conference Series 24, Philadelphia, 1976.

[30] Kreiss, B. and H.-O. Kreiss, *Numerical Methods for Singular Perturbation Problems*, SIAM J. Numer. Anal., 18, 1981, pp.262-276.

[31] Kreiss, H.-O., N.K. Nichols and D.L. Brown, *Numerical Methods for Stiff Two-Point Boundary Value Problems*, SIAM J. Numer. Anal., 23, 1986, pp.325-368.

[32] Lazer, A.C., *Characteristic Exponents and Diagonally Dominant Linear Differential Systems*, J. Math. Anal. Appl., 35, 1971, pp.215-229.

[33] Lentini, M., M.R. Osborne and R.D. Russell, *The close relationship between methods for solving two-point boundary value problems*, SIAM J. Numer. Anal., 22, 1985, pp.280-309.

[34] van Loon, P.M., *Reducing a singular linear two-point BVP to a regular one by means of Riccati transformations*, EUT-report 83-WSK-04, Eindhoven, 1983.

[35] van Loon, P.M., *Riccati Transformations: when and how to use?*, in 'Numerical Boundary Value ODEs', eds. U.M. Ascher and R.D. Russell, Progress in Scientific Computing, 5, Birkhäuser, Boston, 1985.

[36] van Loon, P.M., *A solution method for stiff BVPs*, Computing Centre Note 31, Eindhoven University of Technology, 1986.

[37] van Loon, P.M. and R.M.M. Mattheij, *Stable Continuous Orthonormalisation Techniques for Linear Boundary Value Problems*, J. Austral. Math. Soc. Ser. B, 29, 1988, pp.282-295.

[38] Mattheij, R.M.M., *Characterization of dominant and dominated solutions of linear recursions*, Numer. Math., 35, 1980, pp.421-442.

[39] Mattheij, R.M.M. and G.W.M. Staarink, *An efficient algorithm for solving general linear two point BVP*, SIAM J. Sci. Statist. Comput., 5, 1984, pp.745-763.

[40] Mattheij, R.M.M., *Stability of Block LU-Decompositions of Matrices arising from BVP*, SIAM J. Alg. Disc. Meth.. 5, 1984, pp.314-331.

[41] Mattheij, R.M.M. and R.E. O'Malley, Jr., *On solving boundary value problems for multi-scale systems using asymptotic approximations and multiple shooting*, BIT, 24, 1984, pp.609-622.

[42] Mattheij, R.M.M., *Decoupling and Stability of Algorithms for Boundary Value Problems*, SIAM Rev., 27, 1985, pp.1-44.

[43] Meyer, G.H., *Initial Value Methods for Boundary Value Problems*, Academic Press, London, 1973.

[44] Meyer, G.H., *An Application of the Method of Lines to Multidimensional Free Boundary Problems*, J. Inst. Maths. Applics., 20, 1977, pp.317-329.

[45] Meyer, G.H., *Continuous Orthonormalization for Boundary Value Problems*, J. Comput. Phys., 62, 1986, pp.248-262.

[46] Miranker, W.L., *Numerical Methods for Stiff Equations*, Mathematics and its Applications, 5, D. Reidel Publishing Company, Dordrecht, 1981.

[47] Nelson, P., S. Sagong and I.T. Elder, *Invariant imbedding applied to homogeneous two-point BVPs with a singularity of the first kind*, Appl. Math. and. Comp., 9, 1981, pp.93-110.

[48] O'Malley, R.E., Jr., *Singular Perturbations and Optimal Control*, Lecture Notes in Mathematics 680, Springer-Verlag, Berlin, 1978, pp.170-218.

[49] O'Malley, R.E., Jr., *Introduction to singular perturbations*, Applied Mathematics and Mechanics 14, Academic Press, New York, 1974.

[50] Osborne, M.R., *The Stabilized March is Stable*, SIAM J. Numer. Anal., 16, 1979, pp.923-933.

[51] Osborne, M.R. and R.D. Russell, *The Riccati transformation in the solution of BVPs*, SIAM J. Numer. Anal., 23, 1986, pp.1023-1033.

[52] Reid, W.T., *Riccati differential equations*, Academic Press, New York, 1972.

[53] Scott, M.R., *Invariant Imbedding and its Applications to Ordinary Differential Equations: an introduction*, Applied Mathematics and Computation, Addison-Wesley, 1973.

[54] Scott, M.R. and H.A. Watts, *Computational solution of linear two point boundary value problems via orthonormalization*, SIAM J. Numer. Anal., 14, 1977, pp.40-70.

[55] Shampine, L.F. and M.K. Gordon, *Computer Solutions of Ordinary Differential Equations. The Initial Value Problem.*, W.H. Freeman and Company, San Francisco, 1975.

[56] Söderlind, G. and R.M.M. Mattheij, *Stability and asymptotic estimates in nonautonomous linear differential systems*, SIAM J. Math. Anal., 16, 1985, pp.69-92.

[57] Stewart, G.W., *Error bounds for approximate invariant subspaces of closed linear operators*, SIAM J. Num. Anal., 8, 1971, pp.796-808.

[58] Stewart, G.W., *HQR3 and EXCHNG: Fortran Subroutines for Calculating and Ordering the Eigenvalues of a Real Upper Hessenberg Matrix*, ACM Trans. on Math. Softw., 2, 1976, pp.275-280.

[59] Ström, T., *On logarithmic norms*, SIAM J. Numer. Anal., 12, 1975, pp.741-753.

[60] Taufer, J., *On Factorization Method*, Aplikace Matematiky, 11, 1966, pp.427-450.

[61] Wasow, W., *Asymptotic Expansions for Ordinary Differential Equations*, Pure and Applied Mathematics, Wiley, London, 1965.

[62] Wasow, W., *Linear Turning Point Theory*, Applied Mathematical Sciences, 54, Springer-Verlag New-York Inc., 1985.

[63] Weiss, R., *An analysis of the box and trapezoidal schemes for linear singularly perturbed BVPs*, Math. Comp., 42, 1984, pp.41-67.

Index

MATHEMATICAL CENTRE TRACTS

1 T. van der Walt. *Fixed and almost fixed points.* 1963.

2 A.R. Bloemena. *Sampling from a graph.* 1964.

3 G. de Leve. *Generalized Markovian decision processes, part I: model and method.* 1964.

4 G. de Leve. *Generalized Markovian decision processes, part II: probabilistic background.* 1964.

5 G. de Leve, H.C. Tijms, P.J. Weeda. *Generalized Markovian decision processes, applications.* 1970.

6 M.A. Maurice. *Compact ordered spaces.* 1964.

7 W.R. van Zwet. *Convex transformations of random variables.* 1964.

8 J.A. Zonneveld. *Automatic numerical integration.* 1964.

9 P.C. Baayen. *Universal morphisms.* 1964.

10 E.M. de Jager. *Applications of distributions in mathematical physics.* 1964.

11 A.B. Paalman-de Miranda. *Topological semigroups.* 1964.

12 J.A.Th.M. van Berckel, H. Brandt Corstius, R.J. Mokken, A. van Wijngaarden. *Formal properties of newspaper Dutch.* 1965.

13 H.A. Lauwerier. *Asymptotic expansions.* 1966, out of print; replaced by MCT 54.

14 H.A. Lauwerier. *Calculus of variations in mathematical physics.* 1966.

15 R. Doornbos. *Slippage tests.* 1966.

16 J.W. de Bakker. *Formal definition of programming languages with an application to the definition of ALGOL 60.* 1967.

17 R.P. van de Riet. *Formula manipulation in ALGOL 60, part 1.* 1968.

18 R.P. van de Riet. *Formula manipulation in ALGOL 60, part 2.* 1968.

19 J. van der Slot. *Some properties related to compactness.* 1968.

20 P.J. van der Houwen. *Finite difference methods for solving partial differential equations.* 1968.

21 E. Wattel. *The compactness operator in set theory and topology.* 1968.

22 T.J. Dekker. *ALGOL 60 procedures in numerical algebra, part 1.* 1968.

23 T.J. Dekker, W. Hoffmann. *ALGOL 60 procedures in numerical algebra, part 2.* 1968.

24 J.W. de Bakker. *Recursive procedures.* 1971.

25 E.R. Paërl. *Representations of the Lorentz group and projective geometry.* 1969.

26 European Meeting 1968. *Selected statistical papers, part I.* 1968.

27 European Meeting 1968. *Selected statistical papers, part II.* 1968.

28 J. Oosterhoff. *Combination of one-sided statistical tests.* 1969.

29 J. Verhoeff. *Error detecting decimal codes.* 1969.

30 H. Brandt Corstius. *Exercises in computational linguistics.* 1970.

31 W. Molenaar. *Approximations to the Poisson, binomial and hypergeometric distribution functions.* 1970.

32 L. de Haan. *On regular variation and its application to the weak convergence of sample extremes.* 1970.

33 F.W. Steutel. *Preservation of infinite divisibility under mixing and related topics.* 1970.

34 I. Juhász, A. Verbeek, N.S. Kroonenberg. *Cardinal functions in topology.* 1971.

35 M.H. van Emden. *An analysis of complexity.* 1971.

36 J. Grasman. *On the birth of boundary layers.* 1971.

37 J.W. de Bakker, G.A. Blaauw, A.J.W. Duijvestijn, E.W. Dijkstra, P.J. van der Houwen, G.A.M. Kamsteeg-Kemper, F.E.J. Kruseman Aretz, W.L. van der Poel, J.P. Schaap-Kruseman, M.V. Wilkes, G. Zoutendijk. *MC-25 Informatica Symposium.* 1971.

38 W.A. Verloren van Themaat. *Automatic analysis of Dutch compound words.* 1972.

39 H. Bavinck. *Jacobi series and approximation.* 1972.

40 H.C. Tijms. *Analysis of (s,S) inventory models.* 1972.

41 A. Verbeek. *Superextensions of topological spaces.* 1972.

42 W. Vervaat. *Success epochs in Bernoulli trials (with applications in number theory).* 1972.

43 F.H. Ruymgaart. *Asymptotic theory of rank tests for independence.* 1973.

44 H. Bart. *Meromorphic operator valued functions.* 1973.

45 A.A. Balkema. *Monotone transformations and limit laws.* 1973.

46 R.P. van de Riet. *ABC ALGOL, a portable language for formula manipulation systems, part 1: the language.* 1973.

47 R.P. van de Riet. *ABC ALGOL, a portable language for formula manipulation systems, part 2: the compiler.* 1973.

48 F.E.J. Kruseman Aretz, P.J.W. ten Hagen, H.L. Oudshoorn. *An ALGOL 60 compiler in ALGOL 60, text of the MC-compiler for the EL-X8.* 1973.

49 H. Kok. *Connected orderable spaces.* 1974.

50 A. van Wijngaarden, B.J. Mailloux, J.E.L. Peck, C.H.A. Koster, M. Sintzoff, C.H. Lindsey, L.G.L.T. Meertens, R.G. Fisker (eds.). *Revised report on the algorithmic language ALGOL 68.* 1976.

51 A. Hordijk. *Dynamic programming and Markov potential theory.* 1974.

52 P.C. Baayen (ed.). *Topological structures.* 1974.

53 M.J. Faber. *Metrizability in generalized ordered spaces.* 1974.

54 H.A. Lauwerier. *Asymptotic analysis, part 1.* 1974.

55 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 1: theory of designs, finite geometry and coding theory.* 1974.

56 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 2: graph theory, foundations, partitions and combinatorial geometry.* 1974.

57 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 3: combinatorial group theory.* 1974.

58 W. Albers. *Asymptotic expansions and the deficiency concept in statistics.* 1975.

59 J.L. Mijnheer. *Sample path properties of stable processes.* 1975.

60 F. Göbel. *Queueing models involving buffers.* 1975.

63 J.W. de Bakker (ed.). *Foundations of computer science.* 1975.

64 W.J. de Schipper. *Symmetric closed categories.* 1975.

65 J. de Vries. *Topological transformation groups, 1: a categorical approach.* 1975.

66 H.G.J. Pijls. *Logically convex algebras in spectral theory and eigenfunction expansions.* 1976.

68 P.P.N. de Groen. *Singularly perturbed differential operators of second order.* 1976.

69 J.K. Lenstra. *Sequencing by enumerative methods.* 1977.

70 W.P. de Roever, Jr. *Recursive program schemes: semantics and proof theory.* 1976.

71 J.A.E.E. van Nunen. *Contracting Markov decision processes.* 1976.

72 J.K.M. Jansen. *Simple periodic and non-periodic Lamé functions and their applications in the theory of conical waveguides.* 1977.

73 D.M.R. Leivant. *Absoluteness of intuitionistic logic.* 1979.

74 H.J.J. te Riele. *A theoretical and computational study of generalized aliquot sequences.* 1976.

75 A.E. Brouwer. *Treelike spaces and related connected topological spaces.* 1977.

76 M. Rem. *Associons and the closure statement.* 1976.

77 W.C.M. Kallenberg. *Asymptotic optimality of likelihood ratio tests in exponential families.* 1978.

78 E. de Jonge, A.C.M. van Rooij. *Introduction to Riesz spaces.* 1977.

79 M.C.A. van Zuijlen. *Emperical distributions and rank statistics.* 1977.

80 P.W. Hemker. *A numerical study of stiff two-point boundary problems.* 1977.

81 K.R. Apt, J.W. de Bakker (eds.). *Foundations of computer science II, part 1.* 1976.

82 K.R. Apt, J.W. de Bakker (eds.). *Foundations of computer science II, part 2.* 1976.

83 L.S. van Benthem Jutting. *Checking Landau's "Grundlagen" in the AUTOMATH system.* 1979.

84 H.L.L. Busard. *The translation of the elements of Euclid from the Arabic into Latin by Hermann of Carinthia (?), books vii-xii.* 1977.

85 J. van Mill. *Supercompactness and Wallman spaces.* 1977.

86 S.G. van der Meulen, M. Veldhorst. *Torrix I, a programming system for operations on vectors and matrices over arbitrary fields and of variable size.* 1978.

88 A. Schrijver. *Matroids and linking systems.* 1977.

89 J.W. de Roever. *Complex Fourier transformation and analytic functionals with unbounded carriers.* 1978.

90 L.P.J. Groenewegen. *Characterization of optimal strategies in dynamic games.* 1981.

91 J.M. Geysel. *Transcendence in fields of positive characteristic.* 1979.

92 P.J. Weeda. *Finite generalized Markov programming.* 1979.

93 H.C. Tijms, J. Wessels (eds.). *Markov decision theory.* 1977.

94 A. Bijlsma. *Simultaneous approximations in transcendental number theory.* 1978.

95 K.M. van Hee. *Bayesian control of Markov chains.* 1978.

96 P.M.B. Vitányi. *Lindenmayer systems: structure, languages, and growth functions.* 1980.

97 A. Federgruen. *Markovian control problems; functional equations and algorithms.* 1984.

98 R. Geel. *Singular perturbations of hyperbolic type.* 1978.

99 J.K. Lenstra, A.H.G. Rinnooy Kan, P. van Emde Boas (eds.). *Interfaces between computer science and operations research.* 1978.

100 P.C. Baayen, D. van Dulst, J. Oosterhoff (eds.). *Proceedings bicentennial congress of the Wiskundig Genootschap, part 1.* 1979.

101 P.C. Baayen, D. van Dulst, J. Oosterhoff (eds.). *Proceedings bicentennial congress of the Wiskundig Genootschap, part 2.* 1979.

102 D. van Dulst. *Reflexive and superreflexive Banach spaces.* 1978.

103 K. van Harn. *Classifying infinitely divisible distributions by functional equations.* 1978.

104 J.M. van Wouwe. *Go-spaces and generalizations of metrizability.* 1979.

105 R. Helmers. *Edgeworth expansions for linear combinations of order statistics.* 1982.

106 A. Schrijver (ed.). *Packing and covering in combinatorics.* 1979.

107 C. den Heijer. *The numerical solution of nonlinear operator equations by imbedding methods.* 1979.

108 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science III, part 1.* 1979.

109 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science III, part 2.* 1979.

110 J.C. van Vliet. *ALGOL 68 transput, part I: historical review and discussion of the implementation model.* 1979.

111 J.C. van Vliet. *ALGOL 68 transput, part II: an implementation model.* 1979.

112 H.C.P. Berbee. *Random walks with stationary increments and renewal theory.* 1979.

113 T.A.B. Snijders. *Asymptotic optimality theory for testing problems with restricted alternatives.* 1979.

114 A.J.E.M. Janssen. *Application of the Wigner distribution to harmonic analysis of generalized stochastic processes.* 1979.

115 P.C. Baayen, J. van Mill (eds.). *Topological structures II, part 1.* 1979.

116 P.C. Baayen, J. van Mill (eds.). *Topological structures II, part 2.* 1979.

117 P.J.M. Kallenberg. *Branching processes with continuous state space.* 1979.

118 P. Groeneboom. *Large deviations and asymptotic efficiencies.* 1980.

119 F.J. Peters. *Sparse matrices and substructures, with a novel implementation of finite element algorithms.* 1980.

120 W.P.M. de Ruyter. *On the asymptotic analysis of large-scale ocean circulation.* 1980.

121 W.H. Haemers. *Eigenvalue techniques in design and graph theory.* 1980.

122 J.C.P. Bus. *Numerical solution of systems of nonlinear equations.* 1980.

123 I. Yuhász. *Cardinal functions in topology - ten years later.* 1980.

124 R.D. Gill. *Censoring and stochastic integrals.* 1980.

125 R. Eising. *2-D systems, an algebraic approach.* 1980.

126 G. van der Hoek. *Reduction methods in nonlinear programming.* 1980.

127 J.W. Klop. *Combinatory reduction systems.* 1980.

128 A.J.J. Talman. *Variable dimension fixed point algorithms and triangulations.* 1980.

129 G. van der Laan. *Simplicial fixed point algorithms.* 1980.

130 P.J.W. ten Hagen, T. Hagen, P. Klint, H. Noot, H.J. Sint, A.H. Veen. *ILP: intermediate language for pictures.* 1980.

131 R.J.R. Back. *Correctness preserving program refinements: proof theory and applications.* 1980.

132 H.M. Mulder. *The interval function of a graph.* 1980.

133 C.A.J. Klaassen. *Statistical performance of location estimators.* 1981.

134 J.C. van Vliet, H. Wupper (eds.). *Proceedings international conference on ALGOL 68.* 1981.

135 J.A.G. Groenendijk, T.M.V. Janssen, M.J.B. Stokhof (eds.). *Formal methods in the study of language, part I.* 1981.

136 J.A.G. Groenendijk, T.M.V. Janssen, M.J.B. Stokhof (eds.). *Formal methods in the study of language, part II.* 1981.

137 J. Telgen. *Redundancy and linear programs.* 1981.

138 H.A. Lauwerier. *Mathematical models of epidemics.* 1981.

139 J. van der Wal. *Stochastic dynamic programming, successive approximations and nearly optimal strategies for Markov decision processes and Markov games.* 1981.

140 J.H. van Geldrop. *A mathematical theory of pure exchange economies without the no-critical-point hypothesis.* 1981.

141 G.E. Welters. *Abel-Jacobi isogenies for certain types of Fano threefolds.* 1981.

142 H.R. Bennett, D.J. Lutzer (eds.). *Topology and order structures, part 1.* 1981.

143 J.M. Schumacher. *Dynamic feedback in finite- and infinite-dimensional linear systems.* 1981.

144 P. Eijgenraam. *The solution of initial value problems using interval arithmetic; formulation and analysis of an algorithm.* 1981.

145 A.J. Brentjes. *Multi-dimensional continued fraction algorithms.* 1981.

146 C.V.M. van der Mee. *Semigroup and factorization methods in transport theory.* 1981.

147 H.H. Tigelaar. *Identification and informative sample size.* 1982.

148 L.C.M. Kallenberg. *Linear programming and finite Markovian control problems.* 1983.

149 C.B. Huijsmans, M.A. Kaashoek, W.A.J. Luxemburg, W.K. Vietsch (eds.). *From A to Z, proceedings of a symposium in honour of A.C. Zaanen.* 1982.

150 M. Veldhorst. *An analysis of sparse matrix storage schemes.* 1982.

151 R.J.M.M. Does. *Higher order asymptotics for simple linear rank statistics.* 1982.

152 G.F. van der Hoeven. *Projections of lawless sequences.* 1982.

153 J.P.C. Blanc. *Application of the theory of boundary value problems in the analysis of a queueing model with paired services.* 1982.

154 H.W. Lenstra, Jr., R. Tijdeman (eds.). *Computational methods in number theory, part I.* 1982.

155 H.W. Lenstra, Jr., R. Tijdeman (eds.). *Computational methods in number theory, part II.* 1982.

156 P.M.G. Apers. *Query processing and data allocation in distributed database systems.* 1983.

157 H.A.W.M. Kneppers. *The covariant classification of two-dimensional smooth commutative formal groups over an algebraically closed field of positive characteristic.* 1983.

158 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science IV, distributed systems, part 1.* 1983.

159 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science IV, distributed systems, part 2.* 1983.

160 A. Rezus. *Abstract AUTOMATH.* 1983.

161 G.F. Helminck. *Eisenstein series on the metaplectic group, an algebraic approach.* 1983.

162 J.J. Dik. *Tests for preference.* 1983.

163 H. Schippers. *Multiple grid methods for equations of the second kind with applications in fluid mechanics.* 1983.

164 F.A. van der Duyn Schouten. *Markov decision processes with continuous time parameter.* 1983.

165 P.C.T. van der Hoeven. *On point processes.* 1983.

166 H.B.M. Jonkers. *Abstraction, specification and implementation techniques, with an application to garbage collection.* 1983.

167 W.H.M. Zijm. *Nonnegative matrices in dynamic programming.* 1983.

168 J.H. Evertse. *Upper bounds for the numbers of solutions of diophantine equations.* 1983.

169 H.R. Bennett, D.J. Lutzer (eds.). *Topology and order structures, part 2.* 1983.