J. Austral. Math. Soc. Ser. B 29 (1988), 282–295

STABLE CONTINUOUS ORTHONORMALISATION TECHNIQUES FOR LINEAR BOUNDARY VALUE PROBLEMS

P. M. VAN LOON[1] AND R. M. M. MATTHEIJ[2]

(Received 20 December 1985; revised 5 August 1986 and 10 November 1986)

Abstract

An investigation is made of a hybrid method inspired by Riccati transformations and marching algorithms employing (parts of) orthogonal matrices, both being decoupling algorithms. It is shown that this so-called continuous orthonormalisation is stable and practical as well. Nevertheless, if the problem is stiff and many output points are required the method does not give much gain over, say, multiple shooting.

1. Introduction

Consider the ordinary differential equation (ODE)

$$\frac{dx}{dt} = A(t)x + f(t), \qquad t \in (0,1), \tag{1.1}$$

where $f(t)$ is an n-vector function and $A(t)$ an $n \times n$ matrix function, both being at least continuous. Wherever it turns out to be more practical we shall write $\dot{x}$ rather than $\frac{dx}{dt}$.

We assume that we have *boundary conditions* (BCs)

$$B^0x(0) + B^1x(1) = b, \tag{1.2}$$

($B^0, B^1 \in \mathbf{R}^{n\times n}$ and $b \in \mathbf{R}^n$), by which the solution x of (1.1) is uniquely determined.

[1] Computing Centre, Eindhoven University of Technology, P. O. Box 513, 5600 MB Eindhoven, The Netherlands.

[2] Department of Mathematics and Computing Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands.

There exists a host of methods for obtaining numerical approximations of x satisfying (1.1) and (1.2). In this paper we are interested in two special ones, or rather a hybrid version of them, viz. the *Riccati method* and the *stabilised march*. The former method is attractive because it is able to split (1.1) into both a stable initial value ODE and a stable terminal value ODE, by some appropriate transformation, *before* discretisation.

This gives good hopes that one may handle *stiffness*, which occurs when (1.1) has a large Lipschitz constant, like in initial-value problems after discretisation. The big problem with the Riccati method lies in the fact that such a transformation may not exist on the entire interval (see however [7]). The stabilised march, on the other hand, employs (parts of) orthogonal matrices, is indeed stable and moreover can be performed straightforwardly, but for a discretised problem only (cf. [12]). Since in that method the underlying discretisation has to be done before transformation (or decoupling, cf. [9]), it is clear that it is unsuited for stiff problems.

It is very natural then that various authors, notably Davey [3] and Meyer [11], have tried to combine the virtues of both methods into a class of hybrid methods, to be called *continuous orthonormalisations*.

The main incentive to write this note is that we believe that the properties of continuous orthonormalisation methods (good and bad) become more transparent when viewed as decoupling techniques. In particular this is useful to understand their stability. Their efficiency turns out to be quite a delicate problem. The crucial question here is of course to ask what we have gained compared to other methods, in particular when stiffness problems occur. In principle, one might hope to tackle stiffness problems successfully when the rotational activity of the fundamental solutions belonging to (1.1) is not large with respect to other time scales. Indeed, as was shown in e.g. [9], the activity of decoupling transformations commensurates with rotational activity of these fundamental solutions.

Unfortunately, for the implementation suggested in [11], stiffness does cause problems, in the sense that in the backward sweep the stepsize is dictated by the Lipschitz constant in general. We shall give an explanation for this, and also indicate how to improve the situation when classical invariant imbedding is employed as well. But before that, we give an overview of decoupling transformations consisting of (parts of) orthogonal matrices, both continuous and discrete, and their consequences for actual numerical algorithms. We remark here that there exist other blends of the Riccati method and the stabilised march; in particular we mention [8] and [13], where algorithms are given that have more robustness and efficiency than either one of the two.

This paper is built up as follows. First we describe Lyapunov equations in Section 2, showing that both Riccati transformations and orthogonal transformations may be used for appropriate decoupling; we also briefly discuss the choice

for the initial values. Then in Section 3 we consider discrete analogues, notably multiple shooting and marching techniques, as one might view the algorithm in [11]. Finally we give an instructive numerical example in Section 4 to illustrate the foregoing analysis.

2. Transformations of the ODE

As is well known, transformation of the dependent variable x in (1.1) by a linear nonsingular (time-dependent) transformation T leads to a kinematically similar system ([2], p. 38). For the variable $y := T^{-1}x$ we obtain ([9])

$$\dot{y} = \tilde{A}(t)y + \tilde{f}(t), \qquad t \in (0,1), \tag{2.1}$$

subject to the BCs

$$\tilde{B}^0 y(0) + \tilde{B}^1 y(1) = b, \tag{2.2}$$

where

$$\tilde{A}(t) = T^{-1}(t)A(t)T(t) - T^{-1}(t)\dot{T}(t), \tag{2.3a}$$

$$\tilde{f}(t) = T^{-1}(t)f(t) \tag{2.3b}$$

and

$$\tilde{B}^0 = B^0 T(0), \qquad \tilde{B}^1 = B^1 T(1). \tag{2.3c}$$

Equation (2.3a) is often written in the form

$$\dot{T} = A(t)T - T\tilde{A}(t). \tag{2.4}$$

Hence, (1.1) is transformed into (2.1) by a so-called *Lyapunov equation* for T ((2.4)).

Let Y be a fundamental solution corresponding to (2.1), i.e.,

$$\dot{Y} = \tilde{A}(t)Y, \qquad Y(0) = I_n. \tag{2.5}$$

Then a fundamental solution of the original system (1.1) is obtained by

$$X = TY. \tag{2.6}$$

As was shown in [9], many numerical methods for solving BVPs utilise such a transformation T (or an analogue in the discrete case) that $\tilde{A}$ has a decoupled form, say $\tilde{A}(t)$ is (block) upper triangular for all $t \in (0,1)$.

Under very general—and often prevailing—circumstances, this decoupling naturally induces a splitting of the dynamics into a part that is stable for increasing time and a complementary one that is stable for decreasing time. (For the existence of such a *dichotomy* and its relation to well-conditioning we may refer to [6]). More specifically, there then exists a partitioning of vectors and

matrices

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\begin{matrix} k \\ n-k \end{matrix} \quad \text{and} \quad \tilde{A} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix}\begin{matrix} k \\ n-k \end{matrix}, \tag{2.7}$$

such that (2.1) can be written as

$$\dot{y}_1 = \tilde{A}_{11}(t)y_1 + \tilde{A}_{12}(t)y_2 + \tilde{f}_1(t) \tag{2.8a}$$

$$\dot{y}_2 = \tilde{A}_{22}(t)y_2 + \tilde{f}_2(t), \qquad t \in (0,1). \tag{2.8b}$$

If the system has been decoupled correctly, then (2.8b) will be stable in the forward direction and (2.8a) in the backward direction. Using the invariant imbedding technique, cf. [8], all integrations can even be performed stably in forward direction.

Realising that (2.4) is actually not an ODE for T as such, but rather an equation for both the unknowns T and $\tilde{A}$ simultaneously, we see that we have n^2 degrees of freedom for $2n^2$ variables. Hence, there is a trade-off between requirements imposed on (the form of) $\tilde{A}$ and T. Below we shall consider some special choices for T (and $\tilde{A}$) (Sections 2.1 and 2.2). Moreover we discuss how to find appropriate initial conditions for $T(0)$ (Section 2.3).

2.1. Riccati transformations

By prescribing $\tilde{A}$ to be block upper triangular, i.e. setting $\tilde{A}_{21} \equiv 0$, we may hope that T can be prescribed to have the special form

$$T = \begin{bmatrix} I_k & 0 \\ R & I_{n-k} \end{bmatrix}. \tag{2.9}$$

Let us assume for the moment such a T exists, then this is the well-known *Riccati transformation* (cf. [8], [9]). The matrix R satisfies

$$\dot{R} = A_{21}(t) + A_{22}(t)R - RA_{11}(t) - RA_{12}(t)R, \tag{2.10}$$

(being a Riccati equation for R). It can simply be checked that in this case we have

$$\tilde{A} = \begin{bmatrix} A_{11} + A_{12}R & A_{12} \\ 0 & A_{22} - RA_{12} \end{bmatrix} \tag{2.11}$$

and that Y (see (2.5)) has such a block upper triangular form as well. Hence, $X := TY$ is a fundamental solution of (1.1) with $X(0) = T(0)$.

From this relation we may conclude that

—By (2.9), (2.10) and (2.5) a block LU-decomposition of X is implicitly given.

—The first k columns of $T(t)$ span the same subspace in $\mathbf{R}^n$ as the first k solumns of $X(t)$. To obtain a correctly decoupled system (2.8), these columns should represent (all) solutions of (1.1) which are not significantly growing for decreasing t (see [9]).

$-R(t) = X_{21}(t)X_{11}^{-1}(t)$. Hence, a solution of (2.10) exists as long as X_{11}, the left $k \times k$ upper block of X, is invertible. One might therefore expect, from analogy with matrix decomposition techniques, that some kind of permutation may be necessary (cf. pivoting) when proceeding from 0 to 1; this question has been investigated in [7].

2.2. Orthogonal transformations

A decomposition that does not need pivoting is the factorisation of a matrix into an orthogonal and a (block) upper triangular matrix. If we would require $T(t)$ to be orthonormal, for all t, the process is called *continuous orthonormalisation* ([1]) and we trivially have

$$\dot{T}^{\mathsf{T}}T + T^{\mathsf{T}}\dot{T} = 0 \tag{2.12}$$

(i.e. $T^{\mathsf{T}}\dot{T}$ is skewsymmetric). This reduces the number of degrees of freedom to $\frac{1}{2}n(n-1)$, which is, as will be shown, just sufficient to make $\tilde{A}$ an upper triangular matrix. Hence, we have a unique solution of (2.4) (with respect to the (orthogonal) initial value $T(0)$) subject to the conditions (2.12) and $\tilde{A}$ upper triangular.

A possible construction is as follows. Let $C \in \mathbf{R}^{n\times n}$ be decomposed in

$$C = U^{\mathsf{T}} + D + V, \tag{2.13}$$

where U and V are strictly upper triangular matrices and D is a diagonal matrix. Define the operator Φ by

$$\Phi(C) = U + D + V. \tag{2.14}$$

Note that $\Phi(C)$ is always an upper triangular matrix. We now have

PROPERTY 2.15.

Transformation of the system (1.1) *by* T, *with* $T(t)$ *orthonormal, for all* t, *leads to an upper triangular system if and only if* $\tilde{A}$, *defined by* (2.3a), *satisfies* $\tilde{A} = \Phi(T^{\mathsf{T}}AT)$ (Φ *as defined by* (2.13) *and* (2.14)). (2.15)

PROOF. By (2.3a) and (2.12) we have $\Phi(\tilde{A}) = \Phi(T^{-1}AT - T^{-1}\dot{T}) = \Phi(T^{-1}AT) - \Phi(T^{-1}\dot{T}) = \Phi(T^{\mathsf{T}}AT)$. Moreover, $\tilde{A}$ is upper triangular if and only if $\tilde{A} = \Phi(\tilde{A})$.

Let

$$T = \underset{k}{[T_1}\quad \underset{n-k}{T_2]}n.$$

https://doi.org/10.1017/S0334270000005816 Published online by Cambridge University Press

To make $\tilde{A}$ only a block upper triangular matrix we need the relation (see 2.3a)

$$0 = \tilde{A}_{21} = T_2^{\mathsf{T}}(AT_1 - \dot{T}_1).$$

Note that this implies

$$\dot{T}_1 = AT_1 - T_1C_{11}, \tag{2.16}$$

where $C_{11} \in \mathbf{R}^{k\times k}$ (note that C_{11} depends on the choice of T_1). Furthermore, by the orthonormality condition for T_1,

$$0 = \dot{T}_1^{\mathsf{T}}T_1 + T_1^{\mathsf{T}}\dot{T}_1.$$

So

$$C_{11} + C_{11}^{\mathsf{T}} = T_1^{\mathsf{T}}(A + A^{\mathsf{T}})T_1 \tag{2.17}$$

(cf. [4]). In a similar way we deduce from (2.12)

$$\dot{T}_2 = -A^{\mathsf{T}}T_2 + T_2C_{22}^{\mathsf{T}}, \tag{2.18}$$

where $C_{22} \in \mathbf{R}^{(n-k)\times(n-k)}$ (depending on T_2) must satisfy

$$C_{22} + C_{22}^{\mathsf{T}} = T_2^{\mathsf{T}}(A + A^{\mathsf{T}})T_2. \tag{2.19}$$

For $\tilde{A}$ we thus obtain the matrix

$$\tilde{A} = \begin{bmatrix} C_{11} & T_1^{\mathsf{T}}(A + A^{\mathsf{T}})T_2 \\ 0 & C_{22} \end{bmatrix}. \tag{2.20}$$

Any choice for C_{11} and C_{22} satisfying (2.17) and (2.19), respectively, is possible. For instance:

(i)

$$C_{11} = T_1^{\mathsf{T}}AT_1 \quad \left(\text{i.e. } T_1^{\mathsf{T}}\dot{T}_1 = 0\right) \tag{2.21a}$$

(cf. [3], [11]) and

$$C_{22} = T_2^{\mathsf{T}}AT_2 \left(\text{i.e. } T_2^{\mathsf{T}}\dot{T}_2 = 0\right). \tag{2.21b}$$

(ii) C_{11} and C_{22} symmetric (cf. [2]):

$$C_{11} = \tfrac{1}{2}T_1^{\mathsf{T}}(A + A^{\mathsf{T}})T_1, \tag{2.22a}$$

$$C_{22} = \tfrac{1}{2}T_2^{\mathsf{T}}(A + A^{\mathsf{T}})T_2. \tag{2.22b}$$

(iii) C_{11} and C_{22} upper triangular:

$$C_{11} = \Phi\left(T_1^{\mathsf{T}}AT_1\right), \tag{2.23a}$$

$$C_{22} = \Phi\left(T_2^{\mathsf{T}}AT_2\right) \tag{2.23b}$$

(see Property 2.15). We do not have a preference for any choice, although the first one may be slightly more efficient for computational purposes, as can be seen from

PROPERTY 2.24.

By the choice $C_{11} = T_1^{\mathsf{T}} A T_1$ the quantity $\|\dot{T}_1(t)\|$ is minimal for each t (where we assume $\|\cdot\|$ to be the Frobenius-norm or the 2-norm. (2.24)

The equivalent is true for $C_{22} = T_2^{\mathsf{T}} A T_2$ and $\|\dot{T}_2(t)\|$.

PROOF. This property directly follows from Theorem 8.1–10 of [5], which is also valid for the 2-norm and for nonsymmetric A.

The following result, that justifies the use of initial value integration techniques, applies to any of the choices (i), (ii) or (iii).

PROPERTY 2.25.

Let $\tilde{T} = [\tilde{T}_1 \tilde{T}_2]$ satisfy (2.16)–(2.19) *with $T_j := \tilde{T}_j (j = 1, 2)$. Assume that along $\tilde{T}$ as a function of t there exist positive constants α_1 and α_2 such that, for all t,*

$$\lambda_{\min}\left(C_{11} + C_{11}^{\mathsf{T}}\right) = \lambda_{\min}\left(\tilde{T}_1^{\mathsf{T}}(A + A^{\mathsf{T}})\tilde{T}_1\right) \geqslant 2\alpha_1 \tag{2.25a}$$

and

$$\lambda_{\max}\left(C_{22} + C_{22}^{\mathsf{T}}\right) = \lambda_{\max}\left(\tilde{T}_2^{\mathsf{T}}(A + A^{\mathsf{T}})\tilde{T}_2\right) \leqslant -2\alpha_2. \tag{2.25b}$$

Then

(i) $\lim_{t \to \infty} \tilde{T}(t)^{\mathsf{T}} \tilde{T}(t) = I_n$ (i.e. $\tilde{T}(\infty)$ *is orthonormal*).

If, moreover, $\tilde{T}(0)$ is orthonormal, then

(ii)

$$\textit{span}\ \left(\tilde{T}_1(0)\right) \textit{ defines an unstable solution subspace}. \tag{2.26}$$

PROOF. (i) Observe that by (2.25) the differential equation

$$\dot{Z} = \begin{bmatrix} -C_{11}^{\mathsf{T}} & 0 \\ 0 & C_{22} \end{bmatrix} Z + Z \begin{bmatrix} -C_{11} & 0 \\ 0 & C_{22}^{\mathsf{T}} \end{bmatrix} \tag{2.27}$$

has $Z \equiv 0$ as an asymptotically stable solution. Since $Z := I_n - \tilde{T}^{\mathsf{T}}\tilde{T}$, considered as a function of t, is a solution of (2.27) the required result follows in a straightforward manner.

(ii) Let

$$\dot{Y} = \begin{bmatrix} C_{11} & \tilde{T}_1^{\mathsf{T}}(A + A^{\mathsf{T}})\tilde{T}_2 \\ 0 & C_{22} \end{bmatrix} Y, \qquad Y(0) = I_n$$

(cf. (2.5), (2.6)). For every $c_1 \in \mathbf{R}^k$ it follows from (2.25a) that $\|Y_{11}(t)c_1\| \geqslant \|c_1\| \cdot \exp(\alpha_1 t)$, $t \geqslant 0$. Similarly from (2.25b) we see that, for every $c_2 \in \mathbf{R}^{n-k}$, $\|Y_{22}(t)c_2\| \leqslant \|c_2\| \cdot \exp(-\alpha_2 t)$, $t \geqslant 0$. Hence, $\tilde{T}_1(0)$ induces a k-dimensional manifold of unstable solutions $(\tilde{T}_1(t)Y_{11}(t))$.

REMARK. Observe that the conditions (2.25) imply stability of the transformed system (2.8) in the indicated directions. Moreover, orthonormality of the transformation matrix T can be seen as a stable property. In general, however, (2.25) may not be satisfied. For numerical purposes therefore it seems worthwhile to change the righthand side of (2.16) into $AT_1 - T_1(T_1^{\mathsf{T}}T_1)^{-1}C_{11}$, since then the relation $\frac{d}{dt}(T_1^{\mathsf{T}}T_1) = 0$ is implied by (2.17), even if T_1 is not exactly orthonormal. A similar update for C_{22} can be given. In [3] and [11] it is suggested that we may restrict ourselves to the update $\{\operatorname{diag}(T_1^{\mathsf{T}}T_1)\}^{-1}C_{11}$, where $\operatorname{diag}(T_1^{\mathsf{T}}T_1)$ stands for the diagonal matrix containing the diagonal elements of $T_1^{\mathsf{T}}T_1$.

2.3. Initial values

So far we have not discussed how to find a suitable initial value $T(0)$ for (2.4). There are some fairly generally applicable technbiques for this (cf. [8], [10], [12]). If the BCs are *separated* (for instance occurring in Dirichlet problems) there is a natural choice for $T(0)$, which moreover economises on computational labour (see next section). Assume the BCs are given by

$$n-k \quad \left[B_{21}^0\ B_{22}^0\right]x(0) = b_2 \tag{2.28a}$$

$$k \quad \left[B_{11}^1\ B_{12}^1\right]x(1) = b_1. \tag{2.28b}$$

For a well-conditioned problem it is known (cf. [9]) that a continuous transformation T for which $\|T\|\,\|T^{-1}\|$ is not large and

$$\left[B_{21}^0\ B_{22}^0\right]T(0) = \left[0\ V_{22}^0\right], \tag{2.29}$$

(where $V_{22}^0 \in \mathbf{R}^{(n-k)\times(n-k)}$ is non-singular by the well-posedness), implies stability of (2.8) in the indicated directions.

The importance of the boundedness condition on $\|T\|\,\|T^{-1}\|$ is illustrated by the next

EXAMPLE: consider the BVP

$$\dot{x} = \begin{bmatrix} -10 & 0 \\ 20 & 10 \end{bmatrix} x,$$

subject to the BCs

$$x_2(0) = 0 \quad \text{and} \quad x_1(1) + x_2(1) = 1.$$

This is, as one may easily show, a well-conditioned BVP. The Riccati transformation corresponding to (2.29) is

$$T(t) = \begin{bmatrix} 1 & 0 \\ e^{20t} - 1 & 1 \end{bmatrix}$$

and

$$\tilde{A}(t) = \begin{bmatrix} -10 & 0 \\ 0 & 10 \end{bmatrix}.$$

Hence the system has not been decoupled correctly. (In fact there is no decoupling of increasing and decreasing solutions since both columns of T describe directions of growing solutions.)

This example also illustrates that the stability of (2.8) should be global, not local. Let $z(t) = (e^{-10t}, e^{10t} - e^{-10t})^{\mathsf{T}}$. Then for an orthogonal T we have

$$T_1(t) = z/\|z\|_2 \quad \text{and} \quad C_{11}(t) = (10e^{20t} - 20e^{-20t})/\|z\|_2^2,$$

which moves rapidly from -10 to $+10$ as t goes from 0 to 1. This shows moreover that the differential equation for T_1 (and T_2) may be *stiff*. By Theorem 3.14 of [9] we see that the choice (2.29), with T orthonormal, gives a correct decoupling of the increasing and decreasing modes.

For more general BCs, (2.29) is not useful. However, theoretically we know that for any dichotomic system separated BCs exist such that the resulting BVP is well-conditioned ([6]). The construction of such separated BCs implicitly employs a fundamental solution, which is yet to be computed. Hence, this result is not directly applicable. However, in general, we may still find a reasonable starting guess using the Schur factorisation: if Q is an orthogonal matrix, such that $Q^{\mathsf{T}}A(0)Q$ is (quasi) upper triangular with ordered eigenvalues (cf. [8]), then arguments related to the ones used in subspace iteration (including the QR algorithm), make it very likely that this choice for Q induces the appropriate decoupling.

3. Discrete analogues

In Section 2 we have indicated ways to transform the system (1.1) into a special form. Typically such transformation methods may be seen as continuous analogues of multiple shooting methods (cf. [1], [10]), to be discussed next.

3.1. Multiple shooting

Suppose $(0,1)$ is divided into subintervals (t_i, t_{i+1}), $i = 0, \ldots, m-1$, where $t_0 = 0$ and $t_m = 1$. On each such subinterval let $F^i(t)$ indicate a fundamental solution and $p^i(t)$ a particular solution of (1.1). Then we have a sequence of vectors v^i for which

$$F^{i-1}(t_i)v^{i-1} + p^{i-1}(t_i) = x(t_i) = F^i(t_i)v^i + p^i(t_i), \tag{3.1}$$

$(i = 1, \ldots, m)$. This leads to the one-step recursion (n.b. $\det F^i(t_i) \neq 0$)

$$v^i = \left(F^i(t_i)\right)^{-1}\left\{F^{i-1}(t_i)v^{i-1} + p^{i-1}(t_i) - p^i(t_i)\right\}, \tag{3.2}$$

which is, together with the BCs,

$$B^0F^0(t_0)v^0 + B^1F^m(t_m)v^m = b - B^0p^0(t_0) - B^1p^m(t_m)$$

a discrete analogue of (1.1), (1.2) respectively.

By introducing a sequence of (time-independent) discrete *Lyapunov transformations* $\{T^i\}_{i=0}^m$ we can transform (3.2) into a one-step recursion with incremental matrices of a special form. Defining

$$X^i := \left[F^{i+1}(t_{i+1})\right]^{-1}F^i(t_{i+1})$$

this leads via

$$Y^i := (T^{i+1})^{-1}X^iT^i,$$
$$\tilde{v}^i := (T^i)^{-1}v^i$$

to the recursion

$$\tilde{v}^i = Y^{i-1}\tilde{v}^{i-1} + \left[F^i(t_i)T^i\right]^{-1}\left(p^{i-1}(t_i) - p^i(t_i)\right).$$

For every kind of transformation discussed in Section 2 there is a discrete counterpart $\{T^i\}_{i=0}^m$.

For instance: if all T^i are orthogonal and such that Y^i is upper triangular, then $T^i = T(t_i)$, where T is the solution of (2.4) with $\tilde{A} = \Phi(T^{\mathrm{T}}AT)$ (cf. Property 2.15), subject to $T(0) = T^0$. Moreover, Y^i is equal to $Y(t_{i+1})Y^{-1}(t_i)$ (see (2.5)).

3.2. Marching techniques

The above remarks are also true when the BCs are separated. Assume the BCs are given by

$$n-k \quad \left[B_{21}^0\ B_{22}^0\right]x(0) = b_2 \tag{3.3a}$$

$$k \quad \left[B_{11}^1\ B_{12}^1\right]x(1) = b_1. \tag{3.3b}$$

Then we need to compute just a part of a fundamental solution (marching technique, cf. [12]). Let p be a particular solution of (1.1) satisfying

$$\left[B_{21}^0\ B_{22}^0\right]p(0) = b_2,$$

and X a fundamental solution with

$$\left[B_{21}^0\ B_{22}^0\right]X(0) = [0\ *].$$

Any solution of (1.1) can now be written as

$$x(t) = \begin{bmatrix} X_{11}(t) & X_{12}(t) \\ X_{21}(t) & X_{22}(t) \end{bmatrix}\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} p_1(t) \\ p_2(t) \end{bmatrix}, \qquad t \in (0,1).$$

By the special choice of $p(0)$ and $X(0)$ we obtain $c_2 = 0$. Hence

$$x(t) = \begin{bmatrix} X_{11}(t) \\ X_{21}(t) \end{bmatrix}c_1 + \begin{bmatrix} p_1(t) \\ p_2(t) \end{bmatrix}, \qquad t \in (0,1),$$

and only X_1 (the first k columns of X) has to be computed. If the continuous transformation T is such that $\tilde{A}_{21} \equiv 0$ (cf. (2.3a)) and $T_1(0) = X_1(0)$, then $X_1 = T_1 Y_{11}$, where Y_{11} is the solution of

$$\dot{Y}_{11} = \tilde{A}_{11}(t) Y_{11}, \qquad Y_{11}(0) = I_k$$

(see (2.5)). Observe that the computation of a particular solution still includes the original stability problem. The transformation $y := T^{-1}x$ will not help us much since the coefficients in the ODE

$$\dot{y}_2 = \tilde{A}_{22}(t) y_2 + \tilde{f}_2(t), \qquad t \in (0,1),$$

are depending on T_2 which is unknown, except for the Riccati case.

One possibility to solve this problem is to define $z := (I - T_1 T_1^{\mathsf{T}})p$ and λ by $x = T_1\lambda + z$ (cf. [4], [11]). One can show that z is the solution of

$$\dot{z} = \left[I - T_1(t)T_1^{\mathsf{T}}(t)\right]\left(A(t)z + f(t)\right) - T_1(t)T_1^{\mathsf{T}}(t)A^{\mathsf{T}}(t)z, \tag{3.4a}$$

$t \in (0,1)$, subject to

$$\begin{bmatrix} T_1^{\mathsf{T}}(0) \\ B_{21}^0 \quad B_{22}^0 \end{bmatrix} z(0) = \begin{bmatrix} 0 \\ b_2 \end{bmatrix}, \tag{3.4b}$$

and λ the solution of

$$\dot{\lambda} = C_{11}(t)\lambda + T_1^{\mathsf{T}}(t)\big(\big(A^{\mathsf{T}}(t) + A(t)\big)z(t) + f(t)\big), \tag{3.5a}$$

$t \in (0,1)$, where the value of $\lambda(1)$ is determined by the relation

$$\left[B_{11}^1 \; B_{12}^1\right]\left(T_1(1)\lambda(1) + z(1)\right) = b_1. \tag{3.5b}$$

As one can see, solving (3.5) will be a tedious matter, since λ has to be computed in a backward sweep. For this reason Meyer ([11]) suggests to solve (3.4) and to determine $x(1)$ as in (3.5b). Hereafter the solution is computed by solving the terminal value problem

$$\begin{cases} \dot{x} = A(t)x + f(t), & t \in (0,1) \\ x(1) \text{ known.} \end{cases} \tag{3.6}$$

To control the inherent error, the solution is projected into the solution manifold spanned by z and the columns of T_1, at *a priori* determined points. By this projection the error is not directly damped out, but its direction is such that by further integration it initially will decrease.

Since most of the stability problems are still present in (3.6) (which may lead to a very large number of projections) it seems better to use an invariant imbedding formulation. Let $\{t_i\}$ be the points where output is required and define R_{11}^i and $g^i (i = 0, \ldots, m)$ by the relation

$$\lambda(t_i) = R_{11}^i(t)\lambda(t) + g^i(t), \qquad t \in (t_i, t_{i+1}).$$

Then for $t \in (t_i, t_{i+1})$ we obtain the ODEs

$$\dot{R}_{11}^i = -R_{11}^i C_{11}(t), \qquad R_{11}^i(t_i) = I_k \tag{3.7}$$

and

$$\begin{cases} \dot{g}^i = -R_{11}^i(t)T_1^{\mathsf{T}}(t)\big((A^{\mathsf{T}}(t) + A(t))z(t) + f(t)\big) \\ g^i(t_i) = 0. \end{cases} \tag{3.8}$$

Now all ODEs are stable and can be solved from the left to right. Especially for stiff problems with just a few output points, this formulation may be useful.

4. Example

The following example is very simple, but illustrates quite nicely the features of the previously discussed methods. Let x satisfy

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \omega \begin{bmatrix} -1 & 0 \\ 2 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad t \in (0,1), \tag{4.1a}$$

subject to

$$x_2(0) = 0 \quad \text{and} \quad x_1(1) + x_2(1) = 1. \tag{4.1b}$$

(exact solution: $x(t) = (e^{-\omega(t+1)}, e^{\omega(t-1)} - e^{-\omega(t+1)})^{\mathsf{T}}$). Observe that in this example T_1 is one-dimensional, so C_{11} of (2.16) is unique.

The performance of continuous orthonormalisation combined with the backward sweep (3.6) is hard to check, since there is no adaptive strategy for choosing the projection points. In this example, solutions are in both directions growing like $\exp(\omega t)$. Assume T_1 is determined with an accuracy δ, and x is to be determined within a prescribed tolerance ε. To control the error, the distance Δt between two projection points must then satisfy the relation

$$\delta \cdot \varepsilon \cdot \exp(\omega \Delta t) \leqslant \varepsilon, \tag{4.2}$$

so $\Delta t \leqslant \ln \delta / \omega$. Although this restriction looks quite disastrous, it is less dramatic in practice since the error in T_1 will mainly be in the direction of T_1 itself. Not counting the attempts with too many and/or too few projection points, we obtained the following results (see Table 1):

TABLE 1. Results on the Burroughs B7900 of the Eindhoven University of Technology for solving (4.1) with (2.16) and (3.6). As integrator the Algol-procedure MULTISTEP is used, which is a GEAR-like code. Accuracy: 10^{-6}.

ω	execution time in sec.	number of projection points	by (4.2) expected number of projection points
10	0.49	10	1
10^3	1.35	50	75
10^5	22.39	100	7250

The results of continuous orthonormalisation combined with invariant imbedding are more transparent. In Table 2 the execution times are shown for various values of ω. With II4 the output points were chosen as $t_i = i/4$ ($i = 0, \ldots, 4$) and with II10 as $t_i = i/10$ ($i = 0, \ldots, 10$), see Table 2.

These numbers show that the stiffness of the problem has only a minor influence on the performance of the code (the results became even more accurate as ω increased). This is just what one should expect since, after an initial layer, the stepsize is mainly determined by the smoothness of the outer solution.

TABLE 2. Execution times in sec. on the Burroughs B7900 of the Eindhoven University of Technology for solving (4.1) with (2.16) and (3.7). Integrator: MULTISTEP. Accuracy: 10^{-6}.

	$\omega = 10$	$\omega = 10^3$	$\omega = 10^5$	$\omega = 10^7$
II4	0.39	1.18	1.35	1.30
II10	0.56	3.05	3.34	3.83

5. Conclusion

From the foregoing analysis, we believe that we have demonstrated that continuous transformations may be an alternative for multiple shooting methods. Especially when in both directions the solutions are growing very fast, the merits of an adaptive stiff integrator may be used to gain a lot of computational time. How much can be gained, strongly depends on the rotational activity of the dominant solutions. If this activity is mainly within the dominant solution space a Riccati transformation or formulation (2.21) may be used. The resulting ODEs will then be stiff. If, however, the dominant solution space itself is rapidly rotating, then the Riccati transformation will need too many restarts, and hence is not advisable. The solution of (2.16) will be a rapidly oscillating function and therefore (2.16) is hard to solve. However, for these kinds of problems any solution method will meet difficulties.

The backward sweep (3.6) is not recommendable, since in the nonstiff case multiple shooting will do, and in the stiff case it is much more expensive than the invariant imbedding formulation, since too many projections will be necessary. Moreover, the projection points cannot be determined adaptively.

References

[1] I. Babuska, "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).

[2] W. A. Coppel, "Dichotomies in Stability Theory", *Lecture Notes in Mathematics* **629** (Springer-Verlag, Berlin, 1978).
[3] A. Davey, "An automatic orthonormalization method for solving stiff BVPs", *J. Comput. Phys.* **51** (1983) 343–356.
[4] T. Eirola, "A study of the Back-and-Forth shooting method", Ph.D. Thesis, Helsinki University of Technology, Finland, 1985.
[5] G. H. Golub and C. F. van Loan, *Matrix computations* (North Oxford Academic, Oxford, 1983).
[6] F. De Hoog and R. M. M. Mattheij, "On dichotomy and well-conditioning in boundary value problems", *SIAM J. Numer. Anal.* **24** (1987) 89–105.
[7] H. B. Keller and M. Lentini, "Invariant Imbedding, the Box Scheme and an Equivalence between them", *SIAM J. Numer. Anal.* **19** (1982) 942–962.
[8] P. M. Van Loon, "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).
[9] R. M. M. Mattheij, "Decoupling and Stability of Algorithms for Boundary Value Problems", *SIAM Rev.* **27** (1985) 1–44.
[10] R. M. M. Mattheij and G. W. M. Staarink, "An Efficient Algorithm for solving general linear two point BVP", *SIAM J. Sci. Statist. Comput.* **5** (1984) 745–763.
[11] G. H. Meyer, "Continuous Orthonormalization for Boundary Value Problems", *J. Comput. Phys.* **62** (1986) 248–262.
[12] M. R. Osborne, "The stabilized march is stable", *SIAM J. Numer. Anal.* **16** (1979) 923–933.
[13] M. R. Osborne and R. D. Russell, "The Riccati Transformation in the Solution of Boundary Value Problems", Univ of New Mexico Dept. Maths. Rep., Albuquerque, 1985.