ASC Report No. 24/2009

Normal Forms for Companion Matrices and Contractivity in Inner Product Norms

Winfried Auzinger

Institute for Analysis and Scientific Computing
Vienna University of Technology — TU Wien
www.asc.tuwien.ac.at ISBN 978-3-902627-02-5

Most recent ASC Reports

23/2009 *Harald Woracek*
Existence of Zerofree Functions N-associated to a de Branges Pontryagin Space

22/2009 *Michael Kaltenbäck, Harald Woracek*
Pontryagin Spaces of Entire Functions VI

21/2009 *Michael Kaltenbäck, Harald Woracek*
Pontryagin Spaces of Entire Functions V

20/2009 *Christoph Erath, Stefan Funken, Petra Goldenits, Dirk Praetorius*
Simple Error Estimations for the Galerkin Bem for some Hypersingular Integral Equation in 2D

19/2009 *Daniel Matthes, Ansgar Jüngel, Giuseppe Toscani*
Convex Sobolev Inequalities Derived from Entropy Dissipation

18/2009 *Roberta Bosi, Maria J. Cáceres*
The BGK Model with External Confining Potential: Existence, Long-time Behaviour and Time-periodic Maxwellian Equilibria

17/2009 *Irene M. Gamba, Ansgar Jüngel, Alexis Vasseur*
Global Existence of Solutions to One-Dimensional Viscous Quantum Hydrodynamic Equations

16/2009 *Claire Chainai-Hillairat, Marguerite Gisclon, Ansgar Jüngel*
A Finite-Volume Scheme for the Multidimensional Quantum Drift-Diffusion Model for Semiconductors

15/2009 *Carsten Carstensen, Dirk Praetorius*
Convergence of Adaptive Boundary Element Methods

14/2009 *Franco Fagnola, Lukas Neumann*
Quantum Fokker-Planck Models: Limiting Case in the Lindblad Condition

Institute for Analysis and Scientific Computing
Vienna University of Technology
Wiedner Hauptstraße 8–10
1040 Wien, Austria

E-Mail: `admin@asc.tuwien.ac.at`
WWW: `http://www.asc.tuwien.ac.at`
FAX: +43-1-58801-10196

ISBN 978-3-902627-02-5

Normal forms for companion matrices and contractivity in inner product norms

Winfried Auzinger

Institute for Analysis and Scientific Computing
Vienna University of Technology
Wiedner Hauptstrasse 8–10/E101
A-1040 Wien, Austria, EU

w.auzinger@tuwien.ac.at

Abstract

We study the problem of finding a inner product norm in which a given companion matrix $C \in \mathbb{C}^{n\times n}$ with a [weakly] stable spectrum becomes contractive (or dissipative), via a preferably well-conditioned change of basis. To this end we use a basis transformation related to a rescaled LQ decomposition of the associated Vandermonde matrix which is robust to w.r.t. confluent or non-confluent spectra. For $n = 2$ we give an explicit construction. The transformed, contractive matrix is non-normal in general, and it depends on the distribution of the spectrum in a nonlinear way. This analysis cannot be directly generalized to higher dimension, but it suggests an algebraic/numerical algorithm for a numerically given spectrum. This has been tested for small values of n and appears to be successful.

2000 Mathematics Subject Classification: 15A21, 15A60.

Key words and phrases: companion matrix, weak stability, contractivity.

1 Introduction and overview

The term nonnormality is a placeholder for a rich variety of phenomena in matrix analysis, cf. e.g. [15]. Here our topic is a question nontrivial due to nonnormality, namely to find, for a given matrix $A \in \mathbb{C}^{n\times n}$ with spectrum satisfying a [weak] stability condition, a natural inner product norm in which A becomes a contraction. In principle, one of the well-known equivalent conditions in the Kreiss matrix theorem asserts that an appropriate basis transformation always exists, cf. e.g. [9],[15]. However, the proofs of this fact are not constructive, cf. e.g. the survey paper [13] or the proof given in [12].

In this paper we argue that finding such a transformation, reasonably well-conditioned, is a difficult problem in general, and we provide a partial solution. We restrict ourselves to the special class of companion matrices $C \in \mathbb{C}^{n\times n}$. It is essential to handle non-confluent and confluent spectra in a uniform way, independent of the clustering or multiplicity of eigenvalues. To this end it is favorable to consider companion matrices not in a purely linear algebra setting but to refer to their interpretation in the context of polynomial algebra.

The paper is organized as follows: In Section 2 we review bases in polynomial interpolation, with emphasis on confluent limits and interpretation in terms of orthogonality. In particular, we consider confluent forms of the LU- and LQ-decomposition of Vandermonde matrices. In Section 3 these bases are used to transform a given companion matrix to Hessenberg (or bidiagonal, tridiagonal) forms, which depend on the spectrum in a continuous way (in contrast to the Jordan form). The bidiagonal form is well known; it is considered mainly for the sake of completeness and for motivating use of the alternative Hessenberg or tridiagonal form. In Section 4 the latter used to study the contractivity problem for stable companion matrices. We give a general, explicit construction in terms of the spectrum for dimension $n = 2$, which is already nontrivial. For $n \geq 3$ we discuss the question how to find the appropriate basis transformation by means of an algebraic/numerical algorithm. Our procedure appears to be successful in numerical practice, but we have not been able to give a complete theoretical explanation for this observation. The analogous question for the case of a [weakly] dissipative spectrum is also briefly studied.

Although some simple applications are mentioned, the topic of this paper is mainly theoretical. We stress that our approach taken in Section 3 via maximizing a certain determinant appears to be remarkably successful; trying to explain this observation may be of more general interest. We also note that in [5],[8], the contractivity of stable companion matrices is discussed from a different point of view.

Remark concerning notation: For any $A \in \mathbb{C}^{m \times n}$, A' denotes its ordinary transpose and A^* its Hermitian transpose.[1] For a function f of a real of complex argument, $\dot{f}$ denotes its derivative.

2 Orthogonal polynomial bases in interpolation

Let Π_{n-1} denote the space of complex polynomials of degree $\leq n-1$. Our focus is on orthogonal bases in Π_{n-1}. We assume that n nodes $\eta_1, \ldots, \eta_n \in \mathbb{C}$ are given, not necessarily distinct, and denote

$$\pi(\zeta) := (\zeta - \eta_1) \cdots (\zeta - \eta_n). \tag{2.1}$$

Divided differences of scalar or vector-valued polynomials $u(\zeta)$ w.r.t. the η_k are denoted as

$$u[\eta_1, \ldots, \eta_\ell] \quad \text{or by the shortcut} \quad u_{[1..\ell]}. \tag{2.2}$$

If the η_k are not pairwise distinct, this is to be interpreted in the usual confluent sense.

2.1 Newton-Taylor basis

The monomial basis in Π_{n-1} is denoted by

$$\boldsymbol{m}(\zeta) = (m_0(\zeta), m_1(\zeta), \ldots, m_{n-1}(\zeta))' = (1, \zeta, \ldots, \zeta^{n-1})'. \tag{2.3}$$

The (transposed) Vandermonde matrix associated with the $m_j(\eta_k)$ is [2]

$$V = V(\eta_1, \ldots, \eta_n) = \begin{pmatrix} | & | & & | \\ \boldsymbol{m}(\eta_1) & \boldsymbol{m}(\eta_2) & \ldots & \boldsymbol{m}(\eta_n) \\ | & | & & | \end{pmatrix} = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ \eta_1 & \eta_2 & \cdots & \eta_n \\ \eta_1^2 & \eta_2^2 & \cdots & \eta_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ \eta_1^{n-1} & \eta_2^{n-1} & \ldots & \eta_n^{n-1} \end{pmatrix}. \tag{2.4}$$

The 'Newton-Taylor basis' associated with the given η_k is denoted by[3]

$$\boldsymbol{n}(\zeta) = (n_0(\zeta), n_1(\zeta), \ldots, n_{n-1}(\zeta))', \qquad n_j(\zeta) = \prod_{\ell=1}^{j} (\zeta - \eta_\ell). \tag{2.5}$$

[1] Note that $\|A\|_2 = \|A'\|_2 = \|A^*\|_2$ for all $A \in \mathbb{C}^{n \times n}$.

[2] We are not referring to any confluent regularization of V for the case of multiple η_k.

[3] $\boldsymbol{m}(\zeta)$ may also be called the 'Taylor basis' w.r.t. the node $\eta = 0$; it is the special case of $\boldsymbol{n}(\zeta)$ for $\eta_k \equiv 0$.

It is well known from interpolation theory (cf. e.g. [7]) that, for distinct η_k, the change of basis $\boldsymbol{m}(\zeta) \mapsto \boldsymbol{n}(\zeta)$ is described by the LU-decomposition of V, $V = L\,U$ with

$$L = \begin{pmatrix} | & | & & | \\ \boldsymbol{m}[\eta_1] & \boldsymbol{m}[\eta_1,\eta_2] & \dots & \boldsymbol{m}[\eta_1,\dots,\eta_n] \\ | & | & & | \end{pmatrix} = \begin{pmatrix} 1 & & & & \\ \zeta_{[1]} & 1 & & & \\ \zeta^2_{[1]} & \zeta^2_{[1\cdot\cdot 2]} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ \zeta^{n-1}_{[1]} & \zeta^{n-1}_{[1\cdot\cdot 2]} & \dots & \zeta^{n-1}_{[1\cdot\cdot n-1]} & 1 \end{pmatrix}, \tag{2.6}$$

L unit lower diagonal,

$$U = \begin{pmatrix} | & | & & | \\ \boldsymbol{n}(\eta_1) & \boldsymbol{n}(\eta_2) & \dots & \boldsymbol{n}(\eta_n) \\ | & | & & | \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ & (\eta_2-\eta_1) & (\eta_3-\eta_1) & \dots & (\eta_n-\eta_1) \\ & & \prod_{\ell=1}^{2}(\eta_3-\eta_\ell) & \dots & \prod_{\ell=1}^{2}(\eta_n-\eta_\ell) \\ & & & \ddots & \vdots \\ & & & & \prod_{\ell=1}^{n-1}(\eta_n-\eta_\ell) \end{pmatrix}. \tag{2.7}$$

In the confluent case, V and U have reduced rank, but the basis transformation $\boldsymbol{m}(\zeta) \mapsto \boldsymbol{n}(\zeta)$ represented by $\boldsymbol{m}(\zeta) = L \cdot \boldsymbol{n}(\zeta)$ is always well defined, and identity $V = L\,U$ remains valid, with confluent interpretation of the divided differences defining L.

The Newton-Taylor basis $\boldsymbol{n}(\zeta)$ satisfies the two-term recurrence

$$\zeta\,\boldsymbol{n}(\zeta) = B \cdot \boldsymbol{n}(\zeta) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \pi(\zeta) \end{pmatrix}, \quad B = \begin{pmatrix} \eta_1 & 1 & & & \\ & \eta_2 & 1 & & \\ & & \ddots & \ddots & \\ & & & \eta_{n-1} & 1 \\ & & & & \eta_n \end{pmatrix}, \tag{2.8}$$

with $\pi(\zeta)$ from (2.1), and $\boldsymbol{n}(\zeta)$ is orthonormal w.r.t. the discrete Sobolev product in Π_{n-1},

$$\langle\!\langle u, v \rangle\!\rangle := \sum_{k=1}^{n} \bar{u}[\eta_1,\dots,\eta_k]\, v[\eta_1,\dots,\eta_k]. \tag{2.9}$$

The transition from $\boldsymbol{m}(\zeta)$ to $\boldsymbol{n}(\zeta)$ may be identified with a conventional Gram-Schmidt process w.r.t. $\langle\!\langle \cdot, \cdot \rangle\!\rangle$. For $u, v \in \Pi_{n-1}$ in monomial representation $u(\zeta) = \boldsymbol{u}'\boldsymbol{m}(\zeta)$, $v(\zeta) = \boldsymbol{v}'\boldsymbol{m}(\zeta)$, we have $u[\eta_1,\dots,\eta_k] = (L'\boldsymbol{u})_k$, $v[\eta_1,\dots,\eta_k] = (L'\boldsymbol{v})_k$, and

$$\langle\!\langle u, v \rangle\!\rangle = \boldsymbol{u}^* W \boldsymbol{v}, \quad \text{with} \quad W = (L')^* L'. \tag{2.10}$$

Note that $\langle\!\langle u, v \rangle\!\rangle$ is always a properly positive definite inner product, independent of the multiplicities of the η_k. Multiple occurrence of some η_k corresponds to a version of Hermite interpolation. The Newton-Taylor representation for an interpolation polynomial,

$$u(\zeta) = \sum_{j=0}^{n-1} u[\eta_1,\dots,\eta_{j+1}] n_j(\zeta) = \sum_{j=0}^{n-1} \langle\!\langle u, n_j \rangle\!\rangle n_j(\zeta) \tag{2.11}$$

covers the standard situations, including Lagrange interpolation and Taylor expansion as special cases.

2.2 [Non-]confluent orthogonal ℓ_2-basis

Consider first the non-confluent case of distinct η_k. Let

$$\boldsymbol{q}(\zeta) = (q_0(\zeta), q_1(\zeta), \dots, q_{n-1}(\zeta))' \tag{2.12}$$

denote the graded basis (degree$(q_j) = j$) which is orthonormal w.r.t. the ℓ_2 - inner product

$$\langle u, v\rangle := \sum_{k=1}^{n} \bar{u}(\eta_k) v(\eta_k), \qquad \|u\| := \langle u, u\rangle^{\frac{1}{2}}. \tag{2.13}$$

The change of basis $\boldsymbol{m}(\zeta) \mapsto \boldsymbol{q}(\zeta)$ is represented by the LQ decomposition of V, $V = KQ$, with K lower diagonal and Q unitary. Then, $\boldsymbol{m}(\zeta) = K \cdot \boldsymbol{q}(\zeta)$, and

$$Q = \begin{pmatrix} | & | & & | \\ \boldsymbol{q}(\eta_1) & \boldsymbol{q}(\eta_2) & \dots & \boldsymbol{q}(\eta_n) \\ | & | & & | \end{pmatrix} \quad \text{with} \;\; Q\,Q^* = I, \tag{2.14}$$

cf. e.g. [7]. $\boldsymbol{q}$ is not a monic basis, but a diagonal rescaling yields

$$V = (KD^{-1})(DQ) =: LP \quad \text{with} \;\; D = \mathrm{Diag}(K), \tag{2.15}$$

where the new transformation matrix L is unit lower diagonal, and

$$\boldsymbol{p}(\zeta) = (p_0(\zeta), p_1(\zeta), \dots, p_{n-1}(\zeta))' := D\,\boldsymbol{q}(\zeta) \tag{2.16}$$

is a monic basis which is also orthogonal (but not orthonormal) w.r.t. $\langle\cdot,\cdot\rangle$,

$$P := \begin{pmatrix} | & | & & | \\ \boldsymbol{p}(\eta_1) & \boldsymbol{p}(\eta_2) & \dots & \boldsymbol{p}(\eta_n) \\ | & | & & | \end{pmatrix} \quad \text{with} \;\; PP^* = D^2, \qquad \boldsymbol{m}(\zeta) = L \cdot \boldsymbol{p}(\zeta). \tag{2.17}$$

This process is not well-defined in the confluent case. In the sequel we consider a modification of this orthogonalization procedure which makes sense in general. Assuming now that the η_k are arbitrary, we indicate the Gram-Schmidt process w.r.t. the [semi]-definite form $\langle\cdot,\cdot\rangle$ (with associated [semi]-norm $\|\cdot\|$), which transforms $\boldsymbol{m}(\zeta)$ into $\boldsymbol{p}(\zeta)$ in the general case. For the $p_j(\zeta)$ we use an ansatz which directly yields the coefficients in the associated recurrence.

The case $n = 2$.

It is convenient to consider the special case $n = 2$ first. Let $\hat{\eta} := \frac{1}{2}(\eta_1 + \eta_2)$. In the non-confluent case we have $V = KQ = LP$ with

$$K = \sqrt{2}\begin{pmatrix} 1 & \\ \hat{\eta} & \frac{1}{2}|\eta_1 - \eta_2| \end{pmatrix}, \quad L = \begin{pmatrix} 1 & \\ \hat{\eta} & 1 \end{pmatrix}, \quad P = \begin{pmatrix} 1 & 1 \\ \eta_1 - \hat{\eta} & \eta_2 - \hat{\eta} \end{pmatrix} = \begin{pmatrix} p_0(\eta_1) & p_0(\eta_2) \\ p_1(\eta_1) & p_1(\eta_2) \end{pmatrix}. \tag{2.18}$$

Alternatively, including the confluent case, $\boldsymbol{p}(\zeta) = (p_0(\zeta), p_1(\zeta))'$ is constructed as follows.

[0.] $p_0(\zeta) := m_0(\zeta) = 1$, with $\|p_0\| = \sqrt{2}$.

[1.] Ansatz with parameter γ_1:

$$p_1(\zeta) := (\zeta - \gamma_1)p_0(\zeta) = (\zeta - \gamma_1). \tag{2.19}$$

Requirement $\langle p_0, p_1 \rangle = 0$ yields

$$\gamma_1 = \hat{\eta}, \quad p_1(\zeta) = \zeta - \hat{\eta}. \tag{2.20}$$

Note that $\|p_1\| = 0 \Leftrightarrow \eta_1 = \eta_2$ $(= \hat{\eta})$.

This construction is also well-defined also in the confluent case $\eta_1 = \eta_2$. The monic polynomials

$$p_0(\zeta) = 1, \quad p_1(\zeta) = \zeta - \hat{\eta} \tag{2.21}$$

are linearly independent, and the change of basis $\boldsymbol{m}(\zeta) \mapsto \boldsymbol{p}(\zeta)$ is represented by a unit lower diagonal matrix which we again denote by L,

$$\boldsymbol{m}(\zeta) = L \cdot \boldsymbol{p}(\zeta), \quad L = \begin{pmatrix} 1 & \\ -\gamma_1 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & \\ \hat{\eta} & 1 \end{pmatrix}, \tag{2.22}$$

and $V = L\,P$ with L, P from (2.18) is a properly rescaled LQ-decomposition of V.

In the confluent case, V and P have reduced rank 1, and the inner product $\langle \cdot, \cdot \rangle$ degenerates. But there is of course a natural inner product which is trivially well-defined, definite, and properly scaled: For $u(\zeta) = \boldsymbol{u}'\boldsymbol{p}(\zeta)$, $v(\zeta) = \boldsymbol{v}'\boldsymbol{p}(\zeta) \in \Pi_1$ we define $\langle\!\langle u, v \rangle\!\rangle := \boldsymbol{u}^*\boldsymbol{v}$ $(\boldsymbol{u} = (u_0, u_1)'$, $\boldsymbol{v} = (v_0, v_1)')$. Obviously,

$$\langle\!\langle u, v \rangle\!\rangle = \bar{u}_0 v_0 + \bar{u}_1 v_1 = \bar{u}(\hat{\eta})v(\hat{\eta}) + \dot{\bar{u}}\,\dot{v}, \qquad |\!|\!| u |\!|\!| := \langle\!\langle u, u \rangle\!\rangle^{\frac{1}{2}} \tag{2.23}$$

defines yet another discrete Sobolev product and norm on Π_1, respectively, with $u_0 = u(\hat{\eta}) = \frac{1}{2}(u(\eta_1) + u(\eta_2))$ and $\dot{u} \equiv u[\eta_1, \eta_2]$. By construction, the $p_j(\zeta)$ are orthonormal w.r.t. $\langle\!\langle \cdot, \cdot \rangle\!\rangle$. The basis $\boldsymbol{p}(\zeta)$ may be considered as a 'symmetric version' of the Newton-Taylor basis $\boldsymbol{n}(\zeta)$ which does not depend on a particular ordering of the η_k. Expressed in monomial coordinates, $u(\zeta) = \boldsymbol{u}'\boldsymbol{m}(\zeta)$, $v(\zeta) = \boldsymbol{v}'\boldsymbol{m}(\zeta)$, we have

$$\langle\!\langle u, v \rangle\!\rangle = \boldsymbol{u}^* W \boldsymbol{v}, \quad \text{with} \quad W = (L')^* L' = \begin{pmatrix} 1 & \hat{\eta} \\ \bar{\hat{\eta}} & 1 + |\hat{\eta}|^2 \end{pmatrix}. \tag{2.24}$$

Furthermore, let γ_2 and λ_2 be defined such that

$$(\zeta - \gamma_2)p_1(\zeta) - \lambda_2\, p_0(\zeta) = \pi(\zeta) = (\zeta - \eta_1)(\zeta - \eta_2). \tag{2.25}$$

This gives $\gamma_2 = \hat{\eta}$, $\lambda_2 = \frac{1}{4}(\eta_1 - \eta_2)^2$ With these parameters, the basis $\boldsymbol{p}(\zeta)$ satisfies a recurrence which may be written in the form

$$\zeta\,\boldsymbol{p}(\zeta) = H \cdot \boldsymbol{p}(\zeta) + \begin{pmatrix} 0 \\ \pi(\zeta) \end{pmatrix}, \quad H = \begin{pmatrix} \gamma_1 & 1 \\ \lambda_2 & \gamma_2 \end{pmatrix} = \begin{pmatrix} \hat{\eta} & 1 \\ \frac{1}{4}(\eta_1 - \eta_2)^2 & \hat{\eta} \end{pmatrix}, \quad \hat{\eta} = \tfrac{1}{2}(\eta_1 + \eta_2). \tag{2.26}$$

Outline of general confluent orthogonalization procedure.

In general, we have to take special care for different versions of confluence.

$$s := \text{number of distinct } \eta_k, \tag{2.27}$$

the degree of the minimal polynomial associated with $\pi(\zeta) = (\zeta - \eta_1) \ldots (\zeta - \eta_n)$. By $\hat{\eta} := \frac{1}{n}\sum_{k=1}^{n} \eta_k$ we denote the barycenter of the polygon spanned by the η_k. Let us consider in detail the first steps of the orthogonalization process.

[0.] $p_0(\zeta) := m_0(\zeta) = 1$, with $\|p_0\| = \sqrt{n}$.

[1.] Ansatz:

$$p_1(\zeta) := (\zeta - \gamma_1)p_0(\zeta) = (\zeta - \gamma_1). \tag{2.28}$$

– Requirement $\langle p_0, p_1\rangle = 0$ yields

$$0 = \langle 1, p_1(\zeta)\rangle = \textstyle\sum_{k=1}^{n} p_1(\eta_k) \quad \Rightarrow \quad \gamma_1 = \hat{\eta}, \quad p_1(\zeta) = \zeta - \hat{\eta}. \tag{2.29}$$

Note that $\|p_1\| = 0 \;\Leftrightarrow\; s = 1$, i.e., iff $\eta_1 = \ldots = \eta_n = \hat{\eta}$.

[2.] Ansatz:

$$p_2(\zeta) := (\zeta - \gamma_2)p_1(\zeta) - \lambda_2\, p_0(\zeta). \tag{2.30}$$

– Requirement $\langle p_0, p_2\rangle = 0$ yields

$$\begin{aligned} 0 = \langle 1, p_2(\zeta)\rangle &= \langle 1, \zeta\, p_1(\zeta)\rangle - \gamma_2\langle 1, p_1(\zeta)\rangle - \lambda_2\langle 1, p_0(\zeta)\rangle \\ &= \langle 1, \zeta\, p_1(\zeta)\rangle \;-\; 0 \;-\; n\lambda_2. \end{aligned} \tag{2.31}$$

This uniquely determines λ_2,

$$\lambda_2 = \tfrac{1}{n}\langle 1, \zeta\, p_1(\zeta)\rangle = \textstyle\sum_{k=1}^{n} \tfrac{1}{n}\, p_1(\eta_k)\, \eta_k. \tag{2.32}$$

For $s = 1$ we obtain $\lambda_2 = 0$.

– Requirement $\langle p_1, p_2\rangle = 0$ yields

$$\begin{aligned} 0 = \langle p_1(\zeta), p_2(\zeta)\rangle &= \langle p_1(\zeta), \zeta\, p_1(\zeta)\rangle - \gamma_2\langle p_1(\zeta), p_1(\zeta)\rangle - \lambda_2\langle p_1(\zeta), p_0(\zeta)\rangle \\ &= \langle 1, \zeta|p_1(\zeta)|^2\rangle - \gamma_2\|p_1\|^2 \;+\; 0. \end{aligned} \tag{2.33}$$

This uniquely determines γ_2,

$$\gamma_2 = \frac{\langle 1, \zeta|p_1(\zeta)|^2\rangle}{\|p_1\|^2} = \sum\nolimits_{k=1}^{n} \frac{|p_1(\eta_k)|^2}{\sum_{\ell=1}^{n}|p_1(\eta_\ell)|^2}\,\eta_k \quad \text{if} \quad s > 1. \tag{2.34}$$

Otherwise the natural choice for γ_2 is

$$\gamma_2 = \textstyle\sum_{k=1}^{n} \tfrac{1}{n}\,\eta_k \;\equiv\; \eta_k = \gamma_1 = \hat{\eta}, \quad \text{thus: } \; p_2(\zeta) = (\zeta - \hat{\eta})^2 \; \text{ for } \; s = 1. \tag{2.35}$$

Note that $\|p_2\| = 0 \;\Leftrightarrow\; s \le 2$, because $p_2(\eta_k) \equiv 0$ iff at least $n-1$ of η_k coincide (observing that $p_2(\zeta)$ is monic of degree 2).

[3.] Ansatz:

$$p_3(\zeta) := (\zeta - \gamma_3)p_2(\zeta) - \lambda_3\, p_1(\zeta) - \kappa_3\, p_0(\zeta). \tag{2.36}$$

– Requirement $\langle p_0, p_3\rangle = 0$ yields

$$\begin{aligned} 0 = \langle 1, p_3(\zeta)\rangle &= \langle 1, \zeta\, p_2(\zeta)\rangle - \gamma_3\langle 1, p_2(\zeta)\rangle - \lambda_3\langle 1, p_1(\zeta)\rangle - \kappa_3\langle 1, p_0(\zeta)\rangle \\ &= \langle 1, \zeta\, p_2(\zeta)\rangle \;-\; 0 \;-\; 0 \;-\; n\,\kappa_3. \end{aligned} \tag{2.37}$$

This uniquely determines κ_3,

$$\kappa_3 = \tfrac{1}{n}\langle 1, \zeta\, p_2(\zeta)\rangle = \textstyle\sum_{k=1}^{n} \tfrac{1}{n}\, p_2(\eta_k)\, \eta_k. \tag{2.38}$$

For $s \le 2$ we obtain $\kappa_3 = 0$.

– Requirement $\langle p_1, p_3\rangle = 0$ yields

$$\begin{aligned} 0 = \langle p_1(\zeta), p_3(\zeta)\rangle &= \langle p_1(\zeta), \zeta\, p_2(\zeta)\rangle - \gamma_3\langle p_1(\zeta), p_2(\zeta)\rangle - \lambda_3\langle p_1(\zeta), p_1(\zeta)\rangle - \kappa_3\langle p_1(\zeta), p_0(\zeta)\rangle \\ &= \langle p_1(\zeta), \zeta\, p_2(\zeta)\rangle \;-\; 0 \;-\; \lambda_3\|p_1\|^2 \;-\; 0. \end{aligned} \tag{2.39}$$

This uniquely determines λ_3,

$$\lambda_3 = \frac{\langle p_1(\zeta), \zeta\, p_2(\zeta)\rangle}{\|p_1\|^2} = \sum\nolimits_{k=1}^{n} \frac{\bar{p}_1(\eta_k)p_2(\eta_k)}{\sum_{\ell=1}^{n}|p_1(\eta_\ell)|^2}\,\eta_k \quad \text{if} \quad s > 1. \tag{2.40}$$

Otherwise the natural choice is

$$\lambda_3 = 0 \quad \text{for} \quad s = 1, \tag{2.41}$$

since the multiplicity of the zero of $\bar{p}_1(\zeta)p_2(\zeta)\zeta$ at $\zeta = \hat{\eta}$ is higher than for $|p_1(\zeta)|^2$. For $s = 2$ we also have $\lambda_3 = 0$ due to $p_2(\eta_k) \equiv 0$.

– Requirement $\langle p_2, p_3\rangle = 0$ yields

$$\begin{aligned} 0 = \langle p_2(\zeta), p_3(\zeta)\rangle &= \langle p_2(\zeta), \zeta\, p_2(\zeta)\rangle - \gamma_3\langle p_2(\zeta), p_2(\zeta)\rangle - \lambda_3\langle p_2(\zeta), p_1(\zeta)\rangle - \kappa_3\langle p_2(\zeta), p_0(\zeta)\rangle \\ &= \langle 1, \zeta\, |p_2(\zeta)|^2\rangle - \gamma_3\|p_2\|^2 \;-\; 0 \;-\; 0. \end{aligned} \tag{2.42}$$

This uniquely determines γ_3,

$$\gamma_3 = \frac{\langle 1, \zeta |p_2(\zeta)|^2\rangle}{\|p_2\|^2} = \sum\nolimits_{k=1}^{n} \frac{|p_2(\eta_k)|^2}{\sum_{\ell=1}^{n}|p_2(\eta_\ell)|^2}\,\eta_k \quad \text{if} \quad s > 2. \tag{2.43}$$

Otherwise the natural choice for γ_3 is

$$\gamma_3 = \sum\nolimits_{k=1}^{n} \tfrac{1}{n}\,\eta_k \;\equiv\; \eta_k = \hat{\eta} \quad \text{for} \quad s \le 2. \tag{2.44}$$

Note that $\|p_3\| = 0 \;\Leftrightarrow\; s \le 3$, because $p_3(\eta_k) \equiv 0$ iff at least $n-2$ of the η_k coincide (observing that $p_3(\zeta)$ is monic of degree 3).

- . . .

It is rather obvious how this procedure is to be continued, but the general handling of confluence will have to be done in a systematic way, and a double index notation for the recurrence coefficients has to be used,

$$p_j(\zeta) = (\zeta - \underbrace{h_{jj}}_{=\gamma_j})p_{j-1}(\zeta) - h_{j,j-1}\,p_{j-2}(\zeta) - \ldots - h_{j1}\,p_0(\zeta), \quad j < n. \tag{2.45}$$

The procedure is independent of the ordering of the η_k. In this paper, however, we do not attempt to describe this symbolic algorithm its general form. Of course it may be of interest to show that it is well-defined in general, with the special outcome of a three-term recurrence if the η_k lie on a common line, in particular if they are real, as is to be expected from the theory of orthogonal polynomials (cf. e.g. [4]). An explicit representation of the parameters $h_{j\ell}$ in terms of the η_k becomes quite cumbersome already for $n = 3$, and we will give no details.

The change of basis $\boldsymbol{m}(\zeta) \mapsto \boldsymbol{p}(\zeta)$, where the monic basis

$$\boldsymbol{p}(\zeta) = (p_0(\zeta), p_1(\zeta), \ldots, p_{n-1}(\zeta))' \tag{2.46}$$

is constructed as indicated above, is represented by a unit lower diagonal matrix L,

$$\boldsymbol{m}(\zeta) = L\cdot \boldsymbol{p}(\zeta), \quad L = \begin{pmatrix} 1 & & & \\ -\gamma_1 & 1 & & \\ \gamma_1\gamma_2-\lambda_2 & -\gamma_1-\gamma_2 & 1 & \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}^{-1} = \begin{pmatrix} 1 & & & \\ \gamma_1 & 1 & & \\ \gamma_1^2+\lambda_2 & \gamma_1+\gamma_2 & 1 & \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{2.47}$$

The decomposition $V = L\,P$, with $P = \Big(\, \boldsymbol{p}(\eta_1) \;\big|\; \boldsymbol{p}(\eta_2) \;\big|\; \ldots \;\big|\; \boldsymbol{p}(\eta_n) \,\Big)$ is a rescaled LQ-decomposition of V, the natural extension of the non-confluent decomposition to the general case.

For $u(\zeta) = \boldsymbol{u}^{'}\boldsymbol{p}(\zeta)$, $v(\zeta) = \boldsymbol{v}^{'}\boldsymbol{p}(\zeta) \in \Pi_{n-1}$ let

$$\langle\!\langle u, v \rangle\!\rangle := \boldsymbol{u}^{*}\boldsymbol{v}, \qquad |\!|\!| u |\!|\!| := \langle\!\langle u, u \rangle\!\rangle^{\frac{1}{2}} \tag{2.48}$$

The basis $\boldsymbol{p}(\zeta)$ is orthonormal w.r.t. this inner product. Expressed in monomial coordinates, $u(\zeta) = \boldsymbol{u}^{'}\boldsymbol{m}(\zeta)$, $v(\zeta) = \boldsymbol{v}^{'}\boldsymbol{m}(\zeta)$, we have

$$\langle\!\langle u, v \rangle\!\rangle = \boldsymbol{u}^{*} W \boldsymbol{v}, \quad \text{with} \quad W = (L^{'})^{*}\, L^{'}. \tag{2.49}$$

As for $n = 2$ this may be called a (weighted) discrete Sobolev product; it is uniquely determined by η_k and does not depend on their ordering.

In addition, we complete the above orthogonalization procedure by defining parameters h_{nk} such that

$$(\zeta - \underbrace{h_{nn}}_{=\gamma_n})p_{n-1}(\zeta) - h_{n,n-1}\,p_{n-2}(\zeta) - \ldots - h_{n1}\,p_0(\zeta) = \pi(\zeta) = (\zeta-\eta_1)\cdots(\zeta-\eta_n). \tag{2.50}$$

With all these parameters, the basis $\boldsymbol{p}(\zeta)$ satisfies a recurrence of the general the form

$$\zeta\,\boldsymbol{p}(\zeta) = H\cdot\boldsymbol{p}(\zeta) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \pi(\zeta) \end{pmatrix}, \quad H = \begin{pmatrix} h_{11} & 1 & & & \\ h_{21} & h_{22} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ \vdots & \vdots & & \ddots & 1 \\ h_{n1} & h_{n2} & \ldots & \ldots & h_{nn} \end{pmatrix}, \tag{2.51}$$

where H is lower Hessenberg (or tridiagonal) with unit upper diagonal.

3 Similarity transformation of companion matrices

Consider a Frobenius matrix

$$C = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -c_0 & -c_1 & \ldots & -c_{n-2} & -c_{n-1} \end{pmatrix} \in \mathbb{C}^{n\times n}, \tag{3.1}$$

which is the companion matrix of its associated characteristic polynomial with roots ζ_k, $k = 1\ldots n$ (of arbitrary multiplicities),

$$\chi(\zeta) = \det(\zeta I - C) = \zeta^n + c_{n-1}\zeta^{n-1} + \ldots + c_1\zeta + c_0 = (\zeta-\zeta_1)\cdots(\zeta-\zeta_n), \quad \zeta_k \in \mathbb{C}. \tag{3.2}$$

The normal forms for C described in the sequel depend on the choice of the parameters $\eta_1, \ldots, \eta_n$ which, in principle, are arbitrary complex numbers and need not be pairwise distinct. These normal forms, which transform the lower Hessenberg matrix C into another lower Hessenberg form, are obtained via the basis transformations considered in Section 2. In applications, the η_k typically are given approximations for the characteristic roots ζ_k, and therefore we refer to $\{\eta_1, \ldots, \eta_n\}$ as a 'pseudospectrum' for C, and as before we denote $\pi(\zeta) = (\zeta - \eta_1) \cdots (\zeta - \eta_n)$. For the investigations in Section 4 we will assume that the ζ_k are given, and with $\eta_k :\equiv \zeta_k$ we obtain special normal forms with a simpler structure. We will refer to this as the 'spectral case'.

Frobenius matrices are of relevance in various applications. In the present context the standard interpretation of C in the context of polynomial algebra is convenient, where [non-]confluent situations can be handled in a uniform way: C represents multiplication by ζ mod χ in the complex polynomial ring Π_{n-1} of degree 1 w.r.t. the monomial basis (2.3): For

$$u^\circ(\zeta) := \zeta \circ u(\zeta) := \zeta\, u(\zeta) \mod \chi, \quad u \in \Pi_{n-1}, \tag{3.3}$$

we have

$$\boldsymbol{m}^\circ(\zeta) := \zeta \circ \boldsymbol{m}(\zeta) \equiv C \cdot \boldsymbol{m}(\zeta), \tag{3.4}$$

where $\zeta \circ \boldsymbol{m}(\zeta)$ is to be interpreted componentwise.

For the linear operator $\circ$ or matrix C, respectively, we now apply a basis transformation in Π_{n-1} and this will lead us to a similarity transformation $C = L\,T\,L^{-1}$. As in Section 2 we consider two versions, the first of which is well-known and and mentioned mainly for the sake of completeness.

3.1 Newton-Taylor orthogonalization and Bidiagonal-Frobenius form

For the Newton-Taylor basis $\boldsymbol{n}(\zeta)$ from (2.5) we have $\boldsymbol{m}(\zeta) = L \cdot \boldsymbol{n}(\zeta)$ with L from (2.6), and the two-term recurrence (2.8) holds. This yields

$$\boldsymbol{n}^\circ(\zeta) := \zeta \circ \boldsymbol{n}(\zeta) = \begin{pmatrix} \zeta\, n_0(\zeta) \mod \chi \\ \vdots \\ \zeta\, n_{n-2}(\zeta) \mod \chi \\ \zeta\, n_{n-1}(\zeta) \mod \chi \end{pmatrix} = \begin{pmatrix} \zeta\, n_0(\zeta) \\ \vdots \\ \zeta\, n_{n-2}(\zeta) \\ \zeta\, n_{n-1}(\zeta) - \chi(\zeta) \end{pmatrix}, \tag{3.5}$$

and

$$\boldsymbol{n}^\circ(\zeta) = B \cdot \boldsymbol{n}(\zeta) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \pi(\zeta) - \chi(\zeta) \end{pmatrix}, \quad B \text{ from (2.8)}, \tag{3.6}$$

where

$$\Pi_{n-1} \ni \pi(\zeta) - \chi(\zeta) = \sum_{k=0}^{n-1} -\chi[\eta_1, \ldots, \eta_{k+1}]\, n_k(\zeta). \tag{3.7}$$

This shows

$$\boldsymbol{n}^\circ(\zeta) = L^{-1}\, C\, L \cdot \boldsymbol{n}(\zeta) = T \cdot \boldsymbol{n}(\zeta), \tag{3.8}$$

with

$$T = \begin{pmatrix} \eta_1 & 1 & & & \\ & \eta_2 & 1 & & \\ & & \ddots & \ddots & \\ & & & \eta_{n-1} & 1 \\ -\chi_{[1]} & -\chi_{[1\cdot\cdot 2]} & \cdots & -\chi_{[1\cdot\cdot n-1]} & -\chi_{[1\cdot\cdot n]} + \eta_n \end{pmatrix}. \tag{3.9}$$

The matrix T is of special lower Hessenberg form, the so-called Bidiagonal-Frobenius form, and it is the coefficient matrix associated with the recurrence for the $n_j(\zeta)$ ending up at $\pi(\zeta)$. Identity $C\,L = L\,T$ may also be written in the more conventional, transposed form in terms of the associated coordinate transformation, $C'(L')^{-1} = (L^{-1})'T'$, with T' upper Hessenberg, and where the columns of $(L^{-1})'$ are orthonormal w.r.t. the inner product in $\mathbb{C}^n$ of Newton-Taylor type (cf. (2.9)).

In the spectral case $\eta_k \equiv \zeta_k$, T takes the special bidiagonal form (2.8),

$$T = \begin{pmatrix} \zeta_1 & 1 & & & \\ & \zeta_2 & 1 & & \\ & & \ddots & \ddots & \\ & & & \zeta_{n-1} & 1 \\ & & & & \zeta_n \end{pmatrix} = B. \tag{3.10}$$

For the Taylor case $\eta_k \equiv 0$, T is identical with C.

For various theoretical and numerical applications of this normal form we refer to [1],[2],[3],[6],[14], and references therein. In [6] the bidiagonal form has been used for a quantitative stability analysis of linear multistep methods applied to stiff ODEs. Here the point is that via an appropriate diagonal rescaling, the bidiagonal form can be converted into a contraction in the $\|\cdot\|_\infty$ norm, assuming the ζ_k satisfy a stability condition w.r.t. the unit circle. In Section 4 we will study an analogous question, namely transforming C in such way that a contraction w.r.t. $\|\cdot\|_2$ is obtained. This is a much more difficult problem, and we will base our investigations on the orthogonal basis transformation from Section 2.2, as described in the next section.

Remark. Consider (3.10) and assume that the ζ_k are contained in the complex unit circle, nicely separated, and one of the of modulus close to 1. Then B is diagonalizable with a well-conditioned eigensystem, i.e., the transformation to a contraction is straightforward. If some of the 'inner' ζ_k are close together, this makes no sense. Here one may think of finding a positive diagonal matrix Ω such that $\|\Omega^{-1}B\,\Omega\|_2 < [\leq]\ 1$. Evidently, this must also fail because Ω will necessarily have to be very ill-conditioned. Our approach described in the sequel is based on an alternative to the bidiagonal form which is better adapted to the degree of confluence. For real spectra, for instance, this normal form T will be tridiagonal but not symmetric. The problem of ℓ_2-contractivity will be based on an appropriate diagonal rescaling of T, but we will see that finding the scaling parameters is a nontrivial problem.

3.2 [Non-]confluent ℓ_2-orthogonalization and associated Hessenberg form

In Section 2.2 we have indicated how the Gram-Schmidt process for ℓ_2-orthogonalization works in general. As for the Newton-Taylor case this can be rewritten yielding a transformation of C to another Hessenberg form; this may be called an Arnoldi process applied to C. For later use we first specify the details for the simplest case $n = 2$.

The case $n = 2$.

For

$$C = \begin{pmatrix} 0 & 1 \\ -c_0 & -c_1 \end{pmatrix} \in \mathbb{C}^{2\times 2} \tag{3.11}$$

with characteristic polynomial $\chi(\zeta) = \zeta^2 + c_1\zeta + c_0 = (\zeta - \zeta_1)(\zeta - \zeta_2)$ we have $\boldsymbol{m}^\circ(\zeta) = C \cdot \boldsymbol{m}(\zeta)$, $\boldsymbol{m}(\zeta) = (m_0(\zeta), m_1(\zeta))' = (1, \zeta)'$. The transformed basis $\boldsymbol{p}(\zeta)$ from (2.21) is orthonormal w.r.t. the inner product $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ from (2.23). It satisfies $\boldsymbol{m}(\zeta) = L \cdot \boldsymbol{p}(\zeta)$ with $L = \begin{pmatrix} 1 & 0 \\ \hat\eta & 1 \end{pmatrix}$ from (2.22) (with $\hat\eta = \frac{1}{2}(\eta_1 + \eta_2)$), and the recurrence (2.26) holds. This yields

$$\boldsymbol{p}^\circ(\zeta) := \zeta \circ \boldsymbol{p}(\zeta) = \begin{pmatrix} \zeta\, p_0(\zeta) \mod \chi \\ \zeta\, p_1(\zeta) \mod \chi \end{pmatrix} = \begin{pmatrix} \zeta\, p_0(\zeta) \\ \zeta\, p_1(\zeta) - \chi(\zeta) \end{pmatrix}, \tag{3.12}$$

and

$$\boldsymbol{p}^\circ(\zeta) = H \cdot \boldsymbol{p}(\zeta) + \begin{pmatrix} 0 \\ \pi(\zeta) - \chi(\zeta) \end{pmatrix}, \quad H = \begin{pmatrix} \hat\eta & 1 \\ \frac{1}{4}(\eta_1 - \eta_2)^2 & \hat\eta \end{pmatrix} \text{ from (2.26),} \tag{3.13}$$

where

$$\Pi_1 \ni \pi(\zeta) - \chi(\zeta) = -\chi[\eta_1]p_0(\zeta) - \chi[\eta_1, \eta_2]p_1(\zeta). \tag{3.14}$$

With $\chi[\eta_1] = \chi(\hat\eta) + \frac{1}{4}(\eta_1 - \eta_2)^2$, $\chi[\eta_1, \eta_2] = \dot\chi(\hat\eta) = \text{const.}$ this gives

$$\boldsymbol{p}^\circ(\zeta) = L^{-1}\, C\, L \cdot \boldsymbol{p}(\zeta) = T \cdot \boldsymbol{p}(\zeta), \tag{3.15}$$

with

$$T = \begin{pmatrix} \hat\eta & 1 \\ -\chi(\hat\eta) & \hat\eta - \dot\chi(\hat\eta) \end{pmatrix} = \begin{pmatrix} \langle\!\langle p_0^\circ, p_0 \rangle\!\rangle & \langle\!\langle p_0^\circ, p_1 \rangle\!\rangle \\ \langle\!\langle p_1^\circ, p_0 \rangle\!\rangle & \langle\!\langle p_1^\circ, p_1 \rangle\!\rangle \end{pmatrix}. \tag{3.16}$$

T is the coefficient matrix associated with the recurrence for the $p_j(\zeta)$ ending up at $\chi(\zeta)$. In the spectral case $\eta_k \equiv \zeta_k$, $\pi(\zeta) = \chi(\zeta)$ we have $\hat\eta = \frac{1}{2}(\zeta_1 + \zeta_2)$, $\chi(\hat\eta) = -\frac{1}{4}(\zeta_1 - \zeta_2)^2$, and $\dot\chi(\hat\eta) = 0$. By construction, $T = H$ in this case, see (3.13).

Summing up, we can formulate

Proposition 3.1 *For $n = 2$ and arbitrary $\eta_1, \eta_2 \in \mathbb{C}$ and with $\hat\eta = \frac{1}{2}(\eta_1 + \eta_2)$ we have $C = L\,T\,L^{-1}$, with*

$$L = \begin{pmatrix} 1 & \\ \hat\eta & 1 \end{pmatrix}, \qquad T = \begin{pmatrix} \hat\eta & 1 \\ -\chi(\hat\eta) & \hat\eta - \dot\chi(\hat\eta) \end{pmatrix}. \tag{3.17}$$

If $\{\eta_1, \eta_2\} = \{\zeta_1, \zeta_2\}$ is chosen as the spectrum of C, then

$$T = \begin{pmatrix} \gamma_1 & 1 \\ \lambda_2 & \gamma_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(\zeta_1 + \zeta_2) & 1 \\ \frac{1}{4}(\zeta_1 - \zeta_2)^2 & \frac{1}{2}(\zeta_1 + \zeta_2) \end{pmatrix} = H. \tag{3.18}$$

General procedure.

For a general companion matrix (3.1) with characteristic polynomial (3.2) the construction is analogous. The transformed basis $\boldsymbol{p}(\zeta)$ constructed in Section 2.2 is orthonormal w.r.t. the inner product $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ from

(2.49). It satisfies $\boldsymbol{m}(\zeta) = L \cdot \boldsymbol{p}(\zeta)$ with L from (2.15) or (2.47), and a recurrence of the form (2.51) holds. This yields

$$\boldsymbol{p}^\circ(\zeta) := \zeta \circ \boldsymbol{p}(\zeta) = \begin{pmatrix} \zeta\, p_0(\zeta) \mod \chi \\ \vdots \\ \zeta\, p_{n-2}(\zeta) \mod \chi \\ \zeta\, p_{n-1}(\zeta) \mod \chi \end{pmatrix} = \begin{pmatrix} \zeta\, p_0(\zeta) \\ \vdots \\ \zeta\, p_{n-2}(\zeta) \\ \zeta\, p_{n-1}(\zeta) - \chi(\zeta) \end{pmatrix}, \tag{3.19}$$

and

$$\boldsymbol{p}^\circ(\zeta) = H \cdot \boldsymbol{p}(\zeta) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \pi(\zeta) - \chi(\zeta) \end{pmatrix}, \quad H \text{ from (2.51)}, \tag{3.20}$$

where

$$\Pi_{n-1} \in \pi(\zeta) - \chi(\zeta) = \delta_0\, p_0(\zeta) + \ldots + \delta_{n-1}\, p_{n-1}(\zeta) \tag{3.21}$$

with certain coefficients δ_j depending on the η_k. This gives

$$\boldsymbol{p}^\circ(\zeta) = L^{-1}\, C\, L \cdot \boldsymbol{p}(\zeta) = T \cdot \boldsymbol{p}(\zeta), \tag{3.22}$$

with

$$T = \begin{pmatrix} h_{11} & 1 & & & \\ h_{21} & h_{22} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ \vdots & \vdots & & \ddots & 1 \\ h_{n1} - \delta_0 & h_{n2} - \delta_1 & \ldots & \ldots & h_{nn} - \delta_{n-1} \end{pmatrix}, \quad T_{ij} = \langle\!\langle p^\circ_{i-1}, p_{j-1} \rangle\!\rangle. \tag{3.23}$$

By construction we have $C = L\,T\,L^{-1}$, and T is the coefficient matrix associated with the recurrence for the $p_j(\zeta)$ ending up at $\chi(\zeta)$. In the spectral case $\eta_k \equiv \zeta_k$, $\pi(\zeta) = \chi(\zeta)$ we have $T = H$, see (3.20).

Non-confluent spectral case.

For the non-confluent spectral case, C diagonalizable with distinct eigenvalues ζ_k, the matrix $T = H$ can be obtained algorithmically via the LQ-decomposition $V(\zeta_1, \ldots, \zeta_n) = KQ$ and diagonal rescaling, $V = (KD^{-1})(DQ) = LP$, with $D = \mathrm{Diag}(K)$ invertible and L unit lower diagonal, see (2.15): With $Z := \mathrm{Diag}(\zeta_1, \ldots, \zeta_n)$ we have

$$C = V Z\, V^{-1} = L\, T L^{-1}, \quad T = P\, Z\, P^{-1} = D\, QZQ^{^*} D^{-1}. \tag{3.24}$$

and T is diagonally similar to the normal matrix $QZQ^{^*}$. In the confluent limit this is not well-defined.

Remark. With $L = R\,E$ a polar decomposition of L, we may also write $C = L\,T\,L^{-1} = R\,(E\,T\,E^{^*})\,R^{-1}$ with $R > 0$ and $E\,T\,E^{^*}$ unitarily similar to T.

4 Contractivity, or dissipativity, for stable spectra

We now study the problem of finding a basis transformation, preferably well-conditioned, converting a given companion matrix C with a [weakly] stable spectrum into a ℓ_2-contractive, or ℓ_2-dissipative matrix, respectively. The assumption is that the spectrum of C satifies a [weak] stability condition w.r.t. the closed complex unit circle or the closed complex left half plane.

For C diagonalizable with separated eigenvalues, transformation to a contraction is, in principle, trivial via (3.24), $C = V\,Z\,V^{-1}$. However, $\kappa(V)$ becomes arbitrarily large for eigenvalues clustered together. Our construction below is independent of the distribution of the spectrum and robust w.r.t. confluence. We proceed from the transformed version T of C introduced in Section 3, with the spectral choice $\eta_k \equiv \zeta_k$. Recall that T always well-defined (independent of the distribution of multiplicities of the ζ_k); but an appropriate scaling of T has to be found. Our approach is to seek a diagonal matrix $\Omega > 0$ such that $\Omega^{-1}T\,\Omega$ becomes contractive, or dissipative.

The analysis given below for the special case $n = 2$ shows that the solution is not completely straightforward. We present the explicit solution for $n = 2$. For higher dimension, the search for appropriate scaling parameters leads to a highly nonlinear problem in polynomial algebra. For the general case, we formulate a 'tentative' algorithm which, for a given numerical values of the spectrum, amounts to solving a system of polynomial equations in $n - 1$ unknowns. This gives a set of candidates for the unknown scaling parameters which have to be checked; this involves the solution of Hermitian eigenvalue problems. In extensive numerical tests, in particular for $n = 3$ and $n = 4$, this procedure has proven successful.

4.1 Norm contractivity for spectra in the closed unit circle

The case $n = 2$.

We adopt the notation from Sections 2.2 and 3.2. Assume that C from (3.11) satisfies a weak stability condition w.r.t. unit circle, i.e.,

$$|\zeta_1| \le 1, \;\; |\zeta_2| \le 1, \quad |\zeta_1| < 1 \;\text{ if }\; \zeta_1 = \zeta_2. \tag{4.1}$$

According to Proposition 3.1, T from (3.18) is similar to C, and the transformation matrix L from (3.17) is well-conditioned. We now introduce a scaling parameter $\omega > 0$, unspecified at the moment. Let $\Omega := \mathrm{Diag}(1,\omega)$. We write $\frac{1}{2}(\zeta_1 + \zeta_2) =: \hat{\zeta}$ and define[4]

$$q_0(\zeta) := p_0(\zeta) = 1, \qquad q_1(\zeta) := \omega^{-1} p_1(\zeta) = \frac{(\zeta - \hat{\zeta})}{\omega}\,. \tag{4.2}$$

Furthermore, let

$$\langle\!\langle u, v \rangle\!\rangle_\Omega = \bar{u}_0 v_0 + \omega^2\,\bar{u}_1 v_1 = \bar{u}(\hat{\zeta}) v(\hat{\zeta}) + \omega^2\,\dot{\bar{u}}\,\dot{v}, \qquad |\!|\!| u |\!|\!|_\Omega := \langle\!\langle u, u \rangle\!\rangle^{\frac{1}{2}} \tag{4.3}$$

Then, $\langle\!\langle q_0, q_0 \rangle\!\rangle_\Omega = \langle\!\langle q_1, q_1 \rangle\!\rangle_\Omega = 1$, $\langle\!\langle q_0, q_1 \rangle\!\rangle_\Omega = 0$. W.r.t. to the rescaled basis

$$\Omega\,\boldsymbol{p}(\zeta) \;=:\; \boldsymbol{q}(\zeta) = \left(q_0(\zeta), q_1(\zeta)\right)', \tag{4.4}$$

multiplication by ζ mod χ is now represented by $\boldsymbol{q}^\circ(\zeta) := \zeta \circ \boldsymbol{q}(\zeta) \;\equiv\; T_\Omega \cdot \boldsymbol{q}(\zeta)$, where the entries of

$$T_\Omega := \Omega^{-1} T\,\Omega = \begin{pmatrix} \gamma_1 & \omega \\ \frac{\lambda_2}{\omega} & \gamma_2 \end{pmatrix} = \begin{pmatrix} \hat{\zeta} & \omega \\ \frac{-\pi(\hat{\zeta})}{\omega} & \hat{\zeta} - \dot{\chi}(\hat{\zeta}) \end{pmatrix} = \begin{pmatrix} \langle\!\langle q_0^\circ, q_0 \rangle\!\rangle_\Omega & \langle\!\langle q_0^\circ, q_1 \rangle\!\rangle_\Omega \\ \langle\!\langle q_1^\circ, q_0 \rangle\!\rangle_\Omega & \langle\!\langle q_1^\circ, q_1 \rangle\!\rangle_\Omega \end{pmatrix} \tag{4.5}$$

represent the coefficients in the recurrence for the $q_j(\zeta)$.

Now we wish to choose ω not too small, and at the same time not too large, such that the linear operator $u(\zeta) \mapsto u^\circ(\zeta) = \zeta \circ u(\zeta)$ becomes [strictly] contractive w.r.t. $|\!|\!| \cdot |\!|\!|_\Omega$. For arbitrary

$$u(\zeta) = \boldsymbol{u}'\boldsymbol{q}(\zeta) \in \Pi_1, \quad |\!|\!| u |\!|\!|_\Omega = \|\boldsymbol{u}\|_2, \tag{4.6}$$

[4] In this section, $q_k(\zeta)$ denotes appropriately rescaled versions of the $p_k(\zeta)$, where the scaling parameters are to be determined. They are not identical with the original $q_k(\zeta)$ from Section 2.2.

we have

$$u^\circ(\zeta) = \boldsymbol{u}' T_\Omega\, \boldsymbol{q}(\zeta), \quad |||u^\circ|||_\Omega = \|\boldsymbol{u}^\circ\|_2, \tag{4.7}$$

or equivalently, $u^\circ(\zeta) = \boldsymbol{u}^{\circ\prime} \boldsymbol{q}(\zeta)$ with $\boldsymbol{u}^\circ = T_\Omega' \boldsymbol{u}$. Thus, our contractivity requirement is equivalent to the norm bound[5] $\|T_\Omega'\|_2 \le [<]\, 1$, and this in turn is equivalent to

$$S := \Omega^2 - (T')^* \, \Omega^2 \, T' \ \ \ge [>]\ 0 \ \ !? \tag{4.8}$$

positive semi-definite [or even positive definite], where

$$S = \begin{pmatrix} 1 - |\gamma_1|^2 & -(\bar\gamma_1 \lambda_2) \\ -(\bar\gamma_1 \lambda_2)^- & -|\lambda_2|^2 \end{pmatrix} + \omega^2 \begin{pmatrix} -1 & -\gamma_2 \\ -\bar\gamma_2 & 1 - |\gamma_2|^2 \end{pmatrix}, \tag{4.9}$$

with coefficients $\gamma_1, \gamma_2, \lambda_2$ from (3.18).

Now the idea is to consider the determinant

$$\begin{aligned} \det S &= -\omega^4 + \big((1 - |\gamma_1|^2)(1 - |\gamma_2|^2) - |\gamma_1 \gamma_2|^2 + |\gamma_1 \gamma_2 - \lambda_2|^2\big)\omega^2 - |\lambda_2|^2 \\ &= -\omega^4 + \big((1 - |\hat\zeta|^2)^2 - |\hat\zeta|^4 + |\hat\zeta^2 - (\zeta_1 - \hat\zeta)^2|^2\ \big)\omega^2 - |\zeta_1 - \hat\zeta|^4. \end{aligned} \tag{4.10}$$

This assumes its maximal value for

$$\begin{aligned} \omega^2 &= \tfrac{1}{2}\big((1 - |\hat\zeta|^2)^2 - |\hat\zeta|^4 + |\hat\zeta^2 - (\zeta_1 - \hat\zeta)^2|^2\ \big) \\ &= \tfrac{1}{2}\big(1 - 2\,|\hat\zeta|^2 + |\zeta_1|^2\,|\zeta_2|^2\big) \\ &= \tfrac{1}{2}(1 - |\zeta_1|^2)(1 - |\zeta_2|^2) + \tfrac{1}{4}|\zeta_1 - \zeta_2|^2 \ge 0. \end{aligned} \tag{4.11}$$

With this choice for the ω, (4.10) evaluates to

$$\det S = \omega^4 - |\zeta_1 - \hat\zeta|^4 = \big(\omega^2 - \tfrac{1}{4}|\zeta_1 - \zeta_2|^2\big)\big(\omega^2 + \tfrac{1}{4}|\zeta_1 - \zeta_2|^2\big) \tag{4.12}$$

$$= \tfrac{1}{2}(1 - |\zeta_1|^2)(1 - |\zeta_2|^2)\big(\omega^2 + \tfrac{1}{4}|\zeta_1 - \zeta_2|^2\big) \tag{4.13}$$

$$= \tfrac{1}{4}(1 - |\zeta_1|^2)(1 - |\zeta_2|^2)\,|1 - \zeta_1 \zeta_2|^2. \tag{4.14}$$

Now we check requirement (4.8) for S with ω^2 from (4.11). We consider three different cases of a stable spectrum (in all cases, $|\hat\zeta| < 1$ and $\omega > 0$):

(i) $|\zeta_1| < 1$, $|\zeta_2| < 1$, i.e. $\rho(C) < 1$: Here,

$$\omega^2 \ < \ 1 - |\hat\zeta|^2, \quad \text{i.e. } S_{11} > 0, \quad \text{and} \quad \det S > 0, \quad \text{implying} \quad S > 0. \tag{4.15}$$

(ii) $|\zeta_1| = 1$, $|\zeta_2| < 1$: Here,

$$\omega^2 = \tfrac{1}{4}|\zeta_1 - \zeta_2|^2, \quad \det S = 0, \quad \operatorname{trc} S > 0 \tag{4.16}$$

(the estimate for the trace requires a bit of computation). This implies that the eigenvalues of S must be $\lambda_1 = 0$ and $\lambda_2 > 0$, hence $S \ge 0$ with $\operatorname{rank}(S) = 1$.

(iii) $|\zeta_1| = |\zeta_2| = 1$, with $\zeta_1 \ne \zeta_2$: Here,

$$\omega^2 = \tfrac{1}{4}|\zeta_1 - \zeta_2|^2 = 1 - |\hat\zeta|^2 \quad \text{implies} \quad S = 0. \tag{4.17}$$

[5] For the non-confluent case, the choice $\omega = |\zeta_1 - \hat\zeta|$ gives a normal matrix T_Ω, with $\|T_\Omega\|_2 = \rho(C) \le 1$. However, this is not a proper rescaling: It is undefined in the limit $\zeta_2 \to \zeta_1$, where C is not diagonalizable. For ζ_2 close to ζ_1, the condition number of the scaling matrix Ω tends to infinity. This choice for ω is natural only for $\rho(C) = 1$, see cases (ii) and (iii) below.

(Throughout, rank(S) equals the number of roots ζ_k with $|\zeta_k| = 1$.) Thus we have proved:

Proposition 4.1 *Consider a companion matrix of dimension $n = 2$ with complex spectrum $\{\zeta_1, \zeta_2\}$ satisfying the stability assumption (4.1). With $\hat{\zeta} = \frac{1}{2}(\zeta_1 + \zeta_2)$ and*

$$\omega = \sqrt{\tfrac{1}{2}(1 - |\zeta_1|^2)(1 - |\zeta_2|^2) + \tfrac{1}{4}|\zeta_1 - \zeta_2|^2} > 0 \tag{4.18}$$

we have

$$C = L_\Omega T_\Omega L_\Omega^{-1} \tag{4.19}$$

where

$$L_\Omega = L\,\Omega = \begin{pmatrix} 1 & \\ \hat{\zeta} & \omega \end{pmatrix}, \quad T_\Omega = \begin{pmatrix} \hat{\zeta} & \omega \\ \frac{1}{4}\frac{(\zeta_1 - \zeta_2)^2}{\omega} & \hat{\zeta} \end{pmatrix} \quad \textit{with} \quad \|T_\Omega\|_2 \le 1. \tag{4.20}$$

This also means contractivity of $\circ\colon \Pi_1 \to \Pi_1$ w.r.t. $\langle\!\langle \cdot, \cdot \rangle\!\rangle_\Omega$.

Remark.

- The parameter ω from (4.18) is a measure for 'the distance to instability' of the spectrum $\{\zeta_1, \zeta_2\}$. It vanishes exactly in the limiting (unstable) case $\zeta_1 = \zeta_2$ with $|\zeta_1| = |\zeta_2| = 1$.
- For $\rho(C) = 1$ (cases (ii) and (iii) above), C is diagonalizable. In this case it is easy to verify that T_Ω is normal, $\|T_\Omega\|_2 = 1$. Indeed, up to a scalar factor, $\Omega = \mathrm{Diag}(1, \frac{1}{2}|\zeta_1 - \zeta_2|)$ is identical with Diag(K), K from (2.18), from which we infer $T = Q\,Z\,Q^*$; cf. (3.24). Thus, up to unitary transformation the outcome is equivalent to diagonalization of C, which is quite natural in cases (ii) and (iii). We call T a normalization of C.
- The more interesting case is $\rho(C) < 1$. For $\zeta_1 \ne \zeta_2$, T and T_Ω are not directly related to a diagonalization, or normalization, of C. Here we have $S > 0$ and $\|T_\Omega\|_2 < 1$, but in general, $\|T_\Omega\|_2$ cannot be expressed in a reasonably simple way in terms of the data.

  In the confluent case $\zeta_1 = \zeta_2 = \hat{\zeta}$ we obtain $\omega = \frac{\sqrt{2}}{2}(1 - |\hat{\zeta}|^2)$, and

  $$T = \begin{pmatrix} \hat{\zeta} & 1 \\ 0 & \hat{\zeta} \end{pmatrix}, \quad T_\Omega = \begin{pmatrix} \hat{\zeta} & \frac{\sqrt{2}}{2}(1 - |\hat{\zeta}|^2) \\ 0 & \hat{\zeta} \end{pmatrix}, \tag{4.21}$$

  i.e., T_Ω is a rescaled Jordan form.

Summing up, we see that Proposition 4.1 describes a similarity transformation leading to a contraction which is based on a smooth transition between normalization and Jordan decomposition.

Example: Second order difference equations.

Consider the homogeneous difference equation

$$y_{\nu+2} + c_1\,y_{\nu+1} + c_0\,y_\nu = 0, \quad \nu \ge 0, \tag{4.22}$$

for given y_0, y_1. For the characteristic polynomial $\chi(\zeta) = \zeta^2 + c_1 z + c_0 = (\zeta - \zeta_1)(\zeta - \zeta_2)$ we assume that $\{\zeta_1, \zeta_2\}$ satisfies the stability condition (4.1). With $\boldsymbol{y}_\nu = (y_\nu, y_{\nu+1})'$ this is equivalent to $\boldsymbol{y}_{\nu+1} = C\,\boldsymbol{y}_\nu$ with C from (3.11), or equivalently, $L_\Omega^{-1}\boldsymbol{y}_{\nu+1} = T_\Omega\,L_\Omega^{-1}\boldsymbol{y}_\nu$ with L_Ω, T_Ω from (4.20). Here,

$$L_\Omega^{-1}\boldsymbol{y}_\nu = \begin{pmatrix} y_\nu \\ \frac{1}{\omega}(y_{\nu+1} - \hat{\zeta}\,y_\nu) \end{pmatrix}, \tag{4.23}$$

and Proposition 4.1 asserts that

$$\omega^2 \, \|L_\Omega^{-1} \boldsymbol{y}_\nu\|_2^2 = |\omega \, y_\nu|^2 + |y_{\nu+1} - \hat{\zeta} \, y_\nu|^2 \tag{4.24}$$

is always monotonously decreasing with ν.

A tentative algorithm for the general case.

Let n be arbitrary. For the strictly stable case $\rho(C) < 1$ it is well-known that for any positive definite matrix G, there exists a unique positive definite solution X of the Stein equation $X - (C')^* X \, C' = G$, cf. e.g. [10],[11]. Then, $\|X^{\frac{1}{2}} C' X^{-\frac{1}{2}}\|_2 < 1$. The nontrivial question is what is a 'good' choice for G; we also note the formula for the solution X cannot be directly evaluated. In our approach we use T' instead of C', and we are not prescribing G but force $X = \Omega^2$ to be diagonal and try to compute Ω such that $G = S$ satisfies our needs.

The explicit solution for $n = 2$ given above appears to be quite natural. The proof of Proposition 4.1 was based on maximizing the determinant $\det S$, leading to a linear equation for the scaling parameter ω for $n = 2$. For general dimension n on may think of proceeding in an analogous way, starting from the normal form (3.23) with $\delta_j \equiv 0$, i.e. the spectral case $\eta_k \equiv \zeta_k$. The $\zeta_k \in \mathbb{C}$ are assumed to be given, satisfying a weak stability condition w.r.t. unit circle, i.e.,

$$|\zeta_k| \leq 1, \quad k = 1 \ldots n, \quad \text{where each } \zeta_k \text{ with } |\zeta_k| = 1 \text{ is simple.} \tag{4.25}$$

Analogously as for $n = 2$ we consider

$$T_\Omega := \Omega^{-1} T \, \Omega, \quad \text{with} \ \ \Omega = \mathrm{Diag}(1, \omega_1, \ldots, \omega_{n-1}), \tag{4.26}$$

and we wish to determine parameters $\omega_j > 0$, $j = 1 \ldots n-1$, such that $\|T_\Omega'\|_2 \leq [<] \, 1$. This is equivalent to the requirement

$$S := \Omega^2 - (T')^* \, \Omega^2 \, T' \ \geq [>] \ 0 \ \ !? \tag{4.27}$$

positive semi-definite [or even positive definite].[6]

The basic idea is again to look for a maximum of $\det S$. However, to derive explicit expressions for the entries of T becomes very cumbersome even for $n = 3$. They are nonlinear in the parameters ω_j, and the explicit symbolic procedure which has been used for $n = 2$ cannot be readily generalized to $n > 2$. Therefore we restrict ourselves to the case that numerical values for the ζ_k are given and apply the following 'tentative' algorithm: Since S is Hermitian by construction, the function $\det S =: \varphi(\omega_1, \ldots, \omega_{n-1})$ is a higher order polynomial in the parameters ω_j^2, with real coefficients. Now we consider the system of polynomial equations

$$\tfrac{\partial}{\partial \sigma_j} \, \varphi(\sigma_1, \ldots, \sigma_{n-1}), \quad j = 1 \ldots n-1, \tag{4.28}$$

and determine its solution set by means of a standard algorithm implemented in a computer algebra system. We look for solutions $(\sigma_1, \ldots, \sigma_{n-1})$ with $\sigma_j > 0$ and check the spectrum of S for these cases, inserting $\omega_j^2 = \sigma_j$, hoping to find a solution.

For $n > 2$, $\det S$ is typically an unbounded function in the parameters, and a global maximum does not exist. However, in many cases tested, in particular for $n = 3$ and $n = 4$, it turns out that an appropriate set of parameters $\sigma_j > 0$ is found, where $\det S$ has a local maximum and the spectrum of S is positive (or nonnegative), as required.

[6] As discussed in Section 3.2, the explicit construction of T is nontrivial in a confluent situation. Here we do not discuss this point further but we assume that disctinct numerical values for the ζ_k are given, where T has been obtained via a rescaled LQ-decomposition of the associated Vandermonde matrix.

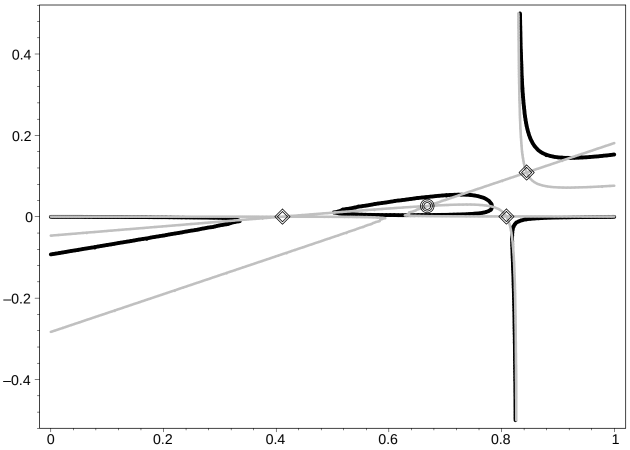

Figure 1: Axes: horizontal = σ_1, vertical = σ_2

The following example for $n = 3$ has been arbitrarily chosen from a collection of a large number of numerical examples which have been treated in the way described above, using Maple 13. Consider the given, stable spectrum $\{\zeta_1, \zeta_2, \zeta_3\} = \{\frac{9}{10}, -\frac{2}{3} + \frac{2}{3}\, i, -\frac{2}{3} + (\frac{2}{3} + \varepsilon)\, i\}$, with ε small. For visualization we choose $\varepsilon = \frac{1}{6}$. We have $\|C\|_2 \approx 2.05$, and for the transformed matrix T, $\|T\|_2 \approx 1.46$. Implementation of the procedure described above in Maple 13 finds a complete solution set of four solution pairs (σ_1, σ_2) in terms of algebraic numbers. For the solution pair $(\sigma_1, \sigma_2) \approx (0.668, 0.027)$ it turns out that $\det S$ has a local maximum, and checking the spectrum of $\operatorname{Re} S$ we obtain $\operatorname{Re} S \geq 0.0036\, I > 0$. Furthermore, $\|T_\Omega\|_2 \approx 0.961$.

A visualization is given in Figure 1. The two hyperbolas correspond to the solution sets of $\frac{\partial \varphi}{\partial \sigma_1} = 0$ and $\frac{\partial \varphi}{\partial \sigma_2} = 0$. Furthermore, the plot shows the contour where $\varphi = \det S \equiv 0$, and the four solution pairs of system $\frac{\partial \varphi}{\partial \sigma_1} = \frac{\partial \varphi}{\partial \sigma_2} = 0$. The solution $(\sigma_1, \sigma_2) \approx (0.618, 0.238)$ is located in the interior of the convex hull of the the other solutions, and $\varphi = \det S$ has a unique local maximum at this point.

For $\varepsilon \to 0$, the matrix T_Ω is neither related to $Z = \operatorname{Diag}(\zeta_1, \zeta_2, \zeta_3)$ nor to a Jordan form of C. It is close to tridiagonal (because the ζ_k approximately lie on a common line) but of course not normal. The condition number of the transformation matrix L_Ω remains bounded for $\varepsilon \to 0$, with a value near 262.

Remark. We believe that, at least for lower dimensions n, the general structure of $\det S$ may be used to argue that a unique local maximum exists for the case $\rho(C) < 1$ ($\rho(C) = 1$ is an exceptional, limiting case). However, already for $n = 3$ the necessary algebra becomes quite involved.

The interesting question is why an appropriate set of parameters is found in this way. The fact that, searching for a diagonal rescaling, local maximization of $\det S$ for S from (4.27) appears to do the job is quite remarkable and may be worth investigating further, possibly also in another context where definiteness is searched for via diagonal scaling.

4.2 Norm dissipativity for spectra in the closed left half plane

The procedure is similar as in Section 4.1.

The case $n = 2$.

Assume that C from (3.11) satisfies a weak stability condition w.r.t. left half plane

$$\operatorname{Re}\zeta_1 \le 0,\ \operatorname{Re}\zeta_2 \le 0, \quad \operatorname{Re}\zeta_1 < 0 \ \text{ if } \ \zeta_1 = \zeta_2. \tag{4.29}$$

Again we wish to choose $\omega > 0$ such that, with $\Omega := \operatorname{Diag}(1, \omega)$, the transformed matrix

$$T_\Omega := \Omega^{-1} T\, \Omega = \begin{pmatrix} \gamma_1 & \omega \\ \frac{\lambda_2}{\omega} & \gamma_2 \end{pmatrix} = \begin{pmatrix} \hat{\zeta} & \omega \\ \frac{1}{4}\frac{(\zeta_1-\zeta_2)^2}{\omega} & \hat{\zeta} \end{pmatrix} \tag{4.30}$$

becomes [strictly] dissipative w.r.t. $|\!|\!| \cdot |\!|\!|_\Omega$, i.e.

$$S \le [<]\ 0 \ \ !? \tag{4.31}$$

negative [semi-]definite, where

$$S := \operatorname{Re} T_\Omega = \tfrac{1}{2}(T_\Omega + T_\Omega^*) = \begin{pmatrix} \operatorname{Re}\,\gamma_1 & \frac{1}{2}(\omega + \frac{\bar{\lambda}_2}{\omega}) \\ \frac{1}{2}(\omega + \frac{\lambda_2}{\omega}) & \operatorname{Re}\,\gamma_2 \end{pmatrix} = \begin{pmatrix} \operatorname{Re}\,\hat{\zeta} & \frac{\omega}{2} + \frac{1}{8}\frac{(\bar{\zeta}_1-\bar{\zeta}_2)^2}{\omega} \\ \frac{\omega}{2} + \frac{1}{8}\frac{(\zeta_1-\zeta_2)^2}{\omega} & \operatorname{Re}\,\hat{\zeta} \end{pmatrix} \tag{4.32}$$

Consider $\tilde{S} := 2\,\omega\, S$. The determinant

$$\det \tilde{S} = -\omega^4 + 2\left(2\,(\operatorname{Re}\,\hat{\zeta})^2 - \operatorname{Re}\lambda_2\right)\omega^2 - |\lambda_2|^2 \tag{4.33}$$

assumes its maximal value for

$$\begin{aligned} \omega^2 &= 2\,(\operatorname{Re}\,\hat{\zeta})^2 - \operatorname{Re}\lambda_2 && (4.34)\\ &= 2\operatorname{Re}\zeta_1 \operatorname{Re}\zeta_2 + \tfrac{1}{4}|\zeta_1 - \zeta_2|^2 \ge 0. && (4.35) \end{aligned}$$

With this choice for ω, (4.33) evaluates to

$$\begin{aligned} \det \tilde{S} = \omega^4 - (\tfrac{1}{4}|\zeta_1 - \zeta_2|^2)^2 &= \left(\omega^2 - \tfrac{1}{4}|\zeta_1 - \zeta_2|^2\right)\left(\omega^2 + \tfrac{1}{4}|\zeta_1 - \zeta_2|^2\right) && (4.36)\\ &= 2\operatorname{Re}\zeta_1 \operatorname{Re}\zeta_2 \left(2\operatorname{Re}\zeta_1 \operatorname{Re}\zeta_2 + \tfrac{1}{2}|\zeta_1 - \zeta_2|^2\right) && (4.37)\\ &= \operatorname{Re}\zeta_1 \operatorname{Re}\zeta_2\ |\zeta_1 + \zeta_2|^2. && (4.38) \end{aligned}$$

Now we check requirement (4.32) for S with ω^2 from (4.34). We consider three different cases of a stable spectrum:

(i) $\operatorname{Re}\zeta_1 < 0,\ \operatorname{Re}\zeta_2 < 0:$ Here, $\omega^2 > 0,\ \zeta_1 + \zeta_2 \ne 0$, and

$$\operatorname{Re}\,\hat{\zeta} < 0, \ \ \text{i.e. } S_{11} < 0, \ \ \text{and} \ \ \det S = 4\,\omega^2 \det \tilde{S} > 0, \quad \text{implying} \ \ S < 0. \tag{4.39}$$

(ii) $\operatorname{Re}\zeta_1 = 0,\ \operatorname{Re}\zeta_2 < 0:$ Here, $\omega^2 = \frac{1}{4}|\zeta_1 - \zeta_2|^2 > 0$, and

$$\det S = 0, \quad \operatorname{trc} S < 0. \tag{4.40}$$

This implies that the eigenvalues of S must be $\lambda_1 = 0$ and $\lambda_2 < 0$, hence $S \le 0$ with $\operatorname{rank}(S) = 1$.

(iii) $\operatorname{Re}\zeta_1 = \operatorname{Re}\zeta_2 = 0$, with $\zeta_1 \ne \zeta_2:$ Here, $\omega^2 = \frac{1}{4}|\zeta_1 - \zeta_2|^2 > 0$, $\operatorname{Re}\,\hat{\zeta} = 0$, and

$$\det S = 0, \quad \operatorname{trc} S = 0, \quad \text{hence} \ \ S = 0. \tag{4.41}$$

Throughout, rank(S) equals the number of roots ζ_k with Re $\zeta_k = 0$. The logarithmic norm of T_Ω, i.e. the rightmost eigenvalue of S is given by

$$\mu_2(T_\Omega) = \mathrm{Re}\ \hat{\zeta} + \big|\,\omega + \frac{\frac{1}{2}(\zeta_1-\zeta_2)^2}{\omega}\big|, \tag{4.42}$$

with $\mu_2(T_\Omega) < 0$ for case (i) and $\mu_2(T_\Omega) = 0$ for cases (ii) and (iii). Thus we have proved:

Proposition 4.2 *Consider a companion matrix of dimension $n = 2$ with complex spectrum $\{\zeta_1, \zeta_2\}$ satisfying the stability assumption (4.29). With $\hat{\zeta} = \frac{1}{2}(\zeta_1 + \zeta_2)$ and*

$$\omega = \sqrt{2\,\mathrm{Re}\,\zeta_1\,\mathrm{Re}\,\zeta_2 + \tfrac{1}{4}|\zeta_1 - \zeta_2|^2} > 0 \tag{4.43}$$

we have

$$C = L_\Omega T_\Omega L_\Omega^{-1} \tag{4.44}$$

where

$$L_\Omega = L\,\Omega = \begin{pmatrix} 1 & \\ \hat{\zeta} & \omega \end{pmatrix}, \quad T_\Omega = \begin{pmatrix} \hat{\zeta} & \omega \\ \frac{1}{4}\frac{(\zeta_1-\zeta_2)^2}{\omega} & \hat{\zeta} \end{pmatrix} \quad \textit{with} \quad \mathrm{Re}\ T_\Omega \le 0. \tag{4.45}$$

Again, the parameter ω from (4.43) is a measure for 'the distance to instability' of the spectrum $\{\zeta_1, \zeta_2\}$. It vanishes exactly in the limiting (unstable) case $\zeta_1 = \zeta_2$ with $\mathrm{Re}\,\zeta_1 = \mathrm{Re}\,\zeta_2 = 0$. Analogous remarks as following Proposition 4.1 apply.

Example: The damped harmonic oscillator.

The purpose of this example is to show that, in the context of a simple ODE problem, Proposition 4.2 automatically provides a 'physically meaningful' dissipation functional.

Consider the second order linear ODE for the free damped harmonic oscillator in the dimensionless variable y,

$$\ddot{y}(t) + 2\gamma\,\dot{y}(t) + \omega_0^2\,y(t) = 0, \tag{4.46}$$

with damping parameter $\gamma \ge 0$ and angular frequency $\omega_0 > 0$. For $\boldsymbol{y}(t) = (y(t), \dot{y}(t))'$ we have

$$\dot{\boldsymbol{y}}(t) = C\,\boldsymbol{y}(t), \quad C = \begin{pmatrix} 0 & 1 \\ -\omega_0^2 & -2\gamma \end{pmatrix}, \tag{4.47}$$

with eigenvalues $\zeta_{1,2} = -\gamma \pm \sqrt{\gamma^2 - \omega_0^2}$ and $\hat{\zeta} = \frac{1}{2}(\zeta_1 + \zeta_2) = -\gamma$. Consider the assertion from Proposition 4.2. In all three cases (over- or underdamping, critical damping) we easily obtain $\omega = \sqrt{\gamma^2 + \omega_0^2}$, and

$$T_\Omega = \begin{pmatrix} \gamma & \sqrt{\gamma^2 + \omega_0^2} \\ \frac{\gamma^2 - \omega_0^2}{\sqrt{\gamma^2+\omega_0^2}} & \gamma \end{pmatrix} \quad \text{with} \quad \mathrm{Re}\ T_\Omega \le \big(\tfrac{\gamma}{\sqrt{\gamma^2+\omega_0^2}} - 1\big)\gamma I =: -\rho I \le 0. \tag{4.48}$$

With $(L_\Omega^{-1}\boldsymbol{y})^{\dot{}} = T_\Omega\,(L_\Omega^{-1}\boldsymbol{y})$ this implies

$$\|L_\Omega^{-1}\boldsymbol{y}(t)\|_2 \le e^{-\rho t}\,\|L_\Omega^{-1}\boldsymbol{y}(0)\|_2, \quad \rho = \big(1 - \tfrac{\gamma}{\sqrt{\gamma^2+\omega_0^2}}\big)\gamma \ge 0. \tag{4.49}$$

In other words,

$$\tilde{E}(y, \dot{y}) := (\gamma^2 + \omega_0^2)\|L_\Omega^{-1}\boldsymbol{y}\|_2^2 = (\gamma^2 + \omega_0^2)\,y^2 + (\dot{y} + \gamma\,y)^2 \tag{4.50}$$

is always a Lyapunov function for the oscillator, $d\tilde{E} \leq 0$ along solution trajectories. In the undamped case, $\tilde{E}$ is identical with the total energy functional $E(y, \dot{y}) = \omega_0^2\, y^2 + \dot{y}^2$ which is conserved, $d\tilde{E} \equiv 0$ for $\gamma = 0$. For $\gamma < 0$ we have $dE < 0$, and $d\tilde{E} < 0$ due to $\rho > 0$, where $\tilde{E} \neq E$. A straightforward calculation shows $d\tilde{E} = -2\gamma E$, i.e., $\tilde{E}(t)$ represents a form of mean energy.

Remark. From numerical experiments we believe that for $n > 2$ and a spectrum satisfying a weak stability condition w.r.t. the closed left half plane, a 'tentative algorithm' will work in a similar way as described in Section 4.1. We are not going into detail here.

References

[1] W. Auzinger, W. Herfort, *A uniform quantitative stiff stability estimate for BDF schemes*, Opuscula Mathematica 26, Vol. 2 (2006), 203-227.

[2] F. Blanchini, *New canonical forms for pole placement*, in: IEEE Proceedings, Vol. 136, No. 6 (1989).

[3] D. Calvetti, L. Reichel, F. Sgallari, *A modified companion matrix based on Newton polynomials*, in: 'Fast Algorithms for Structured Matrices: Theory and Applications' (ed. V. Olshevski), Contemporary Mathematics, Vol. 323, Amer. Math. Soc., Providence, RI (2003), 179–186.

[4] T. S. Chihara, *An Introduction to Orthogonal Polynomials*, Gordon and Breach Science Publishers, 1978.

[5] G. Dahlquist, H. Mingyou, R. LeVeque, *On the uniform power-boundedness of a family of matrices and the applications to one-leg and linear multistep methods*, Numer. Math., Vol. 42 (1983), 1–13.

[6] A. Eder, G. Kirlinger, *A normal form for multistep companion matrices*, Math. Models and Methods in Applied Sciences, Vol. 11, No. 1 (2001), 57–70.

[7] W. Gander, *Change of basis in polynomial interpolation*, Numer. Linear Algebra Appl., Vol. 12, No. 8 (2005), 769–778.

[8] H.-L. Gau, P. Y. Wu, *Companion matrices: reducibility, numerical ranges and similarity to contractions*, Lin. Alg. Appl., Vol. 382 (2004), 127–142.

[9] B. Gustafsson, H.-O. Kreiss, J. Oliger, *Time Dependent Problems and Difference Methods*, J. Wiley & Sons, 1995.

[10] R. Horn, C. Johnson, *Matrix Analyis*, Cambridge University Press, 1990.

[11] P. Lancaster, M. Tismenetsky, *The Theory of Matrices*, second edition with applications, Academic Press, 1985.

[12] S. Ma, *A new proof of the Kreiss matrix theorem* (in Chinese), Math. Numer. Sinica, Vol. 11, No. 1 (1989), 104–106.

[13] J. C. Strikwerda, B. A. Wade, *A survey of the Kreiss matrix theorem for power bounded families of matrices and its extensions*, in: Linear operators (Warsaw, 1994), Banach Center Publ., Vol. 38 (1997), 339–360.

[14] W. Werner, *A generalized companion matrix of a polynomial and some applications*, Lin. Alg. Appl., Vol. 55 (1983), 19–36.

[15] L. N. Trefethen, M. Embree, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*, Princeton University Press, 2005.