

DAEs

Jan Heiland

2021-05-09

Contents

Preface	5
1 Introduction	7
1.1 Examples	7
1.2 Why are DAEs difficult to treat	11
2 Basic Definitions and Notions	13
2.1 Solution Concept	13
2.2 Initial Conditions and Consistency	14
2.3 Additional Remarks	15
3 Linear DAEs with Constant Coefficients	17
3.1 Basic Notions and Definitions	17
3.2 Regularity and Solvability	20
3.3 Solution to the N -DAE, Regularity, and Index of a Matrix Pair .	21
3.4 Existence of Solutions	25
3.5 A Variation of Constant Formula	27
4 Linear DAEs with Time-varying Coefficients	33
5 Numerical Approximation of DAEs	41
6 Construction and Analysis of RKM for nonlinear DAEs	45
6.1 General RKM for Semi-Explicit Strangeness-free DAEs	45
6.2 Collocation RKM for Implicit Strangeness-free DAEs	47
7 Examples	51
7.1 Semi-discrete Navier-Stokes equations	51
8 Exercises	53
8.1 I.I.C.1	53
9 Numerical Analysis and Software Overview	55
9.1 Theory: RKMs and BDF for DAEs	55
9.2 Solvers	56

9.3	Software	57
-----	--------------------	----

Preface

This is a writeup of the introduction and, maybe, some other aspects of my lectures on DAEs at the OVGU Magdeburg.

Fixes and feature requests can be submitted to the github-repo.

Chapter 1

Introduction

Differential-algebraic equations (DAEs) are coupled differential- and algebraic equations. DAEs often describe dynamical processes – here are the *differential equations* – that are subject to constraints: the *algebraic equations*.

Let's start with a few examples.

1.1 Examples

Free fall vs. the pendulum



Figure 1.1: Free fall of a point mass.

Here, the laws of the free fall – a special case of Newton's second law – applies:

force equals mass times acceleration

In 2D, the x , y coordinates of a point of mass m :

$$\begin{aligned} m\ddot{x} &= 0 \\ m\ddot{y} &= -mg \end{aligned}$$

where g is the gravity; see Figure 1.1.

The Pendulum

The same point mass attached to a string.

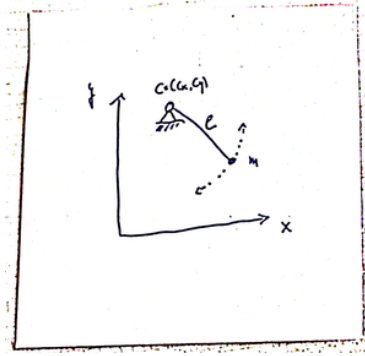


Figure 1.2: A pendulum.

Again, we have `force = mass*acceleration` but also the conditions that the mass moves on a circle:

$$(x(t) - c_x)^2 + (y(t) - c_y)^2 = l^2,$$

where (c_x, c_y) are the coordinates of the center and l is the length of the string; see Figure 1.2.

We use the Lagrangian function to derive the *Euler-Lagrange equations* of motion. For the pendulum, we have the kinetic energy

$$T = \frac{1}{2}m(\dot{x}(t)^2 + \dot{y}(t)^2),$$

the potential

$$U = mgy,$$

and the constraint

$$h = (x(t) - c_x)^2 + (y(t) - c_y)^2 - l^2.$$

Thus, with

$$L := U - T - \lambda h$$

and the requirement that

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} = 0$$

for all of the *generalized coordinates* $q = x, y, \lambda$, one obtains a system of equations:

Generalized coordinate	Equation
$q \leftarrow x$	$m\ddot{x}(t) + 2\lambda(t)(x(t) - c_x) = 0$
$q \leftarrow y$	$m\ddot{y}(t) + mgy + 2\lambda(t)(y(t) - c_y) = 0$
$q \leftarrow \lambda$	$(x(t) - c_x)^2 + (y(t) - c_y)^2 - l^2 = 0$

Example 1.1 (The Pendulum). After an order reduction via the new variables $u := \dot{x}$ and $v = \dot{y}$ the overall system reads

$$\begin{aligned} \dot{x} &= u \\ \dot{y} &= v \\ m\dot{u} &= -2\lambda(x - c_x) \\ m\dot{v} &= -2\lambda(y - c_y) - mgy \\ 0 &= (x - c_x)^2 + (y - c_y)^2 - l^2, \end{aligned} \tag{1.1}$$

where we have omitted the time dependence.

Equation (1.1) is a canonical example for a DAE with combined differential and algebraic equations.

Electrical Circuits

Another class of DAEs arises from the modelling electrical circuits. We consider the example of *charging a conductor through a resistor* as illustrated in Figure 1.3.

We formulate the problem in terms of the potentials x_1, x_2, x_3 , that are assumed to reside in the wires between a source U and a resistor R , the resistor R and the capacitor C , and the capacitor and the source.

A model for the circuit is given through the following principles and considerations.

Model principle	Equation
The source defines the difference in the neighboring potentials:	$x_1 - x_3 - U = 0$
The current I_R that is induced by the potentials neighboring the resistor is defined through <i>Ohm's law</i> :	$I_R = \frac{x_1 - x_2}{R}$

Model principle	Equation
The current I_C that is induced by the potentials neighboring the capacitor is described through:	$I_C = C(\dot{x}_3 - \dot{x}_2)$
Everywhere in the circuit the currents sum up to zero. (This is <i>Kirchhoff's law</i>):	$I_C + I_R =$ $C(\dot{x}_3 - \dot{x}_2) + \frac{x_1 - x_2}{R} = 0$
To fix the potential, one can set a ground potential – here we choose x_3 . (note that so far all equations only consider differences in the potential).	$x_3 = 0$

Example 1.2. Summing all up, the equations that model the circuit are given as

$$\begin{aligned}
 C(\dot{x}_3 - \dot{x}_2) &= -\frac{x_1 - x_2}{R} \\
 0 &= x_1 - x_3 - U \\
 0 &= x_3.
 \end{aligned} \tag{1.2}$$

Navier-Stokes Equations

The Navier-Stokes equations (NSE) are commonly used to model all kind of flows. They describe the evolution of the velocity v of the fluid and the pressure p in the fluid. Note that the flow occupies a spatial domain, say in \mathbb{R}^3 so that v and p are functions both of the time variable t and a space variable ξ :

$$v: (t, \xi) \mapsto v(t, \xi) \in \mathbb{R}^3 \quad \text{and} \quad p: (t, \xi) \mapsto p(t, \xi) \in \mathbb{R}.$$

The NSE:

$$\begin{aligned}
 \frac{\partial v}{\partial t} + (v \otimes \nabla_\xi)v - \Delta_\xi v + \nabla_\xi p &= 0, \\
 \nabla_\xi \cdot v &= 0,
 \end{aligned}$$

with \otimes denoting the outer product and ∇_ξ and Δ_ξ denoting the gradient and the *Laplace* operator. If we only count the derivatives with respect to time, as postulated in the introduction, the NSE can be seen as an (abstract) DAE.

With *dynamical systems*, we focus on the evolution of time. That's why the time derivative is relevant for defining DAEs.

Automatic Modelling or *Engineers vs. Mathematicians*

If a system, say an engine, consists of many interacting processes, it is convenient and common practice to model the dynamics of each particular process and to couple the subprocesses through interface conditions.



Figure 1.3: Electrical circuit with a source, a resistor, and a conductor.

This coupling is done through equating quantities so that the overall model will consist of dynamical equations of the subprocesses and algebraic relations at the interfaces – which makes it a DAE.

In fact, tools like `modelica` for automatic modelling of complex processes do exactly this.

The approach of *automatic modelling* is universal and convenient for engineers. However, the resulting model equations will be DAEs which, as we will see, pose particular problems in their analytical and numerical treatment.

1.2 Why are DAEs difficult to treat

Firstly, DAEs do not have the smoothing properties of ODEs, where the solution is one degree smoother than the right hand side. Secondly, the algebraic constraints are essential for the validity of the model. Thus, a numerical approximation may render the model infeasible.

Non-smooth Solutions

Example 1.3. Consider the equation

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ 0 &= x_2(t) - g(t) \end{aligned}$$

where g can be a nonsmooth function like

$$g(t) = \begin{cases} 0, & \text{if } t < 1 \\ 1, & \text{if } t \geq 1 \end{cases}$$

In this case the solution part $x_1 = \text{const.} + \int_0^t g(s)ds$ will be a smooth function and the solution part $x_2 = g$ will have jumps.

Even worse, the solution of a DAE may depend on derivatives of the right hand sides.

This observation indicates that certain difficulties will arise since

- numerical approximation schemes require smoothness of the solutions
- differentiation is numerically ill-posed unlike numerical integration

Numerical Solution Means Approximation

Imagine the equations (1.1) that describe the pendulum are solved approximately. Then, the algebraic constraint will be violated, i.e. the point mass will leave the circle and the obtained numerical solution becomes infeasible.

Thus, special care has to be taken of the algebraic constraints when the equations of motions are numerically integrated.

Chapter 2

Basic Definitions and Notions

In a very general form, a DAE can be written as

$$F(t, x(t), \dot{x}(t)) = 0 \quad (2.1)$$

with $F: \mathbb{I} \times D_x \times D_{\dot{x}} \rightarrow \mathbb{R}^m$ and with a time interval $\mathbb{I} = [t_0, t_e) \subset \mathbb{R}$ and state spaces $D_x, D_{\dot{x}} \subset \mathbb{R}^n$ and the task to find a function

$$x: \mathbb{I} \rightarrow \mathbb{R}^n$$

with time derivative $\dot{x}: \mathbb{I} \rightarrow \mathbb{R}^n$ such that (2.1) is fulfilled for all $t \in I$.

A dynamical process that evolves in time needs an initial state. Thus, one can expect a unique solution to the DAEs only if an initial value is prescribed

$$x(t_0) = x_0 \in \mathbb{R}^n. \quad (2.2)$$

The form of $F(t, x(t), \dot{x}(t))$ is a very formal way to write down a system of differential and algebraic equations. **X:** Write down the equations of the previous examples in this form – i.e. define suitable functions F , x , and \dot{x} .

2.1 Solution Concept

In order to talk of solutions, we need to define what we understand as a solution.

Definition 2.1.

1. A function $x \in \mathcal{C}^1(\mathbb{I}, \mathbb{R}^n)$ is called a *solution to the DAE* (2.1), if $F(t, x(t), \dot{x}(t)) = 0$ holds for all $t \in \mathbb{I}$.
2. A function $x \in \mathcal{C}^1(\mathbb{I}, \mathbb{R}^n)$ is called a *solution to the initial value problem* (2.1) and (2.2), if, furthermore, $x(t_0) = x_0$ holds.
3. An initial condition (2.2) is called consistent for the DAE (2.1), if there exists at least one solution as defined in 2.

Some remarks

- The requirement that $x \in \mathcal{C}^1$ could be relaxed. Compare Example 1.3, where certain components of the solution were smoother than others.
- Consistency of initial values is a major issue in the treatment of DAEs. See the pendulum...

2.2 Initial Conditions and Consistency

We consider again the equations of motions of the pendulum (Example 1.1)

$$\begin{aligned}\dot{x}(t) &= u(t) \\ \dot{y}(t) &= v(t) \\ m\dot{u}(t) &= -2\lambda(t)(x(t) - c_x) \\ m\dot{v}(t) &= -2\lambda(t)(y(t) - c_y) - mgy(t)\end{aligned}$$

with the constraint

$$0 = (x(t) - c_x)^2 + (y(t) - c_y)^2 - l^2. \quad (2.3)$$

To use this model to predict the time evolution of the system, a starting point needs to be known, say for $t = 0$. This means initial positions and initial velocities:

$$\begin{bmatrix} x(0) \\ y(0) \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} u(0) \\ v(0) \end{bmatrix} = \begin{bmatrix} u_0 \\ v_0 \end{bmatrix}.$$

The constraint (2.3) needs to be fulfilled at all times and also at $t = 0$, which gives the constraint for the initial positions:

$$(x_0 - c_x)^2 + (y_0 - c_y)^2 - l^2 = 0.$$

Moreover, if a constraint $h(x(t), y(t)) = 0$ holds for all t , then, necessarily, $\frac{d}{dt}h = 0$. For the pendulum this means that

$$2(x(t) - c_x)u(t) + 2(y(t) - c_y)v(t) = 0 \quad (2.4)$$

must hold for all t and in particular at $t = 0$ which gives constraints on the initial velocities u_0 and v_0 :

$$2(x_0 - c_x)u_0 + 2(y_0 - c_y)v_0 = 0.$$

Some remarks on consistency, constraints, and derivations:

- The so-called *consistency conditions* on (x_0, y_0, u_0, v_0) have the physical interpretation that the initial positions lie on the prescribed circle and that the velocities are tangent to this circle.
- One can show that the variable λ is completely defined in terms of x and y and their derivatives. Thus, in the formulation (1.1), both in the analysis and in the numerical treatment, there is no need for an initial value for λ . However, as we will see, DAEs can be reformulated as ODEs through differentiation and substitutions. In such an ODE formulation, a necessary initial condition for λ will have to fulfill similar consistency conditions as (x_0, y_0, u_0, v_0) .

Condition (2.4) is an example for a *hidden-constraint* – an algebraic constraint to the system that is not explicit in the original formulation. In theory, condition (2.3) can be replaced by (2.4). Moreover, through differentiation and elimination of constraints, a DAE can be brought into the form of an ODE: in the case of the circuit of Example 1.2 one only needs to replace the constraints by their derivatives:

$$\begin{aligned} C(\dot{x}_3 - \dot{x}_2) &= -\frac{x_1 - x_2}{R} \\ \dot{x}_1 - \dot{x}_3 &= \dot{U} \\ \dot{x}_3 &= 0. \end{aligned} \quad (2.5)$$

Note that (2.5) can be written as $B\dot{x} = Ax + f$ with an invertible matrix B and, thus, is an ODE.

For an ODE there is no constraint on the initial values. However, a solution to (2.5) only solves the original DAE (1.2), if the initial values are consistent with the DAE. In this case, this means $x_3(t_0) = 0$ and $x_1(t_0) - x_3(t_0) = U(t_0)$.

2.3 Additional Remarks

- It just took a single derivation to turn the circuit model into an ODE (2.5). For the *pendulum* this wouldn't be that easy.

- The extend of how much algebraic and differential parts are intertwined is measured by *indices* which is **the classifier** for DAEs.
- There are many indices. We will learn about some of the concepts. But first we will introduce some more theory.

A low index means that differential and algebraic parts are relatively well separated. (The circuit example is of *index 1*). A high index means that the structure is more involved. (The pendulum is of *index 3*).

Chapter 3

Linear DAEs with Constant Coefficients

3.1 Basic Notions and Definitions

Consider the DAE in the form

$$E\dot{x}(t) = Ax(t) + f(t), \quad (3.1)$$

where $E, A \in \mathbb{R}^{m,n}$ and $f \in \mathcal{C}(\mathbb{I}, \mathbb{R}^m)$ with, possibly, an initial condition

$$x(t_0) = x_0 \in \mathbb{R}^n. \quad (3.2)$$

For utmost generality, we consider the case that $m \neq n$, i.e. the number of equations does not meet the number of unknowns, but we will turn to the square case of $m = n$ soon.

3.1.1 Scalings and State Transformations

One can confirm that if x is a solution to (3.1) and $P \in \mathbb{R}^{n,n}$ is invertible, then x is a solution to

$$PE\dot{x}(t) = PAx(t) + Pf(t).$$

This is a scaling of the equations.

Similarly, if $Q \in \mathbb{R}^{n,n}$ is invertible, then $\tilde{x} := Q^{-1}x$ solves

$$EQ\dot{\tilde{x}}(t) = AQ\tilde{x}(t) + f(t).$$

This is a state transformation of the system.

Thus, when talking of solvability of (3.1), one may equivalently consider any regular $Q \in \mathbb{R}^{m,m}$, $P \in \mathbb{R}^{m,m}$ and the scaled and transformed system

$$\tilde{E}\dot{\tilde{x}}(t) = \tilde{A}\tilde{x}(t) + \tilde{f}(t), \quad \tilde{x}(0) = Q^{-1}x_0, \quad (3.3)$$

where $\tilde{E} = PEQ$, $\tilde{A} = PAQ$, $\tilde{f} = Pf$, and $x = Q\tilde{x}$.

To characterize all scalings and state transformations, we define these operations as relations of matrix pairs:

3.1.2 Strong Equivalence and Canonical Forms

Definition 3.1. Two pairs of matrices (E_1, A_1) and (E_2, A_2) , with $E_1, A_1, E_2, A_2 \in \mathbb{R}^{m,n}$, are called *strongly equivalent*, if there exist regular matrices $P \in \mathbb{R}^{m,m}$, $Q \in \mathbb{R}^{n,n}$ such that

$$E_2 = PE_1Q, \quad A_2 = PA_1Q.$$

In this case, we write

$$(E_1, A_1) \sim (E_2, A_2).$$

Lemma 3.1. *The relation \sim defined in Definition 3.1 defines an equivalence relation¹.*

Proof. Exercise. □

For a given equivalence relation on a set, one can define *equivalence classes* by considering all members that are equivalent to each other as basically the same. And for each class one may choose a representative, preferably in *canonical form*, i.e. a form that, e.g.,

1. comes with an simple or characteristic representation and
2. that allows for easy determination or analysis of quantities of interest.

There can be infinitely many canonical forms. For our purposes and for the *strong equivalence* of matrix pairs, we will use the *Kronecker Canonical Form*.

Theorem 3.1. *Let $E, A \in \mathbb{C}^{m,n}$. Then there exist nonsingular matrices $P \in$*

¹Equivalence relation – **RST**. Reflexive: $A \sim A$. Symmetric: $A \sim B$, then $B \sim A$. Transitive: $A \sim B$ and $B \sim C$, then $A \sim C$.

Algorithm for computations exist² but the computation is notoriously ill posed.

3.2 Regularity and Solvability

In what follows we will consider the regular case with, among others, E and A as square matrices.

Like for solving general equation systems, one can expect well posedness for the case that the numbers of equations equals the number of unknowns. Here *squareness* of the system is a necessary condition. For sufficiency one further needs that there are no redundant equations or incompatible equations. This is encoded in the *regularity*.

Definition 3.2. Let $E, A \in \mathbb{C}^{n,n}$. The matrix pair (E, A) is called *regular*, if the *characteristic polynomial* p , defined via

$$p(\lambda) = \det(\lambda E - A),$$

is not identically zero. If such a matrix pair is not regular, it is called *singular*.

Is not identically zero means that there exists a λ_0 such that $p(\lambda_0) = \det(\lambda_0 E - A) \neq 0$, i.e. $\lambda_0 E - A$ is invertible.

Since a characteristic polynomial has but a finite numbers of roots (unless it is the *zero polynomial*), *is not identically zero* means that $p(\lambda) \neq 0$ for all but a few λ .

Next we show, that *regularity* is invariant under *strong equivalence*. This is needed, since we will translate regularity of (E, A) to regularity of the associated DAE and we need to ensure that regular scalings and state transformations do not affect regularity.

Lemma 3.2. Let $E, A \in \mathbb{C}^{n,n}$. If (E, A) is strongly equivalent to a regular matrix pair, then (E, A) is regular.

Proof. Let $E_1, A_1 \in \mathbb{C}^{n,n}$ be similar to (E, A) . By definition, there exist invertible $P, Q \in \mathbb{C}^{n,n}$ such that $\lambda E - A = P(\lambda E_1 - A_1)Q$ for all λ . Thus,

$$\det(\lambda E - A) = \det P \det(\lambda E_1 - A_1) \det Q$$

is not identically zero, since $\det Q$ and $\det P$ are not zero and $\det(\lambda E_1 - A_1)$ is not the zero polynomial. \square

With that we can derive a *canonical form* for *strongly equivalent* matrix pairs.

Theorem 3.2. Let $E, A \in \mathbb{C}^{n,n}$ and (E, A) be regular. Then

$$(E, A) \sim \left(\begin{bmatrix} I_{n_1} & \\ & N \end{bmatrix}, \begin{bmatrix} J & \\ & I_{n_2} \end{bmatrix} \right), \quad (3.4)$$

²Paul v. Dooren The Computation of Kronecker's Canonical Form of a Singular Pencil

3.3. SOLUTION TO THE N-DAE, REGULARITY, AND INDEX OF A MATRIX PAIR 21

- where $n_1, n_2 \in \mathbb{N}$ such that $n = n_1 + n_2$,
- where I_{n_1} and I_{n_2} denote the identity matrices of size $n_1 \times n_1$ and $n_2 \times n_2$, respectively,
- where $J \in \mathbb{C}^{n_1, n_1}$ is in Jordan canonical form,
- and where $N \in \mathbb{C}^{n_2, n_2}$ is a nilpotent matrix.
- Moreover, it is allowed that the one or the other block is not present, i.e., n_1 or n_2 can be zero.

Proof. To be provided. □

Recall that the *Jordan canonical form* can be achieved for any square matrix $M \in \mathbb{C}$ by a similarity transformation.

Definition 3.3 (Nilpotent Matrix). A matrix $M \in \mathbb{C}^{n, n}$ is called *nilpotent*, if there is an integer k such that $M^k = 0$. The smallest such integer *index*, i.e. that $\nu \in \mathbb{N}$ such that $N^\nu = 0$ whereas $N^{\nu-1} \neq 0$ is called the *index of nilpotency* of M .

With the convention that $\mathbf{0}^0 = I$, the zero matrix $\mathbf{0} \in \mathbb{R}^{n, n}$ is of (nilpotency) index 1.

The relation of solvability and regularity of DAEs becomes evident in the canonical form of Theorem 3.2. In fact, it states that through regular scalings and state transforms, any DAE with

$$E\dot{x}(t) = Ax(t) + f(t)$$

with (E, A) regular can be transformed and split into

$$\dot{x}_1(t) = x_1(t) + f_1(t) \tag{3.5}$$

and

$$N\dot{x}_2(t) = x_2(t) + f_2(t), \tag{3.6}$$

i.e

- into an **ODE** (3.5) that already is in *Jordan canonical form*
- and a separated(!) **DAE** (3.6) of a particular type.

Since linear ODEs always have a unique solution for any initial value, solvability of a general linear DAE with constant, regular coefficients will be completely defined by solvability of the special DAE part (3.6).

3.3 Solution to the *N-DAE*, Regularity, and Index of a Matrix Pair

In what follows we will consider the special DAE

$$N\dot{x}(t) = x(t) + f(t) \tag{3.7}$$

with $N \in \mathbb{R}^{n,n}$ nilpotent with ν being the index of nilpotency. For this DAE there is an explicit solution formula:

Lemma 3.3. *Consider (3.7). If $f \in \mathcal{C}^\nu(\mathcal{I}, \mathbb{R}^n)$, $n \geq 1$, where ν is the index of nilpotency of N , then (3.7) has a unique solution given as*

$$x(t) = - \sum_{i=0}^{\nu-1} N^i f^{(i)}(t), \quad (3.8)$$

where $f^{(i)}$ denotes the i -th derivative of f .

Proof. There are a few ways to prove the explicit form (3.8)

1. Bring N into Jordan canonical form and prove the formula for the Jordan blocks of arbitrary size by induction.
2. Write (3.7) as

$$(N \frac{d}{dt} - I)x = f$$

and show that³

$$(N \frac{d}{dt} - I)^{-1} = - \sum_{i=0}^{\nu-1} N^i \frac{d^i}{dt^i}$$

3. We take the direct approach as it can be found in the book by Dai⁴:

Firstly, we observe that

$$x = N\dot{x} - f.$$

Secondly, that (having multiplied by N and differentiated once)

$$N\dot{x} = N^2\ddot{x} - N\dot{f}.$$

And, finally, that (having multiplied k -times by N and differentiated k -times)

$$N^k \dot{x}^{(k)} = N^{k+1} \dot{x}^{(k+1)} - N^k \dot{f}^{(k)}.$$

If one successively replaces $N^k x^{(k)} = N^{k+1} x^{(k+1)} - N^k f^{(k)}$, $k = 1, 2, \dots, \nu - 1$ in

$$x = N\dot{x} - f = N^2\ddot{x} - N\dot{f} - f = \dots = N^\nu x^{(\nu)} - \sum_{i=0}^{\nu-1} N^i f^{(i)},$$

with $N^\nu = 0$, one arrives at formula (3.8).

Since this construction holds for any solution, uniqueness is guaranteed too. \square

[We make three important observations here.](#)

³See the proof of Lemma 2.8 in *Kunkel/Mehrmann*

⁴Dai (1989): *Singular Control Systems*

3.3. SOLUTION TO THE N-DAE, REGULARITY, AND INDEX OF A MATRIX PAIR 23

1. The solution x to (3.7) is uniquely defined without specifying a value at t_0 . Vice versa, an initial value x_0 is consistent if, and only if,

$$x_0 = - \sum_{i=0}^{\nu-1} N^i f^{(i)}(t_0)$$

2. The definition of the solution x requires f to be $\nu - 1$ -times differentiable. In order to be a solution according to Definition 2.1, the function x itself has to be differentiable too. Hence the requirement $f \in \mathcal{C}^\nu(\mathcal{I}, \mathbb{R}^n)$.
3. The index of nilpotency of N defines the necessary smoothness of the right hand side.

The index of nilpotency in the N part of the Weierstrass canonical form is characteristic for a matrix pair and defines an *index* of the associated DAE.

Definition 3.4. Consider a regular matrix pair (E, A) and its Weierstrass canonical form, i.e.

$$(E, A) \sim \left(\begin{bmatrix} I_{n_1} & \\ & N \end{bmatrix}, \begin{bmatrix} J & \\ & I_{n_2} \end{bmatrix} \right).$$

The index ν of nilpotency of N is called the index of the matrix pair (E, A) and we write $\nu = \text{ind}(E, A)$. If N is not present, then we set $\text{ind}(E, A) = 0$.

Furthermore, for a nilpotent matrix N , we will occasionally use the notion $\nu = \text{ind}(N, I)$ to refer to its index of nilpotency. **X:** Is this OK, i.e. consistent?

To be on the safe side and to learn how to handle block structured matrices in the analysis, we confirm that this definition of the index is well-posed, i.e. for any regular matrix (E, A) there is a unique index ν .

If one considers the Weierstrass canonical form as a special case of the Kronecker canonical form, then the statement that *the canonical form is well-posed up to permutations in the order of the blocks* already implies that the index is well defined (since it only depends on the size of the largest nilpotent block but not in which order it appears).

Lemma 3.4. Suppose that the regular matrix pair (E, A) , $E, A \in \mathbb{R}^{n,n}$ has the two canonical forms

$$(E, A) \sim \left(\begin{bmatrix} I_{d_1} & \\ & N_1 \end{bmatrix}, \begin{bmatrix} J_1 & \\ & I_{n-d_1} \end{bmatrix} \right) \sim \left(\begin{bmatrix} I_{d_2} & \\ & N_2 \end{bmatrix}, \begin{bmatrix} J_2 & \\ & I_{n-d_2} \end{bmatrix} \right),$$

where d_1, d_2 are the size of the Jordan blocks J_1, J_2 , respectively. Then $d_1 = d_2 =: d$ and the indices of nilpotency of the nilpotent blocks coincide, i.e. $\text{ind}(N_1, I_{n-d}) = \text{ind}(N_2, I_{n-d})$.

Proof. To show that $d_1 = d_2$, without loss of generality, we can assume that N_i are in Jordan form too. This, in particular, means that they are upper triangular with zeros on the diagonal. (If this was not the case, we can use $N_i = T^{-1}\tilde{N}_i T$ in the arguments below).

Then, we have that the characteristic polynomials

$$\det(\lambda E - A) = \det(\lambda N_i - I_{n-d_i}) \det(\lambda I_{d_i} - J_i) = (-1)^{n-d_i} \det(\lambda I_{d_i} - J_i)$$

are polynomials of degree d_i , $i = 1, 2$.

Since the characteristic polynomials of strongly equivalent matrix pairs only differ by a constant factor (see the proof of Lemma 3.2), this implies that $d_1 = d_2$.

Now we show, that the indices of nilpotency of N_1 and N_2 coincide. If the N -blocks weren't present, there would be nothing to show. So let's assume that they are there.

Let now be $P, Q \in \mathbb{R}^{n,n}$ that invertible matrices that realize the strong equivalence of the canonical forms, i.e.

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} I & \\ & N_2 \end{bmatrix} = \begin{bmatrix} I & \\ & N_1 \end{bmatrix} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \quad (3.9)$$

and

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} J_2 & \\ & I \end{bmatrix} = \begin{bmatrix} J_1 & \\ & I \end{bmatrix} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}, \quad (3.10)$$

where P and Q have been partitioned in line with the canonical forms.

Taking the blockwise matrix product and equating the blocks separately, the following relations are obtained:

Block:	(1,1)	(1,2)	(2,1)	(2,2)
(3.9):	$P_{11} = Q_{11}$	$P_{12}N_2 = Q_{12}$	$P_{21} = N_1Q_{21}$	$P_{22}N_2 = N_1Q_{22}$
(3.10):	$P_{11}J_2 = J_1Q_{11}$	$P_{12} = J_1Q_{12}$	$P_{21}J_2 = Q_{21}$	$P_{22} = Q_{22}$

If we combine the **(2,1)** blocks, we obtain that

$$P_{21} = N_1Q_{21} = N_1P_{21}J_2 = N_1^2P_{21}J_2^2 = N_1^3P_{21}J_2^3 = \dots = 0$$

since N_1 is nilpotent.

Since P is invertible and, because of $P_{21} = 0$, block upper triangular, the blocks on the diagonals P_{11} and P_{22} must be invertible. With $Q_{22} = P_{22}$ the **(2,2)** block of (3.9) implies that

$$N_2 = P_{22}^{-1}N_1P_{22}$$

and, further,

$$N_2^k = P_{22}^{-1} N_1^k P_{22}$$

for all $k \in \mathbb{N}$. So that a power $N_2^k = 0$ if, and only if, $N_1^k = 0$. Consequently, the indices of nilpotency of N_1 and N_2 coincide. \square

3.4 Existence of Solutions

We can now summarize all results and considerations in a theorem.

Theorem 3.3. *Consider the DAE (3.1) with initial condition (3.2),*

$$E\dot{x}(t) = Ax(t) + f(t), \quad x(t_0) = x_0 \in \mathbb{R}^n.$$

Let the pair (E, A) be regular and consider the strongly equivalent DAE

$$\tilde{E}\dot{\tilde{x}}(t) = \tilde{A}\tilde{x}(t) + \tilde{f}(t), \quad \tilde{x}(t_0) = \tilde{x}_0 \in \mathbb{R}^n.$$

with (\tilde{E}, \tilde{A}) in Weierstrass canonical form, i.e.

$$\tilde{E} = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}, \quad \tilde{A} = \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix},$$

and consider the conforming splitting of the transformed variables

$$\tilde{x} = \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix}, \quad \tilde{x}_0 = \begin{bmatrix} \tilde{x}_{0,1} \\ \tilde{x}_{0,2} \end{bmatrix}.$$

Furthermore, let ν be the index of the matrix pair (E, A) .

If $f \in \mathcal{C}^\nu(J, \mathbb{C}^n)$, then

1. *The differential algebraic equation (3.1) is solvable.*
2. *The initial condition x_0 in (3.2) is consistent if, and only if, for the transformed initial condition \tilde{x}_0 it holds that*

$$\tilde{x}_{2,0} = - \sum_{i=0}^{\nu-1} N^i \tilde{f}^{(i)}(t_0)$$

In particular, a consistent initial condition to (3.1) exists.

3. *Every initial value problem (3.1)–(3.2) with a consistent initial condition is uniquely solvable.*

Proof. A summary of the preceding results. \square

From Theorem 3.3 it follows that the regularity of the matrix pair (E, A) implies the existence of a unique solution to the DAE (3.1) with an initial condition (3.2) provided that the initial condition is consistent.

The negation of this statement is a bit diffuse because there are several things that *can go wrong* if the DAE is not regular. Depending on the irregularity there might be infinite many solutions to the initial value problem or no solutions at all to the DAE (even without the initial condition).

The following theorem covers the case of singular or non-square matrix pairs.

Theorem 3.4 (Thm. 2.14 in *Kunkel/Mehrmann*). *Let $E, A \in \mathbb{C}^{m,n}$.*

1. *If $\text{rank}(\lambda E - A) < n$ for all $\lambda \in \mathbb{C}$, then the homogeneous initial value problem*

$$E\dot{x}(t) = Ax(t), \quad x(t_0) = 0,$$

has a nontrivial solution.

2. *If $\text{rank}(\lambda E - A) < m$ for all $\lambda \in \mathbb{C}$, then there exist arbitrarily smooth inhomogeneities f for which the DAE (3.1)*

$$E\dot{x}(t) = Ax(t) + f(t),$$

is not solvable.

Proof. The proof is given for Theorem 2.14 in *Kunkel/Mehrmann* with, however, the second claim being formulated slightly differently. To reduce our formulation to *theirs*, one may identify that columns of $\lambda E - A$ that achieve the maximal rank and split off the redundant columns, e.g., the parts of x associated with them. \square

Note that a nontrivial solution to the *homogeneous* problem, means that existence of **a** solution implies **infinitely many** solutions to the initial value problem. In fact, if x_h is the solution to the homogeneous problem and x_p a solution⁵ to the initial value problem, then $x = x_p + \alpha x_h$ solves the initial value problem for any $\alpha \in \mathbb{R}$.

To illustrate the difficulties with singular or non-square DAEs, consider the example

$$\begin{bmatrix} 0 & 1 \\ & 1 \\ & 0 \end{bmatrix} \frac{d}{dt} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ & 0 \\ & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ f_2 \\ f_3 \end{bmatrix}$$

Here, the first part reads

$$x_1 = \dot{x}_2$$

which defines a solution $x_2 = g$ for any $g \in \mathcal{C}^2$ and a nontrivial solution to the associated *homogeneous* initial value problem if only $g(t) \neq 0$ for some t but $g(t_0) = \dot{g}(t_0) = 0$.

⁵A so-called particular solution. Btw., *all solutions = a particular plus all solutions to the homogeneous problem* is the superposition principle for general linear problems.

The second part reads

$$\begin{aligned}\dot{x}_3 &= f_2 \\ -x_3 &= f_3\end{aligned}$$

which does **not** permit a solution, if $\dot{f}_3 \neq f_2$.

X: find such a g for the first part.

3.5 A Variation of Constant Formula

In this section, we want to derive an explicit solution formula for linear DAEs with constant coefficients $E\dot{x}(t) = Ax(t) + f(t)$ similar to the formula that exist for linear time-invariant (i.e. with constant coefficients) ODEs.

Certainly, a solution formula is given through the transformation to the Weierstrass canonical form and through (3.8). This however is not an explicit solution representation in so far as both the coefficients and the solution x itself had to be undertaken a transformation first.

The outline for deriving the explicit formula

The following derivations and arguments base on two major components

1. Representation of solutions as $x = x_h + x_p$, where
 - x_h describes all solutions to the *homogeneous* problem $E\dot{x} = Ax$ and
 - x_p denotes a (so-called *particular*) solution to $E\dot{x} = Ax + f$.
2. Additive splitting of DAEs into nilpotent and almost ODE parts, in the way that x solves $E\dot{x} = Ax + f$, if, and only if, $x = x_1 + x_2$ where x_1 and x_2 fulfill
 - $\tilde{C}\dot{x}_1 = Ax_1 + f_1$, with \tilde{C} *almost invertible*
 - $\tilde{N}\dot{x}_2 = Ax_2 + f_2$, with \tilde{N} nilpotent.

We start with defining what was meant by *almost invertible*. For that consider the differential equation

$$E\dot{x}(t) = x(t), \quad x(t_0) = x_0$$

with $E \in \mathbb{C}^{n,n}$.

If E was invertible, then the unique solution to (3.5) would be given as

$$x(t) = e^{(t-t_0)E^{-1}}x_0.$$

If E is not invertible, then there exists a matrix $T \in \mathbb{R}^{n,n}$, invertible, that brings E into Jordan canonical form

$$J = TET^{-1} = \begin{bmatrix} C & \\ & N \end{bmatrix}$$

with C invertible and N nilpotent so that (3.5) can be written

$$\begin{bmatrix} C & \\ & N \end{bmatrix} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} I & \\ & I \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}, \quad \begin{bmatrix} x_1(t_0) \\ x_2(t_0) \end{bmatrix} = \begin{bmatrix} x_{1,0} \\ x_{2,0} \end{bmatrix} = Tx_0.$$

Since there is no right hand side, by the formula (3.8) for the special DAE $N\dot{x} = x + f$, we conclude that $x_2 = 0$, so that only the ODE part $C\dot{x}_1 = x_1$ remains and the overall solution writes

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} e^{(t-t_0)C^{-1}} & \\ & 0 \end{bmatrix} \begin{bmatrix} x_{1,0} \\ x_{2,0} \end{bmatrix}.$$

If we define

$$J^D = \begin{bmatrix} C & \\ & N \end{bmatrix}^D := \begin{bmatrix} C^{-1} & \\ & 0 \end{bmatrix},$$

the solution to (3.5) can be expressed as

$$Tx(t) = e^{(t-t_0)J^D}Tx_0$$

or with

$$E^D := TJ^DT^{-1}, \quad e^{(t-t_0)TJ^DT^{-1}} = Te^{(t-t_0)J^D}T^{-1}$$

as

$$x(t) = e^{(t-t_0)E^D}x_0.$$

A few observations

1. The formula looks like the solution formula for the ODE case.
2. In fact, if E is invertible, then $E^D = E^{-1}$ and the formulas coincide.
3. Thus, E^D is a generalized inverse – the so-called *Drazin* inverse.

Definition 3.5. Let $E \in \mathbb{C}^{n,n}$ and $\nu = \text{ind}(E)$. A matrix $X \in \mathbb{C}^{n,n}$ that fulfills

$$EX = XE, \tag{3.11}$$

$$XEX = X, \tag{3.12}$$

$$XE^{\nu+1} = E^\nu, \tag{3.13}$$

is called a *Drazin inverse* of E .

With the following theorem we confirm that a Drazin inverse to a matrix E is unique so that we can write E^D for it.

Theorem 3.5. Every matrix $E \in \mathbb{C}^{n,n}$ has one, and only one, Drazin inverse.

Proof. Uniqueness: Let X_1 and X_2 be two Drazin inverses of E . Then by repeated application of the identities in (3.11)–(3.13) one derives that

$$\begin{aligned} X_1 E X_1 E X_2 &= X_1 E X_2 = X_1 E X_2 E X_2 \\ X_1^2 E^2 X_2 &= \dots = X_1 E X_2 = \dots = X_1 E^2 X_2^2 \\ X_1^{\nu+1} E^{\nu+1} X_2 &= \dots = \dots = X_1 E X_2 = \dots = \dots = X_1 E^{\nu+1} X_2^{\nu+1} \\ X_1^{\nu+1} E^{\nu+1} X_1 &= \dots = \dots = \dots = X_1 E X_2 = \dots = \dots = \dots = X_2 E^{\nu+1} X_2^{\nu+1} \\ X_1 &= \dots = \dots = \dots = \dots = X_1 E X_2 = \dots = \dots = \dots = \dots = X_2, \end{aligned}$$

where in the second last step we used the identities

$$E^{\nu+1} X_1 = X_1 E^{\nu+1} = E^\nu = X_2 E^{\nu+1} = E^{\nu+1} X_2.$$

□

Lemma 3.5. *Let $E, A \in \mathbb{C}^{n,n}$ commuting, i.e. $EA = AE$. Then also*

$$EA^D = A^D E. \quad (3.14)$$

Theorem 3.6. *Let $E \in \mathbb{C}^{n,n}$ with $\nu = \text{ind } E$. Then there exists a unique decomposition*

$$E = \tilde{C} + \tilde{N}$$

with the properties

$$\tilde{C}\tilde{N} = \tilde{N}\tilde{C} = 0, \quad (3.15)$$

$$\tilde{N}^\nu = 0, \quad \tilde{N}^{\nu-1} \neq 0, \quad (3.16)$$

$$\text{ind } \tilde{C} \leq 1, \quad (3.17)$$

and, in particular

$$\tilde{C} = EE^D E, \quad \tilde{N} = E(I - E^D E). \quad (3.18)$$

Proof. 1. Show that such a decomposition with the properties (3.15)–(3.17) also fulfills (3.18), i.e. existence.

2. Show that \tilde{C}, \tilde{N} as in (3.18) are such a decomposition, i.e. uniqueness.

□

Now we can define, how the general DAE can be split *additively* into an *almost* ODE and a particular nilpotent DAE.

Lemma 3.6. *Let $E, A \in \mathbb{C}^{n,n}$ and $f: J \rightarrow \mathbb{C}^n$. If E and A commute, then the system*

$$E\dot{x}(t) = Ax(t) + f(t)$$

is equivalent – in the sense that solutions correspond one-to-one via $x = x_1 + x_2$ – to the system

$$\begin{aligned}\tilde{C}\dot{x}_1(t) &= Ax_1(t) + f_1(t), \\ \tilde{N}\dot{x}_2(t) &= Ax_2(t) + f_2(t),\end{aligned}$$

where

$$x_1 = E^D Ex, \quad x_2 = (I - E^D E)x, \quad (3.19)$$

where

$$f_1 = E^D Ef, \quad f_2 = (I - E^D E)f, \quad (3.20)$$

and where

$$\tilde{N} + \tilde{C} = E$$

are a decomposition as in Theorem 3.6.

This decomposition is used to characterize **all** solutions to the homogeneous problem $E\dot{x} = Ax$.

Theorem 3.7. *Let $E, A \in \mathbb{C}^{n,n}$ commuting, i.e. $EA = AE$, and let (E, A) be regular. Then every solution $x \in \mathcal{C}^1(J, \mathbb{C}^n)$ of $E\dot{x} = Ax$ has the form*

$$x(t) = e^{E^D At} E^D Ev \quad (3.21)$$

for some $v \in \mathbb{C}^n$.

Proof. 1. Confirm directly that $x: t \mapsto e^{E^D At} E^D Ev$ satisfies $E\dot{x} - Ax = 0$. Note that regularity of (E, A) is not needed here.

2. Use regularity and the splitting to show that any solution has this form.

□

It remains to find a particular solution.

Theorem 3.8. *Let $E, A \in \mathbb{C}^{n,n}$ commuting, i.e. $EA = AE$, and let (E, A) be regular. Let $f \in \mathcal{C}^\nu(J, \mathbb{C}^n)$ with $\nu = \text{ind } E$. Then $x \in \mathcal{C}^1(J, \mathbb{C}^n)$ defined by*

$$x(t) = \int_{t_0}^t e^{E^D A(t-s)} E^D f(s) ds - (I - E^D E) \sum_{i=0}^{\nu-1} (EA^D)^i A^D f^{(i)}(t)$$

is a particular solution of $E\dot{x}(t) = Ax(t) + f(t)$.

Theorem 3.9. *Let $E, A \in \mathbb{C}^{n,n}$ commuting, i.e. $EA = AE$, and let (E, A) be regular. Let $f \in \mathcal{C}^\nu(\mathcal{J}, \mathbb{C}^n)$ with $\nu = \text{ind } E$. Then every solution $x \in \mathcal{C}^1(\mathcal{J}, \mathbb{C}^n)$ to $E\dot{x}(t) = Ax(t) + f(t)$ has the representation*

$$x(t) = e^{(t-t_0)E^D A} E^D E v + \int_{t_0}^t e^{E^D A(t-s)} E^D f(s) ds - (I - E^D E) \sum_{i=0}^{\nu-1} (EA^D)^i A^D f^{(i)}(t)$$

for some $v \in \mathbb{C}^n$.

This theorem also defines what is a consistent initial value.

Corollary 3.1. *Let the assumptions of Theorem 3.9 hold. The initial value problem (3.1)–(3.2),*

$$E\dot{x}(t) = Ax(t) + f(t), \quad x(t_0) = x_0,$$

possesses a solution if, and only if, there exists a $v \in \mathbb{C}^n$ such that

$$x_0 = E^D E v - (I - E^D E) \sum_{i=0}^{\nu-1} (EA^D)^i A^D f^{(i)}(t_0).$$

If this is the case, then the solution is unique.

We have derived the solution formula under the assumption that $EA = AE$ which is hardly ever the case.

The following lemma states that the assumption of commutativity is not a restriction for regular matrix pairs.

Lemma 3.7. *Let (E, A) be regular and let $\tilde{\lambda}$ be such that $\tilde{\lambda}E - A$ is invertible. Then*

$$(\tilde{E}, \tilde{A}) := ((\tilde{\lambda}E - A)^{-1}E, (\tilde{\lambda}E - A)^{-1}A) \sim (E, A)$$

and \tilde{E} and \tilde{A} commute.

Chapter 4

Linear DAEs with Time-varying Coefficients

In this section, we consider linear DAEs with *variable* or *time-dependent* coefficients. This means, for matrix-valued functions

$$E \in \mathcal{C}(\mathcal{J}, \mathbb{C}^{m,n}), \quad A \in \mathcal{C}(\mathcal{J}, \mathbb{C}^{m,n})$$

and $f \in \mathcal{C}(\mathcal{J}, \mathbb{C}^m)$, we consider the DAE

$$E(t)\dot{x}(t) = A(t)x(t) + f(t) \quad (4.1)$$

with, possibly, an initial condition

$$x(t_0) = x_0 \in \mathbb{C}^n. \quad (4.2)$$

The same general solution concept applies. Basically x should be differentiable, fulfill the DAE, and, if stated, the initial condition too.

In the constant coefficient case, regularity played a decisive role for the existence and uniqueness of solutions; see, e.g. Section 3.4. Thus it seems natural to extend this concept to the time-varying case, e.g., through requiring that $(E(t), A(t))$ is a regular matrix pair independent of t . However, the following two examples show that this will not work *out of the box*.

Example 4.1. Let E, A be given as

$$E(t) = \begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix}, \quad A(t) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

Then

$$\det(\lambda E(t) - A(t)) = (1 - \lambda t)(1 + \lambda t) + \lambda^2 t^2 \equiv 1,$$

for all $t \in \mathcal{I}$. Still, for every $c \in \mathcal{C}^1(\mathcal{I}, \mathbb{C})$ with $c(t_0) = 0$, the function

$$x: t \mapsto c(t) \begin{bmatrix} t \\ 1 \end{bmatrix}$$

solves the *homogeneous* initial value problem (4.1)–(4.2).

This is an example where the pair (E, A) is regular uniformly with respect to t but still allows for infinitely many solutions to the associated DAE.

Theorem 4.1. *Let $E, A \in \mathbb{C}^{m,n}$ and let*

$$T, Z, T', V \tag{4.3}$$

be

Matrix	as the basis of
T	kernel E
Z	corange $E = \text{kernel } E^H$
T'	cokernel $E = \text{range } E^H$
V	corange($Z^H A T$)

then the quantities

$$r, a, s, d, u, v \tag{4.4}$$

defined as

Quantity	Definition	Name
r	$\text{rank } E$	rank
a	$\text{rank}(Z^H A T)$	algebraic part
s	$\text{rank}(V^H Z^H A T')$	strangeness
d	$r - s$	differential part
u	$n - r - a$	undetermined variables
v	$m - r - a - s$	vanishing equations

are invariant under local equivalence transformations and (E, A) is locally equivalent to the canonical form

$$\left(\begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right), \tag{4.5}$$

where all diagonal blocks are square, except maybe the last one.

Some remarks on the spaces and how the names are derived for the case $E\dot{x} = Ax + f$ with constant coefficients. The ideas are readily transferred to the case with time-varying coefficients.

Let

$$x(t) = Ty(t) + T'y'(t),$$

where y denotes the components of x that evolve in the range of T and y' the respective complement. (Since $[T|T']$ is a basis of \mathbb{C}^n , there exist such y and y' that uniquely define x and vice versa). With T spanning $\ker E$ we find that

$$E\dot{x}(t) = ET\dot{y}(t) + ET'y'(t)$$

so that the DAE basically reads

$$ET'y'(t) = ATy(t) + AT'y'(t) + f,$$

i.e. the components of x defined through y are, effectively, not differentiated. With Z containing exactly those v , for which $v^H E = 0$, it follows that

$$Z^H ET'y'(t) = 0 = Z^H ATy(t) + Z^H AT'y'(t) + Z^H f,$$

or

$$Z^H ATy(t) = -Z^H AT'y'(t) - Z^H f,$$

so that $\text{rank } Z^H AT$ indeed describes the number of purely algebraic equations and variables in the sense that it defines parts of y (which is never going to be differentiated) in terms of algebraic relations (no time derivatives are involved).

With the same arguments and with $V = \text{corange } Z^H AT$, it follows that

$$V^H Z^H AT'y'(t) = -V^H Z^H ATy(t) - V^H Z^H f = -V^H Z^H f,$$

is the part of $E\dot{x} = Ax + f$ in which those components y' that are also differentiated are algebraically equated to a right-hand side. This is the *strangeness* (rather in the sense of *skewness*) of DAEs that variables can be both differential and algebraic. Accordingly, $\text{rank } V^H Z^H AT'$ describes the size of the skewness component.

Outlook: If there is no strangeness, the DAE is called strangeness-free. Strangeness can be eliminated through iterated differentiation and substitution. The needed number of such iterations (that is independent of the size s of the *strange* block here) will define the strangeness index.

Example 4.2. With basic scalings and state transforms, one finds for the coefficients of Example 1.2 that:

$$(E, A) \rightsquigarrow \left(\begin{bmatrix} I_2 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & I_1 \end{bmatrix} \right).$$

We compute the subspaces as defined in (4.3):

Matrix	as the basis of/computed as
$T = \begin{bmatrix} 0 \\ I_1 \end{bmatrix}$	$\text{kernel} \begin{bmatrix} I_2 & 0 \\ 0 & 0 \end{bmatrix}$
$Z = \begin{bmatrix} 0 \\ I_1 \end{bmatrix}$	$\text{corange} \begin{bmatrix} I_2 & 0 \\ 0 & 0 \end{bmatrix} = \text{kernel} \begin{bmatrix} I_2 & 0 \\ 0 & 0 \end{bmatrix}^H$
$T' = \begin{bmatrix} I_2 \\ 0 \end{bmatrix}$	$\text{cokernel} \begin{bmatrix} I_2 & 0 \\ 0 & 0 \end{bmatrix} = \text{range} \begin{bmatrix} I_2 & 0 \\ 0 & 0 \end{bmatrix}^H$
$Z^H A T = I_1$	$\begin{bmatrix} 0 \\ I_1 \end{bmatrix}^H \begin{bmatrix} 0 & 0 \\ 0 & I_1 \end{bmatrix} \begin{bmatrix} 0 \\ I_1 \end{bmatrix}$
$V = 0$	$\text{corange}(Z^H A T) = \text{kernel } I_1^H$
$Z^H A T' = 0_{2 \times 1}$	$\begin{bmatrix} 0 \\ I_1 \end{bmatrix}^H \begin{bmatrix} 0 & 0 \\ 0 & I_1 \end{bmatrix} \begin{bmatrix} I_2 \\ 0 \end{bmatrix}$

and derive the quantities as defined in (4.4):

Name	Value	Derived from
rank	$r = 2$	$\text{rank } E = \text{rank} \begin{bmatrix} I_2 & 0 \\ 0 & 0 \end{bmatrix}$
algebraic part	$a = 1$	$\text{rank } Z^H A T = \text{rank } I_1$
strangeness	$s = 0$	$\text{rank } V^H Z^H A T' = \text{rank } 0_{2 \times 1}$
differential part	$d = 2$	$d = r - s = 2 - 0$
undetermined variables	$u = 0$	$u = n - r - a = 3 - 2 - 1$
vanishing equations	$v = 0$	$v = m - r - a - s = 3 - 2 - 1 - 0$

Example 4.3. With more involved scalings and state transforms, one finds for

the coefficients of the linearized and spatially discretized Navier-Stokes equations (see Exercise I) that:

$$(\mathcal{E}, \mathcal{A}) = \left(\begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A & B^H \\ B & 0 \end{bmatrix} \right) \sim \left(\begin{bmatrix} I_{n_1} & 0 & 0 \\ 0 & I_{n_2} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} & I_{n_1} \\ A_{21} & A_{22} & 0 \\ I_{n_1} & 0 & 0 \end{bmatrix} \right).$$

We compute the subspaces as defined in (4.3):

Matrix	as the basis of/computed as
$T = \begin{bmatrix} 0 \\ 0 \\ I_{n_1} \end{bmatrix}$	kernel $\begin{bmatrix} I_{n_1} & 0 & 0 \\ 0 & I_{n_2} & 0 \\ 0 & 0 & 0 \end{bmatrix}$
$Z = \begin{bmatrix} 0 \\ 0 \\ I_{n_1} \end{bmatrix}$	corange $\begin{bmatrix} I_{n_1} & 0 & 0 \\ 0 & I_{n_2} & 0 \\ 0 & 0 & 0 \end{bmatrix}$
$T' = \begin{bmatrix} I_{n_1} & 0 \\ 0 & I_{n_2} \\ 0 & 0 \end{bmatrix}$	cokernel $\begin{bmatrix} I_{n_1} & 0 & 0 \\ 0 & I_{n_2} & 0 \\ 0 & 0 & 0 \end{bmatrix}$
$Z^H AT = 0_{n_1}$	$\begin{bmatrix} 0 \\ 0 \\ I_{n_1} \end{bmatrix}^H \begin{bmatrix} A_{11} & A_{12} & I_{n_1} \\ A_{21} & A_{22} & 0 \\ I_{n_1} & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ I_{n_1} \end{bmatrix}$
$V = I_{n_1}$	corange($Z^H AT$) = kernel $0_{n_1}^H$
$Z^H AT' = \begin{bmatrix} I_{n_1} & 0_{n_1 \times n_2} \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ I_{n_1} \end{bmatrix}^H \begin{bmatrix} A_{11} & A_{12} & I_{n_1} \\ A_{21} & A_{22} & 0 \\ I_{n_1} & 0 & 0 \end{bmatrix} \begin{bmatrix} I_{n_1} & 0 \\ 0 & I_{n_2} \\ 0 & 0 \end{bmatrix}$

and derive the quantities as defined in (4.4):

Name	Value	Derived from
rank	$r = n_1 + n_2$	$\text{rank } E = \text{rank} \begin{bmatrix} I_{n_1} & 0 & 0 \\ 0 & I_{n_2} & 0 \\ 0 & 0 & 0 \end{bmatrix}$
algebraic part	$a = 0$	$\text{rank } Z^H AT = \text{rank } 0_{n_1}$
strangeness	$s = n_1$	$\text{rank } V^H Z^H AT' = \text{rank} \begin{bmatrix} I_{n_1} & 0_{n_1 \times n_2} \end{bmatrix}$
differential part	$d = n_2$	$d = r - s = (n_1 + n_2) - n_1$
undetermined variables	$u = n_1$	$u = n - r - a = (n_1 + n_2 + n_1) - (n_1 + n_2) - 0$
vanishing equations	$v = 0$	$v = m - r - a - s = (n_1 + n_2 + n_1) - (n_1 + n_2) - n_1$

Theorem 4.2 (see Kunkel/Mehrmann, Thm. 3.9). *Let $E \in \mathcal{C}^l(I, \mathbb{C}^{m,n})$ with $\text{rank } E(t) = r$ for all $t \in I$. Then there exist smooth and pointwise unitary (and, thus, nonsingular) matrix functions U and V , such that*

$$U^H E V = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}$$

with pointwise nonsingular $\Sigma \in \mathcal{C}^l(I, \mathbb{C}^{r,r})$.

Theorem 4.3. *Let $E, A \in \mathcal{C}^l(I, \mathbb{C}^{m,n})$ be sufficiently smooth and suppose that*

$$r(t) = r, \quad a(t) = a, \quad s(t) = s \quad (4.6)$$

for the local characteristic values of $(E(t), A(t))$. Then (E, A) is globally equivalent to the canonical form

$$\left(\begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} \\ 0 & 0 & 0 & A_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right). \quad (4.7)$$

All entries are again matrix functions on I and the last block column in both matrix functions of (4.7) has size $u = n - s - d - a$.

Proof. In what follows, we will tacitly redefine the block matrix entries that appear after the global equivalence transformations. The first step is the continuous

SVD of E ; see Theorem 4.3.

$$\begin{aligned}
(E, A) &\sim \left(\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12}V_1 \\ U_1^H A_{21} & U_1^H A_{22}V_1 \end{bmatrix} - \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \dot{V}_1 \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} I_r & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & I_a & 0 \\ A_{31} & 0 & 0 \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} V_2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11}V_2 & A_{12} & A_{13} \\ A_{21}V_2 & I_a & 0 \\ U_2^H A_{31}V_2 & 0 & 0 \end{bmatrix} - \begin{bmatrix} \dot{I}_r & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{V}_2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} V_{11} & V_{12} & 0 & 0 \\ V_{21} & V_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & A_{13} & A_{14} \\ 0 & A_{22} & A_{23} & A_{24} \\ 0 & A_{32} & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & -A_{32} & I_a & 0 \\ I_s & 0 & 0 & I_a \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & A_{13} & A_{14} \\ 0 & A_{22} & A_{23} & A_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} \\ 0 & A_{22} & 0 & A_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & Q_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12}Q_2 & 0 & A_{14} \\ 0 & A_{22}Q_2 - \dot{Q}_2 & 0 & A_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} \\ 0 & 0 & 0 & A_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right),
\end{aligned}$$

where the final equivalence holds, if Q_2 is chosen as the (unique and pointwise invertible) solution of the linear matrix valued ODE

$$\dot{Q}_2 = A_{22}(t)Q_2, \quad Q_2(t_0) = I_d.$$

Then, A_{22} vanishes because of the special choice of Q_2 and E_{22} becomes I_d after scaling the second block line by Q_2^{-1} . \square

Chapter 5

Numerical Approximation of DAEs

In order to analyse the approximation error of the RKM $(\mathcal{A}, \beta, \gamma)$ applied to a regular linear DAE with constant coefficients

$$E\dot{x} = Ax + f(t).$$

Without loss of generality, we can assume that

- (E, A) is in Kronecker Canonical Form \leftarrow RKM are invariant under equivalence transformation
- $(E, A) = (N, I) \leftarrow$ the *regular part* can be treated by ODE theory
- $E = N = N_\nu$ consists of a single Jordan block \leftarrow otherwise consider each Jordan block separately

Thus, we can consider the special DAE

$$\begin{bmatrix} 0 & 1 & & \\ & 0 & 1 & \\ & & \ddots & \ddots \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix} \dot{x} = x + f(t), \quad (5.1)$$

where

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_\nu(t) \end{bmatrix} \quad \text{and} \quad f(t) = \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_\nu(t) \end{bmatrix}$$

Theorem 5.1. *The local error of an RKM with \mathcal{A} invertible applied to (5.1) behaves like*

$$x(t_{i+1}) - x_{i+1} = \mathcal{O}(h^{\kappa_\nu - \nu + 2} + h^{\kappa_{\nu-1} - \nu + 3} + \dots + h^{\kappa_1 + 1})$$

where κ_j is the maximum number such that

	Condition	range of k
a.)	$\beta^T \mathcal{A}^{-k} e = \beta^T \mathcal{A}^{-j} \gamma^{j-k} / (j-k)!$	$k = 1, 2, \dots, j-1$
b.)	$\beta^T \mathcal{A}^{-j} \gamma^k = k! / (k-j+1)!$	$k = j, j+1, \dots$

for all $k \leq \kappa_j$ and for $j = 1, \dots, \nu$.

Proof. Since we consider the pure consistency error, we can assume that $x_i = x(t_i)$. With that and with the definition of the RKM, the error is given as

$$\tau = x(t_{i+1}) - x_{i+1} = -h \sum_{j=1}^s \beta_j \dot{X}_{ij} + \sum_{k \geq 1} \frac{h^k}{k!} x^{(k)}(t_i).$$

Because of the special structure of the DAE, we can concentrate on the first error component $\tau_1 \leftarrow$ the error component τ_2 is the *first* component of the problem of index $\nu - 1$. For τ_1 we have the formula

$$\tau_1 = x_1(t_{i+1}) - x_{i+1,1} = h\beta^T \sum_{j=1}^{\nu} (h\mathcal{A})^{-j} Z_{ij} + \sum_{k \geq 1} \frac{h^k}{k!} x_1^{(k)}(t_i).$$

One may confirm directly, or by means of the solution formula for $N\dot{x} = x + f$, that the ℓ -th component of x is defined as

$$x_\ell(t) = - \sum_{j=\ell}^{\nu} f_j^{(j-\ell)}(t).$$

The componentwise Taylor expansion of $Z_{i,\ell}$ reads

$$\begin{aligned} Z_{i\ell} &= \begin{bmatrix} x_{i,\ell} + f_\ell(t_i + \gamma_1 h) \\ x_{i,\ell} + f_\ell(t_i + \gamma_2 h) \\ \vdots \\ x_{i,\ell} + f_\ell(t_i + \gamma_s h) \end{bmatrix} = \begin{bmatrix} x_{i,\ell} + f_\ell(t_i) + \sum_{k \geq 1} \frac{h^k}{k!} f_\ell^{(k)}(t_i) \gamma_1^k \\ x_{i,\ell} + f_\ell(t_i) + \sum_{k \geq 1} \frac{h^k}{k!} f_\ell^{(k)}(t_i) \gamma_2^k \\ \vdots \\ x_{i,\ell} + f_\ell(t_i) + \sum_{k \geq 1} \frac{h^k}{k!} f_\ell^{(k)}(t_i) \gamma_s^k \end{bmatrix} \\ &= x_{i,\ell} e + \sum_{k \geq 0} \frac{h^k}{k!} f_\ell^{(k)}(t_i) \gamma^k \end{aligned}$$

With that and with $x_i = x(t_i)$, we expand the error τ_1 as follows:

$$\begin{aligned}
\tau_1 &= \beta^T \sum_{j=1}^{\nu} (h\mathcal{A})^{-j} Z_{ij} + \sum_{k \geq 1} \frac{h^k}{k!} x_1^{(k)}(t_i) \\
&= \beta^T \sum_{j=1}^{\nu} h^{-j+1} \mathcal{A}^{-j} [x_j(t_i)e + \sum_{k \geq 0} \frac{h^k}{k!} f_j^{(k)}(t_i) \gamma^k] \\
&\quad + \sum_{k \geq 1} \frac{h^k}{k!} x_1^{(k)}(t_i) \\
&= \beta^T \sum_{j=1}^{\nu} h^{-j+1} \mathcal{A}^{-j} \left[- \sum_{k=j}^{\nu} f_k^{(k-j)}(t_i) e + \sum_{k \geq 0} \frac{h^k}{k!} f_j^{(k)}(t_i) \gamma^k \right] \\
&\quad - \sum_{k \geq 1} \frac{h^k}{k!} \sum_{j=1}^{\nu} f_j^{(j-1+k)}(t_i) \\
&= - \sum_{j=1}^{\nu} \sum_{k=j}^{\nu} h^{-j+1} \beta^T \mathcal{A}^{-j} e f_k^{(k-j)}(t_i) + \sum_{j=1}^{\nu} \sum_{k \geq 0} \frac{h^{k-j+1}}{k!} \beta^T \mathcal{A}^{-j} \gamma^k f_j^{(k)}(t_i) \\
&\quad - \sum_{k \geq 1} \sum_{j=1}^{\nu} \frac{h^k}{k!} f_j^{(j-1+k)}(t_i),
\end{aligned}$$

which, with $\sum_{j=1}^{\nu} \sum_{k=j}^{\nu} g(j, k) = \sum_{k=1}^{\nu} \sum_{j=1}^k g(j, k) = \sum_{k=1}^{\nu} \sum_{j=1}^k g(k, j)$, becomes

$$\begin{aligned}
\tau_1 &= \sum_{j=1}^{\nu} \left[- \sum_{k=1}^j h^{-k+1} \beta^T \mathcal{A}^{-k} e f_k^{(j-k)}(t_i) \right. \\
&\quad + \sum_{k \geq 0} \frac{h^{k-j+1}}{k!} \beta^T \mathcal{A}^{-j} \gamma^k f_j^{(k)}(t_i) \\
&\quad \left. - \sum_{k \geq 1} \frac{h^k}{k!} f_j^{(j-1+k)}(t_i) \right] \\
&= \sum_{j=1}^{\nu} \left[- \sum_{k=1}^j h^{-k+1} \beta^T \mathcal{A}^{-k} e f_k^{(j-k)}(t_i) \right. \\
&\quad + \sum_{k=0}^{j-1} \frac{h^{k-j+1}}{k!} \beta^T \mathcal{A}^{-j} \gamma^k f_j^{(k)}(t_i) + \sum_{k \geq j} \frac{h^{k-j+1}}{k!} \beta^T \mathcal{A}^{-j} \gamma^k f_j^{(k)}(t_i) \\
&\quad \left. - \sum_{k \geq 1} \frac{h^k}{k!} f_j^{(j-1+k)}(t_i) \right].
\end{aligned}$$

A shift of indices, $\sum_{k=0}^{j-1} g(k) = \sum_{k=1}^j g(j-k)$ and $\sum_{k \geq 1} g(k) = \sum_{k \geq j} g(k-j+1)$,

then gives:

$$\begin{aligned} \tau_1 = \sum_{j=1}^{\nu} & \left[- \sum_{k=1}^j h^{-k+1} \beta^T \mathcal{A}^{-k} e f_k^{(j-k)}(t_i) + \sum_{k=1}^j \frac{h^{-k+1}}{(j-k)!} \beta^T \mathcal{A}^{-j} \gamma^{j-k} f_j^{(j-k)}(t_i) \right. \\ & \left. + \sum_{k \geq j} \frac{h^{k-j+1}}{k!} \beta^T \mathcal{A}^{-j} \gamma^k f_j^{(k)}(t_i) - \sum_{k \geq j} \frac{h^{k-j+1}}{(k-j+1)!} f_j^{(k)}(t_i) \right]. \end{aligned}$$

□

Chapter 6

Construction and Analysis of RKM for nonlinear DAEs

Now we consider RKM for nonlinear DAEs. We start with a DAE in *semi explicit* strangeness-free form and give general results on how to write down a general RKM for it and how to analyse the global error. Then, we consider general strangeness-free nonlinear DAEs and show that a certain class of RKM applies well – namely those that can be constructed by collocation with Lagrange polynomials over the *Radau*, *Lobatto*, or *Gauss* quadrature points.

6.1 General RKM for Semi-Explicit Strangeness-free DAEs

A semi explicit strangeness-free DAE is of the form

$$\dot{x} = f(t, x, y) \tag{6.1}$$

$$0 = g(t, x, y) \tag{6.2}$$

with the Jacobian of g with respect to y , i.e.

$$\partial_y \otimes g(t, x(t), y(t)) =: g_y(t, x(t), y(t)),$$

being invertible for all t along the solution (x, y) .

Some observations:

- this system is strangeness-free
- under certain assumptions, any DAE can be brought into this form
- in the linear case $E\dot{z} = Az + f$, with $z = (x, y)$, the assumptions basically mean that

$$E = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} * & * \\ * & A_{22} \end{bmatrix},$$

with $A_{22}(t)$ invertible for all t .

- The condition g_y invertible means that, locally, one could consider

$$\dot{x} = f(t, x, R(t, x)), \quad \text{with } R \text{ such that } y = R(t, x).$$

However, this is not practical for numerical purposes.

The general strategy to get a suitable formulation of a time discretization of system (6.1)-(6.2) by any RKM is to consider the perturbed version

$$\begin{aligned} \dot{x} &= f(t, x, y), \\ \varepsilon \dot{y} &= g(t, x, y), \end{aligned}$$

which is an ODE, formulate the RKM, and then let $\varepsilon \rightarrow 0$. In the Hairer/Wanner Book, this approach is called ε -embedding.

This is, consider

Table 6.1: RKM applied to semi-explicit DAEs

$$\begin{aligned} x_{i+1} &= x_i + h \sum_{j=1}^s \beta_j \dot{X}_{ij}, & y_{i+1} &= y_i + h \sum_{j=1}^s \beta_j \dot{Y}_{ij}, \\ \dot{X}_{ij} &= f(t_i + \gamma_j h, X_{ij}, Y_{ij}), & \varepsilon \dot{Y}_{ij} &= g(t_i + \gamma_j h, X_{ij}, Y_{ij}), & j &= 1, 2, \dots, s, \\ X_{ij} &= x_i + h \sum_{\ell=1}^s \alpha_{j\ell} \dot{X}_{i\ell}, & Y_{ij} &= y_i + h \sum_{\ell=1}^s \alpha_{j\ell} \dot{Y}_{i\ell}, & j &= 1, 2, \dots, s, \end{aligned} \quad (*)$$

i.e., the RKM applied to an ODE in the variables (x, y) , and replace $(*)$ by

$$\dot{X}_{ij} = f(t_i + \gamma_j h, X_{ij}, Y_{ij}), \quad 0 = g(t_i + \gamma_j h, X_{ij}, Y_{ij}), \quad j = 1, 2, \dots, s.$$

Theorem 6.1 (Kunkel/Mehrmann Thm. 5.16). *Consider a semi-explicit, strangeness-free DAE as in (6.1)-(6.2) with a consistent initial value (x_0, y_0) . The time-discretization by a RKM,*

- *with \mathcal{A} invertible and $\rho := 1 - \beta^T \mathcal{A}^{-1} e$,*
- *applied as in Table 6.1 with $\varepsilon = 0$,*
- *that is convergent of order p for ODEs*
- *and fulfills the Butcher condition $C(q)$ with $q \geq p + 1$*

leads to an global error that behaves like

$$\|\mathfrak{X}(t_N) - \mathfrak{X}_N\| = \mathcal{O}(h^k),$$

where

- $k = p$, *if $\rho = 0$,*
- $k = \min\{p, q + 1\}$, *if $-1 \leq \rho < 1$*
- $k = \min\{p, q - 1\}$, *if $\rho = 1$.*

6.2. COLLOCATION RKM FOR IMPLICIT STRANGENESS-FREE DAES47

If $|\rho| > 1$, then the RKM – applied to (6.1)–(6.2) – does not converge.

Some words on the conditions on p , q , and ρ :

- For *stiffly accurate* methods, $\beta^T \mathcal{A}^{-1} e = 1$ and, thus, $\rho = 0 \rightarrow$ no order reduction for *strangeness free* or *index-1* systems
- For the *implicit midpoint rule* also known as the *1-stage Gauss method*:

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

- the convergence order for ODEs is $p = 2$
- but $1 - \beta^T \mathcal{A}^{-1} e = 1 - 1 \cdot \left(\frac{1}{2}\right)^{-1} 1 = -1$, so that $\min\{p-1, q\} = k \leq 1$, depending on q .
- in fact $k = 1$ as

$$C(q) : \sum_{\ell=1}^s \alpha_{j\ell} \gamma_{\ell}^{\bar{k}-1} = \frac{1}{\bar{k}} \gamma_j^{\bar{k}}, \quad \bar{k} = 1, \dots, q, \quad j = 1, \dots, s$$

in the present case of $s = 1$, $\alpha_{11} = \gamma_1 = \frac{1}{2}$ is fulfilled for $\bar{k} = 1$:

- $\frac{1}{2} = \frac{1}{2}$
- it is not relevant here, but for $\bar{k} = 2$: $\frac{1}{2} \cdot \frac{1}{2} \neq \frac{1}{2} \cdot \frac{1}{4}$

6.2 Collocation RKM for Implicit Strangeness-free DAEs

The general form of a *strangeness-free* DAE is given as

$$\hat{F}_1(t, x, \dot{x}) = 0 \tag{6.3}$$

$$\hat{F}_2(t, x) = 0 \tag{6.4}$$

where the *strangeness-free* or *index-1* assumption is encoded in the existence of *implicit functions* \mathcal{L} , \mathcal{R} such that, with $x = (x_1, x_2)$, the implicit DAE (6.3)–(6.4) is equivalent to the semi-explicit DAE

$$\begin{aligned} \dot{x}_1 &= \mathcal{L}(t, x_1, x_2) \\ 0 &= \mathcal{R}(t, x_1) - x_2 \end{aligned}$$

In what follows we show that a *collocation* approach coincides with certain RKM discretizations so that the convergence analysis of the RKM can be done via approximation theory.

Regression (Collocation): – If one looks for a function $x : [0, 1] \rightarrow \mathbb{R}$ that fulfills $F(x(t)) = 0$ for all $t \in [0, 1]$, one may interpolate x by, say, a polynomial $x_p(t) = \sum_{\ell=0}^k x_{\ell} t^{\ell}$ and determine the $k+1$

coefficients x_ℓ via the solution of the system of (nonlinear) equations $F(x_p(t_\ell)) = 0$, $\ell = 0, 1, \dots, k$, where the $t_\ell \in [0, 1]$ are the $k + 1$ *collocation points*.

Concretely, we parametrize s collocation points via

$$0 < \gamma_1 < \gamma_2 < \dots < \gamma_s = 1 \quad (6.5)$$

and define two sets of *Lagrange polynomials*

$$L_\ell(\xi) = \prod_{j=0, j \neq \ell}^s \frac{\xi - \gamma_j}{\gamma_\ell - \gamma_j} \quad \text{and} \quad \tilde{L}_\ell(\xi) = \prod_{m=1, m \neq \ell}^s \frac{\xi - \gamma_m}{\gamma_\ell - \gamma_m},$$

with $\ell \in \{0, 1, \dots, s\}$.

Let \mathbb{P}_k be the space of polynomials of degree $\leq k - 1$. We define the *collocation polynomial* $x_\pi \in \mathbb{P}_{s+1}$ via

$$x_\pi(t) = \sum_{\ell=0}^s X_{i\ell} L_\ell\left(\frac{t - t_i}{h}\right) \quad (6.6)$$

designed to compute the *stage values* $X_{i\ell}$, where $X_{i0} = x_i$ is already given.

The stage derivatives are then defined as

$$\dot{X}_{ij} = \dot{x}_\pi(t_i + \gamma_j h) = \frac{1}{h} \sum_{\ell=0}^s X_{i\ell} \dot{L}_\ell(\gamma_j). \quad (6.7)$$

To obtain $x_{i+1} = x_\pi(t_{i+1}) = X_{is}$, we require the polynomial to satisfy the DAE (6.3)–(6.4) at the collocation points $t_{ij} = t_i + \gamma_j h$, that is

$$\hat{F}_1(t_i + \gamma_j h, X_{ij}, \dot{X}_{ij}) = 0, \quad \hat{F}_2(t_i + \gamma_j h, X_{ij}) = 0, \quad j = 1, \dots, s. \quad (6.8)$$

Now we show that this collocation defines a RKM discretization of (6.3)–(6.4).

Since $\tilde{L} \in \mathbb{P}_s$, it holds that

$$P_\ell(\sigma) := \int_0^\sigma \tilde{L}_\ell(\xi) d\xi \in \mathbb{P}_{s+1}$$

that is, by *Lagrange* interpolation, it can be written as

$$P_\ell(\sigma) = \sum_{j=0}^s P_\ell(\gamma_j) L_j(\sigma).$$

If we differentiate P_ℓ , we get

$$\dot{P}_\ell(\sigma) = \sum_{j=0}^s P_\ell(\gamma_j) \dot{L}_j(\sigma) = \sum_{j=0}^s \int_0^{\gamma_j} \tilde{L}_\ell(\xi) d\xi \dot{L}_j(\sigma) =: \sum_{j=0}^s \alpha_{j\ell} \dot{L}_j(\sigma)$$

where define simply define

$$\alpha_{j\ell} = \int_0^{\gamma_j} \tilde{L}_\ell(\xi) d\xi.$$

Moreover, by definition of P_ℓ (and the *fundamental theorem of calculus*), it holds that

$$\dot{P}_\ell(\sigma) = \tilde{L}_\ell(\sigma),$$

which gives that $\dot{P}_\ell(\gamma_m) = \delta_{\ell m}$ that is

$$\dot{P}_\ell(\gamma_m) = \sum_{j=1}^s \alpha_{j\ell} \dot{L}_j(\gamma_m) = \begin{cases} 1, & \text{if } \ell = m \\ 0, & \text{otherwise} \end{cases}.$$

for $\ell, m = 1, \dots, s$.

Accordingly, if we define $\mathcal{A} := [\alpha_{j\ell}]_{j,\ell=1,\dots,s} \in \mathbb{R}^{s,s}$ and

$$V := [v_{mj}]_{m,j=1,\dots,s} = [\dot{L}_j(\gamma_m)]_{m,j=1,\dots,s} \in \mathbb{R}^{s,s},$$

it follows that $V = \mathcal{A}^{-1}$.

Moreover, since,

$$\sum_{j=0}^s L_j(\sigma) \equiv 1, \quad \text{so that} \quad \sum_{j=0}^s \dot{L}_j(\sigma) \equiv 0,$$

we have that

$$\sum_{j=0}^s \dot{L}_j(\gamma_m) = 0 = \sum_{j=0}^s v_{mj}$$

and, thus,

$$v_{m0} = - \sum_{j=1}^s \dot{L}_j(\gamma_m) = -e_m^T V e.$$

With these relations we rewrite (6.7) as

$$h \dot{X}_{im} = \sum_{\ell=0}^s X_{i\ell} \dot{L}_\ell(\gamma_m) = v_{m0} x_i + \sum_{\ell=1}^s v_{m\ell} X_{i\ell}.$$

and $h \sum_{m=1}^s \alpha_{\ell m} \dot{X}_{im}$ as

$$\begin{aligned} h \sum_{m=1}^s \alpha_{\ell m} \dot{X}_{im} &= \sum_{m=1}^s \alpha_{\ell m} v_{m0} x_i + \sum_{j,m=1}^s \alpha_{\ell m} v_{mj} X_{ij} \\ &= -e_\ell^T \mathcal{A} V e x_i + \sum_{j=1}^s e_\ell^T \mathcal{A} V e_j X_{ij} \\ &= -x_i + X_{i\ell}, \end{aligned} \tag{6.9}$$

which, together with (6.8), indeed defines a RKM.

Some remarks:

- the preceding derivation shows that the collocation (6.6) and (6.8) is equivalent to the RKM scheme (6.9) and (6.8)
- convergence of these schemes applied to (6.3)–(6.3) is proven in Kunkel/Mehrmann Theorem 5.17
- with fixing $\gamma_s = 1$, the obtained RKM is *stiffly accurate*
- the remaining $s - 1$ γ s can be chosen to get optimal convergence rates \rightarrow RadauIIa schemes
- if also γ_s is chosen optimal in terms of convergence, the Gauss schemes are obtained

Chapter 7

Examples

7.1 Semi-discrete Navier-Stokes equations

By scalings and state transforms, we find that the coefficients of the spatially discretized Navier-Stokes equations are equivalent to:

$$\begin{aligned}
 (\mathcal{E}, \mathcal{A}) &= \left(\begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A & B^H \\ B & 0 \end{bmatrix} \right) \\
 &\sim \left(\begin{bmatrix} M^{-1/2} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A & B^H \\ B & 0 \end{bmatrix} \begin{bmatrix} M^{-1/2} & 0 \\ 0 & I \end{bmatrix} \right) \\
 &\sim \left(\begin{bmatrix} Q^H & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} M^{-1/2} A M^{-1/2} & M^{-1/2} B^H \\ B M^{-1/2} & 0 \end{bmatrix} \begin{bmatrix} Q & 0 \\ 0 & I \end{bmatrix} \right) \\
 &\sim \left(\begin{bmatrix} I & 0 \\ 0 & R^{-H} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} M^{-1/2} A M^{-1/2} & \begin{bmatrix} R \\ 0 \\ 0 \end{bmatrix} \\ \begin{bmatrix} R^H & 0 \end{bmatrix} & 0 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & R^{-1} \end{bmatrix} \right) \\
 &= \left(\begin{bmatrix} I_{n_1} & 0 & 0 \\ 0 & I_{n_2} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} & I_{n_1} \\ A_{21} & A_{22} & 0 \\ I_{n_1} & 0 & 0 \end{bmatrix} \right).
 \end{aligned}$$

where we have used a QR -decomposition:

$$M^{-1/2} B^H = Q \begin{bmatrix} R \\ 0 \end{bmatrix}$$

with unitary Q and invertible R in the third step.

Chapter 8

Exercises

8.1 II.C.1

Let $E, A \in \mathbb{C}^{n,n}$ satisfy $EA = AE$. Then $\ker E \cap \ker A = \{0\}$ implies that (E, A) is regular.

Assume that $\ker A \neq \{0\}$ is of dimension $k \geq 1$. The case that $k = 0$ is trivial, since $\lambda E - A$ is regular for $\lambda = 0$. Let V_0 be the matrix whose columns span $\ker A$ and let V_\perp be the matrix that consists of all eigenvectors of A that are associated with the nonzero eigenvalues.

It holds that,

$$AV_0 = 0 \quad \text{and} \quad AV_\perp = V_\perp L_\perp$$

with an $L_\perp \in \mathbb{C}^{n-k, n-k}$ which is invertible. This is a consequence of V_\perp spanning the A -invariant subspaces with respect to the nonzero eigenvalues.

Because of $ABV_0 = BAV_0 = 0$, it follows that $\text{span } BV_0 \subset \ker A = \text{span } V_0$, i.e., V_0 is a B -invariant subspace which means that there is a $K_0 \in \mathbb{C}^{k,k}$ such that $BV_0 = V_0 K_0$.

Moreover, because of $\ker E \cap \ker A = \{0\}$, the matrix K_0 has no zero eigenvalues. In fact K_0 has the same eigenvalues as $B' := B|_{V_0} : V_0 \rightarrow V_0$, and if B' had a zero eigenvalue this would mean that the associated eigenvector would be in V_0 and, thus, in the kernel of A .

Moreover, since $ABV_\perp = BAV_\perp = BV_\perp L_\perp$ meaning that BV_\perp is in the A -invariant subspace related to the nonzero eigenvalues of A , i.e., $BV_\perp \subset V_\perp$, it follows that V_\perp is a B -invariant subspace and, thus, $BV_\perp = V_\perp K_\perp$ for some matrix $K_\perp \in \mathbb{C}^{n-k, n-k}$.

With $V := [V_0|V_\perp]$ and the observation that V is invertible, since its columns span all of $\mathbb{C}^n = \text{span } V_0 \oplus \text{span } V_\perp$, it follows that

$$\begin{aligned}\lambda E - A &= (\lambda E - A)VV^{-1} = (\lambda E[V_0|V_\perp] - A[V_0|V_\perp])V^{-1} \\ &= ([V_0K_0|V_\perp K_\perp]\lambda - [0|V_\perp L_\perp])V^{-1} \\ &= [V_0|V_\perp] \begin{bmatrix} \lambda K_0 & \\ & \lambda K_\perp - L_\perp \end{bmatrix} V^{-1}\end{aligned}$$

and that

$$\det(\lambda E - A) = \det(\lambda K_0) \det(\lambda K_\perp - L_\perp)$$

is not identically zero, since K_0 and L_\perp are invertible.

Chapter 9

Numerical Analysis and Software Overview

9.1 Theory: RKMs and BDF for DAEs

Table 9.1: Overview of convergence results of RKM/BDF schemes for DAEs

DAEs	
unstructured, linear	$E(t)\dot{x} = A(t)x + f(t)$
semi-linear	$E(t)\dot{x} = f(t, x)$
unstructured	$F(t, \dot{x}, x) = 0$
unstructured, strangeness-free/index-1	$\begin{cases} \hat{F}_1(t, \dot{x}, x) = 0 \\ \hat{F}_2(t, x) = 0 \end{cases}$
semi-explicit, strangeness-free/index-1	$\begin{cases} \dot{x} = f(t, x, y) \\ 0 = g(t, x, y) \end{cases}$
semi-explicit, index-2	$\begin{cases} \dot{x} = f(t, x, y) \\ 0 = g(t, y) \end{cases}$

Description		Reference
a	RKM, linear constant coefficients	KM Thm. 5.12
b	RKM, nonlinear, strangeness-free/index-1, semi-explicit	KM Thm 5.16 / HW Thm. VI.1.1
c	RKM, nonlinear, strangeness-free	KM Thm. 5.18
d	BDF, linear constant coefficients	KM Thm. 5.24

	Description	Reference
e	BDF(\subset MSM), nonlinear, strangeness-free/index-1, semi-explicit	KM Thm. 5.26 (\subset HW Thm. VI.2.1)
f	BDF, nonlinear, strangeness-free/index-1	KM Thm. 5.27
g	RKM, nonlinear, index-2, semi-explicit	HW Ch. VII.4
h	BDF, nonlinear, index-2, semi-explicit	HW Thm. VII.3.5
i	half-explicit RKM, nonlinear, index-2, semi-explicit	HW Thm. VII.6.2
HW	Ernst Hairer, Gerhard Wanner (1996)	<i>Solving ordinary differential equations. II: Stiff and differential-algebraic problems</i>
KM	Peter Kunkel, Volker Mehrmann (2006)	<i>Differential-Algebraic Equations. Analysis and Numerical Solution</i>

9.2 Solvers

As can be seen from the table above, generally usable discretization methods for unstructured DAEs are only there for index-1 problems. However, the solvers GELDA/GENDA include an automated reduction to the strangeness-free form so that they apply for any index; see Lecture Chapter 4++.

9.2.1 Multi purpose

DAEs	Methods	h/p	Language	Remark	Avail
GELDA- μ -*	BDF/RKM	*/	F-77		*/
GENDA- μ -*	BDF	*/	F-77		/
DASSL- ν -1	BDF	*/	F-77	base for <i>Sundials IDA</i> – the base of many DAE solvers	*/
LIMEXsl- ν -1	x-SE-Eul	*/	F-77		/
RADAUsl- ν -1	RKM	*/	F-77		*/

Notes:

Table 9.2: Overview of convergence results of BDF/RKM schemes for DAEs of various index and, possibly, semi-explicit structure. Here, we equate *index-1* and *strangeness-free*. A \cdot indicates that this case is included in a result for a more general case *located* left or above in the table.

Problem / Index	RKM						BDF					
	unstructured			semi-explicit			unstructured			semi-explicit		
	*	2	1	*	2	1	*	2	1	*	2	1
nonlinear			c		g,i	b		f		h	e	
linear TV			\cdot		\cdot	\cdot		\cdot		\cdot	\cdot	
linear CC	a	\cdot	\cdot	\cdot	\cdot	\cdot	d	\cdot	\cdot	\cdot	\cdot	\cdot

Explanation	
DAEs	l-linear, sl-semilinear, nl-nonlinear classification: μ -strangeness index, ν -differentiation index *-includes index reduction
h/p	time step control / order control
availability	code for download / licence provided(*) or other statement(\cdot)
methods	x-SE-Eul: extrapolation based on semiexplicit Euler

9.2.2 Application specific

Furthermore, there are solvers for particularly structured DAEs.

DAEs	Resources
Navier-Stokes (nl-se-2)	See, e.g., Sec. 4.3 of our preprint on definitions of different schemes
Multi-Body (nl-se-3)	See, e.g., the code on Hairer’s homepage

9.3 Software

Many software suits actually wrap SUNDIALS IDA.

DAEs Routines		Method	Remark
Matlab	ode15{i,s}	BDF	
Python			no built-in functionality, DASSL/IDA wrapped in the modules <code>scikit-odes</code> , <code>assimulo</code> , <code>pyDAS</code> , <code>DAEtools</code>
Julia	<code>DifferentialEquations.jl</code>	BDF	calls SUNDIALS IDA