DAEs

Jan Heiland

2019-03-28

Contents

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.

Throughout this lecture, the free variable will be the time t and differentiation will always be considered with respect to t.

1.1 Examples

Free fall vs. the pendulum

Free fall: Newton's second law applies:

```
force = mass*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.

Now **the pendulum**:

The same point mass attached to a string.

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 Langrangian function to derive the equations of motion. For the pendulum, we have the kinetic energy T, the potential U, and the constraint h as

$$T = \frac{1}{2}m(\dot{x}(t)^2 + \dot{y}(t)^2), \quad U = mgy, \quad \text{and} \quad h = (x(t) - c_x)^2 + (y(t) - c_y)^2 - l^2.$$

Thus, from the principle that

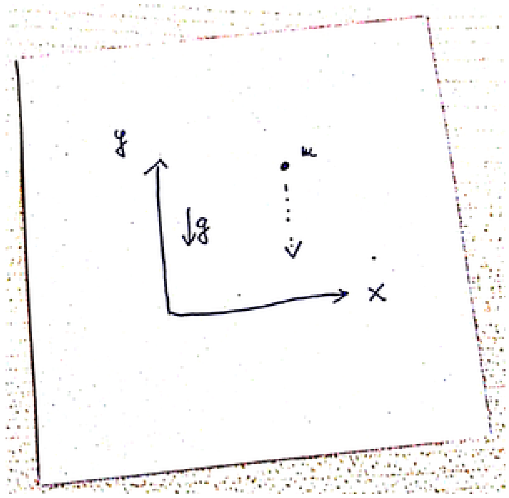

Figure 1.1: Free fall of a point mass.

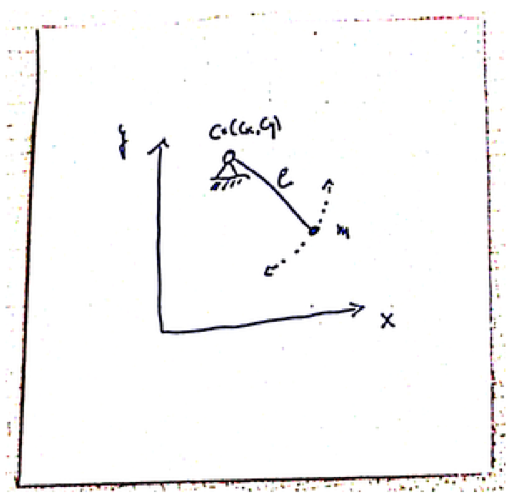

Figure 1.2: A pendulum.

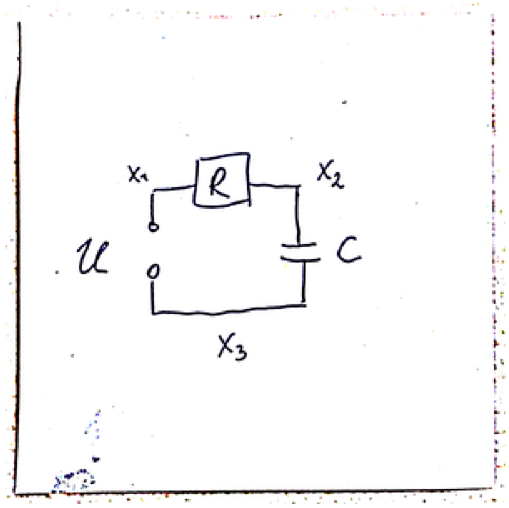

Figure 1.3: Electrical circuit with a source, a resistor, and a conductor.

$$\frac{d}{dt}(\frac{\partial L}{\partial \dot{q}}) - \frac{\partial L}{\partial q} = 0, \quad \text{for} \quad L = U - T - \lambda h \quad \text{and} \quad q = x, y, \lambda$$

one obtains:

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

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

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.

- 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 is defined through *Ohm's law*: $I_R = \frac{x_1 - x_2}{R}$
- the current I_C that is induced by the potentials neighboring the capacitor is described through: $I_C = C(\dot{x}_3 - \dot{x}_2)$
- *Kirchhoff's law* states that everywhere in the circuit the currents sum up to zero: $I_C + I_R = C(\dot{x}_3 - \dot{x}_2) + \frac{x_1 - x_2}{R} = 0$
- finally, we can set a ground potential (note that so far all equations only consider differences in the potential) and we set $x_3 = 0$.

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:

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

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.

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.

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

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

In this case the solution part $x_1 = const. + \int_0^t g(s)ds$ 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

We consider DAEs in the general form

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

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

$$x\colon I \to \mathbb{R}^n$$

with time derivative $\dot{x}\colon I \to \mathbb{R}^n$ 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

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

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 $x \in \mathcal{C}^1(I, \mathbb{R}^n)$ is called a *solution to the DAE* (2.1), if $F(t, x(t), \dot{x}(t)) = 0$ holds for all $t \in I$.

2. A function $x \in \mathcal{C}^1(I, \mathbb{R}^n)$ 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 where 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)

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

with the constraint

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

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

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

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:

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

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 $x_1(t_0) - x_3(t_0) = U(t_0)$.

Chapter 3

Linear DAEs with Constant Coefficients

Definition 3.1. Let $E \in \mathbb{C}^{n,n}$ and $\nu = \operatorname{ind}(E)$. A matrix $X \in \mathbb{C}^{n,n}$ that fulfills

$$EX = XE, \tag{3.1}$$
$$XEX = X, \tag{3.2}$$
$$XE^{\nu+1} = E^{\nu}, \tag{3.3}$$

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 3.1. *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 (3.1)–(3.3) one derives that

$$\begin{aligned}
X_1EX_1EX_2 =& X_1EX_2 = X_1EX_2EX_2\\
X_1^2E^2X_2 = \cdots =& X_1EX_2 = \cdots = X_1E^2X_2^2\\
X_1^{\nu+1}E^{\nu+1}X_2 = \cdots = \cdots =& X_1EX_2 = \cdots = \cdots = X_1E^{\nu+1}X_2^{\nu+1}\\
X_1^{\nu+1}E^{\nu+1}X_1 = \cdots = \cdots = \cdots =& X_1EX_2 = \cdots = \cdots = \cdots = X_2E^{\nu+1}X_2^{\nu+1}\\
X_1 = \cdots = \cdots = \cdots = \cdots =& X_1EX_2 = \cdots = \cdots = \cdots = \cdots = X_2,
\end{aligned}$$

where in the second last step we used the identities

$$E^{\nu+1}X_1 = X_1E^{\nu+1} = E^{\nu} = X_2E^{\nu+1} = E^{\nu+1}X_2.$$

□

Chapter 4

Linear DAEs with Time-varying Coefficients

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

$$T,\ Z,\ T',\ V \tag{4.1}$$

be

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

then the quantities

$$r,\ a,\ s,\ d,\ u,\ v \tag{4.2}$$

defined as

Quantity	*Definition*	*Name*
r	$\operatorname{rank} E$	rank
a	$\operatorname{rank}(Z^H AT)$	algebraic part
s	$\operatorname{rank}(V^H Z^H AT')$	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

$$\left(\begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right), \tag{4.3}$$

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

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

so that the DAE basically reads

$$ET'\dot{y}'(t) = ATy(t) + AT'y'(t) + f,$$

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

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

or

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

so that $\operatorname{rank} Z^H AT$ 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 $V = \operatorname{corange} Z^H AT$, 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, $\operatorname{rank} V^H Z^H AT'$ 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 the size s of the *strange* block here) will define the strangeness index.

Example 4.1. With basic scalings and state transforms, one finds for the coefficients of Example 1.2 that:

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

We compute the subspaces as defined in (4.1):

Matrix	as the basis of/computed as
$T = \begin{bmatrix} 0 \\ I_1 \end{bmatrix}$	kernel $\begin{bmatrix} I_2 & 0 \\ 0 & 0 \end{bmatrix}$
$Z = \begin{bmatrix} 0 \\ I_1 \end{bmatrix}$	corange $\begin{bmatrix} I_2 & 0 \\ 0 & 0 \end{bmatrix} = \text{kernel} \begin{bmatrix} I_2 & 0 \\ 0 & 0 \end{bmatrix}^H$

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

and derive the quantities as defined in (4.2):

Name	Value	Derived from
rank	$r = 2$	$\text{rank}\, E = \text{rank} \begin{bmatrix} I_2 & 0 \\ 0 & 0 \end{bmatrix}$
algebraic part	$a = 1$	$\text{rank}\, Z^H AT = \text{rank}\, I_1$
strangeness	$s = 0$	$\text{rank}\, V^H Z^H AT' = \text{rank}\, 0_{2\times 1}$
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 4.2. 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:

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

We compute the subspaces as defined in (4.1):

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 $\begin{bmatrix} I_{n_1} & 0 & 0 \\ 0 & I_{n_2} & 0 \\ 0 & 0 & 0 \end{bmatrix}$
$Z^H AT = 0_{n_1}$	$\begin{bmatrix} 0 \\ 0 \\ I_{n_1} \end{bmatrix}^H \begin{bmatrix} A_{11} & A_{12} & I_{n_1} \\ A_{21} & A_{22} & 0 \\ I_{n_1} & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ I_{n_1} \end{bmatrix}$
$V = I_{n_1}$	$\text{corange}(Z^H AT) = \text{kernel}\, 0_{n_1}^H$
$Z^H AT' = \begin{bmatrix} I_{n_1} & 0_{n_1 \times n_2} \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ I_{n_1} \end{bmatrix}^H \begin{bmatrix} A_{11} & A_{12} & I_{n_1} \\ A_{21} & A_{22} & 0 \\ I_{n_1} & 0 & 0 \end{bmatrix} \begin{bmatrix} I_{n_1} & 0 \\ 0 & I_{n_2} \\ 0 & 0 \end{bmatrix}$

and derive the quantities as defined in (4.2):

Name	Value	Derived from
rank	$r = n_1 + n_2$	$\operatorname{rank} E = \operatorname{rank} \begin{bmatrix} I_{n_1} & 0 & 0 \\ 0 & I_{n_2} & 0 \\ 0 & 0 & 0 \end{bmatrix}$
algebraic part	$a = 0$	$\operatorname{rank} Z^H AT = \operatorname{rank} 0_{n_1}$
strangeness	$s = n_1$	$\operatorname{rank} V^H Z^H AT' = \operatorname{rank} \begin{bmatrix} I_{n_1} & 0_{n_1 \times n_2} \end{bmatrix}$
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* $\operatorname{rank} E(t) = r$ *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*

$$r(t) = r, \quad a(t) = a, \quad s(t) = s \tag{4.4}$$

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

$$\left(\begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} \\ 0 & 0 & 0 & A_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right). \tag{4.5}$$

All entries are again matrix functions on I and the last block column in both matrix functions of (4.5) *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 continous 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}. □

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

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

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*

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

where κ_j is the maximum number such that

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

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

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:

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

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

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

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:

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

$\square$

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

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

  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 $\varepsilon \to 0$. In the Hairer/Wanner Book, this approach is called *ε-embedding.

This is, consider

Table 6.1: RKM applied to semi-explicit DAEs

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

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

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

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 $\rho = 1$.*

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, $\rho = 0 \to$ 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

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

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

– 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 $x\colon [0, 1] \to \mathbb{R}$ 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 $F(x_p(t_\ell)) = 0$, $\ell = 0, 1, \ldots, k$, where the $t_\ell \in [0, 1]$ are the $k + 1$ *collocation points*.

Concretely, we parametrize s collocation points via

$$0 < \gamma_1 < \gamma_2 < \ldots < \gamma_s = 1 \tag{6.5}$$

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 $\ell \in \{0, 1, \ldots, s\}$.

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

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

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

The stage derivatives are then defined as

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

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

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

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 $\ell, m = 1, \ldots, s$.

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

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

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

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

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 $s-1$ γs 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:

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

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 $L_\perp \in C^{n-k,n-k}$ which is invertible. This 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 $\operatorname{span} BV_0 \subset \ker A = \operatorname{span} V_0$, 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 $B' := B\big|_{V_0} \colon V_0 \to V_0$, 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 $ABV_\perp = BAV_\perp = BVL_\perp$ 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 $\mathbb{C}^n = \operatorname{span} V_0 \oplus \operatorname{span} V_\perp$, 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
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)	*Solving ordinary differential equations. II: Stiff and differential-algebraic problems*

	Description	Reference
KM	Peter Kunkel, Volker Mehrmann (2006)	*Differential-Algebraic Equations. Analysis and Numerical Solution*

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	l-μ-$*$	BDF/RKM	$*/*$	F-77		$*/\cdot$
GENDA	n-μ-$*$	BDF	$*/*$	F-77		$/\cdot$
DASSL	n-ν-1	BDF	$*/*$	F-77	base for *Sundials IDA* – the base of many DAE solvers	$*/$
LIMEX	sl-ν-1	x-SE-Eul	$*/*$	F-77		$/$
RADAU	sl-ν-1	RKM	$*/*$	F-77		$*/$

Notes:

	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.

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 · indicates that this case is included in a result for a more general case *located* left or above in the table.

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

	DAEs	Routines	Method	Remark
Matlab	ind-1	`ode15{i,s}`	BDF	
Python	—			no built-in functionality, DASSL/IDA wrapped in the modules `assimulo`, `pyDAS`, `DAEtools`
Julia	ind-1	`DifferentialEquations.jl`	BDF	calls SUNDIALS IDA