Lecture Notes for the Course
Numerical methods for time-dependent problems

Walter Zulehner
Institute for Computational Mathematics
Johannes Kepler University Linz

Winter Semester 2007/08

Contents

Chapter 1

Introduction

1.1 Examples

1.2 Standard Forms

Part I

Nonstiff Problems

Chapter 2

Euler's Method

See also: Hairer, Nørsett, Wanner, [8], I.7.

Initial value problem (IVP):

$$\begin{aligned} u'(t) &= f(t, u(t)) \\ u(t_0) &= u_0 \end{aligned}$$

Euler's method:

$$u_{j+1} = u_j + \tau_j \, f(t_j, u_j)$$

Theorem 2.1 (Cauchy, 1789-1857, French mathematician)**.** *Let f be continuous on D, $\|f\|$ bounded by A on D, and f satisfy the Lipschitz condition*

$$\|f(t, w) - f(t, v)\| \leq L \, \|w - v\|$$

on D, with

$$D = \{(t, v) \in \mathbb{R} \times \mathbb{R}^n : t_0 \leq t \leq T, \|v - u_0\| \leq b\}.$$

If $T - t_0 \leq b/A$, then we have:

a) For $|\tau| \to 0$, the Euler polygons converge uniformly to a continuous function $u(t)$.

b) $u(t)$ is continuously differentiable and solves (IVP) on $[t_0, T]$.

c) There is no other solution of (IVP) on $[t_0, T]$.

Chapter 3

Explicit Runge-Kutta Methods

See also: Hairer, Nørsett, Wanner, [8], II.1.

Explicit s-stage Runge-Kutta methods:

$$\begin{aligned}
g_1 &= u_0 \\
g_2 &= u_0 + \tau\, a_{21}\, f(t_0, g_1) \\
g_3 &= u_0 + \tau\, [a_{31}\, f(t_0, g_1) + a_{32}\, f(t_0 + c_2\, \tau)] \\
&\vdots \\
g_s &= u_0 + \tau\, [a_{s1}\, f(t_0, g_1) + a_{s2}\, f(t_0 + c_2\, \tau) + \ldots + a_{s,s-1}\, f(t_0 + c_{s-1}\, \tau, g_{s-1}] \\
u_1 &= u_0 + \tau\, [b_1\, f(t_0, g_1) + b_2\, f(t_0 + c_2\, \tau) + \ldots + b_{s-1}\, f(t_0 + c_{s-1}\, \tau, g_{s-1} + b_s\, f(t_0 + c_s\, \tau, g_s)]
\end{aligned}$$

3.1 Order Conditions

See also: Hairer, Nørsett, Wanner, [8], II.2.

Lemma 3.1 (Leibniz' formula)**.**

$$[\tau \cdot \phi(\tau)]^{(q)} \big|_{\tau=0} = q \cdot \phi^{(q-1)}(0).$$

Theorem 3.1. *The Runge-Kutta method is of order 3 iff*

$$\begin{aligned}
\sum_j b_j &= 1 \\
2 \sum_{j,k} b_j a_{jk} &= 1 \\
3 \sum_{j,k,l} b_j a_{jk} a_{jl} &= 1 \\
6 \sum_{j,k,l} b_j a_{jk} a_{kl} &= 1
\end{aligned}$$

Theorem 3.2 (around 1963)**.** *For $p \geq 5$ no explicit Runge-Kutta method exists of order p with $s \leq p$ stages.*

Theorem 3.3 (Butcher, 1965)**.** *For $p \geq 7$ no explicit Runge-Kutta method exists of order p with $s \leq p + 1$ stages.*

Theorem 3.4 (Butcher, 1985)**.** *For $p \geq 8$ no explicit Runge-Kutta method exists of order p with $s \leq p + 2$ stages.*

Chapter 4

Implicit Runge-Kutta Methods

See also: Hairer, Nørsett, Wanner, [8], II.7.

s-stage Runge-Kutta methods:

$$\begin{aligned} g_1 &= u_0 + \tau\ [a_{11}\, f(t_0 + c_1\,\tau, g_1) + \ldots + a_{1s}\, f(t_0 + c_s\,\tau, g_s)] \\ &\vdots \\ g_s &= u_0 + \tau\ [a_{s1}\, f(t_0 + c_1\,\tau, g_1) + \ldots + a_{ss}\, f(t_0 + c_s\,\tau, g_s] \\ u_1 &= u_0 + \tau\ [b_1\, f(t_0 + c_1\,\tau, g_1) + \ldots + b_s\, f(t_0 + c_s\,\tau, g_s)] \end{aligned}$$

Tableau

$$\begin{array}{c|ccc} c_1 & a_{11} & \ldots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \ldots & a_{ss} \\ \hline & b_1 & \ldots & b_s \end{array} \qquad \text{or, shortly,} \qquad \begin{array}{c|c} c & A \\ \hline & b^T \end{array}$$

Definition 4.1. *An s-stage Runge-Kutta method*

1. *is called explicit, if A is a strictly lower triangular matrix,*
2. *is called implicit, otherwise.*

Fixed point forms:

$$g = \Phi(g; t_0, u_0, \tau) \tag{4.1}$$

with $g = (g_j)_{j=1,\ldots,s}$ and

$$\Phi(g; t_0, u_0, \tau) = \left(u_0 + \tau \sum_{j=1}^{s} a_{ij}\, f(t_0 + c_j\,\tau, g_j) \right)_{i=1,\ldots,s}$$

or, equivalently,

$$k = \Psi(k; t_0, u_0, \tau)$$

with $k = (k_j)_{j=1,\ldots,s}$ and

$$\Psi(g; t_0, u_0, \tau) = \left(f(t_0 + c_i\,\tau, u_0 + \tau \sum_{j=1}^{s} a_{ij}\,k_j) \right)_{i=1,\ldots,s}.$$

Theorem 4.1. *Let f be continuous on D, $\|f\|$ bounded by K on D, and f satisfy the Lipschitz condition*

$$\|f(t,w) - f(t,v)\| \le L\,\|w - v\|$$

on D, with

$$D = \{(t,v) \in \mathbb{R} \times \mathbb{R}^n : t_0 \le t \le T, \|v - u_0\| \le b\}.$$

If $t_0 + c_i\,\tau \in [t_0, T]$ for $i = 1, \ldots, s$, $\tau\,\|A\|_\infty\,K \le b$, and $\tau\,\|A\|_\infty\,L < 1$, then there exists a unique solution to the fixed point equations (4.1) in D and the fixed point iteration converges to this solution for any initial guess in D.

4.1 Order of Consistency

See also: Hairer, Wanner, [9], IV.5.

Theorem 4.2. *If the conditions*

$$\sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k} \qquad k = 1, \ldots, p \tag{4.2}$$

$$\sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k} \qquad i = 1, \ldots, s,\ k = 1, \ldots, q \tag{4.3}$$

$$\sum_{i=1}^{s} b_i c_i^{k-1} a_{ij} = \frac{b_j}{k}(1 - c_j^k) \qquad j = 1, \ldots, s,\ k = 1, \ldots, r \tag{4.4}$$

are satisfied with $p \le q + r + 1$, $p \le 2q + 2$, then the method is of order p.

Lemma 4.1. *Assume that $c_1, \ldots, c_s$ are distinct and $p \ge s$. Then*

a) $B(s + \nu)$ and $C(s)$ imply $D(\nu)$.

b) $B(s + \nu)$ and $D(s)$ imply $C(\nu)$.

Theorem 4.3. *The s-stage Gauß method is of order $2s$.*

Theorem 4.4. *Butcher's s-stage Radau I and Radau II methods are of order $2s - 1$.*

Theorem 4.5. *Butcher's s-stage Lobatto III method is of order $2s - 2$.*

Chapter 5

Convergence Analysis for One-Step Methods

See also: Hairer, Nørsett, Wanner, [8], II.3.

A one-step method for solving the initial value problem:

$$\begin{aligned} u'(t) &= f(t, u(t)), \quad t \in I = [t_0, T], \\ u(t_0) &= u_0 \end{aligned}$$

is a method of the form

$$u_{j+1} = u_j + \tau_j \, \phi(t_j, u_j, \tau_j) \quad \text{for } j = 0, \ldots, m-1.$$

The function ϕ is called the increment function.

Example: The Runge-Kutta methods are obviously one-step methods.

The approximations u_j determine a grid function $u_\tau : I_\tau \longrightarrow \mathbb{R}^n$, given by $u_\tau(t_j) = u_j$, where $I_\tau = \{t_0, t_1, \ldots, t_m\}$. The set of all grid functions on the subdivision τ is denoted by X_τ. Notation: For a grid function $v_\tau \in X_j$ the value $v_\tau(t_j)$ will also be denoted by v_j.

The global error (discretization error) $e_\tau \in X_\tau$ is the grid function, given by

$$e_\tau(t_j)(= e_j) = u(t_j) - u_\tau(t_j) = u(t_j) - u_j.$$

The following norm is introduced on X_τ:

$$\|v_\tau\|_{X_\tau} = \max_{j=0,1,\ldots,m} \|v_j\|.$$

Definition 5.1. *A one-step method is called convergent, if*

$$\|e_\tau\|_{X_\tau} \to 0 \quad \textit{for } |\tau| \to 0.$$

If there is a constant $C \geq 0$ such that

$$\|e_\tau\|_{X_\tau} \leq C \, |\tau|^p \quad (\textit{in short: } \|e_\tau\|_{X_\tau} = O(|\tau|^p)),$$

then the one-step method is called convergent of order p.

The one-step method can be written as

$$F_\tau(u_\tau) = 0$$

with the mapping $F_\tau : X_\tau \longrightarrow X_\tau$, given by

$$F_\tau(v_\tau)(t_{j+1}) = \frac{1}{\tau_j}(v_{j+1} - v_j) - \phi(t_j, v_j, \tau_j) \quad \text{for } j = 0, \ldots, m-1$$

and $F_\tau(v_\tau)(t_0) = v_0 - u_0$.

The consistency error (approximation error, local truncation error) $\psi_\tau(u)$ is the grid function, given by

$$\psi_\tau(u)(t_{j+1}) = \frac{1}{\tau_j}(u(t_{j+1}) - u(t_j)) - \phi(t_j, u(t_j), \tau_j)$$

and $\psi(u)_\tau(t_0) = 0$, or, in short

$$\psi_\tau(u) = F_\tau(R_\tau u)$$

with $R_\tau u = u\big|_{I_\tau}$ (restriction operator).

Definition 5.2. *A one-step method is called consistent with the initial-value problem at u, if*

$$\|\psi_\tau(u)\|_{X_\tau} \to 0 \quad \textit{for } |\tau| \to 0.$$

If a constant $C_S \geq 0$ exists such that

$$\|\psi_\tau(u)\|_{X_\tau} \leq C_S\,|\tau|^p \quad (\textit{in short: } \|\psi_\tau(u)\|_{X_\tau} = O(|\tau|^p))\,,$$

then the one-step method is called consistent of order p.

Theorem 5.1. *Let f be continuous. A one-step method with*

$$\max_{j=0,1,m-1} \|\phi(t_j, u(t_j), \tau_j) - f(t_j, u(t_j))\| \to 0 \quad \textit{for } |\tau| \to 0. \tag{5.1}$$

is consistent.

Proof. The statement easily follows from the representation

$$\psi_\tau(u)(t_j + \tau_j) = \left[\frac{1}{\tau_j}\big(u(t_j + \tau_j) - u(t_j)\big) - u'(t_j)\right] + [f(t_j, u(t_j)) - \phi(t_j, u(t_j), \tau_j)]$$

□

The consistency error is directly related to the local error d_τ, given by

$$d_\tau(u)(t_{j+1})(= d_{j+1}(u)) = u(t_{j+1}) - [u(t_j)) + \tau_j \phi(t_j, u(t_j), \tau_j)]$$

and $d_\tau(u)(t_0) = d_0(u) = 0$. Obviously:

$$\psi_{j+1}(u) = \frac{1}{\tau_j} d_{j+1}(u).$$

If we compare the definition of the consistency error, written in the form

$$\frac{1}{\tau_j}\big(u(t_{j+1}) - u(t_j)\big) - \phi(t_j, u(t_j), \tau_j) = \psi_\tau(u)(t_{j+1}) \quad \text{for } j = 0, \ldots, m-1,$$

with the one-step method

$$\frac{1}{\tau_j}\big(u_{j+1} - u_j\big) - \phi(t_j, u_j, \tau_j) = 0 \quad \text{for } j = 0, \ldots, m-1,$$

we see that the exact solutions at the grid points result from the same one-step method perturbed by the consistency error on the right-hand side.

This leads to the more general question: How does the difference $v_\tau - u_\tau$, where the values v_τ is given by

$$\frac{1}{\tau_j}(v_{j+1} - v_j) - \tau_j \phi(t_j, v_j, \tau_j) = y_{j+1}, \quad j = 0, 1, \ldots, m-1 \tag{5.2}$$

with initial value

$$v_0 = u_0 + y_0 \tag{5.3}$$

or, in short,

$$F_\tau(v_\tau) = y_\tau$$

depend on the perturbation y_τ with $y_\tau(t_j) = y_j$.

Lemma 5.1. *If the increment function satisfies the Lipschitz condition*

$$\|\phi(t, w, \tau) - \phi(t, v, \tau)\| \le \Lambda \|w - v\| \quad \textit{for all } t, v, w \textit{ and all } \tau,$$

then the following estimation is satisfied for (5.3), (5.2):

$$\|v_j - u_j\| \le e^{\Lambda (t_j - t_0)} \|y_0\| + \frac{1}{\Lambda}\left(e^{\Lambda (t_j - t_0)} - 1\right) \max_{i=1,2,\ldots,j} \|y_i\|.$$

Proof. In a first step, we consider only the contribution of the perturbation y_0 to the difference $v_j - u_j$, which is $v_j^{(0)} - u_j$, where $v_\tau^{(0)}$ be given by the one-step method

$$v_{j+1}^{(0)} = v_j^{(0)} + \tau_j \phi(t_j, v_j^{(0)}, \tau_j), \quad j = 0, 1, \ldots, m-1$$

with

$$v_0^{(0)} = u_0 + y_0.$$

From

$$v_{j+1}^{(0)} - u_j = v_j^{(0)} - u_j + \tau_j[\phi(t_j, v_j^{(0)}, \tau_j) - \phi(t_j, u_j, \tau_j)]$$

and the Lipschitz condition it follows that

$$\|v_{j+1}^{(0)} - u_j\| \leq (1 + \Lambda\,\tau_j)\|v_j^{(0)} - u_j\| \leq e^{\Lambda\,\tau_j}\|v_j^{(0)} - u_j\|.$$

Hence

$$\|v_j^{(0)} - u_j\| \leq e^{\Lambda\,\tau_{j-1}} e^{\Lambda\,\tau_{j-2}} \cdots e^{\Lambda\,\tau_0}\|v_0^{(0)} - u_j\| = e^{\Lambda\,(t_j - t_0)}\|y_0\|.$$

Next we consider the contribution of the perturbation y_1 to the difference $v_j - u_j$, which is $v_j^{(1)} - v_j^{(0)}$, where $v_\tau^{(1)}$ be given by the one-step method

$$v_{j+1}^{(1)} = v_j^{(1)} + \tau_j \phi(t_j, v_j^{(1)}, \tau_j), \quad j = 1, \ldots, , m-1$$

with

$$v_1^{(1)} = v_1.$$

It follows analogously

$$\|v_j^{(1)} - v_j^{(0)}\| \leq e^{\Lambda\,(t_j - t_1)}\|v_1^{(1)} - v_1^{(0)}\| = e^{\Lambda\,(t_j - t_1)}\tau_0\|y_1\|.$$

In general, we obtain for the contribution of the perturbation y_i, $i = 1, \ldots, j$:

$$\|v_j^{(i)} - v_j^{(i-1)}\| \leq e^{\Lambda\,(t_j - t_i)}\|v_i^{(i)} - v_i^{(i-1)}\| = e^{\Lambda\,(t_j - t_i)}\tau_{i-1}\|y_i\|,$$

where $v_\tau^{(i)}$ is given by the one-step method

$$v_{j+1}^{(i)} = v_j^{(i)} + \tau_j \phi(t_j, v_j^{(i)}, \tau_j), \quad j = i, \ldots, m-1$$

with

$$v_i^{(i)} = v_i.$$

Then, for the difference

$$v_j - u_j = (v_j^{(j)} - v_j^{(j-1)}) + (v_j^{(j-1)} - v_j^{(j-2)}) + \ldots (v_j^{(1)} - v_j^{(0)}) + (v_j^{(0)} - u_j)$$

we obtain the estimate

$$\begin{aligned}
\|v_j - u_j\| &\leq \|v_j^{(j)} - v_j^{(j-1)}\| + \|v_j^{(j-1)} - v_j^{(j-2)}\| + \ldots \|v_j^{(1)} - v_j^{(0)}\| + \|v_j^{(0)} - u_j\| \\
&\leq \left[\tau_{j-1} + e^{\Lambda(t_j - t_{j-1})}\tau_{j-2} + \cdots + e^{\Lambda(t_j - t_1)}\tau_0\right] \max_{i=1,2,\ldots,n} \|y_i\| + e^{\Lambda(t_j - t_0)}\|y_0\| \\
&\leq \left[\int_{t_{j-1}}^{t_j} e^{\Lambda(t_j - t)}\, dt + \int_{t_{j-2}}^{t_{j-1}} e^{\Lambda(t_j - t)}\, dt + \cdots + \int_{t_0}^{t_1} e^{\Lambda(t_j - t)}\, dt\right] \max_{i=1,2,\ldots,n} \|y_i\| \\
&\quad + e^{\Lambda(t_j - t_0)}\|y_0\| = \frac{1}{\Lambda}\left(e^{\Lambda(t_j - t_0)} - 1\right) \max_{i=1,2,\ldots,m} \|y_i\| + e^{\Lambda(t_j - t_0)}\|y_0\|.
\end{aligned}$$

□

Theorem 5.2. *If the increment function satisfies the Lipschitz condition*

$$\|\phi(t,w,\tau)-\phi(t,v,\tau)\| \le \Lambda \|w-v\| \quad \textit{for all } t,v,w \textit{ and all } \tau,$$

then a constant $C \ge 0$ exists with

$$\|v_\tau - u_\tau\|_{X_\tau} \le C \|F_\tau(v_\tau) - F_\tau(u_\tau)\|_{X_\tau} \quad \textit{for all } v_\tau \in X_\tau \textit{ and all } \tau.$$

Proof. For $y_\tau = F_\tau(v_\tau)$ it follows from Lemma 5.1

$$\max_{j=0,1,\dots,m} \|v_j - u_j\| \le C \max_{j=0,1,\dots,m} \|y_j\|$$

with

$$C = e^{\Lambda(T-t_0)} + \frac{1}{\Lambda}\left(e^{\Lambda(T-t_0)} - 1\right).$$

□

Definition 5.3. *A one-step method is called stable at u_τ if a constant $C_S \ge 0$ exists with*

$$\|v_\tau - u_\tau\|_{X_\tau} \le C_S \|F_\tau(v_\tau) - F_\tau(u_\tau)\|_{X_\tau} \quad \textit{for all } v_\tau \in X_\tau \textit{ and all } \tau.$$

Theorem 5.3. *If a one-step method is consistent (of order p) at the exact solution u and stable at the approximate solution u_τ, then the method is convergent (of order p).*

Proof. The restriction $R_\tau u$ of the exact solution u on I_τ satisfies

$$F_\tau(R_\tau u) = \psi_\tau(u).$$

The approximate solution u_τ satisfies

$$F_\tau(u_\tau) = 0.$$

From the stability it follows that

$$\|R_\tau u - u_\tau\|_{X_\tau} \le C_S \|\psi_\tau(u)\|_{X_\tau}.$$

From the consistency it follows

$$\|\psi_\tau(u)\|_{X_\tau} \to 0 \quad \text{for } |\tau| \to 0$$

and, therefore,

$$\|e_\tau\|_{X_\tau} \to 0 \quad \text{for } |\tau| \to 0.$$

From

$$\|\psi_\tau(u)\|_{X_\tau} \le C_A |\tau|^p$$

it follows

$$\|e_\tau\|_{X_\tau} \le C_S C_A |\tau|^p.$$

□

Example: The Runge-Kutta methods are one-step methods with

$$\phi(t,u,\tau)=\sum_{i=1}^{s} b_i f(t+c_i\tau, g_i(t,u,\tau)),$$

where $g(t,u,\tau)$ is the solution of the fixed point equation

$$g=\Phi(g;t,u,\tau).$$

Theorem 5.4. *Let f be continuous and satisfies the Lipschitz condition*

$$\|f(t,v)-f(t,w)\|\le L\,\|v-w\| \quad \textit{for all } t,v,w.$$

If

$$\sum_{j=1}^{s} b_j = 1,$$

then the Runge-Kutta method is consistent with the initial value problem.

Proof. Let $g_i=g_i(t,u(t),\tau)$. From

$$\begin{aligned} g_i-u(t) &= \tau\sum_{j=1}^{s} a_{ij} f(t+c_j\tau, g_j) \\ &= \tau\sum_{j=1}^{s} a_{ij}[f(t+c_j\tau,g_j)-f(t+c_j\tau,u(t)] + \tau\sum_{j=1}^{s} a_{ij} f(t+c_j\tau,u(t)) \end{aligned}$$

it follows

$$\max_{i=1,2,\ldots,s}\|g_i-u(t)\|\le \tau\,L\,\|A\|_\infty \max_{i=1,2,\ldots,s}\|g_i-u(t)\| + \tau\,\|A\|_\infty M$$

with $M=\sup_{s,t\in I}\|f(s,u(t)\|$. Therefore,

$$\max_{i=1,2,\ldots,s}\|g_i-u(t)\|\le \frac{\tau\,\|A\|_\infty M}{1-\tau\,L\,\|A\|_\infty}.$$

From

$$\begin{aligned} \phi(t,u(t),\tau)-f(t,u(t)) &= \sum_{i=1}^{s} b_i[f(t+c_i\tau,g_i)-f(t+c_i\tau,u(t))] \\ &\quad + \sum_{i=1}^{s} b_i[f(t+c_i\tau,u(t))-f(t,u(t))] \end{aligned}$$

it follows

$$\|\phi(t,u(t),\tau)-f(t,u(t))\|\le \frac{\tau\,L\,\|b\|_1\|A\|_\infty M}{1-\tau\,L\,\|A\|_\infty} + \|b\|_1 \max_{i=1,2,\ldots,s}\|f(t+c_i\tau,u(t))-f(t,u(t))\|$$

and, therefore,

$$\max_{j=0,1,\ldots,m-1} \|\phi(t_j, u(t_j), \tau_j) - f(t_j, u(t_j))\| \to 0 \quad \text{for } |\tau| \to 0,$$

since $f(\tau, u(t))$ is uniformly continuous. The rest follows from Lemma 5.1. □

Remark: For explicit Runge-Kutta method the Lipschitz condition is not needed for the proof of consistency. For proving some order of consistency for a Runge-Kutta method it suffices to assume that f is sufficiently smooth.

Under the assumptions of Theorem 5.4 the stability also follows:

Theorem 5.5. *Let f be continuous and satisfies the Lipschitz condition*

$$\|f(t, w) - f(t, v)\| \le L\|w - v\| \quad \textit{for all } t, v, w.$$

Then the Runge-Kutta method is stable.

Proof. Let $g_i^{(1)} = g_i(t, v, h)$ and $g_i^{(2)} = g_i(t, w, h)$. Then

$$\|\phi(t, w, \tau) - \phi(t, v, \tau)\| \le \|b\|_1 \, L \, \|g^{(2)} - g^{(1)}\|.$$

and

$$\|g^{(2)} - g^{(1)}\| \le \|w - v\| + \tau \, L \, \|A\|_\infty \, \|g^{(2)} - g^{(1)}\|,$$

hence

$$\|g^{(2)} - g^{(1)}\| \le \frac{1}{1 - \tau \, L \, \|A\|_\infty} \|w - v\|.$$

This leads to

$$\|\phi(t, w, \tau) - \phi(t, v, \tau)\| \le \frac{L \, \|b\|_1}{1 - \tau \, L \, \|A\|_\infty} \|w - v\|.$$

The rest follows from Theorem 5.2. □

Theorem 5.6. *Let f be continuous and satisfies the Lipschitz condition*

$$\|f(t, w) - f(t, v)\| \le L\|w - v\| \quad \textit{for all } t, v, w.$$

If

$$\sum_{j=1}^{s} b_j = 1,$$

then the Runge-Kutta method is convergent.

Remark: The results of this chapter can also be shown under the local Lipschitz condition

$$\|f(t, v) - f(t, w)\| \le L\|v - w\| \quad \text{for all } (t, v), (t, w) \in U,$$

where $U \subset I \times \mathbb{R}^n$ is a neighborhood of the graph of f, given by $\{(t, f(t, u(t))) : t \in I\}$, and $u(t)$ denotes the exact solution of the initial value problem.

For the local variant of Theorem 5.2 a local Lipschitz condition suffices:

$$\|\phi(t, v, \tau) - \phi(t, w, \tau)\| \le \Lambda\|v - w\| \quad \text{for all } (t, v), (t, w) \in U, \ \tau \le \bar{\tau},$$

if it is additionally assumed that the method is consistent. The stability estimate can be shown for all $v_\tau \in X_\tau$ with $\|F_\tau(v_\tau)\|_{X_\tau} \le \eta$, if η is sufficiently small.

Chapter 6

Practical Computation

See also: Hairer, Nørsett, Wanner, [8], II.4, II.6.

The right choice of the step sizes is of great importance for the efficiency of a one-step method. The aim of a step size control is to achieve a prescribed tolerance of the local error. In the next section two different approaches are discussed how to approximately compute the local error. Subsequently, an automatic step size control is presented. Finally, the question is discussed how to efficiently calculate approximate solutions at prescribed points.

6.1 Error estimation

We assume that the local error can be represented in the following form:

$$u(t_0+\tau)-u_1 = \underbrace{C(t_0,u_0)\,\tau^{p+1}}_{\text{principal error term}} + O(\tau^{p+2}) \tag{6.1}$$

with $C(t_0,u_0)$ independent of the step size τ.

Examples: For the explicit Euler method we have:

$$u(t_0+\tau)-u_1 = \frac{1}{2}(f_t+f_u f)(t_0,u_0)\tau^2 + O(\tau^3).$$

For the explicit midpoint rule we have:

$$u(t_0+\tau)-u_1 = \frac{1}{24}(f_{tt}+2f_{tu}f+f_{uu}f^2+4(f_u f_t+f_u^2 f))(t_0,u_0)\tau^3 + O(\tau^4).$$

In the following we assume that the coefficient $C(t_0,u_0)$ is sufficiently smooth in t_0 and u_0, but we will not make use of the particular form of this coefficient.

6.1.1 Richardson extrapolation

After two steps of the method with step size τ we obtain approximations u_1 and u_2. The error $u(t_0+2\tau)-u_2$ consists of two contributions: the error after the first step $e_1 = u(t_0+\tau)-u_1$, transported to $t_0+2\tau$ by the differential equation: $u(t_0+2\tau)-u^{(1)}(t_0+2\tau)$, where $u^{(1)}(t)$ is the exact solution of the differential equation with initial value $u^{(1)}(t_0+\tau)=u_1$ and the new local error $u^{(1)}(t_0+2\tau)-u_2$. Now

$$u(t_0+2\tau)-u^{(1)}(t_0+2\tau) = e_1 + O(\tau\|e_1\|) = C(t_0,u_0)\tau^{p+1} + O(\tau^{p+2})$$

and

$$u^{(1)}(t_0+2\tau)-u_2 = C(t_0+\tau,u_1)\tau^{p+1} + O(\tau^{p+2}) = C(t_0,u_0)\tau^{p+1} + O(\tau^{p+2}).$$

So

$$u(t_0+2\tau)-u_2 = 2C(t_0,u_0)\tau^{p+1} + O(\tau^{p+2}). \tag{6.2}$$

We start again at (t_0,u_0) and compute one step with step size 2τ leading to the approximation w at $t_0+2\tau$ with

$$u(t_0+2\tau)-w = C(t_0,u_0)(2\tau)^{p+1} + O(\tau^{p+2}). \tag{6.3}$$

Next we use (6.3) and (6.2) in order to eliminate $C(t_0,u_0)$ and obtain a more accurate approximation of the exact solution:

$$u(t_0+2\tau) = \hat{u}_2 + O(\tau^{p+2}) \quad \text{with} \quad \hat{u}_2 = u_2 + \frac{u_2-w}{2^p-1}.$$

This construction can also interpreted in the following way: (6.3) and (6.2) show that the approximations w and u_2 are very close to the values of the polynomial $q(s) = u(t_0+2\tau) - 2\tau C(t_0,u_0)\,s^p$ at $s=2\tau$ and $s=\tau$, respectively. Let $p(s)$ be the interpolation polynomial $p(s)=A+Bs^p$, determined by

$$p(2\tau) = w \quad \text{and} \quad p(\tau) = u_2.$$

It is natural to expect the value $p(0)$ is very close to $q(0)=u(t_0+\tau)$. One easily sees that this extrapolated value $p(0)$ is equal to $\hat{u}_2$.

For the local error we obtain:

$$u(t_0+\tau)-u_2 = \hat{u}_2-u_2+O(\tau^{p+2}) \quad \text{with} \quad \hat{u}_2-u_2 = \frac{u_2-w}{2^p-1},$$

which leads to the following estimation of the local error:

$$err = \frac{1}{2^p-1} max_{i=1,\ldots,n} \frac{|u_{2,i}-w_i|}{d_i},$$

where d_i is an appropriate scaling factor. Typical values: $d_i = 1$ for absolute errors, $d_i = |\hat{u}_{2,i}|$ for componentwise relative errors.

If this step is accepted by the step size control (see later) one continues the calculation at $t_0+2\tau$ either with u_2 or with the better approximation $\hat{u}_2$. The later technique is called local extrapolation. Next we discuss a different method for estimating the local error, which makes the rejection of a step size less expensive.

6.1.2 Embedded Runge-Kutta methods

Consider a Runge-Kutta method of order p. Then, for the local error, we have:

$$u(t_0 + \tau) - u_1 = O(\tau^{p+1}),$$

The basic idea for estimating the local error is to use a second Runge-Kutta method of higher order q, leading to

$$u(t_0 + \tau) - \hat{u}_1 = O(\tau^q).$$

Since $q > p$, it follows that

$$u(t_0 + \tau) - u_1 = \hat{u}_1 - u_1 + O(\tau^q),$$

which leads to the following estimation of the local error:

$$err = max_{i=1,\ldots,n} \frac{|\hat{u}_{1,i} - u_{1,i}|}{d_i}.$$

In order to keep the extra computational costs low we assume that the two Runge-Kutta methods have the same coefficients c and A and differ only in the last row (b and $\hat{b}$). A pair of such methods (called embedded Runge-Kutta methods) are usually represented by one tableau for the coefficients A, b, c with an extra row for $\hat{b}$. For an explicit method the tableau has the form:

$$\begin{array}{c|ccccc} 0 & & & & & \\ c_2 & a_{21} & & & & \\ c_3 & a_{31} & a_{32} & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ c_s & a_{s1} & a_{s2} & \ldots & a_{s,s-1} & \\ \hline & b_1 & b_2 & \ldots & b_{s-1} & b_s \\ \hline & \hat{b}_1 & \hat{b}_2 & \ldots & \hat{b}_{s-1} & \hat{b}_s \end{array}$$

The two approximate solutions are given by

$$u_1 = u_0 + \tau \left(b_1 k_1 + b_2 k_2 + \cdots + b_s k_s\right)$$

and

$$\hat{u}_1 = u_0 + \tau \left(\hat{b}_1 k_1 + \hat{b}_2 k_2 + \cdots + \hat{b}_s k_s\right).$$

Example: We start with a general explicit 3-stage Runge-Kutta method:

$$\begin{array}{c|ccc} 0 & & & \\ c_2 & a_{21} & & \\ c_3 & a_{31} & a_{32} & \\ \hline & b_1 & b_2 & b_3 \\ \hline & \hat{b}_1 & \hat{b}_2 & \hat{b}_s \end{array}$$

The conditions for order 2 for the first method are:

$$\begin{aligned} b_1 + b_2 + b_3 &= 1, \\ b_2 c_2 + b_3 c_3 &= \frac{1}{2}. \end{aligned}$$

The conditions for order 3 for the second method are:

$$\begin{aligned} \hat{b}_1 + \hat{b}_2 + \hat{b}_3 &= 1, \\ \hat{b}_2 c_2 + \hat{b}_3 c_3 &= \frac{1}{2}, \\ \hat{b}_2 c_2^2 + \hat{b}_3 c_3^2 &= \frac{1}{3}, \\ \hat{b}_3 a_{32} c_2 &= \frac{1}{6}. \end{aligned}$$

The choice

$$c_2 = 1, \quad c_3 = \frac{1}{2}, \quad b_3 = 0$$

leads to a so called Runge-Kutta-Fehlberg method abbreviated by RKF 2(3). The symbol 2(3) (in general $p(q)$) means that the basic method is of order 2 (p), the second method used for estimating the local error is a method of order 3 (q). The tableau for RKF 2(3) is given by:

0			
1	1		
1/2	1/4	1/4	
	1/2	1/2	0
	1/6	1/6	4/6

The weights b_i correspond to the trapezoidal rule, the weights $\hat{b}_i$ to Simpson's rule.

Example: Another important example of an embedded explicit Runge-Kutta method was constructed by Dormand and Prince, (in short: DOPRI (4)5), whose tableau is given by:

0							
$\frac{1}{5}$	$\frac{1}{5}$						
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$					
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$				
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$			
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$		
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	
	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0
	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$

Observe that $a_{sj} = b_j$ which additionally reduces the computational work.

It is reasonable to continue the computation not with u_1 but with the more accurate approximation $\hat{u}_1$.

6.2 Step size control

We assume that an estimation err of the local error is available and that

$$err = C\,\tau^{p+1}.$$

The aim is to keep the local error within a given tolerance tol. This leads to an optimal step size τ_{neu}, satisfying the relation

$$tol = C\,\tau_{new}^{p+1}.$$

From these two conditions the unknown constant C can be eliminated and one obtains:

$$\tau_{new} = \tau\,(tol/err)^{1/(p+1)}. \tag{6.4}$$

This motivates the following strategy for a step size selection:

1. One step with a given step size h is computed together with the estimation err for the local error.
2. If $err \leq tol$, the step is accepted and the method is continued with the next step size τ_{neu}, given by (6.4).
3. Otherwise, the step is rejected and the method is restarted with the new step size τ_{neu}, given by (6.4).

In order to be sure that the new step size produces a local error below tol the optimal step size is reduced by a safety factor fac (e.g.: $fac = 0.8$). Additionally, it is reasonable to limit the change of the step size from τ to τ_{neu} by factors $facmax$ for the maximal relative increase and $facmin$ for the maximal decrease, in order to prevent too dramatic changes of step sizes. With these modifications the new formula for the optimal step size becomes

$$\tau_{new} = \tau \cdot \min(facmax, \max(facmin, fac \cdot (tol/err)^{1/(p+1)})).$$

Additionally, it is advisable to set $facmax = 1$ right after a step rejection.

6.3 Dense output

It is often required to compute the approximate solution on a set of prescribed points without interfering with the steps size control. This can be done by so called continuous Runge-Kutta methods: These methods contain a parameter $\theta \in (0, 1]$ and allow the computation of approximations for $u(t_0 + \theta\,\tau)$. For $\theta = 1$ the original Runge-Kutta method is

obtained. For efficiency reasons we assume that the coefficients c and A are independent of θ sind. Only the coefficients of b are allowed to depend on θ. Approximate solutions at prescribed points can be computed without extra function evaluations and with no influence on the step size control.

Example: An explicit 3-stage Runge-Kutta method for approximating $u(t_0 + \theta\,\tau)$ for all $\theta \in (0, 1]$ is of order 3 iff the following conditions are satisfied:

$$\begin{aligned} b_1 + b_2 + b_3 &= \theta, \\ b_2 c_2 + b_3 c_3 &= \frac{\theta^2}{2}, \\ b_2 c_2^2 + b_3 c_3^2 &= \frac{\theta^3}{3}, \\ b_3 a_{32} c_2 &= \frac{\theta^3}{6}. \end{aligned}$$

This is not possible for coefficients c_2, c_3 and a_{32} independent of θ. Instead we require order 3 only for $\theta = 1$ and order 2, otherwise. This guarantees a global error of order 3, also at intermediate points. For $c_2 = 1/2$ and $c_3 = 1$ one obtains the following tableau of a continuous Runge-Kutta method:

0			
$\frac{1}{2}$	$\frac{1}{2}$		
1	-1	2	
	$\theta(1 + \theta(-\frac{3}{2} + \frac{2}{3}\theta))$	$\theta^2(2 - \frac{2}{3}\theta)$	$\theta^2(\frac{3}{2} - \frac{1}{2}\theta)$

Chapter 7

Extrapolation Methods

7.1 Asymptotic Expansions

See also: Hairer, Nørsett, Wanner, [8], II.8.

Theorem 7.1. *Assume that f and ϕ are sufficiently smooth and satisfy the consistency condition $\phi(t,v,0) = f(t,v)$. If the local error $d_\tau(t+\tau) = u(t+\tau) - [u(t) + \tau\,\phi(t,u(t),\tau)]$ possesses an expansion*

$$d_\tau(t+\tau) = d_{p+1}(t)\,\tau^{p+1} + O(\tau^{p+2})$$

with a continuous function $d_{p+1}(t)$, then the global error possesses an expansion

$$u_\tau(t) - u(t) = e_p(t)\,\tau^p + O(\tau^{p+1}),$$

where $e_p(t)$ solves the initial value problem

$$\begin{aligned} e'(t) &= f_u(t,u(t))\,e(t) - d_{p+1}(t) \\ e(0) &= 0. \end{aligned}$$

Theorem 7.2 (Gragg)**.** *Assume that f and ϕ are sufficiently smooth and satisfy the consistency condition $\phi(t,v,0) = f(t,v)$. If the local error $d_\tau(t+\tau) = u(t+\tau) - [u(t) + \tau\,\phi(t,u(t),\tau)]$ possesses an expansion*

$$d_\tau(t+\tau) = d_{p+1}(t)\,\tau^{p+1} + \ldots + d_{q+1} + O(\tau^{q+2})$$

with continuous functions $d_{p+1}(t)$, $\ldots$, $d_{q+1}(t)$, then the global error possesses an expansion

$$u_\tau(t) - u(t) = e_p(t)\,\tau^p + \ldots + e_q(t)\,\tau^q + O(\tau^{q+1}),$$

where $e_p(t)$,$\ldots$,$e_q(t)$ solve initial value problems with

$$e_p(0) = \ldots = e_q(0) = 0.$$

Definition 7.1. *Let $\phi(t, v, \tau)$ be the increment function of a one-step method. The increment function $B = \phi^*(t, A, \tau)$ of the adjoint function is given by the condition*

$$B = A - \tau\, \phi(t + \tau, A, -\tau).$$

Theorem 7.3. *The coefficients A^*, b^*, c^* of the adjoint method of a Runge-Kutta method with coefficients A, b, c are given by*

$$\begin{aligned} a^*_{ij} &= b_{s+1-i} - a_{s+1-i,s+1-j}, \\ b^*_j &= b_{s+1-j}, \\ c^*_i &= 1 - c_{s+1-i}. \end{aligned}$$

Theorem 7.4. *The adjoint method is a method of the same order as the original method. Its principal error term is equal to the principal error term of the original method multiplied by $(-1)^p$.*

Theorem 7.5. *The adjoint method has exactly the same asymptotic expansion for the global error as the original method with τ replaced by $-\tau$.*

Definition 7.2. *A method is symmetric if $\phi^* = \phi$.*

Theorem 7.6. *A Runge-Kutta method is symmetric if*

$$a_{s+1-i,s+1-j} + a_{ij} = b_{s+1-j} = b_i.$$

Theorem 7.7. *If, in addition to the assumptions of Theorem 7.2, the method is symmetric, then*

$$e_r(t) = 0 \quad \textit{for } r \textit{ odd}.$$

7.2 Polynomial Extrapolation

See also: Hairer, Nørsett, Wanner, [8], II.9.

Theorem 7.8. *The values $T_{j,k}$ (as a global extrapolation method) represent a numerical method of order $p + k - 1$.*

Theorem 7.9. *The values $T_{j,k}$ (as a local extrapolation method) represent a numerical method of order $p + k - 1$.*

7.3 The GBS Method

See also: Hairer, Nørsett, Wanner, [8], II.9.

Chapter 8

Multistep Methods

A method is called a multistep method (more precisely a k-step method) if the computation of the next approximate solution u_{j+1} is based on the last approximate solutions $u_j,\ u_{j-1}, \ldots, u_{j-k+1}$. If it is more convenient we will also use the notations u_{j+k} for the new approximate solution and $u_{j+k-1}, \ldots, u_j$ for the previously computed approximate solutions.

In order to perform a k-step method, a starting procedure has to be done first to compute $u_0, u_1, \ldots, u_{k-1}$. The starting procedure can be done, e.g., by using a one-step method (with small step sizes) or by a multistep method with a growing number of steps.

8.1 Classical Linear Multistep Methods

See also: Hairer, Nørsett, Wanner, [8], III.1.

8.1.1 Explicit Adams Methods

We know that

$$u(t_{j+1}) = u(t_j) + \int_{t_j}^{t_{j+1}} f(t, u(t))\ dt \tag{8.1}$$

for a solution u of the ODE

$$u'(t) = f(t, u(t)).$$

The Runge-Kutta methods are based on quadrature rules whose nodes are typically inside the interval $[t_j, t_{j+1}]$ and, therefore, require function evaluations at additional points. If instead the grid points $t_j, t_{j-1}, \ldots, t_{j-k+1}$ are used as nodes for a quadrature rule, no additional function evaluations are required.

One possible strategy is to replace the function $f(t, u(t))$ in (8.1) by an interpolation polynomial. For simplicity we restrict ourselves to the case of equidistant step sizes: The interpolation polynomial with the nodes

$$t_i, \quad i = j-k+1, \ldots, j-1, j,$$

and the values

$$f_i = f(t_i, u_i), \quad i = j-k+1, \ldots, j-1, j,$$

can be written in the following form (Newton's interpolation formula):

$$p(t) = p(t_j + s\,\tau) = \sum_{i=0}^{k-1} (-1)^i \binom{-s}{i} \nabla^i f_j,$$

with

$$\binom{-s}{0} = 1, \quad \binom{-s}{i} = \frac{(-s)(-s-1)\cdots(-s-i+1)}{i!} \quad \text{for } i \geq 1.$$

∇ denotes the backward difference:

$$\nabla f_j = f_j - f_{j-1},$$

whose powers ∇ are given by:

$$\nabla^0 = I, \quad \nabla^{i+1} = \nabla^i \nabla.$$

If $f(t, u(t))$ is replaced by $p(t)$ in (8.1), we obtain the following class of explicit multistep methods (explicit Adams methods, Adams-Bashforth methods):

$$u_{j+1} = u_j + \tau \sum_{i=0}^{k-1} \gamma_i \nabla^i f_j$$

with

$$\gamma_i = (-1)^i \int_0^1 \binom{-s}{i}\, ds.$$

The following table shows a few values of γ_i:

i	0	1	2	3	4	5	6	7	8
γ_i	1	$\frac{1}{2}$	$\frac{5}{12}$	$\frac{3}{8}$	$\frac{251}{720}$	$\frac{95}{288}$	$\frac{19087}{60480}$	$\frac{5257}{17280}$	$\frac{1070017}{3628800}$

The first three Adams-Bashforth methods are:

$$\begin{aligned}
k=1: \quad u_{j+1} &= u_j + \tau\, f_j, \\
k=2: \quad u_{j+1} &= u_j + \tau \left[\frac{3}{2} f_j - \frac{1}{2} f_{j-1}\right], \\
k=3: \quad u_{j+1} &= u_j + \tau \left[\frac{23}{12} f_j - \frac{16}{12} f_{j-1} + \frac{5}{12} f_{j-2}\right].
\end{aligned}$$

For $k = 1$ one obtains Euler's method.

8.1.2 Implicit Adams Methods

In this class of multistep methods, the node t_{j+1} is also used for the interpolation. Then the interpolation polynomial has the form:

$$p^*(t) = p^*(t_j + s\,\tau) = p(t_{j+1} + (s-1)\,\tau) = \sum_{i=0}^{k} (-1)^i \binom{-s+1}{i} \nabla^i f_{j+1}.$$

The corresponding quadrature is given by:

$$u_{j+1} = u_j + \tau \sum_{i=0}^{k} \gamma_i^* \nabla^i f_{j+1} \tag{8.2}$$

with

$$\gamma_i^* = (-1)^i \int_0^1 \binom{-s+1}{i}\, ds.$$

These methods are implicit and require the solution of a (in general nonlinear) system of equations, in order to compute u_{j+1}. For sufficiently small step sizes the solution u_{j+1} exists and the method is well-defined. An approximation for u_{j+1} can be obtained, e.g., by a fixed point iteration for (8.2) or by Newton's method. As an initial guess one can use u_j or the result of one step of the corresponding explicit Adams method. Often, it suffices to perform one step of the explicit Adams method (predictor) followed by one step on an iterative method for the implicit Adams method (corrector).

The following table contains a few values for γ_i^*:

i	0	1	2	3	4	5	6	7	8
γ_i	1	$-\frac{1}{2}$	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{720}$	$-\frac{3}{160}$	$-\frac{863}{60480}$	$-\frac{275}{24192}$	$-\frac{33953}{3628800}$

The first three Adams-Moulton methods are:

$$\begin{aligned}
k=0: \quad u_{j+1} &= u_j + \tau\, f_{j+1}, \\
k=1: \quad u_{j+1} &= u_j + \tau \left[\frac{1}{2} f_{j+1} + \frac{1}{2} f_j\right], \\
k=2: \quad u_{j+1} &= u_j + \tau \left[\frac{5}{12} f_{j+1} + \frac{8}{12} f_j - \frac{1}{12} f_{j-1}\right].
\end{aligned}$$

For $k = 0$ one obtains the implicit Euler method, for $k = 1$ the implicit trapezoidal rule.

8.1.3 Explicit Nyström Methods

These methods are based on the relation

$$u(t_{j+1}) = u(t_{j-1}) + \int_{t_{j-1}}^{t_{j+1}} f(t, u(t))\; dt.$$

Using polynomial interpolation without the node t_{j+1} the following class of explicit methods are obtained:

$$u_{j+1} = u_{j-1} + \tau \sum_{i=0}^{k-1} \kappa_i \nabla^i f_j$$

with

$$\kappa_i = (-1)^i \int_{-1}^{1} \binom{-s}{i} \, ds.$$

The following table contains a few values for κ_i:

i	0	1	2	3	4	5	6
κ_i	2	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{29}{90}$	$\frac{14}{45}$	$\frac{1139}{3780}$

For $k = 1$ and $k = 3$ one obtains the Nyström methods:

$$\begin{aligned} k = 1: \quad u_{j+1} &= u_{j-1} + 2\tau f_j, \\ k = 3: \quad u_{j+1} &= u_{j-1} + \tau \left[\frac{7}{3} f_j - \frac{2}{3} f_{j-1} + \frac{1}{3} f_{j-2} \right]. \end{aligned}$$

For $k = 1$ one obtains the explicit midpoint rule. The case $k = 2$ is identical to the case $k = 1$.

8.1.4 Milne-Simpson Methods

The implicit variants of the Nyström methods are:

$$u_{j+1} = u_{j-1} + \tau \sum_{i=0}^{k} \kappa_i^* \nabla^i f_{j+1}$$

with

$$\kappa_i^* = (-1)^i \int_{-1}^{1} \binom{-s+1}{i} \, ds.$$

The following table contains a few values for κ_i^*:

i	0	1	2	3	4	5	6
κ_i^*	2	-2	$\frac{1}{3}$	0	$-\frac{1}{90}$	$-\frac{1}{90}$	$-\frac{37}{3780}$

The first three Milne-Simpson methods are:

$$\begin{aligned} k = 0: \quad u_{j+1} &= u_{j-1} + 2\tau \, f_{j+1}, \\ k = 1: \quad u_{j+1} &= u_{j-1} + 2\tau \, f_j, \\ k = 2: \quad u_{j+1} &= u_{j-1} + \tau \left[\frac{1}{3} f_{j+1} + \frac{4}{3} f_j + \frac{1}{3} f_{j-1} \right]. \end{aligned}$$

$k = 0$ corresponds to the implicit Euler method with doubled step size, for $k = 1$ one obtains the explicit midpoint rule and for $k = 2$ the Simpson rule.

8.1.5 BDF-Methods

This class of methods is based on numerical differentiation: The interpolation polynomial q with nodes

$$t_i, \quad i = j - k + 1, \ldots, j, j + 1,$$

and values

$$u_i \quad i = j - k + 1, \ldots, j, j + 1,$$

has the following form:

$$q(t) = q(t_j + s\,\tau) = \sum_{i=0}^{k} (-1)^i \binom{-s+1}{i} \nabla^i u_{j+1}.$$

The differential equation

$$u'(t) = f(t, u(t))$$

at $t = t_{j+1}$ is replaced by

$$q'(t_{j+1}) = f_{j+1}.$$

This leads to a multistep method:

$$\sum_{i=0}^{k} \delta_i^* \nabla_i u_{j+1} = \tau\, f_{j+1}$$

with

$$\delta_i^* = (-1)^i \left. \frac{d}{ds} \binom{-s+1}{i} \right|_{s=1}.$$

This method is called backward differencing formula (BDF-method).

The values δ_i^* can be easily calculated:

$$\delta_0^* = 0, \qquad \delta_i^* = \frac{1}{i} \quad \text{for } i \geq 1.$$

The first three BDF-methods are:

$$\begin{aligned}
k = 1: &\quad u_{j+1} - u_j = \tau\, f_{j+1} \\
k = 2: &\quad \frac{3}{2} u_{j+1} - 2u_j + \frac{1}{2} u_{j-1} = \tau\, f_{j+1} \\
k = 3: &\quad \frac{11}{6} u_{j+1} - 3u_j + \frac{3}{2} u_{j-1} - \frac{1}{3} u_{j-2} = \tau\, f_{j+1}
\end{aligned}$$

These methods are implicit.

8.2 Consistency of Linear Multistep Methods

See also: Hairer, Nørsett, Wanner, [8], III.2.

All multistep methods discussed so far are of the following form:

$$\alpha_k u_{j+k} + \alpha_{k-1} u_{j+k-1} + \cdots + \alpha_0 u_j = \tau \left(\beta_k f_{j+k} + \beta_{k-1} f_{j+k-1} + \cdots + \beta_0 f_j\right). \tag{8.3}$$

A multistep method of this form is called a linear multistep method, more precisely, a linear k-step method.

The coefficient α_i and β_i are not uniquely determined by the method. They allow an additional scaling condition, e.g., $\alpha_k = 1$ or $\sum_{i=0}^{k} b_j = 1$.

Let (t, u) be given and let $u(s)$ be the exact solution of the differential equation with $u(t) = u$. The local error of a multistep method is the difference between exact solution and approximate solution:

$$u(t + k\,\tau) - u_\tau(t + k\,\tau),$$

where it is assumed that the starting procedure is exact, i.e.:

$$u(t), u(t + \tau), \ldots, u(t + (k-1)\,\tau)$$

are the initial settings for the computation of $u_\tau(t + k\,\tau)$. The method is called consistent of order p, if

$$u(t + k\,\tau) - u_\tau(t + k\,\tau) = O(\tau^{p+1}).$$

Assume the scaling condition $\sum_{i=0}^{k} b_j = 1$. The approximation error is given by

$$\begin{aligned} \psi_\tau(u)(t + k\,\tau) &= \frac{1}{\tau} \sum_{i=0}^{k} \alpha_i u(t + i\,\tau) - \sum_{i=0}^{k} \beta_i f(t + i\,\tau, u(t + i\,\tau)) \\ &= \frac{1}{\tau} \sum_{i=0}^{k} \alpha_i u(t + i\,\tau) - \sum_{i=0}^{k} \beta_i u'(t + i\,\tau). \end{aligned}$$

We have the following connection between the local error and the approximation error:

Lemma 8.1. *Let f be a continuously differentiable function. Then*

$$u(t + k\,\tau) - u_\tau(t + k\,\tau) = \tau \left[\alpha_k I - \tau\,\beta_k J\right]^{-1} \psi_\tau(u)(t + k\,\tau)$$

with

$$J = \begin{pmatrix} \dfrac{\partial f_1}{\partial u}(t + k\,\tau, \nu_1) \\ \dfrac{\partial f_2}{\partial u}(t + k\,\tau, \nu_2) \\ \vdots \\ \dfrac{\partial f_n}{\partial u}(t + k\,\tau, \nu_n) \end{pmatrix}$$

and $\nu_i = u_\tau(t + k\,\tau) + \delta_i \left[u(t + k\,\tau) - u_\tau(t + k\,\tau)\right]$ for some $\delta_i \in [0, 1]$.

Proof. For one step of the method we obtain

$$\alpha_k u_\tau(t+k\,\tau) - \tau\,\beta_k f(t+kh, u_\tau(t+k\,\tau)) + \sum_{i=0}^{k-1} [\alpha_i u(t+i\,\tau) - \tau\beta_i f(t+i\,\tau, u(t+i\,\tau))] = 0.$$

By definition of the approximation error it follows that

$$\begin{aligned}
\tau\,\psi_\tau(u)(t+k\,\tau) &= \alpha_k\,u(t+k\,\tau) - \tau\,\beta_k\,f(t+k\,\tau, u(t+k\,\tau)) \\
&\qquad - \alpha_k u_\tau(t+k\,\tau) + \tau\,\beta_k f(t+kh, u_\tau(t+k\,\tau)) \\
&= \alpha_k[u(t+k\,\tau) - u_\tau(t+k\,\tau)] \\
&\qquad - \tau\,\beta_k[f(t+k\,\tau, u(t+k\,\tau)) - f(t+k\,\tau, u_\tau(t+k\,\tau))].
\end{aligned}$$

From the mean value theorem it follows that

$$f(t+k\,\tau, u(t+k\,\tau)) - f(t+k\,\tau, u_\tau(t+k\,\tau)) = J[u(t+k\,\tau) - u_\tau(t+k\,\tau)],$$

which implies

$$\tau\,\psi_\tau(u)(t+k\,\tau) = [\alpha_k I - \tau\,\beta_k J](u(t+k\,\tau) - u_\tau(t+k\,\tau)).$$

□

For explicit methods, i.e., $\beta_k = 0$, the relation simplifies to $u(t+k\,\tau) - u_\tau(t+k\,\tau) = (\tau/\alpha_k)\,\psi_\tau(u)(t+k\,\tau)$.

Remark: A multistep method can be written in the following form:

$$F_\tau(u_\tau) = 0$$

with

$$F_\tau(v_\tau)(t+k\,\tau) = \frac{1}{\tau}\sum_{i=0}^{k} \alpha_i v_\tau(t+i\,\tau) - \sum_{i=0}^{k} \beta_i f(t+i\,\tau, v_\tau(t+i\,\tau))$$

and (assuming an ideal starting procedure) $F_\tau(v_\tau)(t_0+i\,\tau) = v_\tau(t_0+i\,\tau) - u(t+i\,\tau)$ for $i = 0, 1, \ldots, k-1$. With these notations we have

$$\psi_\tau(u) = F_\tau(R_\tau u)$$

as for one-step methods. By Lemma 8.1 the method is consistent of order p if

$$\|\psi_\tau(u)\|_{X_\tau} = O(\tau^p).$$

Next the so-called generating polynomials of the multistep method are introduced by

$$\begin{aligned}
\rho(z) &= \alpha_k z^k + \alpha_{k-1} z^{k-1} + \cdots + \alpha_0, \\
\sigma(z) &= \beta_k z^k + \beta_{k-1} z^{k-1} + \cdots + \beta_0.
\end{aligned}$$

We have

Theorem 8.1. *Let f be sufficiently smooth. A linear multistep method of the form* (8.3) *is consistent of order p, if*

$$\sum_{i=0}^{k} \alpha_i = 0 \quad \textit{and} \quad \sum_{i=0}^{k} \alpha_i i^q = q \sum_{i=0}^{k} \beta_i i^{q-1} \quad \textit{for } q = 1, 2, \ldots, p.$$

Proof. We obtain by Taylor expansion:

$$\begin{aligned} \tau\, \psi_\tau(u)(t) &= \sum_{i=0}^{k} \left[\alpha_i \sum_{q=0}^{p} \frac{i^q}{q!} u^{(q)}(t) \tau^q - \beta_i \tau \sum_{r=0}^{p-1} \frac{i^r}{r!} u^{(r+1)}(t) \tau^r \right] + O(\tau^{p+1}) \\ &= \left[\sum_{i=0}^{k} \alpha_i \right] u(t) + \sum_{q=1}^{p} \frac{\tau^q}{q!} \left[\sum_{i=0}^{k} \alpha_i i^q - q \sum_{i=0}^{k} \beta_i i^{q-1} \right] u^{(q)}(t) + O(\tau^{p+1}) \\ &= O(\tau^{p+1}). \end{aligned}$$

□

For the case $p = 1$ the conditions can be written in the form:

$$\rho(1) = 0, \quad \rho'(1) = \sigma(1).$$

Remark: It is easy to show that the Adams-Bashforth methods, the Nyström methods and the BDF-methods are of order k, the Adams-Moulton methods and the Milne-Simpson methods are of order $k + 1$. These classes of multistep methods are exact for the ODEs

$$u'(t) = q t^{q-1}$$

with $q = 0, 1, \ldots, k$ and $q = 0, 1, \ldots, k + 1$, respectively. Hence, with the exact solution $u(t) = t^q$, we obtain:

$$0 = \tau\, \psi_\tau(u)(t + k\,\tau) = \tau^q \left[\sum_{i=0}^{k} \alpha_i i^q - q \sum_{i=0}^{k} \beta_i i^{q-1} \right],$$

which implies the corresponding order.

Remark: The highest attainable order of a k-step method is $2k$.

8.3 Stability of Linear Multistep Methods

See also: Hairer, Nørsett, Wanner, [8], III.3.

For Runge-Kutta methods a Lipschitz condition on f with respect to u is sufficient for stability. For multistep method the situation is more delicate.

Example: The explicit 2-step method of maximum order 3 is given by:

$$u_{j+2} + 4u_{j+1} - 5u_j = \tau\,(4f_{j+1} + 2f_j).$$

If applied to the initial value problem

$$u' = u, \quad u(0) = 1$$

with exact starting procedure $u_0 = 1$ and $u_1 = e^\tau$, the method leads to completely useless results.

We start the discussion of stability for the trivial right-hand side $f = 0$, i.e., for the differential equation

$$u'(t) = 0.$$

Then the method is of the following form:

$$\alpha_k u_{j+k} + \alpha_{k-1} u_{j+k-1} + \cdots + \alpha_0 u_j = 0. \tag{8.4}$$

Theorem 8.2. *Let $\zeta_1, \zeta_2, \ldots, \zeta_l$ be the roots of ρ with multiplicity $m_1, m_2, \ldots, m_l$. Then the general solution of (8.4) is given by*

$$u_j = p_1(j)\,\zeta_1^j + p_2(j)\,\zeta_2^j + \cdots + p_l(j)\,\zeta_l^j,$$

where p_j are arbitrary polynomials of degree $\le m_j - 1$.

Proof. The general solution is obtained as linear combination of the $m_1 + m_2 + \ldots + m_l = k$ particular solutions

$$u_j = \binom{j}{\mu}\zeta^j,$$

where ζ is a root of ρ with multiplicity m and $\mu \le m - 1$. In order to verify that these sequences solve the recurrence relation (8.4), the identity

$$\binom{j+i}{\mu} = \sum_{\nu=0}^{\mu} \binom{j}{\mu-\nu}\binom{i}{\nu}$$

is used. Then we obtain

$$\begin{aligned}
\sum_{i=0}^{k} \alpha_i u_{j+i} &= \sum_{i=0}^{k} \alpha_i \binom{j+i}{\mu}\zeta^{j+i} = \zeta^j \sum_{\nu=0}^{\mu} \binom{j}{\mu-\nu} \sum_{i=0}^{k} \alpha_i \binom{i}{\nu}\zeta^i \\
&= \zeta^j \sum_{\nu=0}^{\mu} \binom{j}{\mu-\nu} \frac{\zeta^\nu}{\nu!} \underbrace{\sum_{i=0}^{k} \alpha_i\, i(i-1)\cdots(i-\nu+1)\zeta^{i-\nu}}_{=\rho^{(l)}(\zeta)=0} = 0.
\end{aligned}$$

□

The k coefficients of the polynomials $p_1(j)$, $p_2(j)$, $\ldots$, $p_l(j)$ are uniquely determined by prescribing the k values u_0, u_1, $\ldots$, u_{k-1} of the starting phase. It immediately follows from the last theorem that the sequence $(u_j)_{j\in\mathbb{N}_0}$, generated by the linear multistep method, is bounded for arbitrary initial phase if and only if the roots ζ of ρ satisfy the following condition:

$$|\zeta| \leq 1 \quad \text{and} \quad |\zeta| = 1 \quad \text{only if } \zeta \text{ is simple.} \tag{8.5}$$

This leads to the following definition:

Definition 8.1. *The multistep method* (8.3) *is called 0–stable, if all roots ζ of ρ satisfy the condition* (8.5).

For the explicit and the implicit Adams methods we have:

$$\rho(z) = z^k - z^{k-1}.$$

0 is a root of multiplicity $(k-1)$, 1 is a simple root. Hence, the methods are 0–stable.

For the explicit Nyström methods and the Milne-Simpson methods we have:

$$\rho(z) = z^k - z^{k-2}.$$

0 is a root of multiplicity $(k-2)$, 1 and -1 are simple roots. The methods are 0-stable.

The analysis of the 0–stability of the BDF-methods is more difficult. It can be shown that these methods are 0-stable for $k \leq 6$ and not 0–stable for $k \geq 7$.

Theorem 8.3 (The first Dahlquist barrier)**.** *The order of a 0–stable k-step method satisfies:*

$$p \leq \begin{cases} k+2 & \text{if } k \text{ is odd,} \\ k+1 & \text{if } k \text{ is even,} \\ k & \text{if } \beta_k/\alpha_k \leq 0. \end{cases}$$

8.4 Convergence of Linear Multistep Methods

See also: Hairer, Nørsett, Wanner, [8], III.4.

Let f be continuous and satisfies the Lipschitz condition

$$\|f(t,w) - f(t,v)\| \leq L\,\|w - v\| \quad \text{for all } t, v, w.$$

Then there is a unique solution $u_{j+k} = \phi(t_j, u_{j+k-1}, u_{j+k-2}, \ldots, u_j, \tau)$ of

$$u_{j+k} + \sum_{i=0}^{k-1} \alpha'_i u_{j+j} = \tau\,\beta'_k\,f(t_j + k\,\tau, u_{j+k}) + \tau \sum_{i=0}^{k-1} \beta'_i, f_{j+i},$$

for arbitrary values t_j, $u_{j+k-1}, u_{j+k-2}, \ldots, u_j$ and sufficiently small τ, where

$$\alpha'_i = \frac{\alpha_i}{\alpha_k}, \quad \beta'_i = \frac{\beta_i}{\alpha_k}.$$

With

$$\psi(t_j, u_{j+k-1}, u_{j+k-2}, \ldots, u_j, \tau) = \beta'_k \, f(t_j + k\,\tau, \phi(t_j, u_{j+k-1}, u_{j+k-2}, \ldots, u_j, \tau)) + \sum_{i=0}^{k-1} \beta'_i \, f_{j+i}$$

the multistep method can be written in the form

$$u_{j+k} = -\sum_{i=0}^{k-1} \alpha'_i \, u_{j+i} + \tau \, \psi(t_j, u_{j+k-1}, u_{j+k-2}, \ldots, u_j, \tau).$$

Let

$$U_j = \begin{pmatrix} u_{j+k-1} \\ u_{j+k-2} \\ \vdots \\ u_j \end{pmatrix}, \quad A = \begin{pmatrix} -\alpha'_{k-1} & -\alpha'_{k-2} & \ldots & -\alpha'_0 \\ 1 & 0 & \ldots & 0 \\ & \ddots & \ddots & \vdots \\ & & 1 & 0 \end{pmatrix}, \quad e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Then the multistep method can be written as a one-step method

$$U_{j+1} = (A \otimes I)\, U_j + \tau \, \Phi(t_j, U_j, \tau) \tag{8.6}$$

with

$$\Phi(t, U, \tau) = (e_1 \otimes I)\psi(t, U, \tau).$$

Here, $C \otimes D$ denotes the Kronecker product (tensor product) of two matrices:

$$C \otimes D = \begin{pmatrix} c_{11}\, D & c_{12}\, D & \cdots & c_{1n}\, D \\ c_{21}\, D & c_{22}\, D & \cdots & c_{2n}\, D \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1}\, D & c_{m2}\, D & \cdots & c_{mn}\, D \end{pmatrix}.$$

Remark: We have

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$$

for arbitrary matrices A, B, C, D with suitable dimensions.

Let $u(t)$ be the exact solution of the initial value problem and set

$$U(t) = \begin{pmatrix} u(t + (k-1)\,\tau) \\ u(t + (k-2)\,\tau) \\ \vdots \\ u(t) \end{pmatrix}.$$

The approximate solution at $t+\tau$, which is obtained by (8.6) starting with $U(t)$ is denoted by $U_\tau(t+\tau)$. Then the first component of

$$U(t+\tau) - U_\tau(t+\tau)$$

is the local error of the multistep method, the other components all vanish. If the multistep method is consistent of order p it follows

$$\|U(t+\tau) - U_\tau(t+\tau)\| = O(\tau^{p+1}).$$

Lemma 8.2. *Assume that the multistep method is 0-stable. Then there is a vector norm on $(\mathbb{R}^n)^k$, such that*

$$\|A \otimes I\| \leq 1.$$

for the corresponding matrix norm.

Proof. The roots ζ of ρ are the eigenvalue of A with eigenvector $(\zeta^{k-1}, \zeta^{k-2}, \ldots, \zeta, 1)^T$. Therefore, there is a transformation matrix T with

$$T^{-1}AT = \left(\begin{array}{ccc|ccccc} \zeta_1 & & & & & & & \\ & \ddots & & & & & & \\ & & \zeta_l & & & & & \\ \hline & & & \zeta_{l+1} & \delta_l & & & \\ & & & & \ddots & \ddots & & \\ & & & & & \ddots & \delta_{k-1} & \\ & & & & & & \zeta_k & \end{array}\right)$$

with $|\zeta_i| = 1$ for $i = 1, \ldots, l$ and $|\zeta_i| < 1$, $\delta_i \in \{0, 1\}$ for $i = l+1, \ldots, k$. Then

$$T_\varepsilon^{-1}AT_\varepsilon = \left(\begin{array}{ccc|ccccc} \zeta_1 & & & & & & & \\ & \ddots & & & & & & \\ & & \zeta_l & & & & & \\ \hline & & & \zeta_{l+1} & \varepsilon\,\delta_l & & & \\ & & & & \ddots & \ddots & & \\ & & & & & \ddots & \varepsilon\,\delta_{k-1} & \\ & & & & & & \zeta_k & \end{array}\right) = J$$

with $T_\varepsilon = TD_\varepsilon$ and the diagonal matrix $D_\epsilon = \mathrm{diag}(1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{k-1})$. We choose $\varepsilon > 0$ sufficiently small, such that $|\varepsilon\delta_i| < 1 - |\zeta_i|$ for all $i = l+1, \ldots, k$.

We define the following norm in $(\mathbb{R}^n)^k$: $\|x\| = \|(T_\varepsilon^{-1} \otimes I)x\|_\infty$. Then

$$\begin{aligned} \|(A \otimes I)x\| &= \|(T_\varepsilon^{-1} \otimes I)(A \otimes I)x\|_\infty = \|((T_\varepsilon^{-1}A) \otimes I)x\|_\infty \\ &= \|((JT_\varepsilon^{-1}) \otimes I)x\|_\infty = \|(J \otimes I)(T_\varepsilon^{-1} \otimes I)x\|_\infty \\ &\leq \underbrace{\|(J \otimes I)\|_\infty}_{\leq 1} \underbrace{\|(T_\varepsilon^{-1} \otimes I)x\|_\infty}_{= \|x\|} \end{aligned}$$

□

Remark: A linear multistep method is 0–stable if and only if there exists a constant C such that

$$\|A^j\| \leq C \quad \text{for all } j \in \mathbb{N}.$$

Theorem 8.4. *Assume that f satisfies the Lipschitz condition*

$$\|f(t,w) - f(t,v)\| \le \Lambda \|w - v\| \quad \textit{for all } t, v, w.$$

If the linear multistep method is consistent of order p and 0-stable, then the method is convergent of order p.

Proof. From the Lipschitz condition for f it easily follows a Lipschitz condition for Φ:

$$\|\Phi(t,W,\tau) - \Phi(t,V,\tau)\| \le \Lambda \|W - V\| \quad \text{for all } t, V, W \text{ and for sufficiently small } \tau.$$

Together with Lemma 8.2 stability follows. The rest is obtained analogously to the convergence proof of one-step methods. □

8.5 Variable Step Size Multistep Methods

See also: Hairer, Nørsett, Wanner, [8], III.5.

The presented classes of linear multistep methods can be extended to variable step sizes. We will discuss this for the explicit and implicit Adams methods.

By Newton's interpolation formula we have:

$$p(t) = \sum_{i=0}^{k-1} \prod_{l=0}^{i-1} (t - t_{j-l}) f[t_j, t_{j-1}, \ldots, t_{j-i}]$$

where the divided differences $f[t_j, t_{j-1}, \ldots, t_{j-l}]$ are recursively defined by

$$\begin{aligned} f[t_j] &= f_j, \\ f[t_j, t_{j-1}, \ldots, t_{j-l}] &= \frac{f[t_j, t_{j-1}, \ldots, t_{j-l+1}] - f[t_{j-1}, t_{j-1}, \ldots, t_{j-l}]}{t_j - t_{j-l}}. \end{aligned}$$

Therefore,

$$p(t) = \sum_{i=0}^{k-1} \prod_{l=0}^{i-1} \frac{t - t_{j-l}}{t_{j+1} - t_{j-l}} \Phi_i^*(j) \quad \text{with} \quad \Phi_i^*(j) = \prod_{l=0}^{i-1} (t_{j+1} - t_{j-l}) f[t_j, t_{j-1}, \ldots, t_{j-l}].$$

This leads to the explicit Adams method

$$u_{j+1} = u_j + \int_{t_j}^{t_{j+1}} p(t)\, dt = u_j + \tau_j \sum_{i=0}^{k-1} g_i(j) \Phi_i^*(j)$$

with

$$g_i(j) = \frac{1}{\tau_j} \int_{t_j}^{t_{j+1}} \prod_{l=0}^{i-1} \frac{t - t_{j-l}}{t_{j+1} - t_{j-l}}\, dt.$$

For constant step sizes these expressions reduce to:

$$g_i(j) = \gamma_i, \quad \Phi_i^*(j) = \Delta^i f_j.$$

For the interpolation polynomial of the implicit Adams methods we have:

$$p^*(t) = p(t) + \prod_{l=0}^{k-1}(t - t_{j-l}) f[t_{j+1}, t_j \ldots, t_{j-k+1}].$$

Therefore

$$u_{j+1} = u_j + \int_{t_j}^{t_{j+1}} p^*(t)\ dt = p_{j+1} + \tau_j\, g_k(j)\Phi_k(j+1)$$

with the approximation obtained by the explicit Adams method

$$p_{j+1} = u_j + \tau_j \sum_{i=0}^{k-1} g_i(j)\Phi_i^*(j)$$

and

$$\Phi_k(j+1) = \prod_{l=0}^{k-1}(t_{j+1} - t_{j-l})\ f[t_{j+1}, t_j, \ldots, t_{j-k+1}].$$

The values $\Phi_i(j), \Phi_i^*(j)$ and $g_i(j)$ can be calculated by appropriate recurrence relations in j and i, see Hairer, Wanner, Nørsett [8].

General variable step size multistep methods

The different classes of linear multistep methods with variable step sizes can be written in the form

$$u_{j+k} + \sum_{i=0}^{k-1} \alpha_{ij} u_{j+i} = \tau_{j+k-1} \sum_{i=0}^{k} \beta_{ij} f_{j+i}.$$

The coefficients are now functions in $\omega_i = \tau_i/\tau_{i-1}$ for $i = k+1, \ldots, j+k-1$:

$$\alpha_{ij} = \alpha_i(\omega_{j+1}, \ldots, \omega_{j+k-1}), \quad \beta_{ij} = \beta_i(\omega_{j+1}, \ldots, \omega_{j+k-1}).$$

Example: The implicit Adams methods for $k = 2$ can be written in the form:

$$u_{j+1} = u_j + \frac{\tau_j}{6(1+\omega_j)}\big((3+2\omega_j)\, f_{j+1} + (3+\omega_j)(1+\omega_j)\, f_j - \omega_l^2\, f_{j-1}\big).$$

Order of consistency

The method is called consistent of order p if the local error at t_{j+k} is $O(\tau_j^{p+1})$.

Theorem 8.5. *If*

$$q(t_{j+k}) + \sum_{i=0}^{k-1} \alpha_{ij}\, q(t_{j+i}) = \tau_{j+k-1} \sum_{i=0}^{k} \beta_{ij}\, q'(t_{j+i})$$

for all polynomials q of degree $\le p$, if the ratios τ_i/τ_j are bounded for $i = j+1, \ldots, j+k-1$ and if the coefficients α_{ij} and β_{ij} are bounded, then the method is consistent of order p.

For the explicit and implicit Adams methods the coefficients α_{ij} and β_{ij} are bounded if

$$\tau_j/\tau_{j-1} \le \Omega$$

for some constant Ω. Under this condition the explicit Adams methods are of order k and the implicit Adams methods are of order $k+1$.

Stability

If applied to the trivial differential $u' = 0$ the general multistep method can be written in the form

$$U_{j+1} = A_j\, U_j.$$

with

$$U_j = \begin{pmatrix} u_{j+k-1} \\ u_{j+k-2} \\ \vdots \\ u_j \end{pmatrix}, \quad A_j = \begin{pmatrix} -\alpha_{k-1,j} & -\alpha_{k-2,j} & \ldots & -\alpha_{0,j} \\ 1 & 0 & \ldots & 0 \\ & \ddots & \ddots & \vdots \\ & & 1 & 0 \end{pmatrix}.$$

A general linear multistep method is called 0–stable, if there is a constant C such that

$$\|A_{j+l} A_{j+l-1} \cdots A_j\| \le C$$

for all j and $l \ge 0$.

Theorem 8.6. *Assume that*

1. $1 + \sum_{i=0}^{k-1} \alpha_{ij} = 0$;
2. *the coefficients $\alpha_i(\omega_{j+1}, \ldots, \omega_{j+k-1})$ and $\beta_i(\omega_{j+1}, \ldots, \omega_{j+k-1})$ are continuous in a neighborhood of $(1, 1, \ldots, 1)$;*
3. *all roots ζ of*

$$z^k + \sum_{i=0}^{k-1} \alpha_i(1, \ldots, 1)\, z^i = 0$$

lie inside the open unit disc $|\zeta| < 1$ with the exception $\zeta_1 = 1$.

Then there exist real numbers ω, Ω with $0 < \omega < 1 < \Omega$ such that the method is 0–stable, if

$$\omega \le \omega_j \le \Omega \quad \textit{for all } j.$$

Convergence

Consistency and stability imply convergence.

8.6 Practical Implementation and Comparison

See also: Hairer, Nørsett, Wanner, [8], III.7.

In the following several implementation issues are discussed. As an example we consider only the class of implicit Adams methods.

8.6.1 Predictor–corrector methods

The computation of the new approximate solution u_{j+1} is typically performed iteratively:

(P) Predictor step: We start with the result of the corresponding explicit Adams method:

$$u_{j+1}^{(p)} = p_{j+1}.$$

(E) Evaluation: The actual approximate solution $u_{j+1}^{(a)}$ for u_{j+1} is used for the approximate evaluation

$$f_{j+1}^{(a)} = f(t_{j+1}, u_{j+1}^{(a)})$$

for f_{j+1}.

(C) Corrector step: From

$$u_{j+1}^{(c)} = p_{j+1} + \tau_j \, g_k(j) \Phi_i^{(a)}(j+1)$$

an improved approximate solution is obtained, where $\Phi_i^{(a)}(j+1)$ is obtained from $\Phi_i(j+1)$ by replacing f_{j+1} by the actual approximate solution $f_{j+1}^{(a)}$.

The second and third step is repeatedly l times, leading to a predictor-corrector method, denoted by the symbols $P(EC)^l E$ or $P(EC)^l$. Typically, $l = 1$ or $l = 2$.

8.6.2 Order and step size control

By using Newton's interpolation formula a change in k and, therefore, a change of the order is relatively easy to do. This offers the possibility of a combined order and step size control. In contrast to extrapolation methods, the number of function evaluation does not change with the order.

The principle of a combined order and step size control is discussed for the example of the implicit Adams methods: An estimation of the local error for u_{j+1}, obtained by the

k-step Adams method, by using the approximate solution $\hat{u}_{j+1}$, obtained the $k+1$-step Adams method:

$$\begin{aligned} u(t_{j+1}) - u_{j+1} &\approx \hat{u}_{j+1} - u_{j+1} \\ &= \tau_j \left(g_{k+1}(j) - g_k(j)\right)\Phi_{k+1}(j+1) \\ &\approx \tau_j \left(g_{k+1}(j) - g_k(j)\right)\Phi^{(p)}_{k+1}(j+1) \\ &\approx \tau_j \, \gamma^*_{k+1}\Phi^{(p)}_{k+1}(j+1). \end{aligned}$$

8.6.3 Comparison of the methods

Chapter 9

Numerical Methods for Second-Order Differential Equations

See also: Hairer, Nørsett, Wanner, [8], II.14.

Part II

Stiff Problems

Chapter 10

One-Sided Lipschitz Conditions

See also: Hairer, Wanner, [9], IV.12.

Definition 10.1. *A differential equation*

$$u' = f(t, u) \tag{10.1}$$

is dissipative if

$$(f(t, w) - f(t, v), w - v) \leq 0 \quad \textit{for all } t, v, w.$$

Lemma 10.1. *Let f be continuous and satisfy the one-sided Lipschitz condition*

$$(f(t, w) - f(t, v), w - v) \leq \nu \|w - v\|^2 \quad \textit{for all } t, v, w.$$

Then, for any two solutions $v(t)$ and $w(t)$ of (10.1), we have

$$\|w(t) - v(t)\| \leq e^{\nu (t - t_0)} \|w(t_0) - v(t_0)\| \quad \textit{for all } t \geq t_0.$$

Definition 10.2. *A one-step method is contractive if*

$$\|w_{j+1} - v_{j+1}\| \leq \|w_j - v_j\| \quad \textit{for all } j.$$

Chapter 11

A-Stability

11.1 The Stability Function

See also: Hairer, Wanner, [9], IV.3.

Definition 11.1. *The stability function $R(z)$ of a Runge-Kutta method is given by*

$$R(z) = 1 + z\, b^T (I - z\, A)^{-1} e.$$

Lemma 11.1.

$$R(z) = \frac{P(z)}{Q(z)} \quad \textit{with} \quad P(z) = \det(I - z\, A + z\, e b^T) \quad \textit{and} \quad Q(z) = \det(I - z\, A).$$

Definition 11.2. *The stability domain S of a Runge-Kutta method is given by*

$$S = \{z \in \mathbb{C} | R(z)| \leq 1\}.$$

Definition 11.3. *A Runge-Kutta method is called A-stable if*

$$\mathbb{C}^- \subset S$$

with $\mathbb{C}^- = \{z \in \mathbb{C} : \operatorname{Re} z \leq 0\}$.

Lemma 11.2. *No explicit Runge-Kutta method is A-stable.*

Definition 11.4. *A Runge-Kutta method is called L-stable if it is A-stable and*

$$\lim_{z \to \infty} R(z) = 0.$$

Lemma 11.3. *If an implicit Runge-Kutta method with nonsingular matrix A satisfies one of the following two conditions:*

a) $a_{sj} = b_j$ *for* $j = 1, \ldots s$,

b) $a_{i1} = b_1$ *for* $i = 1, \ldots, s$,

then $R(\infty) = 0$.

11.2 Padé Approximation of the Exponential Function

See also: Hairer, Wanner, [9], IV.3, IV.4.

Theorem 11.1. *If a Runge-Kutta method is of order p, then*

$$e^z - R(z) = O(z^{p+1}) \quad \text{for } z \to 0.$$

Theorem 11.2. *Of an explicit Runge-Kutta method is of order p, then*

$$R(z) = 1 + z + \frac{1}{2!}\, z^2 + \ldots + \frac{1}{p!}\, z^p + O(z^{p+1}).$$

Theorem 11.3. *Let $j, k \in \mathbb{N}_0$. The (k, j)-Padé approximation to e^z, given by*

$$R_{kj}(z) = \frac{P_{kj}(z)}{Q_{kj}(z)},$$

where

$$P_{kj}(z) = 1 + \frac{k}{j+k}\, z + \frac{k(k-1)}{(j+k)(j+k-1)}\, \frac{z^2}{2!} + \ldots + \frac{k(k-1)\ldots 1}{(j+k)(j+k-1)\ldots(j+1)}\, \frac{z^k}{k!}$$

and

$$Q_{kj}(z) = 1 - \frac{j}{k+j}\, z + \frac{j(j-1)}{(k+j)(k+j-1)}\, \frac{z^2}{2!} + \ldots + (-1)^j \frac{j(j-1)\ldots 1}{(k+j)(k+j-1)\ldots(k+1)}\, \frac{z^j}{j!}$$

($Q_{kj}(z) = P_{jk}(-z)$), is the unique rational approximation to e^z of order $j + k$, such that the degrees of numerator and denominator are k and j, respectively:

$$e^z - R_{jk}(z) = O(z^{j+k+1}).$$

Theorem 11.4. *Assume that $R_{jk}(z)$ is the stability function of a Runge-Kutta method. Then the method is A-stable if and only if $k \le j \le k + 2$.*

Theorem 11.5. *The s-stage Gauß method is of order $2s$. Its stability function is $R_{s,s}(z)$ and the method is A-stable.*

Theorem 11.6. *The s-stage Radau IA method and the s-stage Radau IIA method are of order $2s - 1$. Their stability function is $R_{s-1,s}(z)$ and the methods are A-stable.*

Theorem 11.7. *The s-stage Lobatto IIIA, IIIB, and IIC methods are of order $2s-2$. The stability function of the Lobatto IIIA and IIIB methods is $R_{s-1,s-1}(z)$, the stability function of the Lobatto IIIC method is $R_{s-2,s}(z)$. All these methods are A-stable.*

11.3 Linear Systems of ODEs with Constant Coefficients

See also: Hairer, Wanner, [9], IV.2, IV.11.

Linear system

$$\begin{aligned} u'(t) &= Ju(t) + f(t), \\ u(0) &= 0 \end{aligned}$$

with constant matrix $J \in \mathbb{R}^{n \times n}$.

Theorem 11.8. *If*

$$(Jv, v) \leq 0 \quad \textit{for all } v \in \mathbb{C}^n$$

and if the Runge-Kutta method is A-stable, then the method is contractive for all $\tau > 0$.

11.4 General Dissipative Problems

See also: Hairer, Wanner, [9], IV.12.

Definition 11.5. *A one-step method is called B-stable if*

$$\|\hat{u}_1 - u_1\| \leq \|\hat{u}_0 - u_0\|$$

for all $\tau > 0$ *and all dissipative problems, i.e., for all* f *with*

$$(f(t, w) - f(t, v), w - v) \leq 0 \quad \textit{for all } t, v, w.$$

Theorem 11.9. *B-stability implies A-stability.*

Definition 11.6. *A Runge-Kutta method is called algebraically stable if*

1. $b_i \geq 0$ *for all* $i = 1, \ldots, s$.
2. $M = (m_{ij})$ *with* $m_{ij} = b_i a_{ij} + b_j a_{ji} - b_i b_j$ *is positive semi-definite.*

Theorem 11.10. *An algebraically stable Runge-Kutta method is B-stable.*

11.5 Practical Implementation

See also: Hairer, Wanner, [9], IV.8.

11.6 Multistep Methods for Stiff Problems

See also: Hairer, Wanner, [9], V.1.

General linear k-step method:

$$\alpha_k\, u_{j+k} + \alpha_{k-1}\, u_{j+k-1} + \ldots + \alpha_0\, u_k = \tau\, [\beta_k\, f_{j+k} + \beta_{k-1}\, f_{j+k-1} + \ldots + \beta_0\, f_k]$$

If applied to the model problem

$$u' = \lambda u,$$

we obtain

$$(\alpha_k - \mu\,\beta_k) + (\alpha_{k-1} - \mu\,\beta_{k-1}) + \ldots + (\alpha_0 - \mu\,\beta_0) = 0$$

with $\mu = \tau\,\lambda$.

Definition 11.7. *The set*

$$\begin{aligned} S = \{\mu \in \mathbb{C} :\ & \textit{all roots } \zeta(\mu) \textit{ of } \rho(z) - \mu\sigma(z) \textit{ satisfy} \\ & \textit{either } |\zeta(\mu)| < 1 \textit{ or } (|\zeta(\mu)| = 1 \textit{ and } \zeta(\mu) \textit{ is a simple root})\} \end{aligned}$$

is the stability domain of the linear k-step method.

Definition 11.8. *A linear k-step method is called A-stable if $\mathbb{C}^- \subset S$.*

Theorem 11.11 (The second Dahlquist barrier)**.** *An A-stable linear multistep method must be of order $p \leq 2$. The implicit trapezoidal rule is that A-stable method of this class, which has the smallest principal error term.*

Definition 11.9. *A method is called $A(\alpha)$-stable if $\mathbb{C}_\alpha \subset S$ with*

$$\mathbb{C}_\alpha = \{z \in \mathbb{C} : |\arg(-z)| < \alpha, z \neq 0\}.$$

Definition 11.10. *A linear k-step method is called G-stable if there exists a symmetric and positive definite matrix $G \in \mathbb{R}^{k\times k}$ such that*

$$\|\hat{U}_{j+1} - U_{j+1}\|_G \leq \|\hat{U}_j - U_j\|_G$$

for all $\tau > 0$ and all dissipative problems, i.e., for all f with

$$(f(t,w) - f(t,v), w - v) \leq 0 \quad \textit{for all } t, v, w.$$

Part III

Differential-Algebraic Problems

Chapter 12

Index and Classification of DAEs

See also: Hairer, Wanner, [9], VII.1.

12.1 Linear DAEs with Constant Coefficients

Linear systems with constant coefficients:

$$Bu'(t) + Au(t) = f(t) \tag{12.1}$$

Special case explicit ODEs: $B = I$.

Definition 12.1. *1. The expression $A + \lambda B$ as a function in $\lambda \in \mathbb{C}$ is called a matrix pencil.*

2. A matrix pencil is called regular if $\det[A + \lambda B] \not\equiv 0$.

Let P and Q be non-singular matrices. By multiplying with P and using the transformation $u(t) = Qv(t)$ we obtain

$$PBQv'(t) + PAQv(t) = Pf(t).$$

Example: Explicit ODEs:
Jordan canonical (or normal) form:

$$A = QJQ^{-1} \quad \text{with } J = \operatorname{diag}(J_1, \ldots, J_k),$$

where the $\mu_i \times \mu_i$-matrices J_i are of the form

$$J_i = \begin{pmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \lambda_i & 1 \\ & & & \lambda_i \end{pmatrix}.$$

With $P = Q^{-1}$ we obtain:

$$PAQ = J \quad PIQ = I.$$

and

$$v'(t) + Jv(t) = g(t)$$

with $v(t) = Q^{-1}u(t)$ and $g(t) = Pf(t)$.

The system of differential equations consists of systems of the form

$$\begin{aligned} w_1'(t) + \lambda\, w_1(t) + w_2(t) &= g_1(t), \\ &\vdots \\ w_{\nu-1}'(t) + \lambda\, w_{\nu-1}(t) + w_\nu(t) &= g_{\nu-1}(t), \\ w_\nu(t) + \lambda\, w_\nu(t) &= g_\nu(t). \end{aligned}$$

Hence

$$\begin{aligned} w_\nu(t) &= w_\nu(0)e^{-\lambda t} + \int_0^t g_\nu(s)e^{\lambda(s-t)}\, ds, \\ w_{\nu-1}(t) &= [w_{\nu-1}(0) - w_\nu(0)\, t]\, e^{-\lambda t} + \int_0^t [g_{\nu-1}(s) - g_\nu(s)\, s]\, e^{\lambda(s-t)}\, ds, \\ &\vdots \\ w_1(t) &= \left[w_1(0) - w_2(0)\, t + w_3(0)\frac{t^2}{2} - \ldots + (-1)^{\nu-1}w_\nu(0)\frac{t^{\nu-1}}{(\nu-1)!}\right] e^{-\lambda t}, \\ &\quad + \int_0^t \left[g_1(s) - g_2(s)\, s + g_3(s)\frac{s^2}{2} - \ldots + (-1)^{\nu-1}g_\nu(s)\frac{s^{\nu-1}}{(\nu-1)!}\right] e^{\lambda(s-t)}\, ds. \end{aligned}$$

Therefore, the following stability estimate results:

$$\|w(t)\| \le C\, \left[\|w(0)\| + \int_0^t \|g(s)\|\, ds\right]$$

with a constant $C > 0$ for all $t \in [0, T]$. In terms of the original quantities:

$$\|u(t)\| \le C\, \left[\|u(0)\| + \int_0^t \|f(s)\|\, ds\right].$$

Theorem 12.1 (Weierstraß, Kronecker)**.** *Let $A + \lambda B$ be a regular matrix pencil. Then there exist matrices P and Q such that*

$$PAQ = \begin{pmatrix} J & 0 \\ 0 & I \end{pmatrix}, \quad PBQ = \begin{pmatrix} I & 0 \\ 0 & N \end{pmatrix},$$

where J is a Jordan canonical form, $N = \operatorname{diag}(N_1, \ldots, N_m)$ with $\nu_i \times \nu_i$-matrices N_i, given by

$$N_i = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & 0 \end{pmatrix}.$$

For

$$Q^{-1}u(t) = v(t) = \begin{pmatrix} y(t) \\ z(t) \end{pmatrix}$$

we obtain

$$\begin{aligned} y'(t) + Jy(t) &= g(t), && (12.2) \\ Nz'(t) + z(t) &= h(t). && (12.3) \end{aligned}$$

The system (12.2) is an explicit ODE and possesses a unique solution for arbitrary initial values $y(0)$, see above.

The system (12.3) consists of systems of the form

$$\begin{aligned} w_2'(t) + w_1(t) &= h_1(t), \\ &\vdots \\ w_\nu'(t) + w_{\nu-1}(t) &= h_{\nu-1}(t), \\ w_\nu(t) &= h_\nu(t). \end{aligned}$$

Hence

$$\begin{aligned} w_\nu(t) &= h_\nu(t), \\ w_{\nu-1}(t) &= h_{\nu-1}(t) - h_\nu'(t), \\ &\vdots \\ w_1(t) &= h_1(t) - h_2'(t) + h_3''(t) - \ldots + (-1)^{\nu-1} h_\nu^{(\nu-1)}(t). \end{aligned}$$

So, it is uniquely solvable (without initial values $z(0)$) and

$$\|w(t)\| \le C \left[\|h(t)\| + \|h'(t)\| + \ldots + \|h^{(\nu-1)}(t)\| \right].$$

In summary, in terms of the original quantities

$$\begin{aligned} \|u(t)\| \le C \Bigg[\|u(0)\| &+ \int_0^t \|f(s)\| \, ds \\ &+ \max_{0\le s\le t} \|f(s)\| + \max_{s\in[0,t]} \|f'(s)\| + \ldots + \max_{s\in[0,t]} \|f^{(\nu-1)}(s)\| \Bigg]. \end{aligned}$$

The highest derivative is of order $\nu - 1$ with $\nu = \max\{\nu_i | 1 \le i \le m\}$.

Definition 12.2. *The index ν of a linear system of DAEs with constant coefficients is given by*

$$\nu = \max_{1\le i\le m} \nu_i.$$

$\nu = 0$: explicit ODEs. There is a unique solution for arbitrary initial values $u(0)$. The solution can be estimated by the initial data and the L^1-norm of the right hand side.

$\nu = 1$: $N = 0$. The transformed problem consists of an explicit ODE and a purely an algebraic problem. For estimating the solution we additionally need the L^∞-norm of the right hand side.

$\nu > 1$: higher index DAEs, hidden constraints. For estimating the solution we additionally need the L^∞-norm of derivatives of the right hand side.

Equivalent definition of ν:

$$N^{\nu-1} \neq 0 \quad \text{and} \quad N^\nu = 0.$$

The matrix N is called nilpotent with index Index ν.

12.2 Differentiation Index and Perturbation Index

General implicit ODE:

$$F(t, u(t), u'(t)) = 0 \tag{12.4}$$

Definition 12.3. *The implicit ODE (12.4) has differentiation index ν_d if $m = \nu_d$ is the smallest integer such that the system*

$$\begin{aligned} F(t, u(t), u'(t)) &= 0, \\ \frac{d}{dt} F(t, u(t), u'(t)) &= 0, \\ &\vdots \\ \frac{d^m}{dt^m} F(t, u(t), u'(t)) &= 0, \end{aligned}$$

or, in short,

$$G(t, u, u', w) = 0 \quad \textit{with } w = (u'', \ldots, u^{(m+1)}),$$

allows us to extract an explicit ODE

$$u'(t) = f(t, u(t))$$

by purely algebraic manipulations.

Consider implicit ODEs

$$F(t, u(t), u'(t)) = 0$$

with non-singular $F_{u'}$: $\nu_d = 1$

Consider the linear system of DAEs with constant coefficients

$$\begin{aligned} w_2'(t) + w_1(t) &= h_1(t), \\ w_3'(t) + w_2(t) &= h_2(t), \\ &\vdots \\ w_\nu'(t) + w_{\nu-1}(t) &= h_{\nu-1}(t), \\ w_\nu(t) &= h_\nu(t). \end{aligned}$$

If the first equation is differentiated once, the second twice, and so on, we obtain

$$\begin{aligned} w_2''(t) + w_1'(t) &= h_1'(t), \\ w_3'''(t) + w_2''(t) &= h_2''(t), \\ &\vdots \\ w_\nu^{(\nu)}(t) + w_{\nu-1}^{(\nu-1)}(t) &= h_{\nu-1}^{(\nu-1)}(t), \\ w_\nu^{(\nu)}(t) &= h_\nu^{(\nu)}(t). \end{aligned}$$

Hence

$$w_1'(t) = h_1'(t) - h_2''(t) + h_3'''(t) - \ldots + (-1)^{\nu-1} h_\nu^{(\nu)}(t).$$

Therefore: $\nu_d = \nu$.

Consider the semi-explicit DAE

$$\begin{aligned} y'(t) &= f(t, y(t), z(t)), \\ 0 &= g(t, y(t), z(t)) \end{aligned}$$

with non-singular matrix $g_z(t, y, z)$. If the second equation is differentiated once, we obtain

$$g_z(t, y(t), z(t))z'(t) + g_y(t, y(t), z(t))y'(t) + g_t(t, y(t), z(t)) = 0,$$

Hence

$$\begin{aligned} y'(t) &= f(t, y(t), z(t)), \\ z'(t) &= -g_z(t, y(t), z(t))^{-1}[g_y(t, y(t), z(t))f(t, y(t), z(t)) + g_t(t, y(t), z(t))]. \end{aligned}$$

Therefore: $\nu_d = 1$. The original DAE is called a Hessenberg index-1 system.

Consider a semi-explicit DAE of the form

$$\begin{aligned} y'(t) &= f(t, y(t), z(t)), \\ 0 &= g(t, y(t)) \end{aligned}$$

with non-singular matrix $g_y(t, y)\, f_z(t, y, z)$. If the second equation is differentiated once, we obtain

$$g_y(t, y(t))y'(t) + g_t(t, y(t)) = 0$$

and, therefore, the new (hidden) constraint

$$g_y(t, y(t))f(t, y(t), z(t)) + g_t(t, y(t)) = 0.$$

If this constraint is differentiated once, we obtain an explicit system of ODEs, since $g_y(t, y(t))f_z(t, y(t), z(t))$ is non-singular. Therefore: $\nu_d = 2$. It a called a Hessenberg index-2 system.

Similarly one can show that the semi-explicit DAE

$$\begin{aligned} x'(t) &= f(t, x(t), y(t), z(t)), \\ y'(t) &= g(t, x(t), y(t)), \\ 0 &= h(t, y(t)) \end{aligned}$$

with non-singular matrix $h_y(t,y)g_x(t,x,y)f_z(t,x,y,z)$: $\nu_d = 3$. It a called a Hessenberg index-3 system.

The last two examples are special cases of Hessenberg index-m DAEs:

$$\begin{aligned} x_1'(t) &= f_1(t, x_1(t), x_2(t), \ldots, x_{m-1}(t), x_m(t)), \\ x_2'(t) &= f_2(t, x_1(t), x_2(t), \ldots, x_{m-1}(t)), \\ &\vdots \\ x_i'(t) &= f_i(t, x_{i-1}(t), x_i(t), \ldots, x_{m-1}(t)), \\ &\vdots \\ x_{m-1}'(t) &= f_{m-1}(t, x_{m-2}(t), x_{m-1}(t)), \\ 0 &= f_m(t, x_{m-1}(t)) \end{aligned}$$

with a non-singular

$$\frac{\partial f_m}{\partial x_{m-1}} \frac{\partial f_{m-1}}{\partial x_{m-2}} \cdots \frac{\partial f_2}{\partial x_1} \frac{\partial f_1}{\partial x_m}.$$

Obviously: $\nu_d = m$.

Linear Hessenberg index-m DAEs with constant coefficients:

$$\begin{pmatrix} I & 0 & \cdots & \cdots & 0 \\ 0 & I & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & I & 0 \\ 0 & \cdots & \cdots & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ \vdots \\ x_m' \end{pmatrix} + \begin{pmatrix} A_{11} & \cdots & \cdots & A_{1,m-1} & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2,m-1} & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & A_{m-1,m-1} & \vdots \\ 0 & \cdots & 0 & A_{m,m-1} & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ \vdots \\ f_m \end{pmatrix}$$

with

$$A_{m,m-1}\, A_{m-1,m-2} \cdots A_{21}\, A_{1m} \text{ non-singular.}$$

By reordering we obtain

$$\begin{pmatrix} 0 & I & \cdots & \cdots & 0 \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 & I \\ 0 & \cdots & \cdots & 0 & 0 \end{pmatrix} \begin{pmatrix} x_m' \\ x_1' \\ \vdots \\ \vdots \\ x_{m-1}' \end{pmatrix} + \begin{pmatrix} A_{1m} & A_{11} & \cdots & \cdots & A_{1,m-1} \\ 0 & A_{21} & A_{22} & \cdots & A_{2,m-1} \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & A_{m-1,m-1} \\ 0 & 0 & \cdots & 0 & A_{m,m-1} \end{pmatrix} \begin{pmatrix} x_m \\ x_1 \\ \vdots \\ \vdots \\ x_{m-1} \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ \vdots \\ f_m \end{pmatrix}$$

Hence

$$Nx' + Rx = f$$

with a nilpotent matrix N of index m and a non-singular upper block-triangular matrix R, which easily imply $\nu = m$.

Example: Constrained mechanical systems

$$\begin{aligned} M(q)\ddot{q} &= f(q,\dot{q}) - G(q)^T\lambda \\ 0 &= g(q) \end{aligned}$$

written as a system of first order

$$\begin{aligned} M(q)\dot{u} &= f(q,u) - G(q)^T\lambda \\ \dot{q} &= v \\ 0 &= g(q) \end{aligned}$$

are DAEs of index 3, if the matrix $M(q)$ is nonsingular and $G(q)$ has full rank equal to the number of rows. If the first equation is multiplied by $M(q)^{-1}$, the system becomes a Hessenberg index-3 DAE with $x = u$, $y = q$, $z = \lambda$, and $h_y(t,y)g_x(x,y)f_z(x,y,z) = S(q) = -G(q)M(q)^{-1}G(q)^T$.

If the constraint is differentiated once, we obtain

$$\begin{aligned} M(q)\dot{u} &= f(q,u) - G(q)^T\lambda \\ \dot{q} &= u \\ 0 &= G(q)u \end{aligned}$$

This system is a DAE of index 2, if the matrix $M(q)$ is nonsingular and $G(q)$ has full rank equal to the number of rows. If the first equation is multiplied by $M(q)^{-1}$, the system becomes a Hessenberg index-2 DAE with $y = (u,q)$, $z = \lambda$, and $g_y(y)\, f_z(y,z) = S(q) = -G(q)M(q)^{-1}G(q)^T$.

If the constraint is differentiated twice, we obtain

$$\begin{aligned} \dot{q} &= u \\ \begin{pmatrix} M(q) & G(q)^T \\ G(q) & 0 \end{pmatrix} \begin{pmatrix} \dot{u} \\ \lambda \end{pmatrix} &= \begin{pmatrix} f(q,u) \\ -g_{qq}(q)(u,u) \end{pmatrix} \end{aligned}$$

This system is a DAE of index 1, if the matrix $M(q)$ is nonsingular and $G(q)$ has full rank equal to the number of rows. If the second part of the system is multiplied by the inverse of the 2-by-2 block matrix, the system becomes a Hessenberg index-1 DAE with $y = (u,q)$, $z = \lambda$, and $g_z(y,z) = S(q) = -G(q)M(q)^{-1}G(q)^T$.

Consider the GGL-formulation (after Gear, Gupta, and Leimkuhler):

$$\begin{aligned} M(q)\dot{u} &= f(q,u) - G(q)^T\lambda \\ \dot{q} &= u - G(q)^T\mu \\ 0 &= G(q)u \\ 0 &= g(q) \end{aligned}$$

If the matrix $M(q)$ is nonsingular and $G(q)$ has full rank equal to the number of rows, this system is equivalent to the index-2 formulation and has also index 2. If the first equation is multiplied by $M(q)^{-1}$, the system becomes a Hessenberg index-2 DAE with $y = (u, q)$, $z = (\lambda, \mu)$, and

$$\begin{aligned} g_y(y) f_z(y,z) &= \begin{pmatrix} G(q) & g_{qq}(q)u \\ 0 & G(q) \end{pmatrix} \begin{pmatrix} -M(q)^{-1}G(q)^T & 0 \\ 0 & -G(q)^T \end{pmatrix} \\ &= \begin{pmatrix} S(q) & -g_{qq}(q)(u, G(q)^T .) \\ 0 & -G(q)G(q)^T \end{pmatrix}. \end{aligned}$$

Consider the general implicit ODEs:

$$F(t, u(t), u'(t)) = 0. \tag{12.5}$$

Definition 12.4. *The implicit ODE (12.5) has perturbation index ν_p with respect to a solution $u(t)$, $t \in [0,T]$ if $m = \nu_p$ is the smallest integer such that, for all functions $\hat{u}(t)$ with*

$$F(t, \hat{u}(t), \hat{u}'(t)) = \delta(t),$$

there exists a constant $C > 0$ with

$$\begin{aligned} \|\hat{u}(t) - u(t)\| \;\leq\; C \Bigg[& \|\hat{u}(0) - u(0)\| \\ & + \int_0^t \|\delta(s)\|\, ds + \max_{0\leq s\leq t} \|\delta(s)\| + \max_{s\in[0,t]} \|\delta'(s)\| + \ldots + \max_{s\in[0,t]} \|\delta^{(m-1)}(s)\| \Bigg] \end{aligned}$$

for all $t \in [0,T]$ and all sufficiently small perturbations δ.

For

- linear systems of DAEs with constant coefficients,
- DAEs in Hessenberg form.

it follows that

$$\nu_p = \nu_d.$$

See also Gear [3].

Chapter 13

Numerical Methods for Implicit ODEs

13.1 Runge-Kutta Methods

Consider the implicit ODE

$$F(t, u(t), u'(t)) = 0$$

Runge-Kutta methods:

$$u_{j+1} = u_j + \tau_j \sum_{k=1}^{s} b_k k_{jk}$$

with

$$F(t_{ji}, U_{ji}, k_{ji}) = 0 \quad \text{and} \quad U_{ji} = u_j + \tau_j \sum_{k=1}^{s} a_{ik} k_{jk},$$

where $t_{ji} = t_j + c_i \tau_j$. So

$$F(t_{ji}, u_j + \tau_j \sum_{k=1}^{s} a_{ik} k_{jk}, k_{ik}) = 0.$$

In particular, we obtain for the linear DAE with constant coefficients

$$Nz'(t) + z(t) = h(t)$$

with $N^\nu = 0$, $N^{\nu-1} \neq 0$ for $\nu \geq 1$ the linear system

$$Nl_{ni} + z_j + \tau_j \sum_{k=1}^{s} a_{ik} l_{jk} = h(t_{ji}), \quad i = 1, \ldots, s,$$

for the values l_{ji}, which can be interpreted as approximations to $z'(t_{ji})$. That means

$$\left[\begin{pmatrix} N & 0 & \cdots & 0 \\ 0 & N & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & N \end{pmatrix} + \tau_j \begin{pmatrix} a_{11}\, I & a_{12}\, I & \cdots & a_{1s}\, I \\ a_{21}\, I & a_{22}\, I & \cdots & a_{2s}\, I \\ \vdots & \vdots & \ddots & \vdots \\ a_{s1}\, I & a_{s2}\, I & \cdots & a_{ss}\, I \end{pmatrix}\right] \begin{pmatrix} l_{j1} \\ l_{j2} \\ \vdots \\ l_{js} \end{pmatrix} = \begin{pmatrix} h(t_{j1}) - z_j \\ h(t_{j2}) - z_j \\ \vdots \\ h(t_{js}) - z_j \end{pmatrix}$$

or, in short,

$$[I \otimes N + \tau_j\, A \otimes I] \begin{pmatrix} l_{j1} \\ l_{j2} \\ \vdots \\ l_{js} \end{pmatrix} = \begin{pmatrix} h(t_{j1}) - z_j \\ h(t_{j2}) - z_j \\ \vdots \\ h(t_{js}) - z_j \end{pmatrix}.$$

In order to obtain a well-defined method the matrix $I \otimes N + \tau_j\, A \otimes I$ must be non-singular, which is true if and only if A is non-singular.

Proof. If A is singular, then there is a vector $y \neq 0$ with $Ay = 0$. Since N is singular, there is a vector $z \neq 0$ with $Nz = 0$. Then: $(I \otimes N + \tau_n\, A \otimes I)(y \otimes z) = y \otimes Nz + \tau_n\, Ay \otimes z = 0$.

Assume that A is non-singular. Because of

$$I \otimes N + \tau_n\, A \otimes I = [\tau_n\, A \otimes I][I \otimes I + \frac{1}{\tau_n}\, A^{-1} \otimes N].$$

it follows that

$$\begin{aligned} [\tau_n\, A \otimes I + I \otimes N]^{-1} &= [I \otimes I + \frac{1}{\tau_n}\, A^{-1} \otimes N]^{-1} [\tau_n\, A \otimes I]^{-1} \\ &= \sum_{k=0}^{\nu-1} \frac{(-1)^k}{\tau_n^k} \left[A^{-1} \otimes N\right]^k \frac{1}{\tau_n} \left[A^{-1} \otimes I\right] \\ &= \sum_{k=0}^{\nu-1} \frac{(-1)^k}{\tau_n^{k+1}} \left[A^{-(k+1)} \otimes N^k\right]. \end{aligned}$$

□

The requirement that A is non-singular, excludes the class of explicit Runge-Kutta methods.

Remark:

1. The Gauß methods, the Radau IA methods, the Radau IIA methods, and the LobattoIIIC methods have a non-singular coefficient matrix $A = (a_{ij})$.

2. For the Lobatto IIIA methods, the first row of $A = (a_{ij})$ vanishes. Nevertheless, it can be shown that these methods are also suitable for DAEs, since A is of the form

$$A = \left(\begin{array}{c|c} 0 & 0 \\ \hline \underline{a} & \underline{A} \end{array}\right)$$

with a non-singular matrix $\underline{A}$ if properly modified: Set $l_{j1} = l_{j-1,s}$ and determine $(l_{j2}, \cdots, l_{js})$ from the reduced system obtained by ignoring the first equation. This approach requires an initial value for l_{01}, e.g. $l_{01} = z'(0)$.

3. For the Lobatto IIIB methods, the last column of $A = (a_{ij})$ vanishes, these methods are not appropriate methods for DAEs.

13.2 BDF-Methods

Consider the implicit ODE

$$F(t, u(t), u'(t)) = 0.$$

BDF-methods:

$$F(t_{j+k}, u_{j+k}, \frac{1}{\tau} \sum_{i=1}^{k} \frac{1}{i} \nabla^i u_{j+k}) = 0.$$

In particular, we obtain for the linear DAE with constant coefficients

$$Nz'(t) + z(t) = h(t)$$

with $N^\nu = 0$, $N^{\nu-1} \neq 0$ for $\nu \geq 1$

$$N \frac{1}{\tau} \sum_{i=1}^{k} \frac{1}{i} \nabla^i z_{j+k} + z_{j+k} = h(t_{j+k}).$$

So

$$z_{j+k} = \left(I + N \frac{1}{\tau} \sum_{i=1}^{k} \frac{1}{i} \nabla^i \right)^{-1} h(t_{j+k}) = \sum_{l=0}^{\nu-1} (-1)^l N^l \left(\frac{1}{\tau} \sum_{i=1}^{k} \frac{1}{i} \nabla^i \right)^l h(t_{j+k}).$$

This shows the method is defined and z_{j+k} depends only on the values of the right-hand side $h(t)$ at t_{n+k} and at the previous $(\nu - 1)\, k$ grid points.

For the exact solution we have a similar representation:

$$z(t) = \left(I + N \frac{d}{dt} \right)^{-1} h(t) = \left(\sum_{i=0}^{\nu-1} (-1)^i N^i \frac{d^i}{dt^i} \right) h(t) = \sum_{i=0}^{\nu-1} (-1)^i N^i h^{(i)}(t).$$

Chapter 14

Hessenberg Index-1 DAEs

Consider the Hessenberg index-1 DAE:

$$\begin{aligned} y'(t) &= f(t, y(t), z(t)), && (14.1)\\ 0 &= g(t, y(t), z(t)) && (14.2) \end{aligned}$$

where

$$g_z(t, y, z) \quad \text{non-singular in a neighborhood of the solution.} \qquad (14.3)$$

Then, by the implicit function theorem, we have locally:

$$z(t) = G(t, y(t)).$$

Reduced problem:

$$y'(t) = f(t, y(t), G(t, y(t)). \qquad (14.4)$$

Any method appropriate for explicit ODEs can be applied to (14.4). This approach is called indirect approach or state space form method.

A Runge-Kutta method applied to (14.4):

$$y_{j+1} = y_j + \tau_j \sum_{k=1}^{s} b_k k_{jk}$$

with

$$k_{ji} = f(t_{ji}, Y_{ji}, G(t_{ji}, Y_{ji}))$$

and

$$Y_{ji} = y_j + \tau_j \sum_{k=1}^{s} a_{ik} k_{jk}.$$

For z_{j+1} one obtains

$$z_{j+1} = G(t_{j+1}, y_{j+1}).$$

By introducing

$$Z_{ji} = G(t_{ji}, Y_{ji})$$

the method can be written in the form

$$y_{j+1} = y_j + \tau_j \sum_{k=1}^{s} b_k k_{jk}, \quad 0 = g(t_{j+1}, y_{j+1}, z_{j+1})$$

with

$$Yk_{ji} = f(t_{ji}, Y_{ji}, Z_{ji}), \quad 0 = g(t_{ji}, Y_{ji}, Z_{ji})$$

and

$$Y_{ji} = y_j + \tau_j \sum_{k=1}^{s} a_{ik} Y'_{jk}.$$

A BDF-method applied to (14.4):

$$\sum_{i=1}^{k} \frac{1}{i} \nabla^i y_{j+i} = \tau \, f(t_{j+k}, y_{j+k}, G(t_{j+k}, y_{j+k}))$$

and

$$z_{j+1} = G(t_{j+1}, y_{j+1}).$$

One obtains

$$\begin{aligned} \sum_{i=1}^{k} \frac{1}{i} \nabla^i y_{j+i} &= \tau \, f(t_{j+k}, y_{j+k}, z_{j+k}), \\ 0 &= g(t_{j+k}, y_{j+k}, z_{j+k}). \end{aligned}$$

This corresponds exactly to the approach of chapter 13.

An alternative approach for constructing a method for the semi-explicit DAE (14.1), (14.2) is called direct approach or ε-embedding method. (14.1), (14.2) is considered as limit case $\varepsilon = 0$ of the explicit singularly perturbed ODE

$$\begin{aligned} y'(t) &= f(t, y(t), z(t)), & (14.5) \\ \varepsilon \, z'(t) &= g(t, y(t), z(t)). & (14.6) \end{aligned}$$

for which any method appropriate for explicit ODEs could, at least in principle, be considered. Subsequently, we set $\varepsilon = 0$.

If a Runge-Kutta method is applied to (14.5), (14.6), we obtain

$$y_{j+1} = y_j + \tau_j \sum_{k=1}^{s} b_k k_{jk}, \quad z_{j+1} = z_j + \tau_j \sum_{k=1}^{s} b_k l_{jk}$$

with

$$k_{ji} = f(t_{ji}, Y_{ji}, Z_{ji}), \quad \varepsilon \, l_{ji} = g(t_{ji}, Y_{ji}, Z_{ji})$$

and

$$Y_j = y_j + \tau_j \sum_{k=1}^{s} a_{ik} k_{jk}, \quad Z_j = z_j + \tau_j \sum_{k=1}^{s} a_{ik} l_{jk}.$$

The limit $\varepsilon \to 0$ leads to the following method:

$$y_{j+1} = y_j + \tau_j \sum_{k=1}^{s} b_k k_{jk}, \quad z_{j+1} = z_j + \tau_j \sum_{k=1}^{s} b_k l_{jk}$$

with

$$k_{ji} = f(t_{ji}, Y_{ji}, Z_{ji}), \quad 0 = g(t_{ji}, Y_{ji}, Z_{ji})$$

and

$$Y_{ji} = y_j + \tau_j \sum_{k=1}^{s} a_{ik} k_{jk}, \quad Z_{ji} = z_j + \tau_j \sum_{k=1}^{s} a_{ik} l_{jk}.$$

This corresponds exactly to the approach of chapter 13.

Observe, however, the new approximate solutions (y_{j+1}, z_{j+1}) need not necessarily satisfy the algebraic constraint

$$0 = g(t_{j+1}, y_{j+1}, z_{j+1}),$$

despite the fact that:

$$0 = g(t_j, y_j, z_j).$$

If

$$c_s = 1, \quad b_j = a_{sj}, \quad j = 1, 2, \ldots, s, \tag{14.7}$$

(the method is called stiffly accurate) the direct and the indirect approach coincide: We have $y_{j+1} = Y_{js}$ and $z_{j+1} = Z_{js}$. Therefore, the new approximate solutions satisfy the constraints since the intermediate values satisfy the constraints.

Remark: The Radau IIA methods, the Lobatto IIIA methods, and the Lobatto IIIC methods satisfy (14.7).

If a BDF-method is applied to (14.5), (14.6), one obtains

$$\begin{aligned} \sum_{i=1}^{k} \frac{1}{i} \nabla^i y_{j+k} &= \tau\, f(t_{j+k}, y_{j+k}, z_{j+k}), \\ \varepsilon \left(\sum_{i=1}^{k} \frac{1}{i} \nabla^i z_{j+k} \right) &= \tau\, g(t_{j+k}, y_{j+k}, z_{j+k}). \end{aligned}$$

The limit $\varepsilon \to 0$ leads to the following method:

$$\begin{aligned} \sum_{i=1}^{k} \frac{1}{i} \nabla^i y_{j+k} &= \tau\, f(t_{j+k}, y_{j+k}, z_{j+k}), \\ 0 &= g(t_{j+k}, y_{j+k}, z_{j+k}). \end{aligned}$$

Here, the direct approach coincides with the indirect approach. As a consequence, the following estimate for the global error holds:

$$\|y_j - y(t_j)\| = O(\tau^k)$$

and

$$\|z_j - z(t_j)\| = \|G(t_j, y_j) - G(t_j, y(t_j))\| = O(\|y_j - y(t_j)\|) = O(\tau^k)$$

for all $j = k, k+1, \ldots$, if the initial values satisfy

$$\|y_j - y(t_j)\| = O(\tau^k), \qquad \text{for all } j = 0, \ldots, k-1.$$

14.1 Direct Approach for Runge-Kutta Methods

For a Runge-Kutta method

$$y_{j+1} = y_j + \tau_j \sum_{k=1}^{s} b_k k_{jk}, \quad z_{j+1} = z_j + \tau_j \sum_{k=1}^{s} b_k l_{jk}$$

with

$$k_{ji} = f(t_{ji}, Y_{ji}, Z_{ji}) \quad 0 = g(t_{ji}, Y_{ji}, Z_{ji})$$

and

$$Y_{ji} = y_j + \tau_j \sum_{k=1}^{s} a_{ik} k_{jk}, \quad Z_{ji} = z_j + \tau_j \sum_{k=1}^{s} a_{ik} l_{jk}$$

one obtains

$$z_{j+1} = z_j + \tau_j \, [b^T \otimes I] l_j$$

and

$$Z_j = e \otimes z_j + \tau_j \, [A \otimes I] l_j$$

with

$$Y_j = \begin{pmatrix} Y_{j1} \\ Y_{j2} \\ \vdots \\ Y_{js} \end{pmatrix}, \quad k_j = \begin{pmatrix} k_{j1} \\ k_{j2} \\ \vdots \\ k_{js} \end{pmatrix}, \quad Z_j = \begin{pmatrix} Z_{j1} \\ Z_{j2} \\ \vdots \\ Z_{js} \end{pmatrix}, \quad l_j = \begin{pmatrix} l_{j1} \\ l_{j2} \\ \vdots \\ l_{js} \end{pmatrix}.$$

The last equation implies:

$$l_j = \frac{1}{\tau_j} [A^{-1} \otimes I] \, (Z_j - e \otimes z_j) = \frac{1}{\tau_j} [A^{-1} \otimes I] Z_j - \frac{1}{\tau_j} [A^{-1} e \otimes z_j].$$

Hence

$$\begin{aligned} z_{j+1} &= z_j + \tau_j \, [b^T \otimes I] l_j = z_j + [b^T \otimes I] \left([A^{-1} \otimes I] Z_j - [A^{-1} e \otimes z_j] \right) \\ &= (1 - b^T A^{-1} e) z_j + [b^T A^{-1} \otimes I] Z_j. \end{aligned}$$

Therefore, we obtain the following form of the Runge-Kutta method:

$$\begin{aligned} y_{j+1} &= y_j + \tau_j \sum_{k=1}^{s} b_k f(t_{jk}, Y_{jk}, Z_{jk}) \\ z_{j+1} &= (1 - b^T A^{-1} e) z_k + [b^T A^{-1} \otimes I] Z_j \end{aligned}$$

where Y_j and Z_j are given by the system of equations

$$\begin{aligned} Y_{ji} &= y_j + \tau_j \sum_{k=1}^{s} a_{ik} f(t_{jk}, Y_{jk}, Z_{jk}), && (14.8) \\ 0 &= g(t_{ji}, Y_{ji}, Z_{ji}). && (14.9) \end{aligned}$$

The existence of a locally unique solution follows from the implicit function theorem: If

$$y_j - y(t_j) = O(\bar{\tau}_{j-1}) \quad \text{and} \quad z_j - z(t_j) = O(\bar{\tau}_{j-1}) \quad \text{with} \quad \bar{\tau}_i = \max_{0 \le l \le i} \tau_i,$$

then (14.3) implies the existence of a locally unique solution $\bar{z}_j = G(t_j, y_j)$ to the equation

$$0 = g(t_j, y_j, \bar{z}_j).$$

Therefore, $Y_{ji} = y_j$ and $Z_{ji} = \bar{z}_j$ are the locally unique solution to the system (14.8), (14.9) for $\tau_j = 0$. The Jacobian with respect to Y_{ji} and Z_{ji} at $\tau_j = 0$ has the following form:

$$\begin{pmatrix} I \otimes I & 0 \\ I \otimes g_y(t_j, y_j, \bar{z}_j) & I \otimes g_z(t_j, y_j, \bar{z}_j) \end{pmatrix}$$

and, therefore, is non-singular. From the implicit function theorem the existence of a locally unique solution to the system (14.8), (14.9) follows for sufficiently small step sizes τ_j, and

$$Y_{ji} - y_j = O(\tau_j) \quad \text{and} \quad Z_{ji} - \bar{z}_j = O(\tau_j).$$

Theorem 14.1. *Assume that the system* (14.1), (14.2) *satisfy the condition* (14.3) *in a neighborhood of the exact solution and the initial values* (y_0, z_0) *are consistent, i.e.:* $g(0, y_0, z_0) = 0$. *Consider a Runge-Kutta method of order* p, *of stage order* q, *i.e.:* $C(q)$ *is satisfied, whose coefficient matrix* A *in non-singular. Then the global error satisfies*

$$y_j - y(t_j) = O(\bar{\tau}_{j-1}^p), \quad z_j - z(t_j) = O(\bar{\tau}_{j-1}^r)$$

with

1. $r = p$ *if (14.7) holds,*
2. $r = \min(p, q+1)$ *if* $|R(\infty)| < 1$,
3. $r = \min(p-1, q)$ *if* $|R(\infty)| = 1$

Remark: For $R(\infty) = -1$ and constant step sizes one can show an improved result in the last case with $r = \min(p, q+1)$.

Stage order q and order of accuracy p for explicit ODEs, p for the differential variable y and r of the algebraic variable r for Hessenberg index-1 DAEs:

method	explicit ODE		DAE Index 1		
	q	p		p	r
Gauß	s	$2s$	s odd:	$2s$	$s+1$
			s even:	$2s$	s
Radau IA	$s-1$	$2s-1$		$2s-1$	s
Radau IIA	s	$2s-1$		$2s-1$	$2s-1$
Lobatto IIIA	s	$2s-2$		$2s-2$	$2s-2$
Lobatto IIIC	$s-1$	$2s-2$		$2s-2$	$2s-2$

Order of accuracy r for the algebraic variable z:

method	$s=1$	$s=2$	$s=3$	$s=4$
Gauß	1(2)	2	3(4)	4
Radau IA	1	2	3	4
Radau IIA	1	3	5	7
Lobatto IIIA		2	4	6
Lobatto IIIC		2	4	6

System of equations to be solved in each step, is of dimension $s \cdot n$:

$$\begin{aligned} Y_{ji} &= y_j + \tau_j \sum_{k=1}^{s} a_{ik} f(t_{jk}, Y_{jk}, Z_{jk}), \\ 0 &= g(t_{ji}, Y_{ji}, Z_{ji}) \end{aligned}$$

This system is solved by the simplified Newton method with the Jacobian at $Y_{ji} = y_j$ and $Z_{ji} = z_j$:

$$\begin{pmatrix} I \otimes I - \tau\, A \otimes f_y(t_j, y_j, z_j) & -\tau\, A \otimes f_z(t_j, y_j, z_j) \\ I \otimes g_y(t_j, y_j, z_j) & I \otimes g_z(t_j, y_j, z_j) \end{pmatrix}$$

By multiplying the first block row by $(\tau\, A)^{-1} \otimes I$, one obtains

$$\begin{pmatrix} (\tau\, A)^{-1} \otimes I - I \otimes f_y(t_j, y_j, z_j) & -I \otimes f_z(t_j, y_j, z_j) \\ I \otimes g_y(t_j, y_j, z_j) & I \otimes g_z(t_j, y_j, z_j) \end{pmatrix}$$

Using

$$T^{-1} A T = \Lambda$$

with $\Lambda = \mathrm{diag}(\Lambda_1, \ldots, \Lambda_s)$ the system can be further transformed to

$$\begin{pmatrix} (\tau\, \Lambda)^{-1} \otimes I - I \otimes f_y(t_j, y_j, z_j) & -I \otimes f_z(t_j, y_j, z_j) \\ I \otimes g_y(t_j, y_j, z_j) & I \otimes g_z(t_j, y_j, z_j). \end{pmatrix}$$

which consists of sub-matrices of the form

$$\begin{pmatrix} (\tau\, \Lambda_i)^{-1}\, I - f_y(t_j, y_j, z_j) & -f_z(t_j, y_j, z_j) \\ g_y(t_j, y_j, z_j) & g_z(t_j, y_j, z_j). \end{pmatrix}$$

In this case the original linear system of dimension $s \cdot n$ is reduced to s linear systems of dimension n. The computational costs of a direct solver like Gaußian elimination reduces from $4(sn)^3/3 = 4s^3n^3/3$ to $4sn^3/3$ elementary operations.

In case of complex eigenvalues Λ_i complex arithmetic is necessary.

Example: Consider the 3-stage Radau IIA method (RADAU5). A has a pair of complex conjugate eigenvalues $\Lambda_1 = \alpha + i\beta$, $\Lambda_2 = \alpha - i\beta$ and one real eigenvalue $\Lambda_3 = \gamma$. LU-decompositions must be computed only for Λ_1 and Λ_3. The $LU - decomposition$ for Λ_2 is the complex conjugate of the LU-decomposition for Λ_2.

Chapter 15

Hessenberg Index-2 DAEs

Consider a Hessenberg index-2 DAE (without loss of generality in autonomous form):

$$y'(t) = f(y(t), z(t)), \tag{15.1}$$
$$0 = g(y(t)) \tag{15.2}$$

with

$$g_y(y) f_z(y, z) \quad \text{non-singular in a neighborhood of the solution.} \tag{15.3}$$

Hidden algebraic constraint:

$$0 = g_y(y(t))\, f(y(t), z(t)). \tag{15.4}$$

Summary of results for some classes of Runge-Kutta methods:

Stage order q and order of accuracy p for explicit ODEs, p for the differential variable y and r of the algebraic variable r for Hessenberg index-2 DAEs:

method	explicit ODE		DAE index 2		
	q	p		p	r
Gauß	s	$2s$	s odd:	$s+1$	$s-1$
			s even:	s	$s-2$
Radau IA	$s-1$	$2s-1$		s	$s-1$
Radau IIA	s	$2s-1$		$2s-1$	s
Lobatto IIIA	s	$2s-2$	s odd:	$2s-2$	$s-1$
			s even:	$2s-2$	s
Lobatto IIIC	$s-1$	$2s-2$		$2s-2$	$s-1$

Order of accuracy r for the algebraic variable z:

method	$s=1$	$s=2$	$s=3$	$s=4$
Gauß			1(2)	2
Radau IA		1	2	3
Radau IIA	1	2	3	4
Lobatto IIIA		1(2)	2	3(4)
Lobatto IIIC		1	2	3

15.1 BDF-Methods

BDF-method applied to (15.1), (15.2):

$$\sum_{i=1}^{k} \frac{1}{i} \nabla^i y_{j+k} = \tau\, f(y_{j+k}, z_{j+k}), \tag{15.5}$$

$$0 = g(y_{j+k}). \tag{15.6}$$

or, in general notation

$$\sum_{i=0}^{k} \alpha_i y_{j+i} = \tau\, f(y_{j+k}, z_{j+k}),$$

$$0 = g(y_{j+k}).$$

Theorem 15.1. *Assume that*

$$y_{j+i} = y(t_{j+i}) + O(\tau) \quad \textit{and} \quad g(y_{j+i}) = 0 \quad \textit{for } i = 0, 1, \ldots, k-1.$$

Then the system (15.5), (15.6) *has a locally unique solution*

$$y_{j+k} = y(t_{j+k}) + O(\tau), \quad z_{j+k} = z(t_{j+k}) + O(\tau)$$

for sufficiently small step sizes.

Let y_{j+k} and z_{j+k} denote the approximate solutions after one step of the BDF-method with initial values $y(t_{j+i})$, $z(t_{j+i})$ for $i = 0, 1, \ldots, k-1$. For the local error

$$d_{y,j+k} = y_{j+k} - y(t_{j+k}), \quad d_{z,j+k} = z_{j+k} - z(t_{j+k})$$

we have

Theorem 15.2.

$$d_{y,j+k} = O(\tau^{k+1}), \quad d_{z,j+k} = O(\tau^k).$$

Theorem 15.3. *For initial values satisfying*

$$\|y_i - y(t_i)\| = O(\tau^{k+1}) \qquad \textit{for all } i = 0, \ldots, k-1,$$

we have for the global error

$$\|y_i - y(t_i)\| = O(\tau^k), \quad \|z_i - z(t_i)\| = O(\tau^k) \qquad \textit{for all } i = k, k+1, \ldots.$$

Bibliography

[1] Uri M. Ascher and Linda R. Petzold. *Computer methods for ordinary differential equations and differential-algebraic equations.* Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics, 1998.

[2] K.E. Brenan, S.L. Campbell, and L.R. Petzold. *Numerical solution of initial-value problems in differential-algebraic equations.* Classics in Applied Mathematics. 14. Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics, 1996.

[3] Stephen L. Campbell and C.William Gear. The index of general nonlinear DAEs. *Numer. Math.*, 72(2):173–196, 1995.

[4] Peter Deuflhard and Folkmar Bornemann. *Numerische Mathematik. 2: Gewöhnliche Differentialgleichungen.* de Gruyter Lehrbuch. Berlin: de Gruyter, 2002.

[5] Griepentrog, Eberhard and März, Roswitha. *Differential-algebraic equations and their numerical treatment.* Teubner-Texte zur Mathematik, Bd. 88. Leipzig: BSB B. G. Teubner Verlagsgesellschaft, 1986.

[6] R.D. Grigorieff. *Numerik gewöhnlicher Differentialgleichungen. Band 1: Einschrittverfahren.* Teubner Studienbücher. Stuttgart: B. G. Teubner, 1972.

[7] Rolf Dieter Grigorieff and Hans Joachim Pfeiffer. *Numerik gewöhnlicher Differentialgleichungen. Band 2: Mehrschrittverfahren.* Teubner-Studienbücher: Mathematik. Stuttgart: B.G. Teubner, 1977.

[8] Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving ordinary differential equations. I: Nonstiff problems. 2nd rev. ed.* Springer Series in Computational Mathematics. 8. Berlin: Springer, 1993.

[9] Ernst Hairer and Gerhard Wanner. *Solving ordinary differential equations. II: Stiff and differential-algebraic problems. 2nd rev. ed.* Springer Series in Computational Mathematics. 14. Berlin: Springer, 1996.

[10] März, Roswitha. Numerical methods for differential algebraic equations. *Acta Numerica*, pages 141–198, 1992.

[11] L.F. Shampine and M.K. Gordon. *Computer solution of ordinary differential equations. The initial value problem.* San Francisco: W. H. Freeman & Comp., 1975.