ELSEVIER

Applied Numerical Mathematics 34 (2000) 231–252

APPLIED NUMERICAL MATHEMATICS

www.elsevier.nl/locate/apnum

The implicit Euler method for the numerical solution of singular initial value problems ☆

Othmar Koch *, Peter Kofler, Ewa B. Weinmüller

Department of Applied Mathematics and Numerical Analysis, Vienna University of Technology, Vienna, Austria

Abstract

The solvability of a certain class of singular nonlinear initial value problems is discussed. Particular attention is paid to the structure of initial conditions necessary for a bounded solution to exist. The implicit Euler rule applied to approximate the solution of the singular system is shown to be stable and to retain its classical convergence order. Moreover, the asymptotic error expansion for the global error of the above approximation is proven to have the classical structure. Finally, experimental results showing the feasibility of the approximation obtained by the Euler method to serve as a basic method for the acceleration technique known as the Iterated Defect Correction are presented.

Keywords: Ordinary differential equations; Initial value problems; Singularity of the first kind; Existence and uniqueness theory; Stability and convergence of the implicit Euler method; Asymptotic error expansion for the global error of the Euler solution; Iterated Defect Correction for singular initial value problems

1. Introduction

Mathematical models of numerous applications from physics, chemistry and mechanics (e.g., Thomas–Fermi differential equation, Ginzburg–Landau equation, singular Sturm–Liouville eigenvalue problems, Emden–Fowler equations, problems in shell buckling) take the form of systems of time-dependent partial differential equations subject to initial/boundary conditions. For the investigation of stationary solutions many of these models can be reduced to singular systems of ordinary differential equations, especially when—due to symmetries in the geometry of the problem and in the problem data—polar, cylindrical or spherical coordinates can be used.

Here we investigate initial value problems of the form

$$z'(t) = \frac{M(t)}{t} z(t) + f\big(t, z(t)\big), \quad t \in (0, 1], \tag{1a}$$

☆ This project was supported in part by the Austrian Research Fund (FWF) grant P-12507-MAT.

* Corresponding author.

0168-9274/00/$20.00
PII: S0168-9274(99)00130-0

$$B_0 z(0) = \beta, \tag{1b}$$

$$z \in C[0,1], \tag{1c}$$

where z and f are vector-valued functions of dimension n, M is a smooth $n \times n$ matrix, B_0 is an $m \times n$ matrix and β is a vector of dimension $m \leqslant n$.

We first establish the existence and uniqueness theory for (1). Especially, we describe the structure of (most general) linear initial conditions necessary and sufficient for a solution of (1) to exist. We also examine the smoothness properties of z and show that they depend not only on the smoothness of M and f but also on the eigenvalue structure of $M(0)$. Finally, we estimate z near $t = 0$ in order to show the local behavior of z close to the singularity.

Next, we study convergence of the implicit Euler rule applied on an equidistant grid with the step size h to find an approximation for the solution of (1). It turns out that standard techniques based on local stability and consistency concepts fail, and therefore we prove that the Euler method converges of order $\mathrm{O}(h)$ (as in the classical case) by explicitly inverting the associated discrete operators.

For the numerical solution obtained by the Euler method, $v_i \approx z(t_i),\ i = 0, \ldots, N$, we derive an asymptotic error expansion of the following form:

$$v_i = z(t_i) + \sum_{j=1}^{5} h^j e_j(t_i) + r_i, \quad i = 0, \ldots, N, \tag{2}$$

where $e_j \in C[0,1]$ are smooth functions and $r_i = \mathrm{O}(h^6)$. The smoothness properties of f and M from (1) which are sufficient for such an expansion to hold are given and it is shown that (2) will have an arbitrary length provided problem data is appropriately smooth.

The motivation for the analysis presented here is that in general, standard one step methods do not retain their classical order of convergence when they are applied to solve singular problems, see [7] for the behavior of Runge–Kutta methods [1]. Therefore, it can be more efficient to use a low order method first to obtain a basic approximation and then improve it by one of the well-known acceleration techniques (e.g., Iterated Defect Correction). An asymptotic error expansion for the basic approximation, cf. (2), will certainly be a crucial tool in the theoretical proof of the performance of such a procedure, see [2–4].

2. Preliminaries

Throughout the paper, the following notation is used. We denote by $\mathbb{C}^n$ the space of complex-valued vectors of dimension n and use $|\cdot|$ to denote the maximum norm in $\mathbb{C}^n$,

$$|x| = \left|(x_1, x_2, \ldots, x_n)^{\mathrm{T}}\right| = \max_{1 \leqslant i \leqslant n} |x_i|.$$

$C_n^p[0,1]$ is the space of complex vector-valued functions which are p times continuously differentiable on $[0,1]$. For every function $y \in C_n^0[0,1]$ we define the maximum norm,

$$\|y\| = \max_{0 \leqslant t \leqslant 1} \left|y(t)\right|.$$

We will also use the maximum norm restricted to the interval $[0, \delta],\ \delta > 0$,

$$\|y\|_\delta = \max_{0 \leqslant t \leqslant \delta} \left|y(t)\right|.$$

[1] Runge–Kutta methods of higher order may show an order reduction down to $\mathrm{O}(h^2)$.

$C^p_{n\times n}[0,1]$ is the space of complex-valued $n\times n$ matrices with columns in $C^p_n[0,1]$. For a matrix $A=(a_{ij})^n_{i,j=1}$, $A\in C^0_{n\times n}[0,1]$, $\|A\|$ is the induced norm,

$$\|A\| = \max_{0\leqslant t\leqslant 1}\big|A(t)\big| = \max_{0\leqslant t\leqslant 1}\left(\max_{1\leqslant i\leqslant n}\sum_{j=1}^{n}|a_{ij}(t)|\right).$$

Where there is no confusion we will delete the subscripts n and $n\times n$ and call $C=C[0,1]=C^0[0,1]$. For a mapping f, $\mathcal{R}(f)$ denotes the range of f.

For the numerical analysis, we define equidistant grids of the form

$$\Delta_h=(t_0,t_1,\ldots,t_N),$$

where $t_i=ih$, $i=0,\ldots,N$, $h=1/N$, and grid vectors

$$u_h=(u_0,\ldots,u_N).$$

The norm on the space of grid vectors is defined as

$$\|u_h\|_h=\max_{0\leqslant k\leqslant N}|u_k|.$$

For a continuous function $x(t)\in C[0,1]$, we denote by R_h the projection onto the space of grid vectors,

$$R_h(x)=\big(x(t_0),\ldots,x(t_N)\big).$$

3. Analytical results

Analytical properties of (1) have been studied in full detail in [8]. In this section, results crucial for the subsequent analysis will be recapitulated for the reader's convenience.

3.1. Linear problems with constant coefficient matrix

We begin the analysis with the discussion of the following linear problem with a constant coefficient matrix M:

$$z'(t)=\frac{M}{t}z(t)+f(t),\quad t\in(0,1], \tag{3a}$$

$$B_0z(0)=\beta, \tag{3b}$$

$$z\in C[0,1], \tag{3c}$$

where B_0 is an $m\times n$-matrix and $\beta\in\mathbb{C}^m$, $m\leqslant n$.

Here, we only examine the case where the real parts of the eigenvalues of M are nonpositive since otherwise the problem (3) may not be uniquely solvable for any initial condition (3b). As an example, consider the scalar problem

$$z'(t)=\frac{1}{t}z(t),$$

whose general solution is $z(t)=ct$, $c\in\mathbb{C}$. Clearly, any initial condition posed at $t=0$ fails to determine c, and consequently a unique solution of the above equation. For a complete discussion, see [8].

First, we transform (3a) to a simpler form. Let J be the Jordan canonical form of M, $J = E^{-1}ME$, where E is the matrix of the generalized eigenvectors of M. With $v(t) := E^{-1}z(t)$ and $g(t) := E^{-1}f(t)$ we obtain the (almost uncoupled) system

$$v'(t) = \frac{J}{t}v(t) + g(t). \tag{4}$$

To simplify matters, we assume $J \in \mathbb{C}^{n\times n}$ to consist of only one box,

$$J = \begin{pmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \lambda & 1 \\ 0 & & & \lambda \end{pmatrix}, \quad \lambda = \sigma + \mathrm{i}\rho \in \mathbb{C}. \tag{5}$$

Every solution of (4) has the form

$$v(t) = \Phi(t)c + \Phi(t)\int_1^t \Phi^{-1}(\tau)g(\tau)\,\mathrm{d}\tau, \tag{6}$$

where $c \in \mathbb{C}^n$ is an arbitrary vector and

$$\Phi(t) = t^J := \exp\bigl(J\ln(t)\bigr)$$

is the fundamental solution matrix which satisfies the following matrix initial value problem:

$$\Phi'(t) = \frac{J}{t}\Phi(t), \quad \Phi(1) = I, \qquad t \in (0,1]. \tag{7}$$

For the proof of (6), see [1].

It is easy to see that the fundamental solution matrix has the form

$$t^J = t^\lambda \begin{pmatrix} 1 & \ln(t) & \dfrac{\ln(t)^2}{2} & \dots & \dfrac{\ln(t)^{n-1}}{(n-1)!} \\ 0 & 1 & \ln(t) & \dots & \dfrac{\ln(t)^{n-2}}{(n-2)!} \\ 0 & \ddots & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ln(t) \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix}. \tag{8}$$

From the structure of this matrix, it is clear that the solution $v(t)$ given by (6) is not continuous on $[0,1]$, in general. Nevertheless, suitably prescribed initial conditions yield a solution $v \in C[0,1]$. The structure of such initial conditions will be discussed in detail in Lemmas 2 and 3, where we treat the cases [2] $\sigma < 0$ and $\lambda = 0$, respectively.

The proofs of those lemmas heavily rely on the following result:

Lemma 1. *In J from* (5) *assume either $\sigma < 0$ or $\lambda = 0$. Then for*

$$u(t) := t\int_0^1 s^{-J} f(st)\,\mathrm{d}s$$

[2] We exclude the case of purely imaginary eigenvalues of M.

the following estimate holds:

$$\left|u(t)\right| \leqslant \text{const}\, t \|f\|_\delta, \quad t \in [0,\delta].$$

Lemma 2. *Let all eigenvalues of M have negative real parts. Then for every $f \in C^p[0,1]$, $p \geqslant 0$, there exists a unique solution $z \in C$ of* (3a)*. This solution has the form*

$$z(t) = t \int_0^1 s^{-M} f(st)\,\mathrm{d}s, \tag{9}$$

and satisfies $z(0) = 0$. Moreover, $z \in C^{p+1}[0,1]$ and the following estimates hold:

$$\left|z(t)\right| \leqslant \text{const}\, t \|f\|, \tag{10}$$

$$\left|z'(t)\right| \leqslant \text{const}\, \|f\|. \tag{11}$$

Lemma 3. *Let all eigenvalues of M be zero. Denote by R the projection onto the eigenspace of M and by $\widetilde{R}$ the $n \times r$ matrix consisting of the linearly independent columns of R. If $m = r$, the $r \times r$ matrix $B_0\widetilde{R}$ is nonsingular, and $f \in C^p[0,1]$, $p \geqslant 0$, then there exists a unique solution z of* (3)*. This solution has the form*

$$z(t) = \gamma + t \int_0^1 s^{-M} f(st)\,\mathrm{d}s, \tag{12}$$

where $z(0) = \gamma = \widetilde{R}(B_0\widetilde{R})^{-1}\beta \in \ker(M)$. Moreover, $z \in C^{p+1}[0,1]$ and the following estimates hold:

$$\left|z(t)\right| \leqslant \text{const}\, t \|f\| + \left|\widetilde{R}\left(B_0\widetilde{R}\right)^{-1}\beta\right|, \tag{13}$$

$$\left|z'(t)\right| \leqslant \text{const}\, \|f\|. \tag{14}$$

Obviously, Lemmas 2 and 3 can be combined to obtain the result for a general M. We state this result in the next theorem.

Theorem 4. *Let the $r \times r$ matrix $B_0\widetilde{R}$ be nonsingular. Then for every $f \in C^p[0,1]$, $p \geqslant 0$, and any vector $\beta \in \mathbb{C}^r$, there is a unique solution $z \in C^{p+1}[0,1]$ of* (3)*. This solution has the form*

$$z(t) = \widetilde{R}\left(B_0\widetilde{R}\right)^{-1}\beta + t \int_0^1 s^{-M} f(st)\,\mathrm{d}s. \tag{15}$$

Furthermore,

$$\left|z(t)\right| \leqslant \text{const}\, t \|f\| + \left|\widetilde{R}\left(B_0\widetilde{R}\right)^{-1}\beta\right|, \tag{16}$$

$$\left|z'(t)\right| \leqslant \text{const}\, \|f\|. \tag{17}$$

3.2. Linear problems with variable coefficient matrix

We consider the linear problem

$$z'(t) = \frac{M(t)}{t} z(t) + f(t), \quad t \in (0,1], \tag{18a}$$

$$B_0 z(0) = \beta, \tag{18b}$$

$$M(0) z(0) = 0, \tag{18c}$$

where $M(t)$ has the form

$$M(t) = M + t\,\mathring{C}(t), \quad \mathring{C} \in C[0,1], \tag{19}$$

which means that (18a) is equivalent to

$$z'(t) = \frac{M}{t} z(t) + \mathring{C}(t) z(t) + f(t), \quad t \in (0,1]. \tag{20}$$

Theorem 5. *If $B_0 \widetilde{R}$ is nonsingular, then for every f, $\mathring{C} \in C^p$, $p \geqslant 0$, there exists a unique solution z of* (18)*. This solution satisfies $z \in C^{p+1}[0,1]$. Moreover, the following estimates hold*:

$$|z(t)| \leqslant \text{const}\, t\big(\|f\|_\delta + |\widetilde{R}(B_0\widetilde{R})^{-1}\beta|\big) + |\widetilde{R}(B_0\widetilde{R})^{-1}\beta|, \tag{21}$$

$$|z'(t)| \leqslant \text{const}\, \big(\|f\|_\delta + |\widetilde{R}(B_0\widetilde{R})^{-1}\beta|\big). \tag{22}$$

3.3. Nonlinear problems

Finally, we investigate the nonlinear problem

$$z'(t) = \frac{M(t)}{t} z(t) + f\big(t, z(t)\big), \quad t \in (0,1], \tag{23a}$$

$$B_0 z(0) = \beta, \tag{23b}$$

$$M(0) z(0) = 0, \tag{23c}$$

where $M(t)$ is given by (19) and $f(t,z)$ is Lipschitz-continuous with respect to z on a suitably chosen domain. In the nonlinear case, we assume all quantities to be real.

Theorem 6. *Assume $f \in C^p([0,1] \times \mathbb{R}^n)$, $\mathring{C} \in C^p[0,1]$, $p \geqslant 0$, and let $f(t,z)$ satisfy a Lipschitz-condition with respect to z on $[0,1] \times \mathbb{R}^n$. Moreover, let the matrix $B_0\widetilde{R}$ be nonsingular. Then, there exists a unique solution z of* (23)*, which satisfies $z \in C^{p+1}[0,1]$.*

Proof. We first prove the result on the subinterval $[0,\delta]$. Then, standard arguments yield the extension to the whole interval.

Clearly, solving (23) on $[0,\delta]$ is equivalent to finding a fixed point of the nonlinear operator $(\mathcal{K}F_\gamma z)(t) := (\mathcal{K}Fz)(t) + \gamma$, or equivalently, solving the nonlinear integral equation

$$z(t) = (\mathcal{K}Fz)(t) + \gamma, \quad t \in [0,\delta], \tag{24}$$

where

$$(\mathcal{K}Fz)(t) = t\int_0^1 s^{-M}\big(\overset{\circ}{C}(st)z(st) + f\big(st, z(st)\big)\big)\,\mathrm{d}s$$

and

$$\gamma = \widetilde{R}\big(B_0\widetilde{R}\big)^{-1}\beta.$$

Using the Lipschitz condition for f we conclude from Lemma 1 that for a sufficiently small δ the operator $\mathcal{K}F_\gamma$ is contracting with a constant $L < 1$ on $C[0,\delta]$. Consequently, there exists a unique solution z of (24) and $z \in C[0,\delta]$. The smoothness result follows by substituting (24) into (23a). □

We now estimate z and z'. From

$$\|\mathcal{K}F_\gamma z\|_\delta - \|\mathcal{K}F_\gamma 0\|_\delta \leqslant \|\mathcal{K}F_\gamma z - \mathcal{K}F_\gamma 0\|_\delta \leqslant L\|z\|_\delta$$

we obtain

$$\|z\|_\delta \leqslant \frac{1}{1-L}\Big(|\gamma| + D\max_{\tau\in[0,\delta]}\big|f(\tau,0)\big|\Big) =: r.$$

Consequently, we can estimate f on the bounded domain given below:

$$U := [0,\delta]\times\big\{z\in\mathbb{R}^n\colon\ |z|\leqslant r\big\}$$

with

$$F_\delta := \max_{(t,z)\in U}\big|f(t,z)\big|.$$

Using this bound and Lemma 1 we finally have, for $t\in[0,\delta]$,

$$\big|z(t)\big| \leqslant \text{const}\, t\big(F_\delta + \big|\widetilde{R}\big(B_0\widetilde{R}\big)^{-1}\beta\big|\big) + \big|\widetilde{R}\big(B_0\widetilde{R}\big)^{-1}\beta\big|, \tag{25}$$

$$\big|z'(t)\big| \leqslant \text{const}\,\big(F_\delta + \big|\widetilde{R}\big(B_0\widetilde{R}\big)^{-1}\beta\big|\big). \tag{26}$$

4. Convergence of the implicit Euler method

Define an operator $F: C^1[0,1]\to C[0,1]$,

$$F(x) := \begin{pmatrix} x'(t) - \dfrac{1}{t}M(t)x(t) - f\big(t,x(t)\big), & t\in(0,1] \\ x(0)-\zeta_0 \end{pmatrix}, \tag{27}$$

and assume that the form and properties of $M(t)$, $f(t,z)$ and the initial value ζ_0 is specified by Theorem 6. We use the implicit Euler method to approximate the solution of $F(z)=0$.

For a grid vector x_h define the operator

$$F_h(x_h) := \begin{pmatrix} \dfrac{x_{i+1}-x_i}{h} - \dfrac{M(t_{i+1})}{t_{i+1}}x_{i+1} - f(t_{i+1},x_{i+1}), & i=0,\dots,N-1 \\ x_0-\zeta_0 \end{pmatrix}. \tag{28}$$

Clearly, the approximation z_h for the solution of $F(z) = 0$ solves the nonlinear scheme $F_h(z_h) = 0$. Our aim is to show that the Euler method retains its classical convergence order, or equivalently, that the global error $\varepsilon_h := R_h(z) - z_h$ satisfies

$$\|\varepsilon_h\|_h = \mathrm{O}(h), \quad h \to 0.$$

The defining system for ε_h reads

$$\begin{pmatrix} \dfrac{\varepsilon_{i+1} - \varepsilon_i}{h} - \dfrac{M(t_{i+1})}{t_{i+1}}\varepsilon_{i+1} - l_{i+1}, & i = 0, \ldots, N-1 \\ \varepsilon_0 - l_0 \end{pmatrix} = 0,$$

where $l_i = F_h(R_h(z))_i - f(t_i, z_i) + f(t_i, z(t_i))$, $i = 1, \ldots, N$, $l_0 = 0$, and we prove the convergence result by discussing consistency and stability of the associated discrete operators on a suitably chosen interval $[0, \delta]$, $0 < \delta \leqslant 1$.

4.1. Consistency

Lemma 7. *Assume that the solution z of $F(z) = 0$ satisfies $z \in C^2[0, 1]$. Then,*

$$\big\|F_h\big(R_h(z)\big)\big\|_h = \mathrm{O}(h), \quad h \to 0.$$

Proof. Using Taylor expansion, we obtain

$$\begin{aligned} \big|F_h\big(R_h(z)\big)_{i+1}\big| &= \left|\frac{z(t_{i+1}) - z(t_i)}{h} - z'(t_{i+1})\right| = h\left|-\int_0^1 (1-\tau)z''(t_{i+1} - \tau h)\,\mathrm{d}\tau\right| \\ &\leqslant \frac{h}{2}\max_{\theta\in[t_i, t_{i+1}]}\big|z''(\theta)\big| = \mathrm{O}(h), \quad i = 0, \ldots, N-1, \end{aligned}$$

and the result follows on noting that $F_h(R_h(z))_0 = z(0) - \zeta_0 = 0$. □

4.2. Stability

We first show the stability result for the linear problem with a constant coefficient matrix and then generalize the statement for the linear case with $M(t) = M(0) + \overset{\circ}{C}(t)$ and for the nonlinear case.

4.2.1. Constant coefficient matrix

Consider the problem

$$F^{(1)}(z) := \begin{pmatrix} z'(t) - \dfrac{M}{t}z(t) - f(t), & t \in (0, 1] \\ z(0) - \zeta_0 \end{pmatrix} = 0, \tag{29}$$

and the Euler scheme for its numerical solution,

$$F_h^{(1)}(z_h) := \begin{pmatrix} \dfrac{z_{i+1} - z_i}{h} - \dfrac{M}{t_{i+1}}z_{i+1} - f_{i+1}, & i = 0, \ldots, N-1 \\ z_0 - \zeta_0 \end{pmatrix} = 0, \tag{30}$$

where $f_i = f(t_i),\ i = 1, \ldots, N$, and M is a constant $n \times n$ matrix [3]. For the global error we have

$$\begin{pmatrix} \dfrac{\varepsilon_{i+1} - \varepsilon_i}{h} - \dfrac{M}{t_{i+1}}\varepsilon_{i+1} - l_{i+1}, \quad i = 0, \ldots, N-1 \\ \varepsilon_0 - l_0 \end{pmatrix} = 0, \tag{31}$$

where $l_h = F_h^{(1)}(R_h(z)),\ l_0 = 0$.

Motivated by the technique used in the investigation of the analytical problem, we transform (30), and consequently (31), using the Jordan decomposition of M, $J = E^{-1}ME$. With $v_h := E^{-1}\varepsilon_h := (E^{-1}\varepsilon_0, \ldots, E^{-1}\varepsilon_N)$ and $g_h := E^{-1}l_h$ the problem (31) reduces to

$$\begin{pmatrix} \dfrac{v_{i+1} - v_i}{h} - \dfrac{J}{t_{i+1}}v_{i+1} - g_{i+1}, \quad i = 0, \ldots, N-1 \\ v_0 \end{pmatrix} = 0. \tag{32}$$

Note that for a fixed i this is a system of n scalar equations relating the components of $v_i = (v_{i;1}, \ldots, v_{i;n})$ and v_{i+1}. Moreover, each equation is in one of the four following forms:

$$\begin{pmatrix} \dfrac{v_{i+1;j} - v_{i;j}}{h} - \dfrac{\lambda}{t_{i+1}}v_{i+1;j} - g_{i+1;j}, \quad i = 0, \ldots, N-1 \\ v_{0;j} \end{pmatrix} = 0, \tag{33}$$

$$\begin{pmatrix} \dfrac{v_{i+1;j} - v_{i;j}}{h} - \dfrac{1}{t_{i+1}}(\lambda v_{i+1;j} + v_{i+1;j+1}) - g_{i+1;j}, \quad i = 0, \ldots, N-1 \\ v_{0;j} \end{pmatrix} = 0, \tag{34}$$

$$\begin{pmatrix} \dfrac{v_{i+1;j} - v_{i;j}}{h} - g_{i+1;j}, \quad i = 0, \ldots, N-1 \\ v_{0;j} \end{pmatrix} = 0, \tag{35}$$

$$\begin{pmatrix} \dfrac{v_{i+1;j} - v_{i;j}}{h} - \dfrac{1}{t_{i+1}}v_{i+1;j+1} - g_{i+1;j}, \quad i = 0, \ldots, N-1 \\ v_{0;j} \end{pmatrix} = 0. \tag{36}$$

Before estimating v_h in terms of g_h we need two technical results stated below.

Lemma 8. *Let $\lambda = \sigma + \mathrm{i}\kappa \in \mathbb{C}$ with $\sigma = \Re(\lambda) > 0$. For $j \geqslant k \geqslant 1$, define*

$$z_{kj}(\lambda) := \begin{cases} 1, & k = j, \\ \displaystyle\prod_{l=k}^{j-1}\left(1 + \frac{\lambda}{l}\right)^{-1}, & 1 \leqslant k < j,\ j = 2, 3, \ldots. \end{cases}$$

Then, there exist constants $\eta = \eta(\lambda) > 0$ and $C \geqslant 1$ such that

$$\left|z_{kj}(\lambda)\right| \leqslant C\left(\frac{k}{j}\right)^{\eta}, \quad 1 \leqslant k \leqslant j,\ j = 1, 2, \ldots. \tag{37}$$

[3] Again, all eigenvalues of M are assumed to have only nonpositive real parts.

Proof. The estimate for z_{kj} follows from

$$|z_{kj}(\lambda)| = \left|\prod_{l=k}^{j-1}\left(\frac{l+\lambda}{l}\right)^{-1}\right| = \left|\prod_{l=k}^{j-1}\frac{l}{l+\lambda}\right| = \left|\frac{\Gamma(j)}{\Gamma(j+\lambda)}\frac{\Gamma(k+\lambda)}{\Gamma(k)}\right|$$
$$= \left|j^{-\lambda}\left(1+\mathrm{O}\left(\frac{1}{j}\right)\right)k^{\lambda}\left(1+\mathrm{O}\left(\frac{1}{k}\right)\right)\right| \leqslant C\left(\frac{k}{j}\right)^{\eta},$$

on noting that

$$\Gamma(z+1) = z\Gamma(z), \quad \forall z \in \mathbb{C},$$

and taking into account the following asymptotic behavior of the function Γ (cf. [11]):

$$\frac{\Gamma(s+a)}{\Gamma(s+b)} = s^{a-b}\left(1+\mathrm{O}\left(\frac{1}{s}\right)\right), \quad s \to \infty,\ \Re(s) > 0. \qquad \square$$

The proof of the next lemma follows in a straightforward manner using an integral as a bound for the sum.

Lemma 9. *Let $h > 0$, $t_j := jh$, $k > j \geqslant 1$ and $\gamma \in \mathbb{R}$, then*

$$\sum_{l=j}^{k-1} h t_l^{\gamma-1} \leqslant \begin{cases} \text{const } |t_k^{\gamma} - t_j^{\gamma}|, & \gamma \neq 0, \\ \text{const } \ln\left(\dfrac{t_k}{t_j}\right), & \gamma = 0. \end{cases} \tag{38}$$

We now estimate the solutions of (33)–(36), see Lemmas 10–13, respectively. All proofs in these lemmas are shown in a fairly similar manner. In order to avoid repetitions, we carry out only the second in some detail. Motivated by the linear problem with a variable coefficient matrix and by the nonlinear problem, we provide in each case an additional uniform bound for the solution on grids defined on the subinterval $[0, \delta]$, $\delta \leqslant 1$.

Lemma 10. *Consider the scheme* (33). *Its solution $v_{h;j}$ has the form*

$$v_{i;j} = \sum_{l=1}^{i}\prod_{k=l}^{i}\left(1-\frac{h\lambda}{t_k}\right)^{-1} h g_{l;j} =: \sum_{l=1}^{i} z_{l,i+1}(-\lambda) h g_{l;j}, \quad i = 1, \ldots, N, \tag{39}$$

and satisfies the following estimate:

$$|v_{i;j}| \leqslant \text{const } t_i \max_{1\leqslant l\leqslant N} |g_{l;j}| \tag{40}$$
$$\leqslant \text{const } \delta \|g_{h;j}\|_h, \quad i = 0, \ldots, N. \tag{41}$$

Lemma 11. *Consider the scheme* (34). *Then*

$$v_{i;j} = \sum_{l=1}^{i}\prod_{k=l}^{i}\left(1-\frac{h\lambda}{t_k}\right)^{-1} h \widetilde{g}_{l;j} =: \sum_{l=1}^{i} z_{l,i+1}(-\lambda) h \widetilde{g}_{l;j}, \quad i = 1, \ldots, N, \tag{42}$$

where $\widetilde{g}_{i;j} := t_i^{-1} v_{i;j+1} + g_{i;j}$, $i = 1, \ldots, N$. If $v_{h;j+1}$ can be bounded by

$$|v_{i;j+1}| \leqslant \text{const } t_i \max_{1\leqslant k\leqslant N} |g_{k;j+1}|, \quad i = 0, \ldots, N,$$

then the following estimates hold:

$$|v_{i;j}| \leqslant \text{const}\, t_i \max_{1\leqslant k\leqslant N} \max\{|g_{k;j}|, |g_{k;j+1}|\} \tag{43}$$

$$\leqslant \text{const}\, \delta \max\{\|g_{h;j}\|_h, \|g_{h;j+1}\|_h\}, \quad i = 0, \ldots, N. \tag{44}$$

Proof. The representation (42) can be derived using the representation

$$v_{i+1;j} = \left(1 - \frac{h\lambda}{t_{i+1}}\right)^{-1} (v_{i;j} + h\widetilde{g}_{i+1;j})$$

and induction. The estimates follow from

$$\begin{aligned} |v_{i;j}| &\leqslant \text{const} \left(\sum_{l=1}^{i} \frac{t_l^\eta}{t_{i+1}^\eta} h t_l^{-1} |v_{l;j+1}| + \sum_{l=1}^{i} \frac{t_l^\eta}{t_{i+1}^\eta} h |g_{l;j}|\right) \\ &\leqslant \text{const} \left(\frac{1}{t_{i+1}^\eta} \sum_{l=1}^{i} h t_l^\eta \max_{1\leqslant k\leqslant N} |g_{k;j+1}| + \frac{1}{t_{i+1}^\eta} \sum_{l=1}^{i} h t_l^\eta \max_{1\leqslant k\leqslant N} |g_{k;j}|\right) \\ &\leqslant \text{const}\, t_i \max_{1\leqslant k\leqslant N} \max\{|g_{k;j}|, |g_{k;j+1}|\}. \qquad \square \end{aligned}$$

Lemma 12. *Consider the scheme* (35). *Its solution* $v_{h;j}$ *has the form*

$$v_{i;j} = \sum_{l=1}^{i} h g_{l;j}, \quad i = 1, \ldots, N, \tag{45}$$

and satisfies the following estimate:

$$|v_{i;j}| \leqslant \text{const}\, t_i \max_{1\leqslant k\leqslant N} |g_{k;j}| \tag{46}$$

$$\leqslant \text{const}\, \delta \|g_{h;j}\|_h, \quad i = 0, \ldots, N. \tag{47}$$

Lemma 13. *Consider the scheme* (36). *Its solution* $v_{h;j}$ *reads*

$$v_{i;j} = \sum_{l=1}^{i} h \widetilde{g}_{l;j}, \quad i = 0, \ldots, N-1, \tag{48}$$

where $\widetilde{g}_{i;j} := t_i^{-1} v_{i;j+1} + g_{i;j}$, $i = 1, \ldots, N$. *If*

$$|v_{i;j+1}| \leqslant \text{const}\, t_i \max_{1\leqslant k\leqslant N} |g_{k;j+1}|, \quad i = 0, \ldots, N,$$

then

$$|v_{i;j}| \leqslant \text{const}\, t_i \max_{1\leqslant k\leqslant N} \max\{|g_{k;j}|, |g_{k;j+1}|\} \tag{49}$$

$$\leqslant \text{const}\, \delta \max\{\|g_{h;j}\|_h, \|g_{h;j+1}\|_h\}, \quad i = 0, \ldots, N. \tag{50}$$

The convergence result for the implicit Euler method applied to solve (29) is a simple consequence of the results of Lemmas 7 and 10–13.

Theorem 14. *Consider the scheme* (30) *with the linear operator* $F_h^{(1)}$. *For every* $f \in C$ *and every starting value* ζ_0 *this scheme,* $F_h^{(1)}(z_h) = 0$, *has a unique solution* z_h. *If the solution* z *of the underlying analytical problem* $F^{(1)}(z) = 0$, *cf.* (29), *satisfies* $z \in C^2[0,1]$, *then the global error of the solution of* (30) *satisfies*

$$\|\varepsilon_h\|_h = \mathrm{O}(h), \quad h \to 0.$$

4.2.2. Variable coefficient matrix

Now we consider the linear problem with variable coefficient matrix of the form (19),

$$F^{(2)}(z) := \begin{pmatrix} z'(t) - \dfrac{M(0)}{t} z(t) - \overset{\circ}{C}(t) z(t) - f(t), & t \in (0,1] \\ z(0) - \zeta_0 \end{pmatrix} = 0. \tag{51}$$

The associated discrete operator equation reads

$$F_h^{(2)}(z_h) := \begin{pmatrix} \dfrac{z_{i+1} - z_i}{h} - \dfrac{M(0)}{t_{i+1}} z_{i+1} - \overset{\circ}{C}(t_{i+1}) z_{i+1} - f_{i+1}, & i = 0, \ldots, N-1 \\ z_0 - \zeta_0 \end{pmatrix} = 0. \tag{52}$$

We first have to show that (52) has a unique solution z_h. Define $x_h = G_h(y_h)$ as the solution of

$$\begin{pmatrix} \dfrac{x_{i+1} - x_i}{h} - \dfrac{M(0)}{t_{i+1}} x_{i+1} - \overset{\circ}{C}(t_{i+1}) y_{i+1} - f_{i+1}, & i = 0, \ldots, N-1 \\ x_0 - \zeta_0 \end{pmatrix} = 0. \tag{53}$$

Then the solution of (52) is equivalent to finding a fixed point of G_h. We use the Banach Fixed Point Theorem to show the existence of such a fixed point. We first note that $v_h = G_h(x_h) - G_h(y_h)$ solves

$$\begin{pmatrix} \dfrac{v_{i+1} - v_i}{h} - \dfrac{M(0)}{t_{i+1}} v_{i+1} - l_{i+1}, & i = 0, \ldots, N-1 \\ v_0 \end{pmatrix} = 0, \tag{54}$$

where $l_i = \overset{\circ}{C}(t_i)(x_i - y_i)$, $i = 1, \ldots, N$. Using estimates derived in the previous section we obtain

$$|v_i| \leqslant \text{const}\, t_i \max_{1 \leqslant k \leqslant N} |l_k| \leqslant \text{const}\, \delta \max_{1 \leqslant k \leqslant N} |x_k - y_k|, \quad i = 0, \ldots, N,$$

and this implies that on the interval $[0, \delta]$ with a sufficiently small δ, G_h is a contraction on the space of grid vectors u_h with $u_0 = \zeta_0$. Thus, (52) has a unique solution on $[0, \delta]$.

To prove stability, we note that the global error ε_h satisfies the equation

$$\begin{pmatrix} \dfrac{\varepsilon_{i+1} - \varepsilon_i}{h} - \dfrac{M(0)}{t_{i+1}} \varepsilon_{i+1} - \overset{\circ}{C}(t_{i+1}) \varepsilon_{i+1} - l_{i+1}, & i = 0, \ldots, N-1 \\ \varepsilon_0 - l_0 \end{pmatrix} = 0, \tag{55}$$

where $l_h = F_h^{(2)}(R_h(z))$, $l_0 = 0$. The existence of a unique solution ε_h of (55) follows by the argument used for (52). Denote by $L < 1$ the Lipschitz-constant of the operator G_h. Then

$$\|\varepsilon_h\|_h - \|G_h(0)\|_h = \|G_h(\varepsilon_h)\|_h - \|G_h(0)\|_h \leqslant \|G_h(\varepsilon_h) - G_h(0)\|_h \leqslant L\|\varepsilon_h\|_h$$
$$\Rightarrow \|\varepsilon_h\|_h \leqslant \frac{1}{1-L}\|G_h(0)\|_h.$$

Also, $w_h = G_h(0)$ is the solution of

$$\begin{pmatrix} \dfrac{w_{i+1} - w_i}{h} - \dfrac{M(0)}{t_{i+1}} w_{i+1} - l_{i+1}, \quad i = 0, \ldots, N-1 \\ w_0 \end{pmatrix} = 0, \tag{56}$$

and satisfies

$$\|w_h\|_h \leqslant \text{const}\, \delta \|l_h\|_h.$$

Consequently,

$$\|\varepsilon_h\|_h \leqslant \text{const}\, \delta \|l_h\|_h$$

on a grid defined on $[0, \delta]$. This together with consistency yields the convergence result formulated in the next theorem.

Theorem 15. *Consider the scheme* (52) *with the operator* $F_h^{(2)}$. *For every* f, $\overset{\circ}{C} \in C$ *and every starting value* ζ_0 *there exists a unique solution* z_h *of* (52). *If the solution* z *of the analytical problem* (52) *is in* $C^2[0, 1]$, *then the global error satisfies*

$$\|\varepsilon_h\|_h = \mathrm{O}(h), \quad h \to 0.$$

Proof. Clearly, the statement holds for grids defined on the interval $[0, \delta]$ by means of the previous considerations. Since on $[\delta, 1]$ the problem is regular, classical theory can be applied to extend the result. □

4.2.3. Nonlinear problems

Finally, we consider the nonlinear problem

$$F(z) = \begin{pmatrix} z'(t) - \dfrac{M(0)}{t} z(t) - \overset{\circ}{C}(t) z(t) - f\big(t, z(t)\big), \quad t \in (0, 1] \\ z(0) - \zeta_0 \end{pmatrix} = 0. \tag{57}$$

The associated numerical scheme is $F_h(z_h) = 0$, where

$$F_h(z_h) = \begin{pmatrix} \dfrac{z_{i+1} - z_i}{h} - \dfrac{M(0)}{t_{i+1}} z_{i+1} - \overset{\circ}{C}(t_{i+1}) z_{i+1} - f(t_{i+1}, z_{i+1}), \quad i = 0, \ldots, N-1 \\ z_0 - \zeta_0 \end{pmatrix}. \tag{58}$$

We assume that $f(t, z)$ has a continuous and bounded Fréchet-derivative (denoted by $D_2 f(t, z)$) with respect to z on $[0, 1] \times \mathbb{R}^n$. To show the unique solvability of $F_h(z_h) = 0$ on a suitable interval $[0, \delta]$, we define a mapping $x_h = G_h(y_h)$ as the solution of

$$\begin{pmatrix} \dfrac{x_{i+1} - x_i}{h} - \dfrac{M(0)}{t_{i+1}} x_{i+1} - \overset{\circ}{C}(t_{i+1}) y_{i+1} - f(t_{i+1}, y_{i+1}), \quad i = 0, \ldots, N-1 \\ x_0 - \zeta_0 \end{pmatrix} = 0.$$

We now use the representation

$$f(t_i, x_i) - f(t_i, y_i) = \int_0^1 D_2 f\big(t_i, y_i + \tau(x_i - y_i)\big)\,\mathrm{d}\tau (x_i - y_i) =: D_i(x_i - y_i)$$

to show that G_h is a contraction on the space of grid vectors on $[0, \delta]$, provided that δ is sufficiently small.

The global error, ε_h, satisfies

$$\begin{pmatrix} \dfrac{\varepsilon_{i+1} - \varepsilon_i}{h} - \dfrac{M(0)}{t_{i+1}}\varepsilon_{i+1} - \big(\overset{\circ}{C}(t_{i+1}) + D_{i+1}\big)\varepsilon_{i+1} - l_{i+1}, \quad i = 0, \ldots, N-1 \\ \varepsilon_0 \end{pmatrix} = 0,$$

where $l_h = F_h(R_h(z))$. We now view the operator $x_h = G_h(y_h)$ as the solution of the system

$$\begin{pmatrix} \dfrac{x_{i+1} - x_i}{h} - \dfrac{M(0)}{t_{i+1}}x_{i+1} - \big(\overset{\circ}{C}(t_{i+1}) + D_{i+1}\big)y_{i+1} - l_{i+1}, \quad i = 0, \ldots, N-1 \\ x_0 \end{pmatrix} = 0,$$

and conclude

$$\begin{aligned} &\|\varepsilon_h\|_h - \big\|G_h(0)\big\|_h \leqslant \big\|G_h(\varepsilon_h) - G_h(0)\big\|_h \leqslant L\|\varepsilon_h\|_h \\ &\quad \Rightarrow \|\varepsilon_h\|_h \leqslant \frac{1}{1-L}\big\|G_h(0)\big\|_h = \mathrm{O}(h), \quad h \to 0. \end{aligned}$$

The convergence result required for the interval $[\delta, 1]$ follows by the classical theory.

Theorem 16. *Consider the system* (58) *with the nonlinear operator* $F_h(z_h)$. *For every* $f(t, z) \in C([0, 1] \times \mathbb{R}^n)$ *which has a bounded Fréchet-derivative with respect to* z *on* $[0, 1] \times \mathbb{R}^n$, $\overset{\circ}{C} \in C$ *and every starting value* ζ_0, $F_h(z_h) = 0$ *has a unique solution* z_h. *If the solution* z *of the underlying analytical problem* (57) *satisfies* $z \in C^2[0, 1]$, *then the global error of the approximate solution satisfies*

$$\|\varepsilon_h\|_h = \mathrm{O}(h), \quad h \to 0.$$

5. Asymptotic error expansions

An asymptotic expansion for the global error of the approximation to the solution of (23) obtained by the implicit Euler rule is a crucial tool used in the proof of the convergence results for the Iterated Defect Correction Method (IDeC), see Section 6. This convergence result is still "work in progress" and therefore we restrict our attention here to the special case of an expansion of order five which on one hand is sufficient to explain the experimental results for the IDeC procedure presented in the next section, and on the other hand makes it intuitively clear that under suitable assumptions on the problem data the length of the expansion can be extended to an arbitrary order.

Let v_h be a solution of $F_h(v_h) = 0$,

$$F_h(v_h) = \begin{pmatrix} \dfrac{v_{i+1} - v_i}{h} - \dfrac{M(t_{i+1})}{t_{i+1}}v_{i+1} - f(t_{i+1}, v_{i+1}), \quad i = 0, \ldots, N-1 \\ v_0 - \zeta_0 \end{pmatrix}. \tag{59}$$

For the global error of v_h on the grid Δ_h we make an ansatz of the form

$$v_i - z(t_i) = \sum_{j=1}^{5} h^j e_j(t_i) + r_i, \quad i = 0, \ldots, N,$$

where $e_j(t)$ are appropriately smooth functions, and r_h is the remainder term. More precisely, we have to prove that such smooth functions exist and the remainder term shows the proper asymptotic quality,

$$\|r_h\|_h = \mathrm{O}\big(h^6\big), \quad h \to 0.$$

Let $v(t)$ be a smooth interpolant of the values v_h, then we can rewrite the ansatz and obtain

$$v(t) = z(t) + \sum_{j=1}^{5} h^j e_j(t) + r(t), \quad t \in [0,1]. \tag{60}$$

Let us assume $f, \overset{\circ}{C} \in C^6[0,1]$, cf. (23), then it follows from Theorem 6 that $z \in C^7[0,1]$. We first evaluate (60) at t_i and t_{i+1} and substitute into (59). Local Taylor expansion at t_{i+1} for all involved functions and equating coefficients of equal powers of h yield the following variational equations for e_j, $j = 1, \ldots, 5$:

$$\begin{aligned}
&e_1'(t) - \frac{1}{t}M(t)e_1(t) = D_2 f\big(t, z(t)\big)e_1(t) + \tfrac{1}{2}z''(t), \quad t \in (0,1],\\
&e_1(0) = 0,\\
&e_2'(t) - \frac{1}{t}M(t)e_2(t) = D_2 f\big(t, z(t)\big)e_2(t) + \tfrac{1}{2}D_2^2 f\big(t, z(t)\big)e_1^2(t) + \tfrac{1}{2}e_1''(t) - \tfrac{1}{6}z^{(3)}(t), \quad t \in (0,1],\\
&e_2(0) = 0,\\
&e_3'(t) - \frac{1}{t}M(t)e_3(t)\\
&\quad = D_2 f\big(t, z(t)\big)e_3(t) + D_2^2 f\big(t, z(t)\big)e_1(t)e_2(t) + \tfrac{1}{6}D_2^3 f\big(t, z(t)\big)e_1^3(t) + \tfrac{1}{2}e_2''(t)\\
&\qquad - \tfrac{1}{6}e_1^{(3)}(t) + \tfrac{1}{24}z^{(4)}(t), \quad t \in (0,1],\\
&e_3(0) = 0,\\
&e_4'(t) - \frac{1}{t}M(t)e_4(t)\\
&\quad = D_2 f\big(t, z(t)\big)e_4(t) + D_2^2 f\big(t, z(t)\big)\big(e_1(t)e_3(t) + \tfrac{1}{2}e_2^2(t)\big) + \tfrac{1}{2}D_2^3 f\big(t, z(t)\big)e_1^2(t)e_2(t)\\
&\qquad + \tfrac{1}{24}D_2^4 f\big(t, z(t)\big)e_1^4(t) + \tfrac{1}{2}e_3''(t) - \tfrac{1}{6}e_2^{(3)}(t) + \tfrac{1}{24}e_1^{(4)}(t) - \tfrac{1}{120}z^{(5)}(t), \quad t \in (0,1],\\
&e_4(0) = 0,\\
&e_5'(t) - \frac{1}{t}M(t)e_5(t)\\
&\quad = D_2 f\big(t, z(t)\big)e_5(t) + D_2^2 f\big(t, z(t)\big)\big(e_1(t)e_4(t) + e_2(t)e_3(t)\big)\\
&\qquad + \tfrac{1}{2}D_2^3 f\big(t, z(t)\big)\big(e_1^2(t)e_3(t) + e_1(t)e_2^2(t)\big) + \tfrac{1}{6}D_2^4 f\big(t, z(t)\big)e_1^3(t)e_2(t)\\
&\qquad + \tfrac{1}{120}D_2^5 f\big(t, z(t)\big)e_1^5(t) + \tfrac{1}{2}e_4''(t) - \tfrac{1}{6}e_3^{(3)}(t) + \tfrac{1}{24}e_2^{(4)}(t) - \tfrac{1}{120}e_1^{(5)}(t) + \tfrac{1}{720}z^{(6)}(t), \quad t \in (0,1],\\
&e_5(0) = 0.
\end{aligned}$$

Here, $D_2^k f(t,z)$ denotes the kth Fréchet-derivative of $f(t,z)$ with respect to the second argument z (for example, $D_2^2 f(t,z)y^2$ is the bilinear mapping defined by the second derivative of f with vector y as both its arguments). For $z \in C^7$, Theorem 6 implies the existence of unique solutions $e_j \in C^{7-j}$, $j = 1, \ldots, 5$.

For the analysis of the remainder term r_h we define a function

$$g(t_i, r_i) := \int_0^1 D_2 f\left(t_i, z(t_i) + \sum_{j=1}^{5} h^j e_j(t_i) + \tau r_i\right) \mathrm{d}\tau \cdot r_i,$$

which satisfies $g(t,0) = 0$ for all t. Moreover, we assume that $g(t,r)$ has a continuous and bounded Fréchet-derivative with respect to r.

The remainder r_h solves the difference scheme

$$\begin{pmatrix} \dfrac{r_{i+1} - r_i}{h} - \dfrac{1}{t_{i+1}} M(t_{i+1}) r_{i+1} - g(t_{i+1}, r_{i+1}) - l_{i+1} \\ r_0 \end{pmatrix} = 0, \tag{61}$$

with $i = 0, \ldots, N-1$, where $l_h = \mathrm{O}(h^6)$ if $f \in C^6$. Let us define an operator $x_h = G_h(y_h)$ as the solution of

$$\begin{pmatrix} \dfrac{x_{i+1} - x_i}{h} - \dfrac{1}{t_{i+1}} M(0) x_{i+1} - \mathring{C}(t_{i+1}) y_{i+1} - g(t_{i+1}, y_{i+1}) - l_{i+1} \\ x_0 \end{pmatrix} = 0, \tag{62}$$

where $i = 0, \ldots, N-1$.

It follows from Theorem 16 that G_h is a contraction on the space of grid vectors $u = (0, u_1, \ldots, u_N)$ on $[0, \delta]$ with a constant $L < 1$, provided that δ is sufficiently small. Also,

$$\|r_h\|_h \leqslant \frac{1}{1-L} \big\|G_h(0)\big\|_h \quad \text{and} \quad \big\|G_h(0)\big\|_h \leqslant \text{const } \|l_h\|_h.$$

Thus,

$$\|r_h\|_h = \mathrm{O}\big(h^6\big), \quad h \to 0,$$

and the result follows. Obviously, if f and $\mathring{C}$ are sufficiently smooth, expansion (60) can be extended to an arbitrary order.

6. Iterated Defect Correction (IDeC)

An effective technique for the numerical solution of ODEs is the Iterated Defect Correction method, originally proposed as a method for the estimation of the global error of Runge–Kutta methods. Here, we are interested in the performance of the IDeC procedure, based on the implicit Euler method. The idea of the IDeC method is to obtain a basic solution (in our case by the implicit Euler scheme) and gradually improve its accuracy in the course of a specially designed iteration. The existence and the structure of the asymptotic error expansion for the basic solution, see Section 5, suggests that in each step of the iteration the convergence order can be improved by $\mathrm{O}(h)$ until a certain convergence order, $\mathrm{O}(h^{m_{\max}})$, is reached. The power $m_{\max}$ depends on the length of the asymptotic error expansion for the basic solution and on

technical details of the IDeC procedure which we will specify later. The aim of this section is to verify this convergence behavior experimentally.

6.1. The IDeC method

The IDeC procedure based on Zadunaisky's idea, see [15], has been successfully applied to solve classical second order boundary value problems [4] and its performance in this context has been theoretically investigated in [2–4]. These results do not carry over to the case of singular problems directly, however. Therefore we chose an experimental approach to collect evidence for the justification to use IDeC for singular problems and to create a basis for a theoretical proof of the convergence properties to follow.

We now briefly discuss some important features of this acceleration technique. We consider initial value problems of the form [5]

$$z'(t) = F\big(t, z(t)\big), \quad t \in (0, 1], \tag{63a}$$

$$z(0) = \beta. \tag{63b}$$

We assume that we know the approximate solution, $z_h^{(0)} := z_h$, obtained by the implicit Euler rule, and denote by $p^{(0)}(t)$ the polynomial of degree N interpolating the values of $z_h^{(0)}$,

$$p^{(0)}(t_i) = z_i^{(0)}, \quad i = 0, \ldots, N.$$

Using this polynomial we construct a neighboring problem associated with (63) and solved by $p^{(0)}(t)$,

$$z'(t) = F\big(t, z(t)\big) + d^{(0)}(t), \quad t \in (0, 1], \tag{64a}$$

$$z(0) = p^{(0)}(0) = \beta, \tag{64b}$$

where

$$d^{(0)}(t) := \big(p^{(0)}\big)'(t) - F\big(t, p^{(0)}(t)\big).$$

We now solve (64) by the same numerical method (implicit Euler rule) and obtain an approximate solution $p_h^{(0)}$ for $p^{(0)}(t)$. This means that for the solution of the neighboring problem (64) we know the global error which we can use to estimate the unknown error of the original problem (63),

$$\varepsilon_h = R_h(z) - z_h \approx \delta_h^{(0)} := R_h\big(p^{(0)}\big) - p_h^{(0)} = z_h^{(0)} - p_h^{(0)}. \tag{65}$$

Zadunaisky gave the following heuristic argument for his method to work: If the values z_h are good approximations for the values of the solution $R_h(z)$ at the grid points, then the polynomial $p^{(0)}(t)$ is a good approximation for the solution $z(t)$ itself. Consequently, the defect $d^{(0)}(t)$ is small and hence the neighboring problem (64) and the original problem (63) are closely related. This implies that the global error of the solution of (64) is closely related to the global error of the solution of (63), and therefore the estimate (65) shall provide some dependable information about its size.

[4] Unfortunately, it fails to show its advantageous behavior when it is applied to solve singular boundary value problems, see [5].

[5] In case of singular problems, we may think of $F(t, z(t))$ being the right-hand side of (1a).

Having the estimate for the global error of the solution $z_h^{(0)}$ we are able to improve this solution by setting

$$z_h^{(1)} := z_h^{(0)} + \delta_h^{(0)} = z_h^{(0)} + \big(R_h(p^{(0)}) - p_h^{(0)}\big).$$

We use these values to define a new interpolating polynomial $p^{(1)}(t)$ by requiring $p^{(1)}(t_i) = z_i^{(1)},\ i = 0, \ldots, N$, and the associated defect

$$d^{(1)}(t) := \big(p^{(1)}\big)'(t) - F\big(t, p^{(1)}(t)\big).$$

Clearly, the next neighboring problem reads

$$z'(t) = F\big(t, z(t)\big) + d^{(1)}(t), \quad t \in (0, 1], \tag{66a}$$

$$z(0) = \beta, \tag{66b}$$

and we solve it by the Euler method to obtain the approximation $p_h^{(1)}$ which is used to correct the basic solution again,

$$z_h^{(2)} := z_h^{(0)} + \delta_h^{(1)} = z_h^{(0)} + \big(R_h(p^{(1)}) - p_h^{(1)}\big).$$

Clearly, the procedure can be continued in the above manner.

For obvious reasons one does not use one interpolating polynomial for the whole interval in practice. Instead, a piecewise polynomial function, composed of polynomials of (moderate) degree m is defined to specify the neighboring problem. Due to classical theory, see [4], $m_{\max} = m$ holds provided that the remainder term is at least $\mathrm{O}(h^{m+1})$, which means that in general the choice of m determines the final level of accuracy reached during the iteration.

6.2. Experimental results

In this section we present experimental results illustrating the performance of the IDeC method when applied to solve singular problems. Two examples have been chosen to show the typical behavior observed for all models.

In the tables we denote by $\varepsilon_h^{(k)},\ k = 1, \ldots, 5$, the (exact) global error of the solution $z_h^{(k)}$ obtained in the kth step of the iteration and by $\varepsilon_h^{(0)}$ the error of the basic solution. Also, convergence orders and error constants are listed. The theory suggests that for a sufficiently small h the relation $\|\varepsilon_h^{(k)}\|_h \approx ch^p$ holds, where c is a constant independent of h. We use this relation to compute the order of convergence $p^{(k)}$ and the error constant $c^{(k)}$ by comparing the errors on two subsequent grids.

In all experiments polynomial degree $m = 5$ was used and hence we expect to see the following order sequence:

$$\big\|\varepsilon_h^{(k)}\big\|_h = \big\|R_h(z) - z_h^{(k)}\big\|_h = \mathrm{O}\big(h^{k+1}\big), \quad k = 0, \ldots, 4.$$

All experiments have been carried out in double precision, relative machine accuracy 10^{-16} using an Intel Pentium processor.

Example 1. We consider the problem, see [14],

$$y''(t) = -\frac{2}{t}y'(t) - n^2\cos(nt) - \frac{2}{t}n\sin(nt), \quad t \in (0, 1],$$

$$y(0) = 2, \qquad y'(0) = 0,$$

with $n = 3$ and exact solution $y(t) = 1 + \cos(nt)$.

Table 1
Convergence of the IDeC method; Example 1

h	$\|\varepsilon_h^{(0)}\|_h$	$p^{(0)}$	$c^{(0)}$	$\|\varepsilon_h^{(1)}\|_h$	$p^{(1)}$	$c^{(1)}$
$\frac{1}{5}$	8.5	0.970	$-4.0 \cdot 10^{+01}$	5.0	1.924	$-1.1 \cdot 10^{+01}$
$\frac{1}{5} \cdot 2^{-1}$	4.3	0.982	$-4.1 \cdot 10^{+01}$	$1.3 \cdot 10^{-01}$	1.985	$-1.2 \cdot 10^{+01}$
$\frac{1}{5} \cdot 2^{-2}$	$2.1 \cdot 10^{-01}$	0.990	$-4.2 \cdot 10^{+01}$	$3.3 \cdot 10^{-01}$	1.995	$-1.3 \cdot 10^{+01}$
$\frac{1}{5} \cdot 2^{-3}$	$1.1 \cdot 10^{-01}$	0.995	$-4.3 \cdot 10^{+01}$	$8.3 \cdot 10^{-02}$	1.997	$-1.3 \cdot 10^{+01}$
$\frac{1}{5} \cdot 2^{-4}$	$5.5 \cdot 10^{-01}$	0.997	$-4.3 \cdot 10^{+01}$	$2.0 \cdot 10^{-03}$	1.999	$-1.3 \cdot 10^{+01}$
$\frac{1}{5} \cdot 2^{-5}$	$2.7 \cdot 10^{-02}$	0.998	$-4.4 \cdot 10^{+01}$	$5.2 \cdot 10^{-03}$	1.999	$-1.3 \cdot 10^{+01}$
$\frac{1}{5} \cdot 2^{-6}$	$1.3 \cdot 10^{-02}$	0.999	$-4.4 \cdot 10^{+01}$	$1.3 \cdot 10^{-04}$	1.999	$-1.3 \cdot 10^{+01}$
$\frac{1}{5} \cdot 2^{-7}$	$6.9 \cdot 10^{-02}$	0.999	$-4.4 \cdot 10^{+01}$	$3.2 \cdot 10^{-04}$	1.999	$-1.3 \cdot 10^{+01}$
$\frac{1}{5} \cdot 2^{-8}$	$3.4 \cdot 10^{-02}$	0.999	$-4.4 \cdot 10^{+01}$	$8.1 \cdot 10^{-05}$	1.999	$-1.3 \cdot 10^{+01}$
$\frac{1}{5} \cdot 2^{-9}$	$1.7 \cdot 10^{-03}$	0.999	$-4.4 \cdot 10^{+01}$	$2.0 \cdot 10^{-06}$	1.999	$-1.3 \cdot 10^{+01}$
h	$\|\varepsilon_h^{(2)}\|_h$	$p^{(2)}$	$c^{(2)}$	$\|\varepsilon_h^{(3)}\|_h$	$p^{(3)}$	$c^{(3)}$
$\frac{1}{5}$	$1.0 \cdot 10^{-01}$	3.275	$-1.9 \cdot 10^{+01}$	$1.1 \cdot 10^{-01}$	4.526	$-1.7 \cdot 10^{+02}$
$\frac{1}{5} \cdot 2^{-1}$	$1.0 \cdot 10^{-02}$	2.792	$-6.4 \cdot 10^{+01}$	$5.0 \cdot 10^{-02}$	4.053	$-5.7 \cdot 10^{+02}$
$\frac{1}{5} \cdot 2^{-2}$	$1.5 \cdot 10^{-03}$	2.916	$-9.3 \cdot 10^{+01}$	$3.0 \cdot 10^{-04}$	3.988	$-4.7 \cdot 10^{+02}$
$\frac{1}{5} \cdot 2^{-3}$	$1.9 \cdot 10^{-04}$	2.959	$-1.0 \cdot 10^{+01}$	$1.9 \cdot 10^{-05}$	3.986	$-4.6 \cdot 10^{+02}$
$\frac{1}{5} \cdot 2^{-4}$	$2.5 \cdot 10^{-05}$	2.979	$-1.2 \cdot 10^{+01}$	$1.2 \cdot 10^{-06}$	3.991	$-4.7 \cdot 10^{+02}$
$\frac{1}{5} \cdot 2^{-5}$	$3.2 \cdot 10^{-05}$	2.989	$-1.2 \cdot 10^{+01}$	$7.6 \cdot 10^{-07}$	3.995	$-4.8 \cdot 10^{+02}$
$\frac{1}{5} \cdot 2^{-6}$	$4.0 \cdot 10^{-06}$	2.994	$-1.3 \cdot 10^{+01}$	$4.7 \cdot 10^{-08}$	3.997	$-4.9 \cdot 10^{+02}$
$\frac{1}{5} \cdot 2^{-7}$	$5.1 \cdot 10^{-07}$	2.997	$-1.3 \cdot 10^{+01}$	$2.9 \cdot 10^{-10}$	3.982	$-4.4 \cdot 10^{+02}$
h	$\|\varepsilon_h^{(4)}\|_h$	$p^{(4)}$	$c^{(4)}$	$\|\varepsilon_h^{(5)}\|_h$	$p^{(5)}$	$c^{(5)}$
$\frac{1}{5}$	$1.5 \cdot 10^{-02}$	3.993	$-9.6 \cdot 10^{+01}$	$1.5 \cdot 10^{-02}$	3.987	$-9.5 \cdot 10^{+01}$
$\frac{1}{5} \cdot 2^{-1}$	$9.7 \cdot 10^{-03}$	5.105	$-1.2 \cdot 10^{+02}$	$9.8 \cdot 10^{-03}$	5.104	$-1.2 \cdot 10^{+02}$
$\frac{1}{5} \cdot 2^{-2}$	$2.8 \cdot 10^{-05}$	5.058	$-1.0 \cdot 10^{+02}$	$2.8 \cdot 10^{-05}$	5.079	$-1.1 \cdot 10^{+02}$
$\frac{1}{5} \cdot 2^{-3}$	$8.5 \cdot 10^{-06}$	5.024	$-9.5 \cdot 10^{+02}$	$8.4 \cdot 10^{-06}$	5.044	$-1.0 \cdot 10^{+02}$
$\frac{1}{5} \cdot 2^{-4}$	$2.6 \cdot 10^{-08}$	5.012	$-9.0 \cdot 10^{+02}$	$2.5 \cdot 10^{-08}$	5.023	$-9.2 \cdot 10^{+02}$
$\frac{1}{5} \cdot 2^{-5}$	$8.1 \cdot 10^{-09}$	5.007	$-8.8 \cdot 10^{+02}$	$7.8 \cdot 10^{-09}$	5.018	$-9.0 \cdot 10^{+02}$

We use the linear transformation $z(t) := (y(t), ty'(t))^{\mathrm{T}}$ to transform the second order system to the equivalent first order form and obtain

Table 2
Convergence of the IDeC method; Example 2

h	$\|\varepsilon_h^{(0)}\|_h$	$p^{(0)}$	$c^{(0)}$	$\|\varepsilon_h^{(1)}\|_h$	$p^{(1)}$	$c^{(1)}$
$\frac{1}{5}$	$2.4\cdot10^{-02}$	0.834	$-9.2\cdot10^{-01}$	$7.7\cdot10^{-02}$	1.835	$-1.4\cdot10^{-01}$
$\frac{1}{5}\cdot2^{-1}$	$1.3\cdot10^{-02}$	0.921	$-1.1\cdot10^{-01}$	$2.1\cdot10^{-03}$	1.918	$-1.7\cdot10^{-01}$
$\frac{1}{5}\cdot2^{-2}$	$7.1\cdot10^{-02}$	0.960	$-1.2\cdot10^{-01}$	$5.7\cdot10^{-03}$	1.959	$-2.0\cdot10^{-01}$
$\frac{1}{5}\cdot2^{-3}$	$3.6\cdot10^{-02}$	0.980	$-1.3\cdot10^{-01}$	$1.4\cdot10^{-04}$	1.979	$-2.1\cdot10^{-01}$
$\frac{1}{5}\cdot2^{-4}$	$1.8\cdot10^{-03}$	0.990	$-1.4\cdot10^{-01}$	$3.7\cdot10^{-04}$	1.989	$-2.2\cdot10^{-01}$
$\frac{1}{5}\cdot2^{-5}$	$9.3\cdot10^{-03}$	0.995	$-1.4\cdot10^{-01}$	$9.4\cdot10^{-05}$	1.994	$-2.3\cdot10^{-01}$
$\frac{1}{5}\cdot2^{-6}$	$4.7\cdot10^{-03}$	0.997	$-1.4\cdot10^{-01}$	$2.3\cdot10^{-06}$	1.997	$-2.3\cdot10^{-01}$
$\frac{1}{5}\cdot2^{-7}$	$2.3\cdot10^{-04}$	0.998	$-1.4\cdot10^{-01}$	$5.9\cdot10^{-06}$	1.998	$-2.4\cdot10^{-01}$
$\frac{1}{5}\cdot2^{-8}$	$1.1\cdot10^{-04}$	0.999	$-1.5\cdot10^{-01}$	$1.4\cdot10^{-07}$	1.999	$-2.4\cdot10^{-01}$
$\frac{1}{5}\cdot2^{-9}$	$5.8\cdot10^{-04}$	0.999	$-1.5\cdot10^{-01}$	$3.7\cdot10^{-07}$	1.999	$-2.4\cdot10^{-01}$
h	$\|\varepsilon_h^{(2)}\|_h$	$p^{(2)}$	$c^{(2)}$	$\|\varepsilon_h^{(3)}\|_h$	$p^{(3)}$	$c^{(3)}$
$\frac{1}{5}$	$1.6\cdot10^{-03}$	2.924	$-1.8\cdot10^{-01}$	$9.5\cdot10^{-03}$	3.330	$-2.0\cdot10^{-01}$
$\frac{1}{5}\cdot2^{-1}$	$2.1\cdot10^{-04}$	2.918	$-1.7\cdot10^{-01}$	$9.5\cdot10^{-04}$	3.525	$-3.1\cdot10^{+00}$
$\frac{1}{5}\cdot2^{-2}$	$2.8\cdot10^{-05}$	2.731	$-1.0\cdot10^{-01}$	$8.2\cdot10^{-05}$	3.770	$-6.6\cdot10^{+00}$
$\frac{1}{5}\cdot2^{-3}$	$4.3\cdot10^{-05}$	2.868	$-1.6\cdot10^{-01}$	$6.0\cdot10^{-06}$	3.889	$-1.0\cdot10^{+00}$
$\frac{1}{5}\cdot2^{-4}$	$5.9\cdot10^{-06}$	2.934	$-2.2\cdot10^{-01}$	$4.0\cdot10^{-07}$	3.946	$-1.3\cdot10^{+00}$
$\frac{1}{5}\cdot2^{-5}$	$7.7\cdot10^{-07}$	2.967	$-2.6\cdot10^{-01}$	$2.6\cdot10^{-09}$	3.973	$-1.5\cdot10^{+00}$
$\frac{1}{5}\cdot2^{-6}$	$9.8\cdot10^{-08}$	2.983	$-2.9\cdot10^{-01}$	$1.6\cdot10^{-10}$	3.974	$-1.5\cdot10^{+00}$
h	$\|\varepsilon_h^{(4)}\|_h$	$p^{(4)}$	$c^{(4)}$	$\|\varepsilon_h^{(5)}\|_h$	$p^{(5)}$	$c^{(5)}$
$\frac{1}{5}$	$8.1\cdot10^{-03}$	4.954	$-2.3\cdot10^{+00}$	$5.4\cdot10^{-03}$	5.558	$-4.1\cdot10^{+01}$
$\frac{1}{5}\cdot2^{-1}$	$2.6\cdot10^{-05}$	5.079	$-3.1\cdot10^{+00}$	$1.1\cdot10^{-05}$	4.921	$-9.5\cdot10^{+00}$
$\frac{1}{5}\cdot2^{-2}$	$7.7\cdot10^{-06}$	5.095	$-3.3\cdot10^{+01}$	$3.7\cdot10^{-06}$	5.141	$-1.8\cdot10^{+00}$
$\frac{1}{5}\cdot2^{-3}$	$2.2\cdot10^{-08}$	5.084	$-3.1\cdot10^{+01}$	$1.0\cdot10^{-08}$	5.102	$-1.6\cdot10^{+00}$
$\frac{1}{5}\cdot2^{-4}$	$6.6\cdot10^{-09}$	4.995	$-2.1\cdot10^{+00}$	$3.1\cdot10^{-10}$	5.080	$-1.4\cdot10^{+00}$

$$z'(t) = \frac{1}{t}\begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} z(t) + \begin{pmatrix} 0 \\ -tn^2\cos(nt) - 2n\sin(nt) \end{pmatrix}, \tag{67a}$$

$$z(0) = \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \tag{67b}$$

whose exact solution is $z(t) = (1+\cos(nt), -nt\sin(nt))^{\mathrm{T}}$. The results for this model can be found in Table 1. As expected, we observe the classical order sequence, $\mathrm{O}(h)$, $\mathrm{O}(h^2)$, $\mathrm{O}(h^3)$, $\mathrm{O}(h^4)$ and $\mathrm{O}(h^5)$.

Example 2. Finally, we study the nonlinear "Emden-Differential Equation", cf. [12],

$$y''(t) = -\frac{2}{t}y'(t) - y^5(t), \quad t \in (0,1],$$
$$y(0) = 1, \qquad y'(0) = 0,$$

with $y(t) = 1/\sqrt{1+t^2/3}$, or equivalently,

$$z'(t) = \frac{1}{t}\begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} z(t) - t\begin{pmatrix} 0 \\ z_1^5(t) \end{pmatrix}, \quad t \in (0,1], \tag{68a}$$

$$z(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{68b}$$

with

$$z(t) = \begin{pmatrix} \dfrac{1}{\sqrt{1+t^2/3}} \\ -\dfrac{t^2}{3\sqrt{1+t^2/3}^3} \end{pmatrix}.$$

The numerical results for this problem are listed in Table 2 and we can see that the IDeC iteration again shows its classical convergence behavior.

References

[1] E.A. Coddington, N. Levison, Theory of Ordinary Differential Equations, McGraw-Hill, New York, 1955.
[2] R. Frank, The method of iterated defect-correction and its application to two-point boundary value problems, Part I, Numer. Math. 25 (1976) 409–419.
[3] R. Frank, The method of iterated defect-correction and its application to two-point boundary value problems, Part II, Numer. Math. 27 (1977) 407–420.
[4] R. Frank, C. Überhuber, Iterated defect correction for Runge–Kutta methods, Report No. 14/75, Department of Applied Mathematics and Numerical Analysis, University of Technology Vienna, 1975.
[5] F. Frommlet, E. Weinmüller, Asymptotic error expansions for singular BVP's, MAS, submitted.
[6] F.R. de Hoog, R. Weiss, Difference methods for boundary value problems with a singularity of the first kind, SIAM J. Numer. Anal. 13 (1976) 775–813.
[7] F.R. de Hoog, R. Weiss, The application of Runge–Kutta schemes to singular initial value problems, Math. Comp. 44 (1985) 93–103.
[8] O. Koch, P. Kofler, E. Weinmüller, Analysis of singular initial and terminal value problems, Technical Report No. 125/99, Department of Applied Mathematics and Numerical Analysis, University of Technology Vienna, 1999.
[9] O. Koch, P. Kofler, E. Weinmüller, On the initial value problems for systems of ordinary first and second order differential equations with a singularity of the first kind, SIAM J. Numer. Anal., submitted.
[10] P. Kofler, E. Weinmüller, Numerical treatment of singular boundary and initial value problems, in: Proceedings of RAAM 1996, Kuwait.
[11] Y.L. Luke, The Special Functions and their Approximations, Academic Press, New York, 1969.
[12] R.D. Russel, L.F. Shampine, Numerical methods for singular boundary value problems, SIAM J. Numer. Anal. 12 (1975) 13–36.
[13] E. Weinmüller, On the boundary value problems for systems of ordinary second order differential equations with a singularity of the first kind, SIAM J. Math. Anal. 15 (1984) 287–307.

[14] E. Weinmüller, A difference method for a singular boundary value problem of second order, Math. Comp. 42 (1984) 441–464.
[15] P.E. Zadunaisky, On the estimation of errors propagated in the numerical integration of ordinary differential equations, Numer. Math. 27 (1976) 21–39.