Numerische Mathematik manuscript No.
(will be inserted by the editor)

Asymptotically Correct Error Estimation for Collocation Methods Applied to Singular Boundary Value Problems

Othmar Koch

Vienna University of Technology
e-mail: othmar@othmar-koch.org

Received: date / Revised version: date

Summary We discuss an a posteriori error estimate for collocation methods applied to boundary value problems in ordinary differential equations with a singularity of the first kind. As an extension of previous results we show the asymptotical correctness of our error estimate for the most general class of singular problems where the coefficient matrix is allowed to have eigenvalues with positive real parts. This requires a new representation of the global error for the numerical solution obtained by piecewise polynomial collocation when applied to our problem class.

Key words Boundary value problems – singularity of the first kind – collocation methods – a posteriori error estimate – asymptotical correctness

Mathematics Subject Classification (2000): 65L05

1 Introduction

We deal with the numerical solution of singular boundary value problems of the form

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

$$B_0 z(0) + B_1 z(1) = \beta, \tag{1b}$$

where z is an n-dimensional real function, M is a smooth $n \times n$ matrix and f is an n-dimensional smooth function on a suitable domain. B_0 and

B_1 are constant matrices which are subject to certain restrictions for a well-posed problem. The analytical properties of (1) have been discussed in [12]. We will recapitulate the most important results in §2, where we focus on the most general boundary conditions which guarantee well-posedness of the problem. Moreover, we will extend the results to problems where f is defined as a piecewise continuous function and derive a representation of the solution which will be useful later on in our discussion.

Our interest in the numerical solution of boundary value problems with a singularity of the first kind (1) is backed by numerous applications from physics ([9], [15]), mechanics ([10]) or ecology ([19]). In this paper, we treat the most general problem class, where the spectrum of the matrix $M(0)$ contains both eigenvalues with positive and negative real parts. This case is of particular importance in physical applications, see for instance [5], [7], [22], [25].

To compute the numerical solution of (1), we use collocation at an even number m of collocation points spaced equidistantly in the interior of every collocation interval. Our decision to use collocation was motivated by its advantageous convergence properties for (1), while in the presence of a singularity other high order methods show order reductions and become inefficient (see for example [14]). Here, we will show that the convergence order of collocation methods with polynomials of degree $\leq m$ is at least equal to the stage order m[1]. One of the reasons why we concentrate on collocation at an even number of equidistant points is that in general, we cannot expect to observe superconvergence (cf. [6]) when collocation is applied to (1). At most, a convergence order of $O(|\ln(h)|^{n_0-1}h^{m+1})$, for some positive integer n_0, holds for a method of stage order m, see [13].

This paper is intended to complete the analysis of a new a posteriori error estimate for collocation methods applied to general boundary value problems with a singularity of the first kind (1). This estimate is based on the defect correction principle (see for example [21]) and was first introduced in [4], where an analysis of the convergence properties of the error estimate is given for regular problems. In [2], we could prove analogous results for a restricted class of singular problems, where we assumed that the eigenvalues of the coefficient matrix $M(0)$ have no positive real parts. In that case, a shooting argument can be used to analyze the collocation solution. Convergence results from [13] finally yield a basis for the discussion of collocation and the error estimate. In the case of a general spectrum of $M(0)$, which we consider here, this approach is not feasible. In order to prove the analogous results, we have to derive a new representation of the global error of collocation methods applied to (1), where we make use

[1] This is an extension of results from [13] and [2] to the most general class of singular problems (1).

of estimates given in [24] for collocation applied to second order boundary value problems with a singularity of the first kind. Our analysis of collocation schemes is given in §3. With these prerequisites, we show in §4 that the error of the error estimate as compared with the exact global error is $O(|\ln(h)|^{n_0-1}h^{m+1})$ uniformly in t. This means that our error estimate is asymptotically correct when no superconvergence is observed for the collocation solution. This is the case for our choice of collocation points, and moreover we cannot expect superconvergence for singular problems in general anyway, see above.

The collocation method and error estimate analyzed in this paper were also implemented in a MATLAB code designed especially to solve boundary value problems with a singularity of the first kind. Our error estimate yields a reliable basis for a mesh selection procedure which enables an efficient computation of the numerical solution [3]. A description of the code and experimental evidence of its advantageous properties are given in [1].

1.1 Notation

Throughout the paper, the following notation is used. We denote by $\mathbb{R}^n$ the space of real vectors of dimension n and use $|\cdot|$ to denote the maximum norm in $\mathbb{R}^n$. For an interval $[a,b]$, $C^p[a,b]$ is the space of real vector-valued functions or real matrices which are p times continuously differentiable on $[a,b]$ (we usually write $C[a,b] := C^0[a,b]$). For functions $y \in C[0,b]$, where $0 < b \leq 1$, we define the maximum norm,

$$\|y\|_b := \max_{0\leq t\leq b} |y(t)|.$$

In the case where $b = 1$ we omit the subscript to avoid confusion. For a matrix $A = (a_{ij})_{i,j=1}^n$, $A \in C[0,b]$, $\|A\|_b$ is the induced norm,

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

The subscript is again omitted for $b = 1$.

For the numerical analysis, we define meshes

$$\Delta := (\tau_0, \tau_1, \ldots, \tau_N), \tag{2}$$

and

$$h_i := \tau_{i+1} - \tau_i, \quad J_i := [\tau_i, \tau_{i+1}], \quad i = 0, \ldots, N-1, \quad \tau_0 = 0,\ \tau_N = 1. \tag{3}$$

For reasons of simplicity, we restrict the discussion to equidistant meshes,

$$h_i = h, \quad i = 0, \dots, N-1.$$

However, the results also hold for nonuniform meshes which have a limited variation in the stepsizes, cf. [13], [24]. On Δ, we define corresponding grid vectors

$$u_\Delta := (u_0, \dots, u_N) \in \mathbb{R}^{(N+1)n}. \tag{4}$$

The norm on the space of grid vectors is given by

$$\|u_\Delta\|_\Delta := \max_{0\le k\le N} |u_k|. \tag{5}$$

For a continuous function $y \in C[0,1]$, we denote by R_Δ the pointwise projection onto the space of grid vectors,

$$R_\Delta(y) := (y(\tau_0), \dots, y(\tau_N)). \tag{6}$$

For collocation, m points $t_{i,j},\ j = 1, \dots, m$, are inserted in each subinterval J_i. We choose the same distribution of collocation points in every subinterval, thus yielding the (fine) grid[2]

$$\Delta^m = \{t_{i,j} = \tau_i + \rho_j h, \quad i = 0, \dots, N-1,\ j = 1, \dots, m\}, \tag{7}$$

with

$$0 < \rho_1 < \rho_2 \cdots < \rho_m \le 1. \tag{8}$$

For reasons of convenience, we define $\rho_{m+1} := 1$. We restrict ourselves to grids where $\rho_1 > 0$ to avoid a special treatment of the singular point $t = 0$. For the analysis of the stability of collocation methods in §3, we allow $\rho_m = 1$. In the discussion of the error estimate, we consider only equidistant collocation points, where

$$\rho_j := \frac{j}{m+1}, \quad j = 1, \dots, m. \tag{9}$$

For a grid Δ^m, u_{Δ^m}, $\|\cdot\|_{\Delta^m}$ and R_{Δ^m} are defined analogously to (4)–(6).

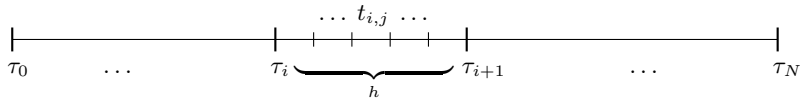

Fig. 1. The computational grid

[2] For convenience, we denote τ_i by $t_{i,0} \equiv t_{i-1,m+1},\ i = 1, \dots, N$.

2 Analytical results

First, we discuss the analytical properties of linear boundary value problems with a singularity of the first kind,

$$z'(t) = \frac{M(t)}{t} z(t) + f(t), \quad t \in (0,1], \tag{10a}$$

$$B_0 z(0) + B_1 z(1) = \beta. \tag{10b}$$

Throughout, we assume $M \in C^1[0,1]$. Consequently, we can rewrite $M(t)$ and obtain

$$M(t) = M(0) + tC(t) \tag{11}$$

with a continuous matrix $C(t)$.

It was shown in [12] that the form of the boundary conditions (10b) which guarantee that (10) has a unique, continuous solution depends on the spectral properties of the coefficient matrix $M(0)$. To avoid fundamental modes of (10a) which have the form $\cos(\sigma \ln(t)) + \mathrm{i}\sin(\sigma \ln(t))$, we assume that zero is the only eigenvalue of $M(0)$ on the imaginary axis.

Now, let S denote a projection onto the invariant subspace which is associated with eigenvalues of $M(0)$ which have positive real parts, and R a projection onto the kernel of $M(0)$. Finally, define

$$P := S + R, \quad Q := I - P, \tag{12}$$

where I denotes the identity matrix in $\mathbb{R}^n$. Later on in our discussion, we will also use $\tilde{S},\ \tilde{R},\ \tilde{P}$ and $\tilde{Q}$ for matrices consisting of maximal sets of linearly independent columns of the respective projections.

In [12] it was shown that the boundary value problem (10) is well-posed iff the boundary conditions (10b) can equivalently be written as

$$(Q+R)z(0) = Rz(0) = \gamma \in \ker(M(0)), \tag{13a}$$

$$Sz(1) = S\eta, \quad \eta \in \mathbb{R}^n. \tag{13b}$$

Remark 2.1 Note that (13a) could also be written equivalently as

$$Qz(0) = 0, \quad \tilde{B}_0 Rz(0) + \tilde{B}_1 Rz(1) = \tilde{\beta}, \tag{14}$$

with suitable matrices $\tilde{B}_0,\ \tilde{B}_1 \in \mathbb{R}^{r\times n}$ and $\tilde{\beta} \in \mathbb{R}^r$. This does not change our arguments, however, so for reasons of simplicity we use (13a).

It follows from the variation of constant formula (see for example [8]) that, for any $0 < b \leq 1$, every solution z of (10a) satisfies the integral equation

$$z(t) = \left(\frac{t}{b}\right)^{M(0)} z(b) + t^{M(0)} \int_b^t \tau^{-M(0)} \left(C(\tau)z(\tau) + f(\tau)\right)\, d\tau. \tag{15}$$

For sufficiently small b, (15) can be shown to have a unique continuous solution on $[0,b]$, and classical theory yields the existence of a unique solution of (10) on $[0,1]$.

For the analysis of the nonlinear case, i. e., when $f = f(t,z)$ in (10), we make the following assumptions:

1. Equation (1) has an isolated solution $z \in C[0,1] \cap C^1(0,1]$. This means that
$$u'(t) = \frac{M(t)}{t}u(t) + A(t)u(t), \quad t \in (0,1],$$
$$B_0u(0) + B_1u(1) = 0,$$
where
$$A(t) := D_2f(t,z(t)) := \frac{\partial f(t,z)}{\partial z}(t,z(t)),$$
has only the trivial solution.
With the solution z and a $\rho > 0$ we associate the spheres
$$S_\rho(z(t)) := \{y \in \mathbb{R}^n : |z(t) - y| \le \rho\}$$
and the tube
$$T_\rho := \{(t,y) : t \in [0,1],\ y \in S_\rho(z(t))\}.$$
2. $f(t,z)$ is continuously differentiable with respect to z, and $\frac{\partial f(t,z)}{\partial z}$ is continuous on T_ρ.

For this situation, the following smoothness properties hold, for a proof see [12]:

Theorem 2.1 *Let f be p times continuously differentiable on T_ρ and $M \in C^{p+1}[0,1]$. Then*

1. *$z \in C^{p+1}(0,1]$.*
2. *If all the eigenvalues of $M(0)$ have nonpositive real parts, then $z \in C^{p+1}[0,1]$.*
3. *Let σ_+ denote the smallest of the positive real parts of the eigenvalues of $M(0)$ and n_0 the dimension of the largest Jordan box associated with the eigenvalue zero in the Jordan canonical form of $M(0)$. Then the following statements hold:*
   - *For $p < \sigma_+ < p+1$, $|z^{(p+1)}(t)| \le \text{const}\, t^{\sigma_+ - p - 1}(|\ln(t)|^{n_0-1}+1)$.*
   - *For $\sigma_+ = p+1$, $|z^{(p+1)}(t)| \le \text{const}\,(|\ln(t)|^{n_0}+1)$.*
   - *For $\sigma_+ > p+1$, $z \in C^{p+1}[0,1]$.*

Motivated by the last result, we will assume throughout this paper that $\sigma_+ > 1$ to ensure that $z \in C^1[0,1]$. Note that if this assumption is not satisfied, we can transform the equation (1a) by letting $t \to t^\lambda$, $1 > \lambda > 0$, whence $\sigma_+ \to \sigma_+/\lambda$. Thus, the assumption $\sigma_+ > 1$ imposes no restriction of generality.

2.1 Extension to piecewise problems

Now, we extend the analytical results from [12] to the case of piecewise problems (we restrict ourselves to the linear case and write the boundary conditions in their form (13a) and (13b))

$$y_i'(t) = \frac{M(t)}{t} y_i(t) + f_i(t), \quad t \in J_i,\ i = 0, \ldots, N-1, \tag{16a}$$
$$y_i(\tau_{i+1}) = y_{i+1}(\tau_{i+1}), \quad i = 0, \ldots, N-2, \tag{16b}$$
$$(Q+R)y_0(0) = \gamma =: \tilde{R}\alpha_1, \tag{16c}$$
$$Sy_{N-1}(1) = S\eta =: \tilde{S}\alpha_2, \tag{16d}$$

where $f(t) := f_i(t),\ t \in J_i,\ i = 0, \ldots, N-1$ is a piecewise continuous function, $f_i \in C(J_i)$. We will show the existence of a continuous solution $y(t) = y_i(t),\ t \in J_i,\ i = 0, \ldots, N-1$, discuss its smoothness and derive estimates which are useful for the analysis of the global error of collocation methods, see §3.2.

We use a contraction argument to show the existence of a unique solution of (16). To this end, we define a mapping $\mathcal{R} : C[0,1] \to C[0,1]$ by requiring that $y = \mathcal{R}(x)$ is the solution of the linear (piecewise) boundary value problem (see (11))

$$y_i'(t) = \frac{M(0)}{t} y_i(t) + C(t)x(t) + f_i(t), \quad t \in J_i,\ i = 0, \ldots, N-1, \tag{17a}$$
$$y_i(\tau_{i+1}) = y_{i+1}(\tau_{i+1}), \quad i = 0, \ldots, N-2, \tag{17b}$$
$$(Q+R)y_0(0) = Ry(0) = \gamma, \tag{17c}$$
$$Sy_{N-1}(1) = S\eta. \tag{17d}$$

Similarly as in (15) (see also [18]), and considering that (see (12))

$$I = S + R + Q,$$

we show that, for a suitable $c \in \mathbb{R}^n$, y can be written as

$$\begin{aligned} y(t) = \mathcal{R}(x) &= t^{M(0)}c + t^{M(0)} \sum_{l=N-1}^{i+1} \int_{\tau_{l+1}}^{\tau_l} \tau^{-M(0)} \left(C(\tau)x(\tau) + f_l(\tau)\right)\, d\tau \\ &\quad + t^{M(0)} \int_{\tau_{i+1}}^{t} \tau^{-M(0)} \left(C(\tau)x(\tau) + f_i(\tau)\right)\, d\tau \\ &= t^{M(0)}Sc + t^{M(0)}R\left(c - \int_0^1 \tau^{-M(0)} \left(C(\tau)x(\tau) + f(\tau)\right)\, d\tau\right) \\ &\quad + t^{M(0)}Q\left(c - \int_0^1 \tau^{-M(0)} \left(C(\tau)x(\tau) + f(\tau)\right)\, d\tau\right) \\ &\quad + t\int_0^1 (Q+R)s^{-M(0)} \left(C(st)x(st) + f(st)\right)\, ds \\ &\quad + t^{M(0)} \int_1^t S\tau^{-M(0)} \left(C(\tau)x(\tau) + f(\tau)\right)\, d\tau, \quad t \in J_i. \end{aligned} \tag{18}$$

The vector c is then determined from the boundary conditions (17c) and (17d). We conclude that

$$Q\left(c - \int_0^1 \tau^{-M(0)} \left(C(\tau)x(\tau) + f(\tau)\right)\, d\tau\right) = 0, \tag{19}$$

and moreover

$$Sy(1) = Sc = S\eta = \tilde{S}\alpha_2, \tag{20}$$

and

$$Ry(0) = R\left(c - \int_0^1 \tau^{-M(0)} \left(C(\tau)x(\tau) + f(\tau)\right)\, d\tau\right) = \tilde{R}\alpha_1, \tag{21}$$

since

$$t^{M(0)}R = R, \quad t \in [0,1],$$

see for example [17]. Thus, the constant c is uniquely determined from (19), (20) and (21), and we conclude

$$\begin{aligned} y(t) &= t^{M(0)}(\tilde{S}\alpha_2 + \tilde{R}\alpha_1) \\ &\quad + t\int_0^1 (Q+R)s^{-M(0)} \left(C(st)x(st) + f(st)\right)\, ds \\ &\quad + t^{M(0)} \int_1^t S\tau^{-M(0)} \left(C(\tau)x(\tau) + f(\tau)\right)\, d\tau. \end{aligned} \tag{22}$$

We may now assume $x \in C[b,1],\ 0 < b < 1$, to be a known, fixed function (we may later justify this assumption, since classical theory implies that a

solution of (16a) with suitable boundary conditions exists on $[b,1]$). Consequently, it follows from [17, Lemma 2.2] and [17, Lemma 2.5] that the right-hand side of (22) is a contraction for $t \in [0,b]$ for sufficiently small $b > 0$, that is

$$\|\mathcal{R}(x_1) - \mathcal{R}(x_2)\|_b \le L \|x_1 - x_2\|_b, \quad L < 1. \tag{23}$$

Thus, there exists a unique, continuous solution $y = \mathcal{R}(y)$ of (17) which satisfies the estimate

$$\begin{aligned} &\|\mathcal{R}(y)\|_b - \|\mathcal{R}(0)\|_b \le \|\mathcal{R}(y) - \mathcal{R}(0)\|_b \le L \|y\|_b \Rightarrow \\ &\|y\|_b \le \text{const } (|\gamma| + |\eta| + \|f\|). \end{aligned} \tag{24}$$

If we assume that $\sigma_+ > 1$ (cf. Theorem 2.1), we can substitute (22) into (17a) and conclude that y' is piecewise continuous on [0,1], and moreover

$$\left\|y'\right\|_b \le \text{const } (|\gamma| + |\eta| + \|f\|). \tag{25}$$

Note that we can use classical theory to extend the results to the whole interval $[0,1]$.

Finally, we discuss the smoothness of higher derivatives of the solution y. To this end, we consider the representation of y' resulting from the substitution of (22) into (17a). Clearly,

$$t^{M(0)-I}\tilde{S}\alpha_2 \in C^p[0,1] \quad \text{if } \sigma_+ > p+1, \tag{26}$$

$$\lim_{t\to 0} t^{M(0)-I}\tilde{S}\alpha_2 = 0. \tag{27}$$

Now, for $t \in J_i$ consider

$$\begin{aligned} \zeta(t) &:= \int_0^1 (Q+R)s^{-M(0)} \left(C(st)x(st) + f(st)\right)\, ds \\ &= \sum_{l=0}^{i-1} \int_{\tau_l/t}^{\tau_{l+1}/t} (Q+R)s^{-M(0)} \left(C(st)x(st) + f_l(st)\right)\, ds \\ &\quad + \int_{\tau_i/t}^{1} (Q+R)s^{-M(0)} \left(C(st)x(st) + f_i(st)\right)\, ds. \end{aligned} \tag{28}$$

It is straightforward to compute

$$\begin{aligned} \zeta'(t) = (Q+R) &\int_0^1 s^{I-M(0)} \left(C'(st)x(st) + C(st)x'(st) + f'(st)\right)\, ds \\ &- (Q+R)\frac{1}{t}\sum_{l=0}^{i-1} \Bigg(\left(\frac{\tau_{l+1}}{t}\right)^{I-M(0)} \left(C(\tau_{l+1})x(\tau_{l+1}) + f_l(\tau_{l+1})\right) \\ &- \left(\frac{\tau_l}{t}\right)^{I-M(0)} \left(C(\tau_l)x(\tau_l) + f_l(\tau_l)\right) \Bigg) \\ &+ (Q+R)\frac{1}{t}\left(\frac{\tau_i}{t}\right)^{I-M(0)} \left(C(\tau_i)x(\tau_i) + f_i(\tau_i)\right). \end{aligned} \tag{29}$$

Consequently, for $M \in C^2[0,1],\ x, f_i \in C^1(J_i)$, $\zeta'(t)$ can be estimated as

$$\|\zeta'\|_{\tau_1} \leq \text{const } \left(|\eta| + |\gamma| + \|f\| + \|f'\|\right), \tag{30}$$

$$\|\zeta'\|_{\tau_{i+1}} \leq \text{const } \left((1+1/h)(|\eta| + |\gamma| + \|f\|) + \|f'\|\right), \quad i \geq 1, \tag{31}$$

since in the latter case we can use $1/t \leq 1/h$ for $t \geq \tau_1$ and estimate (30) on the first interval. Note that the $1/h$ terms in (31) are present only because of the jump discontinuities in $f(t)$. For smooth f, estimates analogous to (30) hold on the whole interval.

Finally, we analyze the last term of $y(t)/t$ in the representation (22), defining

$$\varphi(t) := t^{M(0)-I} \int_1^t S\tau^{-M(0)} \left(C(\tau)x(\tau) + f(\tau)\right)\, d\tau$$

similarly as in [17, Lemma 2.6]. It turns out that estimates analogous to (30) and (31) hold for φ'. We will not repeat the calculations here.

The considerations up to this point put us in a position to estimate y''. For higher derivatives of y, we proceed analogously, where we require sufficient (piecewise) smoothness of $M,\ x_i$ and f_i and we assume that σ_+ is sufficiently large (note that the smoothness of $x(t) = y(t)$ is concluded successively in every step from the results of the last step).

Altogether we have proven the following theorem:

Theorem 2.2 *Assume $f_i \in C^p(J_i),\ M \in C^{p+1}[0,1]$ and $\sigma_+ > p+1$. Then there exists a unique, continuous solution $y(t) = y_i(t),\ t \in J_i,\ i = 0, \ldots, N-1$, of* (16)*. Moreover, $y_i \in C^{p+1}(J_i)$ holds and y satisfies the estimates*

$$\|y\| \leq \text{const } (|\gamma| + |\eta| + \|f\|), \tag{32}$$

$$\left\|y^{(p+1)}\right\| \leq \text{const } \sum_{k=0}^{p} h^{k-p} \left(|\gamma| + |\eta| + \left\|f^{(k)}\right\|\right). \tag{33}$$

3 Collocation methods

In this section, we analyze collocation with continuous, piecewise polynomial functions of degree $\leq m$. First, in §3.1 we give existence results and estimates for linear initial and terminal value problems. We proceed by generalyzing these results to systems of linear and nonlinear boundary value problems. The convergence analysis is postponed to §3.2, where we will derive a new representation of the global error of collocation for (1). Motivated by the discussion in §2, we restrict ourselves to boundary conditions of the form (13).

Let us denote by B_m the Banach space of continuous, piecewise polynomial functions $q \in \mathbb{P}_m$ of degree $\leq m,\ m \in \mathbb{N}$ (m is called the *stage order* of the method), equipped with the norm $\|\cdot\|$. As an approximation for the exact solution z of (1), we define an element of B_m which satisfies the differential equation (1a) at a finite number of points and which is subject to the same boundary conditions. Thus, we are seeking a function $p(t) = p_i(t),\ t \in J_i,\ i = 0, \ldots, N-1$, in B_m which satisfies

$$p'(t_{i,j}) = \frac{M(t_{i,j})}{t_{i,j}} p(t_{i,j}) + f(t_{i,j}, p(t_{i,j})),$$
$$i = 0, \ldots, N-1,\ j = 1, \ldots, m, \tag{34a}$$
$$(Q+R)p(0) = \gamma = \tilde{R}\alpha_1, \tag{34b}$$
$$Sp(1) = S\eta = \tilde{S}\alpha_2. \tag{34c}$$

We consider collocation on general grids Δ^m (see (7)), subject to the restriction $\rho_1 > 0$.

3.1 Existence and stability results

For the discussion of the existence, uniqueness and stability of the solution of collocation schemes, we consider general collocation grids where we permit $\rho_m = 1$, see (8). The first result we give was already employed in [2] for the analysis of collocation schemes. The result holds for problems where the matrix $M(0)$ has no eigenvalues with positive real parts. In that case, problem (1), and consequently (34), can equivalently be written as an initial value problem. We will show later that general boundary value problems can be rewritten in a way such that the following lemma can still be used. The proof of this result is given in [2, Theorem 4.2].

Lemma 3.1 *Assume that all eigenvalues of $M(0)$ have nonpositive real parts. For $\mu, \alpha \in \{0, 1\}$ and arbitrary constants $c_{i,j}$, there exists a unique $p \in B_m$ which satisfies*

$$p'(t_{i,j}) = \frac{M(0)}{t_{i,j}} p(t_{i,j}) + \frac{M(0)^\mu}{t_{i,j}^\alpha} c_{i,j},$$
$$i = 0, \ldots, N-1,\ j = 1, \ldots, m, \tag{35a}$$
$$p(0) = \gamma = \tilde{R}\alpha_1. \tag{35b}$$

Furthermore,

$$\|p\|_{\tau_{i+1}} \leq \text{const} \left(|\gamma| + C_i \tau_{i+1}^{1-\alpha} |\ln(h)|^{(\alpha(n_0-\mu))_+} \right), \quad i = 0, \ldots, N-1, \tag{36}$$

where n_0 is the dimension of the largest Jordan block of $M(0)$ corresponding to the eigenvalue 0,

$$(x)_+ := \begin{cases} x, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

and

$$C_i := \max_{\substack{l=0,\dots,i \\ j=1,\dots,m}} |c_{l,j}|.$$

The next lemma is concerned with terminal value problems, where all the eigenvalues of $M(0)$ have positive real parts.

Lemma 3.2 *Assume that all eigenvalues of $M(0)$ have positive real parts. For $\alpha \in \{0,1\}$ and arbitrary constants $c_{i,j}$, there exists a unique $p \in B_m$ which, for any $0 < b \leq 1$, satisfies*

$$p'(t_{i,j}) = \frac{M(0)}{t_{i,j}} p(t_{i,j}) + \frac{1}{t_{i,j}^{\alpha}} c_{i,j}, \qquad i = 0,\dots,N-1,\ j = 1,\dots,m, \tag{37a}$$

$$p(b) = \eta. \tag{37b}$$

Furthermore,

$$\begin{aligned} \|p\|_{\tau_{i+1}} &\leq \operatorname{const} \left(|\eta| + \sum_{l=N}^{i} \left(\frac{\tau_{i+1}}{\tau_{l+1}} \right)^{\nu} h \tau_{l+1}^{-\alpha} C_{N-1} \right. \\ &\quad \left. + \sum_{l=N}^{i} \left(\frac{\tau_{i+1}}{\tau_{l+1}} \right)^{\nu} h \tau_l^{\sigma_+ - 1} |\eta| \right) \\ &\leq \begin{cases} \operatorname{const} \left(|\eta| + \tau_{i+1} \left| \left(\frac{\tau_{i+1}}{b} \right)^{\nu-1} - 1 \right| C_{N-1} \right), & \text{for } \alpha = 0, \\ \operatorname{const} \left(|\eta| + C_{N-1} \right), & \text{for } \alpha = 1, \end{cases} \\ &\leq \operatorname{const} \left(|\eta| + \tau_{i+1}^{1-\alpha} C_{N-1} \right) \quad i = 0,\dots,N-1, \end{aligned} \tag{38}$$

where σ_+ is the same as in Theorem 2.1. Since we have assumed that $\sigma_+ > 1$, we may choose $\nu > 1$.

Proof The proof follows from results given in [24]. In the latter paper, second order problems with a singularity of the first kind are discussed. However, the collocation equations for these problems can be transformed to an equivalent first order formulation. The results from [24, Lemma 3.4] which we use here are originally derived for the first order formulation and can thus be employed for our purpose. Note that from a close inspection of the arguments in [24, pp. 1092–1094] we can conclude that actually it is possible to use $\nu = \sigma_+$. For this case, the estimate from [24, Lemma 3.4] can be improved to (38) if [24, Lemma 3.2] is appropriately used. □

Next, we show that the linear collocation scheme

$$p'(t_{i,j}) = \frac{M(0)}{t_{i,j}} p(t_{i,j}) + C(t_{i,j}) p(t_{i,j}) + f(t_{i,j}),$$
$$i = 0, \ldots, N-1,\ j = 1, \ldots, m, \tag{39a}$$
$$(Q+R)p(0) = \gamma = \tilde{R}\alpha_1, \tag{39b}$$
$$Sp(1) = S\eta = \tilde{S}\alpha_2 \tag{39c}$$

has a unique solution and derive estimates for this solution.

As in §2, we rewrite (39) as a fixed point problem. Thus, let $p = \mathcal{K}(q)$ be defined for $q \in B_m$ as the solution of

$$p'(t_{i,j}) = \frac{M(0)}{t_{i,j}} p(t_{i,j}) + C(t_{i,j}) q(t_{i,j}) + f(t_{i,j}),$$
$$i = 0, \ldots, N-1,\ j = 1, \ldots, m, \tag{40a}$$
$$(Q+R)p(0) = \gamma = \tilde{R}\alpha_1, \tag{40b}$$
$$Sp(1) = S\eta = \tilde{S}\alpha_2. \tag{40c}$$

Since (40) defines a collocation problem with constant coefficient matrix, the boundary value problem can be decoupled in the following sense: Let us denote by J the Jordan canonical form of $M(0)$, and by E the associated matrix of generalized eigenvectors of $M(0)$. The transformation $p(t) \to Ep(t)$ yields

$$p'(t_{i,j}) = \frac{J}{t_{i,j}} p(t_{i,j}) + D(t_{i,j}) q(t_{i,j}) + g(t_{i,j}),$$
$$i = 0, \ldots, N-1,\ j = 1, \ldots, m, \tag{41a}$$
$$V^{(L)} p(0) = E^{-1}\tilde{R}\alpha_1, \tag{41b}$$
$$V^{(R)} p(1) = E^{-1}\tilde{S}\alpha_2, \tag{41c}$$

where $D(t_{i,j}) := E^{-1}C(t_{i,j}),\ g(t_{i,j}) := E^{-1}f(t_{i,j})$, and

$$J = \begin{pmatrix} J^{(L)} & 0 \\ 0 & J^{(R)} \end{pmatrix}, \quad V^{(L)} := \begin{pmatrix} I^{(L)} & 0 \\ 0 & 0 \end{pmatrix}, \quad V^{(R)} := \begin{pmatrix} 0 & 0 \\ 0 & I^{(R)} \end{pmatrix}.$$

Here, $J^{(L)}$ is the Jordan block of dimension $\operatorname{rank}(Q) + \operatorname{rank}(R)$ associated with the eigenvalues with nonpositive real parts, $J^{(R)}$ is associated with the eigenvalues with positive real parts and has dimension $\operatorname{rank}(S)$, and $I^{(L)}$, $I^{(R)}$ are unit matrices of dimensions $\operatorname{rank}(Q) + \operatorname{rank}(R)$ and $\operatorname{rank}(S)$, respectively, where $Q,\ R$ and S denote the projections from (12). Note that the last $\operatorname{rank}(S)$ components of $E^{-1}\tilde{R}\alpha_1$ and the first $\operatorname{rank}(Q) + \operatorname{rank}(R)$ components of $E^{-1}\tilde{S}\alpha_2$ are zero.

Thus, we may consider separately an initial value problem, where the associated coefficient matrix has eigenvalues with nonpositive real parts

only, and a terminal value problem where the coefficient matrix has only eigenvalues with positive real parts. We use a contraction argument to show that (41) has a unique solution: $v := \mathcal{K}(q_1) - \mathcal{K}(q_2)$ is implicitly defined as the solution of the collocation problem

$$\begin{aligned} v'(t_{i,j}) &= \frac{J}{t_{i,j}} v(t_{i,j}) + D(t_{i,j})(q_1(t_{i,j}) - q_2(t_{i,j})), \\ &\qquad\qquad i = 0, \dots, N-1,\ j = 1, \dots, m, \\ V^{(L)} v(0) + V^{(R)} v(1) &= 0. \end{aligned}$$

Now, Lemma 3.1 implies that

$$\|V^{(L)} v\|_{b^{(L)}} \leq \text{const}\, b^{(L)} \|D\|_{b^{(L)}} \|q_1 - q_2\|_{b^{(L)}} = L^{(L)} \|q_1 - q_2\|_{b^{(L)}},$$

and $L^{(L)} < 1$ if $0 < b^{(L)} \leq 1$ is sufficiently small, since $\|D\| < \infty$. For the terminal value problem, we use classical theory to show the existence of a unique collocation solution on the interval $[b^{(R)}, 1]$, and for sufficiently small $b^{(R)}$, Lemma 3.2 shows that

$$\|V^{(R)} v\|_{b^{(R)}} \leq \text{const}\, b^{(R)} \|D\|_{b^{(R)}} \|q_1 - q_2\|_{b^{(R)}} = L^{(R)} \|q_1 - q_2\|_{b^{(R)}},$$

and $L^{(R)} < 1$ for $0 < b^{(R)} \leq 1$ sufficiently small. Altogether we have proven that

$$\|\mathcal{K}(q_1) - \mathcal{K}(q_2)\|_b \leq L \|q_1 - q_2\|_b, \quad L < 1,$$

where $b := \min\left\{b^{(L)}, b^{(R)}\right\}$, and $\mathcal{K}$ is a contraction on $[0, b]$.

We can now use the same arguments as in the proof of (24) to show that

$$\|p\| \leq \text{const}\ \left(|\gamma| + |\eta| + \|f\|\right). \tag{42}$$

Using this last estimate, and considering that $\tau_{i+1} \leq 1$, we can even conclude that

$$\|p\|_{\tau_{i+1}} \leq \text{const}\ \left(|\gamma| + |\eta| + \tau_{i+1}\|f\|\right). \tag{43}$$

We will postpone the discussion of the higher derivatives of p to §3.2.

Remark 3.1 With the same arguments we can also prove that

$$\begin{aligned} p'(t_{i,j}) &= \frac{M(t_{i,j})}{t_{i,j}} p(t_{i,j}) + \frac{M(0)}{t_{i,j}} f(t_{i,j}), \\ &\qquad\qquad i = 0, \dots, N-1,\ j = 1, \dots, m, && \text{(44a)} \\ (Q+R)p(0) &= \gamma = \tilde{R}\alpha_1, && \text{(44b)} \\ Sp(1) &= S\eta = \tilde{S}\alpha_2. && \text{(44c)} \end{aligned}$$

has a unique solution p which satisfies

$$\|p\| \leq \text{const}\,(|\gamma| + |\eta| + |\ln(h)|^{n_0 - 1}\|f\|). \tag{45}$$

This estimate will be important in the analysis of our error estimate in §4.

To show that the nonlinear scheme (34) has a (locally) unique solution, we can proceed analogously as in [2, Theorem 4.4] using the estimate (42). We do not repeat the whole line of argument here, but merely point out that the only modifications are related to the different form of the boundary conditions (13) we consider for the general case here. [2, (4.23)] now reads[3]

$$z(t) - p_{\mathrm{ref}}(t) = \tilde{S}O(h^m) + tO(h^m).$$

Due to (27), the estimate [2, (4.24)] holds in this case as well. We may also conclude that under the same assumptions, Newton's method converges quadratically for the computation of p.

The convergence properties of the (locally) unique collocation solution p will be discussed in §3.2, where we will derive a representation for the global error of p.

3.2 Representation of the global error

Here, we derive a representation for the global error of the collocation solution p at the grid points $t_{i,j},\ i = 0, \ldots, N-1,\ j = 1, \ldots, m+1$. This representation of the error will be useful in the analysis of our error estimate in §4, so for technical reasons we restrict ourselves to grids where $\rho_m < 1$ (as for example in (9)). We make the ansatz

$$\begin{aligned} p(t_{i,j}) &= z(t_{i,j}) + e(t_{i,j})h^m + r(t_{i,j}), \\ &\quad i = 0, \ldots, N-1,\ j = 1, \ldots, m+1, \end{aligned} \tag{46}$$

where z is the exact solution of (1), and $e \in C[0,1],\ r \in B_m$ are to be determined. We would like to stress that (46) does not yield an *asymptotical error expansion* in the classical sense (cf. [20]), since it will turn out that $e = O(h^m),\ r = O(h^m)$. However, the form we choose in the representation (46) will be convenient in the analysis of our error estimate in §4.

In order to derive relations for e and r, we rewrite the collocation equations (34a): Let

$$\Omega(t) := \prod_{k=1}^{m+1} (t - \rho_k), \tag{47}$$

[3] Note that if we had used (14) instead of (13a), we would obtain

$$z(t) - p_{\mathrm{ref}}(t) = \tilde{S}O(h^m) + \tilde{R}O(h^m) + tO(h^m).$$

where ρ_k are defined in (8) (recall that $\rho_{m+1} := 1$). The Lagrange polynomials associated with the abscissae $\rho_1, \dots, \rho_{m+1}$ are defined as

$$L_k(t) = \frac{\Omega(t)}{\Omega'(\rho_k)(t-\rho_k)}, \quad k = 1, \dots, m+1. \tag{48}$$

Using

$$w_{j,k} := L_k'(\rho_j), \quad j,k = 1, \dots, m+1,$$

we can now write

$$p_i'(t_{i,j}) = \frac{1}{h} \sum_{k=1}^{m+1} w_{j,k} p_i(t_{i,k}), \tag{49}$$

since p_i is a polynomial of degree $\leq m$.

As an auxiliary consideration, observe that

$$\frac{1}{h} \sum_{k=1}^{m+1} \frac{(t_{i,k} - t_{i,j})^l}{l!} w_{j,k} = q_l'(t_{i,j}),$$

for $i = 0, \dots, N-1,\ j = 1, \dots, m,\ l = 0, \dots, m+1$, where $q_l(t),\ t \in J_i$ are the polynomials of degree $\leq m$ interpolating the functions

$$g_l(t) := \frac{(t - t_{i,j})^l}{l!}, \quad t \in J_i,\ l = 0, \dots, m+1,$$

at the points $t_{i,k},\ k = 1, \dots, m+1$, respectively. Clearly, $q_l(t) = g_l(t),\ t \in J_i$ holds for $l = 0, \dots, m$. Consequently,

$$q_l'(t_{i,j}) = g_l'(t_{i,j}) = \begin{cases} 1, & l = 1, \\ 0, & l = 0, 2, 3, \dots, m. \end{cases} \tag{50}$$

On noting that $g_{m+1}'(t_{i,j}) = 0,\ g_{m+1}^{(m+1)}(\xi) = 1,\ \xi \in J_i$, and $g_{m+1}^{(m+2)}(t) = 0,\ t \in J_i$, we conclude for $l = m+1$ that

$$\begin{aligned} q_{m+1}'(t_{i,j}) &= g_{m+1}'(t_{i,j}) - \frac{h^{m+1}}{(m+1)!} \frac{d}{dt} \left(\Omega\left(\frac{t - \tau_i}{h} \right) \right) \Bigg|_{t=t_{i,j}} g_{m+1}^{(m+1)}(\xi) \\ &= -\frac{1}{(m+1)!} \Omega'(\rho_j) h^m, \end{aligned} \tag{51}$$

see [11]. Now, we derive defining relations for the quantities from (46). Formal Taylor expansion about $t_{i,j}$ yields

$$z(t_{i,k}) = \sum_{l=0}^{m+1} \frac{(t_{i,k} - t_{i,j})^l}{l!} z^{(l)}(t_{i,j}) + O(h^{m+2}) \|z^{(m+2)}\|, \tag{52}$$

$$e(t_{i,k}) = e(t_{i,j}) + e'(t_{i,j})(t_{i,k} - t_{i,j}) + O(h^2) \|e''\|, \tag{53}$$

where $\|z^{(m+2)}\| = O(1)$ for sufficiently smooth data in (1), see Theorem 2.1.

Substitution of (46) into (34a) in the form using the equality (49) and taking into account the Taylor expansions (52) and (53) we obtain (recall that $D_2 f(t,z)$ denotes the partial derivative w. r. t. the second argument of a function f)

$$\begin{aligned}
&\sum_{l=0}^{m+1} z^{(l)}(t_{i,j})\frac{1}{h}\sum_{k=1}^{m+1}\frac{(t_{i,k}-t_{i,j})^l}{l!}w_{j,k} \\
&\qquad+\sum_{l=0}^{1} e^{(l)}(t_{i,j})\frac{1}{h}\sum_{k=1}^{m+1}\frac{(t_{i,k}-t_{i,j})^l}{l!}w_{j,k}h^m \\
&\qquad+\frac{1}{h}\sum_{k=1}^{m+1} w_{j,k}r(t_{i,k}) + (1+\|e''\|)O(h^{m+1}) \\
&= z'(t_{i,j}) + e'(t_{i,j})h^m + r'(t_{i,j}) \\
&\qquad-\frac{1}{(m+1)!}\Omega'(\rho_j)z^{(m+1)}(t_{i,j})h^m + (1+\|e''\|)O(h^{m+1}) \\
&= \frac{M(t_{i,j})}{t_{i,j}}\left(z(t_{i,j}) + e(t_{i,j})h^m + r(t_{i,j})\right) + f(t_{i,j}, p(t_{i,j})) \\
&= \frac{M(t_{i,j})}{t_{i,j}}\left(z(t_{i,j}) + e(t_{i,j})h^m + r(t_{i,j})\right) + f(t_{i,j}, z(t_{i,j})) \\
&\qquad+\int_0^1 D_2 f(t_{i,j}, z(t_{i,j}) + e(t_{i,j})h^m + \tau r(t_{i,j}))\, d\tau r(t_{i,j}) \\
&\qquad+D_2 f(t_{i,j}, z(t_{i,j}))e(t_{i,j})h^m + O(h^{2m}).
\end{aligned} \tag{54}$$

Since (54) must hold for all $h \le h_0$ with suitable $h_0 > 0$, we can use the same line of reasoning as in the derivation of classical asymptotical error expansions (see for example [20]) to collect terms in the following way: taking into account the terms vanishing because z is the exact solution of (1), the terms with factors h^m yield for e the relations (clearly, both e and r satisfy homogeneous boundary conditions)

$$e'(t_{i,j}) = \frac{\hat{M}(t_{i,j})}{t_{i,j}}e(t_{i,j}) + \frac{1}{(m+1)!}\Omega'(\rho_j)z^{(m+1)}(\tau_i), \qquad i = 0,\dots,N-1,\ j = 1,\dots,m, \tag{55a}$$

$$(Q+R)e(0) = 0, \tag{55b}$$

$$Se(1) = 0, \tag{55c}$$

where

$$\hat{M}(t) := M(t) + tD_2 f(t, z(t)),$$

and we note that

$$z^{(m+1)}(t_{i,j})h^m = z^{(m+1)}(\tau_i)h^m + O(h^{m+1})$$

for sufficiently smooth z. Finally, we collect all remaining terms in relations for r as

$$\begin{aligned} r'(t_{i,j}) &= \frac{M(t_{i,j})}{t_{i,j}} r(t_{i,j}) + (1+\|e''\|)O(h^{m+1}) \\ &\quad + \int_0^1 D_2 f(t_{i,j}, z(t_{i,j}) + e(t_{i,j})h^m + \tau r(t_{i,j}))\, d\tau r(t_{i,j}), \\ &\qquad i = 0,\dots,N-1,\ j = 1,\dots,m, \end{aligned} \tag{56a}$$

$$(Q+R)r(0) = 0, \tag{56b}$$

$$Sr(1) = 0. \tag{56c}$$

Now, in order to find a function $e \in C[0,1]$ which satisfies (55), we rewrite the defining equations in the form of a differential equation for $t \in (0,1]$,

$$e'(t) = \frac{\hat{M}(t)}{t} e(t) + \frac{1}{(m+1)!} g(t), \quad t \in (0,1], \tag{57a}$$

$$(Q+R)e(0) = 0, \tag{57b}$$

$$Se(1) = 0, \tag{57c}$$

where $g = g_i(t),\ t \in J_i,\ i = 0,\dots,N-1$, is a suitable, piecewise polynomial function which satisfies $g_i(t_{i,j}) = \Omega'(\rho_j) z^{(m+1)}(\tau_i),\ i = 0,\dots,N-1,\ j = 1,\dots,m$. For example, we could choose the unique interpolant $g \in B_{m-1}$, or

$$g_i(t) = z^{(m+1)}(\tau_i)\Omega'\left(\frac{t-\tau_i}{h}\right), \quad t \in J_i. \tag{58}$$

From Theorem 2.2 we conclude that (57) has a unique solution $e \in C[0,1]$, which is piecewise smooth, and for sufficiently smooth $\hat{M}$

$$\|e^{(p+1)}\| \le \text{const} \sum_{k=0}^{p} h^{k-p} \left\| g^{(k)} \right\| = O(h^{-p}), \quad p = 0,\dots,m-1 \tag{59}$$

due to $\|g_i^{(k)}\| = O(h^{-k}),\ k = 0,\dots,m-1$. Finally note that we can use the representation (22) to show that for $\sigma_+ > 1$ (see Theorem 2.1 and the subsequent remark concerning this assumption)

$$\begin{aligned} \left| \frac{M(0)}{t} e(t) \right| &\le \left| M(0) \int_0^1 (Q+R) s^{-M(0)} \left(C(st)e(st) + f(st) \right)\, ds \right| \\ &\quad + \left| M(0) t^{M(0)-I} \int_1^t S\tau^{-M(0)} \left(C(\tau)e(\tau) + f(\tau) \right)\, d\tau \right| \\ &\le \text{const}\,. \end{aligned}$$

Finally, we discuss the solution of (56). If we assume that $D_2 f$ is bounded in $[0,1] \times \mathbb{R}^n$, (56) can be analyzed using the same methods as for general linear collocation problems in §3.1. We can conclude that a unique solution $r \in B_m$ exists and r can be estimated as (note that $\|e''\| h^{m+1} = O(h^m)$ from (59))

$$\|r\|_{\tau_{i+1}} \leq \tau_{i+1} O(h^m), \quad i = 0, \ldots, N-1, \tag{60}$$

see (43). Substitution of (60) into (56a) additionally yields

$$\|r'\| \leq O(h^m), \tag{61}$$

on noting that r' is a piecewise polynomial interpolant of $r'(t_{i,j})$ of degree $\leq m-1$ (cf. [11]). From a well known result for polynomial interpolation, it follows finally that

$$\|r^{(k+1)}\| \leq O(h^{m-k}), \quad k = 1, \ldots, m-1. \tag{62}$$

Using the results we have proven for the functions e and r, we can now formulate the following theorem.

Theorem 3.1 *Assume that $M \in C^{m+2}[0,1]$, f is $m+1$ times continuously differentiable in $[0,1] \times \mathbb{R}^n$ with $D_2 f$ bounded in that domain and $\sigma_+ > m+2$. Then the collocation scheme* (34) *has a unique solution $p \in B_m$ in a neighborhood of an isolated solution $z \in C^{m+2}[0,1]$ of* (1)*. This solution can be computed using Newton's method, which converges quadratically. Moreover, $R_{\Delta^m}(p)$ can be represented in the form* (46)*, with a function $e \in C[0,1]$ which is piecewise smooth and satisfies*

$$\left| \frac{M(0)}{t} e(t) \right| \leq \text{const}\,, \quad t \in [0,1] \tag{63}$$

$$\left\| e^{(k+1)} \right\| = O(h^{-k}), \quad k = 0, \ldots, m-1. \tag{64}$$

Similarly, the function $r \in B_m$ satisfies

$$\|r\|_{\tau_{i+1}} \leq \tau_{i+1} O(h^m), \quad i = 0, \ldots, N-1, \tag{65}$$

$$\|r^{(k+1)}\| = O(h^{m-k}), \quad k = 0, \ldots, m-1. \tag{66}$$

Altogether, we conclude that

$$\|p - z\| = O(h^m), \tag{67}$$

$$\left| \frac{M(0)}{t} (p(t) - z(t)) \right| = O(h^m), \quad t \in [0,1], \tag{68}$$

$$\left\| p^{(k+1)} - z^{(k+1)} \right\| = O(h^{m-k}), \quad k = 0, \ldots, m-1, \tag{69}$$

$$\left| p'(t) - \frac{M(t)}{t} p(t) - f(t, p(t)) \right| = O(h^m), \quad t \in [0,1]. \tag{70}$$

Proof The estimate (70) is a simple consequence of (69) for $k = 0$, (68) and (67). □

Note that the condition $\sigma_+ > m+2$ does not impose a restriction of generality, see also the remark following Theorem 2.1. Furthermore, if $\sigma_+ \leq m+2$ we cannot in general guarantee that $z \in C^{m+2}$ (Theorem 2.1), and thus we cannot expect to observe the desired convergence orders in this case anyway and the restriction $\sigma_+ > m+2$ is thus natural in this context.

4 The error estimate

In this section, we use the results of Theorem 3.1 to show that an error estimate, originally introduced in [4], is asymptotically correct for collocation with an even number of equidistant collocation points. Similar results were shown for regular problems in [4], and in [2] for problems with a singularity of the first kind, where the spectrum of $M(0)$ was restricted to eigenvalues with non-positive real parts. The analysis of the latter case is analogous to the situation we consider here, if we use the results derived for collocation methods from Theorem 3.1, where no restriction on the spectrum of $M(0)$ was imposed. Accordingly, we only give a brief description of the error estimate and refer the reader to [2] for the technical details of the proof. It is only necessary to replace the estimates for collocation methods given in [2] with the results from Theorem 3.1, and use the estimate (45).

For our error estimate, the numerical solution p obtained by collocation is used to define a 'neighboring problem' to (1). The original and neighboring problems are solved by the backward Euler method at the points $t_{i,j},\ i = 0,\dots,N-1,\ j = 1,\dots,m+1$. This yields the grid vectors $\xi_{i,j}$ and $\pi_{i,j}$ as the solutions of the following schemes, subject to boundary conditions (13),

$$\frac{\xi_{i,j}-\xi_{i,j-1}}{t_{i,j}-t_{i,j-1}} = \frac{M(t_{i,j})}{t_{i,j}}\xi_{i,j} + f(t_{i,j},\xi_{i,j}), \quad \text{and} \tag{71a}$$

$$\frac{\pi_{i,j}-\pi_{i,j-1}}{t_{i,j}-t_{i,j-1}} = \frac{M(t_{i,j})}{t_{i,j}}\pi_{i,j} + f(t_{i,j},\pi_{i,j}) + \bar{d}_{i,j}, \tag{71b}$$

where $\bar{d}_{i,j}$ is a defect term defined by

$$\bar{d}_{i,j} := \frac{p(t_{i,j})-p(t_{i,j-1})}{t_{i,j}-t_{i,j-1}} - \sum_{k=1}^{m+1}\alpha_{j,k}\left(\frac{M(t_{i,k})}{t_{i,k}}p(t_{i,k}) + f(t_{i,k},p(t_{i,k}))\right). \tag{72}$$

Here, the coefficients $\alpha_{j,k}$ are chosen in such a way that the quadrature rules given by

$$\frac{1}{t_{i,j}-t_{i,j-1}}\int_{t_{i,j-1}}^{t_{i,j}}\varphi(\tau)\,d\tau \approx \sum_{k=1}^{m+1}\alpha_{j,k}\varphi(t_{i,k})$$

have precision $m+1$.

In the next theorem, we state the result that the difference $\xi_{\Delta^m}-\pi_{\Delta^m}$ is an asymptotically correct estimate for the global error of the collocation solution, $R_{\Delta^m}(z)-R_{\Delta^m}(p)$.

Theorem 4.1 *Assume that the singular boundary value problem* (1) *has an isolated (sufficiently smooth*[4]*) solution* z *and satisfies the assumptions of* Theorem 3.1. *Then, provided that* h *is sufficiently small, the following estimate holds:*

$$\|(R_{\Delta^m}(z)-R_{\Delta^m}(p))-(\xi_{\Delta^m}-\pi_{\Delta^m})\|_{\Delta^m} = O(|\ln(h)|^{n_0-1}h^{m+1}), \quad (73)$$

with n_0 *specified in* Lemma 3.1.

5 Numerical examples

To illustrate the theory, we first consider the following linear problem:

$$z'(t) = \frac{1}{t}\begin{pmatrix}0&1\\2&6\end{pmatrix}z(t) - \begin{pmatrix}0\\4k^4t^5\sin(k^2t^2)+10t\sin(k^2t^2)\end{pmatrix}, \quad (74a)$$

$$\begin{pmatrix}0&1\\0&0\end{pmatrix}z(0)+\begin{pmatrix}0&0\\1&0\end{pmatrix}z(1) = \begin{pmatrix}0\\\sin(k^2)\end{pmatrix}. \quad (74b)$$

The exact solution of this test problem reads

$$z(t) = (t^2\sin(k^2t^2), 2k^2t^4\cos(k^2t^2)+2t^2\sin(k^2t^2))^T.$$

In Table 1, we used $k=5$. Note that the eigenvalues of $M(0)$ are $3\pm\sqrt{11}$.

The computations were carried out with the subroutines from our MATLAB code `sbvp` (cf. [1]) on fixed, equidistant grids. For the purpose of determining the empirical convergence orders the mesh adaptation strategy was disabled. The tests were performed in IEEE double precision with EPS $\approx 1.11\cdot 10^{-16}$. In Table 1, we give the exact global error $\mathrm{err}_{\mathrm{coll}} := \|R_{\Delta^m}(z)-R_{\Delta^m}(p)\|_{\Delta^m}$ of the collocation solution for the respective stepsize h, and the convergence order p_{coll} computed from the errors for two consecutive stepsizes. Moreover, the error of the error estimate with respect to the exact global error, $\mathrm{err}_{\mathrm{est}} := \|(R_{\Delta^m}(z)-R_{\Delta^m}(p))-$

[4] In fact, we require $z\in C^{m+2}[0,1]$.

$(\xi_{\Delta^m} - \pi_{\Delta^m})\|_{\Delta^m}$, is recorded, together with the associated empirical convergence order p_{est}. In accordance with the theoretical results from §§3–4, convergence orders $O(h^4)$ for collocation and $O(h^5)$ for $\mathrm{err}_{\mathrm{est}}$ are observed for the choice $m = 4$.

Table 1. Convergence orders of collocation and error estimate for (74) ($m = 4$)

h	$\mathrm{err}_{\mathrm{coll}}$	p_{coll}	$\mathrm{err}_{\mathrm{est}}$	p_{est}
2^{-4}	1.1876e+00		3.4628e−01	
2^{-5}	6.1802e−02	4.26	7.3370e−03	5.56
2^{-6}	3.6828e−03	4.07	2.4145e−04	4.93
2^{-7}	2.2746e−04	4.02	7.4780e−06	5.01
2^{-8}	1.4174e−05	4.00	2.3533e−07	4.99
2^{-9}	8.8522e−07	4.00	7.3218e−09	5.01

As a second test example, we consider a nonlinear problem from [16], see also [23]:

$$z'(t) = \frac{1}{t}\begin{pmatrix} 0&0&1&0\\ 0&0&0&1\\ 1&0&0&0\\ 0&1&0&0 \end{pmatrix} z(t) + \begin{pmatrix} 0\\ 0\\ \beta t^2 + z_1(t)z_2(t)\\ t^2 - z_1^2(t) \end{pmatrix}, \tag{75a}$$

$$\begin{pmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&0&0\\ 0&0&0&0 \end{pmatrix} z(0) + \begin{pmatrix} 0&0&0&0\\ 0&0&0&0\\ 1&0&0&0\\ 0&-1/3&0&1 \end{pmatrix} z(1) = \begin{pmatrix} 0\\ 0\\ 1\\ 0 \end{pmatrix}, \tag{75b}$$

with $\beta = 5000$. Since the exact solution $z = (z_1, \ldots, z_4)$ of this problem is not known, we use a reference solution on a very fine grid to determine the empirical orders of $\mathrm{err}_{\mathrm{coll}}$ and $\mathrm{err}_{\mathrm{est}}$. The results are given in Table 2.

Table 2. Convergence orders of collocation and error estimate for (75) ($m = 4$)

h	$\mathrm{err}_{\mathrm{coll}}$	p_{coll}	$\mathrm{err}_{\mathrm{est}}$	p_{est}
2^{-4}	3.5016e−02		6.0759e−03	
2^{-5}	2.3394e−03	3.90	2.0141e−04	4.91
2^{-6}	1.4830e−04	3.98	6.3615e−06	4.98
2^{-7}	9.2822e−06	4.00	1.9924e−07	5.00
2^{-8}	5.7990e−07	4.00	6.2215e−09	5.00
2^{-9}	3.6224e−08	4.00	1.9313e−10	5.01

Both examples presented in this section illustrate the asymptotical correctness of the error estimate analyzed in this paper.

6 Conclusions

In this paper, we have analyzed an a posteriori error estimate for singular boundary value problems based on the defect correction principle. This represents an extension of previous results for the most general class of problems with a singularity of the first kind, where the leading coefficient matrix may have both eigenvalues with negative and positive real parts. In order to derive a bound for the error of the error estimate as compared with the exact global error of collocation, we have derived a new representation of the global error of collocation methods, showing that the error is at least $O(h^m)$ if polynomials of degree m are used to define the basic numerical scheme. The analysis of the error estimate finally revealed that the error of the error estimate is $O(|\ln(h)|^{n_0-1}h^{m+1})$ with some positive integer n_0. Thus, for our generic choice of an even number of equidistant collocation points, the error estimate is asymptotically correct and may serve as a sound basis for adaptive mesh selection.

References

1. W. Auzinger, G. Kneisl, O. Koch, and E. Weinmüller. A collocation code for boundary value problems in ordinary differential equations. *Numer. Algorithms*, 33:27–39, 2003.
2. W. Auzinger, O. Koch, and E. Weinmüller. Analysis of a new error estimate for collocation methods applied to singular boundary value problems. To appear in SIAM J. Numer. Anal. Also available at `http://www.math.tuwien.ac.at/~inst115/preprints.htm/`.
3. W. Auzinger, O. Koch, and E. Weinmüller. Efficient mesh selection for collocation methods applied to singular BVPs. To appear in J. Comput. Appl. Math. Also available at `http://www.math.tuwien.ac.at/~inst115/preprints.htm/`.
4. W. Auzinger, O. Koch, and E. Weinmüller. Efficient collocation schemes for singular boundary value problems. *Numer. Algorithms*, 31:5–25, 2002.
5. E. Badralexe and A.J. Freeman. Eigenvalue equation for a general periodic potential and its multipole expansion solution. *Phys. Rev. B*, 37(3):1067–1084, 1988.
6. C. de Boor and B. Swartz. Collocation at Gaussian points. *SIAM J. Numer. Anal.*, 10:582–606, 1973.
7. T.W. Carr and T. Erneux. Understanding the bifurcation to traveling waves in a class-b laser using a degenerate Ginzburg-Landau equation. *Phys. Rev. A*, 50:4219–4227, 1994.
8. E. Coddington and N. Levison. *Theory of Ordinary Differential Equations*. McGraw-Hill, New York, 1955.
9. F.M. Fernandez and J.F. Ogilvie. Approximate solutions to the Thomas-Fermi equation. *Phys. Rev. A*, 42:149–154, 1990.
10. M. Gräff, R. Scheidl, H. Troger, and E. Weinmüller. An investigation of the complete post-buckling behavior of axisymmetric spherical shells. *ZAMP*, 36:803–821, 1985.

11. F. B. Hildebrand. *Introduction to Numerical Analysis*. McGraw-Hill, New York, 2^{nd} edition, 1974.
12. F.R. de Hoog and R. Weiss. Difference methods for boundary value problems with a singularity of the first kind. *SIAM J. Numer. Anal.*, 13:775–813, 1976.
13. F.R. de Hoog and R. Weiss. Collocation methods for singular boundary value problems. *SIAM J. Numer. Anal.*, 15:198–217, 1978.
14. F.R. de Hoog and R. Weiss. The application of Runge-Kutta schemes to singular initial value problems. *Math. Comp.*, 44:93–103, 1985.
15. D. Jinqiao, H.V. Ly, and E.S. Titi. The effect of nonlocal interactions on the dynamics of the Ginzburg-Landau equation. *Z. Ang. Math. Phys.*, 47:432–455, 1996.
16. H. Keller and A. Wolfe. On the nonunique equilibrium states and buckling mechanism of spherical shells. *J. Soc. Indust. Applied Math.*, 13:674–705, 1965.
17. O. Koch, P. Kofler, and E. Weinmüller. Initial value problems for systems of ordinary first and second order differential equations with a singularity of the first kind. *Analysis*, 21:373–389, 2001.
18. O. Koch and E. Weinmüller. Iterated Defect Correction for the solution of singular initial value problems. *SIAM J. Numer. Anal.*, 38(6):1784–1799, 2001.
19. O. Koch and E. Weinmüller. Analytical and numerical treatment of a singular initial value problem in avalanche modeling. *Appl. Math. Comput.*, 148(2):561–570, 2003.
20. H. J. Stetter. *Analysis of Discretization Methods for Ordinary Differential Equations*. Springer-Verlag, Berlin-Heidelberg-New York, 1973.
21. H. J. Stetter. The defect correction principle and discretization methods. *Numer. Math.*, 29:425–443, 1978.
22. P. Tholfsen and H. Meissner. Cylindrically symmetric solutions of the Ginzburg-Landau equations. *Phys. Rev.*, 169:413–416, 1968.
23. 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.*, 15:287–307, 1984.
24. E. Weinmüller. Collocation for singular boundary value problems of second order. *SIAM J. Numer. Anal.*, 23:1062–1095, 1986.
25. Chin-Yu Yeh, A.-B. Chen, D.M. Nicholson, and W.H. Butler. Full-potential Korringa-Kohn-Rostoker band theory applied to the Mathieu potential. *Phys. Rev. B*, 42(17):10976–10982, 1990.