BIT Numerical Mathematics
2003, Vol. 43, No. 1, pp. 001–018
0006-3835/03/4301-0001 $16.00

LOCAL ERROR ESTIMATES FOR MODERATELY SMOOTH PROBLEMS: PART I – ODEs AND DAEs

THORSTEN SICKENBERGER[1], EWA WEINMÜLLER[2] and RENATE WINKLER[3] *

[1,3] *Institut für Mathematik, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin, Germany. email: sickenberger/winkler@math.hu-berlin.de*

[2] *Institut für Analysis und Scientific Computing, Technische Universität Wien, Wiedner Hauptstrasse 8-10, A-1040 Wien, Austria. email: e.weinmueller@tuwien.ac.at*

Abstract.

The paper consists of two parts. In the first part, we propose a procedure to estimate local errors of low order methods applied to solve initial value problems in ordinary differential equations (ODEs) and index 1 differential-algebraic equations (DAEs). Based on the idea of Defect Correction we develop local error estimates for the case when the problem data is only moderately smooth. Numerical experiments illustrate the performance of the mesh adaptation based on the error estimation developed in this paper. In the second part of the paper, we will consider the estimation of local errors in context of stochastic differential equations with small noise.

AMS subject classification (2000): 65L06, 65L80, 65L50, 65L05.

Key words: Local error estimation, Step-size control, Adaptive methods, Initial Value Problems, Differential-algebraic equations, Defect Correction.

1 Introduction.

In this paper we are interested in the design of error estimates for the local errors arising during the numerical integration of classical ODEs and DAEs. Our main concern is to deal with only moderate smoothness of the problem data and the solution. Our motivation for considering this type of difficulty are applications in electrical circuit simulation, where the models often contain data with poor smoothness. In a consecutive paper dealing with stochastic differential equations and stochastic differential algebraic equations, we will focus our attention on the case of small noise, where the dominant part of the local error still exhibits deterministic behavior. The results derived here will provide the necessary techniques for this further development and therefore we decided to discuss them in a detailed and comprehensive manner.

Our ideas originate from the well-known principle of Defect Correction which can be utilized to estimate local and global errors of discretization schemes in the

*The first author acknowledges support by the BMBF-project 03RONAVN, the second author support by the Austrian Science Fund Project P17253, and the third author support by the DFG Research Center MATHEON in Berlin.

context of both, initial and boundary value problems in ODEs. Defect Correction also constitutes the acceleration technique called Iterated Defect Correction (IDeC). An overview of the Defect Correction method in which we motivate and describe the ideas behind it, is given in Section 2. In Section 3 we show how this technique can be used to provide a reliable error estimate for systems of ODEs. The essential feature of the error estimation procedure is that it should not require more smoothness of the solution than the numerical solver itself. In Section 4 we discuss several modifications of this technique in context of DAEs. In both cases, we are mainly interested to cover the case of only very limited smoothness. We propose a step-size control algorithm based on the local error estimation in Section 5 and finally, we report on numerical experiments illustrating the performance of the code in Section 6.

2 Principles of Defect Correction.

For the reader's convenience, we now give an overview of the Defect Correction technique, referring to the literature for further details. Our aim is not only to describe the main ideas but also to show the importance of the way how the defects are evaluated, because this is a crucial point when it matters how much smoothness we need to require for the method to work. By only choosing a proper way of defect evaluation, we can show that method performs well with less smoothness requirements posed on the analytical solution. We first deal with initial value problems for ODEs of the form

$$x'(t) = f(t, x(t)), \;\; t \in \mathcal{J}, \quad x(t_0) = x_0, \tag{2.1}$$

where $\mathcal{J} = [t_0, t_{end}]$, $x\colon \mathcal{J} \to \mathbb{R}^n$, $f\colon \mathcal{J} \times \mathbb{R}^n \to \mathbb{R}^n$, and assume that (2.1) has a unique solution $x = x(\cdot; t_0, x_0)$.

Before presenting possible realizations of the Defect Correction approach, we briefly discuss its main principle, cf. [31]. Let us first define the *original problem*,

$$F(u) = 0 \tag{2.2}$$

whose exact solution will be denoted by $u^\star$. We can view (2.2) as the analytical problem (2.1) but also as an *ideal*, defect defining discretization scheme of high order. Moreover, let

$$\widetilde{F}(u) = 0 \tag{2.3}$$

specify a *discretization scheme* with the exact solution $\widetilde{u}$. We usually identify (2.3) with a basic solver of low order. Finally, we assume that an approximation v for $u^\star$ is given and note that v does not necessarily need to be a solution of (2.3), and consequently, $\widetilde{F}(v) \neq 0$, in general.

In order to estimate the *global error* of v, we proceed as follows: We first calculate the defect of v by

$$F(v) =: d \tag{2.4}$$

and then solve the perturbed problem (2.3),

$$\widetilde{F}(h) = d. \tag{2.5}$$

Now, we have

$$\underbrace{v - u^\star}_{\text{error}} = F^{-1}F(v) - F^{-1}F(u^\star) \approx \widetilde{F}^{-1}d - \widetilde{F}^{-1}0 = \underbrace{h - \widetilde{u}}_{\text{estimate}} . \tag{2.6}$$

In case that v solves (2.3), $\widetilde{F}(v) = 0$, $v = \widetilde{u}$ holds and the global error is related to the difference of solutions of the perturbed and unperturbed problems, $\widetilde{F}(h) = d$ and $\widetilde{F}(v) = 0$,

$$\underbrace{v - u^\star}_{\text{error}} \approx \underbrace{h - v}_{\text{estimate}} .$$

We argue similarly to estimate the local truncation error defined by $\widetilde{F}(u^\star)$. In this case, in addition to $d = F(v)$, we also require $\widetilde{d}$,

$$\widetilde{F}(v) =: \widetilde{d}.$$

This yields,

$$\underbrace{\widetilde{F}(u^\star)}_{\text{error}} = \widetilde{F}(u^\star) - F(u^\star) \approx \widetilde{F}(v) - F(v) = \underbrace{\widetilde{d} - d}_{\text{estimate}} \tag{2.7}$$

and in case that v is a solution of (2.3), $\widetilde{F}(v) = 0$, and thus $\widetilde{d} = 0$, we obtain

$$\underbrace{\widetilde{F}(u^\star)}_{\text{error}} = \underbrace{-d}_{\text{estimate}} . \tag{2.8}$$

Clearly, this general discussion may be useful for a better understanding of the Defect Correction method but the specifications are not precise enough to put the method to work in practice. This question will be addressed in the next sections.

2.1 Estimate for the global error, IDeC.

Since its introduction in the 1970's, cf. [12], [31], [33], the idea of IDeC has been successfully applied to various classes of differential equations. The method is carried out in the following way: Compute a simple, basic approximation and form its defect w.r.t. the given ODE via a piecewise interpolant. This defect is used to define a neighboring problem whose exact solution is known. Solving the neighboring problem with the basic discretization scheme yields a global error estimate. This can be used to construct an improved approximation, and the procedure can be iterated. The fixed point of such an iterative process corresponds to a certain collocating solution. Let

$$\Gamma := \{t_0 < t_1 < \ldots < t_i < \cdots < t_N = t_{end}\} \tag{2.9}$$

be a partition of the interval $\mathcal{J}$. We denote the length of the subinterval $[t_{i-1}, t_i]$ by $h_i = t_i - t_{i-1}$, $i = 1, \ldots, N$. Let $\mathbf{h}$ be the maximal step-size of Γ, $\mathbf{h} := \max_{1 \le i \le N} h_i$. For the subsequent analysis we assume that the step-size $\mathbf{h}$ is sufficiently small to guarantee the convergence of the involved numerical schemes and the asymptotic regime of the error estimates.

For the IVP (2.1) the IDeC procedure can be realized as follows: An approximate solution $x_\Gamma^{[0]} = (x_0^{[0]}, x_1^{[0]}, \ldots, x_i^{[0]}, \ldots, x_N^{[0]})$ is obtained by some discretization method on the grid Γ. For simplicity assume that $x_\Gamma^{[0]}$ has been computed by the backward Euler scheme (BEUL),

$$\frac{x_i - x_{i-1}}{h_i} = f(t_i, x_i), \quad i = 1, \ldots, N. \tag{2.10}$$

Using the polynomial $p^{[0]}(t)$ of degree $\le N$ which interpolates the values of $x_\Gamma^{[0]}$, $p^{[0]}(t_i) = x_i$, $i = 0, \ldots, N$, we construct a neighboring problem

$$x'(t) = f(t, x(t)) + d^{[0]}(t), \quad x(t_0) = x_0, \tag{2.11}$$

where $d^{[0]}(t)$ denotes the defect w.r.t. (2.1),

$$d^{[0]}(t) := \frac{d}{dt} p^{[0]}(t) - f(t, p^{[0]}(t)). \tag{2.12}$$

We now solve (2.11) using the same numerical method as before to obtain an approximation $p_\Gamma^{[0]}$ for the exact solution $p^{[0]}(t)$ of (2.11). Note that for (2.11) we know the global error given by $p_\Gamma^{[0]} - R_\Gamma p^{[0]}$, where R_Γ denotes the restriction operator $[t_0, t_{\text{end}}] \to \Gamma$. Assuming $x_\Gamma^{[0]}$ to be a good approximation for $R_\Gamma x$ and therefore $p_\Gamma^{[0]}$ to be a good approximation for $x(t)$, we may expect $d^{[0]}(t)$ to be small and the problems (2.1) and (2.11) to be closely related. Consequently, the global error for the neighboring problem (2.11) should provide a good estimate for the unknown error of the original problem (2.1)

$$\underbrace{x_\Gamma^{[0]} - R_\Gamma x}_{\text{error}} \approx \underbrace{p_\Gamma^{[0]} - R_\Gamma p^{[0]}}_{\text{error estimate}},$$

cf. (2.6). The approximation $p_\Gamma^{[0]} - R_\Gamma p^{[0]}$ of the global error of the original problem can now be used to improve the numerical solution of (2.1),

$$x_\Gamma^{[1]} := x_\Gamma^{[0]} - (p_\Gamma^{[0]} - R_\Gamma p^{[0]}). \tag{2.13}$$

The values $x_\Gamma^{[1]}$ are used to define a new interpolating polynomial $p^{[1]}(t)$ by requiring $p^{[1]}(t_i) = x_i^{[1]}$ and $p^{[1]}(t)$ defines a new neighboring problem analogous to (2.11). This procedure can be continued iteratively in an obvious manner. In practice one does not use one interpolating polynomial for the whole interval $[t_0, t_{\text{end}}]$. Instead, piecewise functions composed of polynomials of (moderate)

degree m are used to specify the neighboring problem. For sufficiently smooth data functions $f(t,x)$ it can be shown that the approximations $x_\Gamma^{[\nu]}$ satisfy

$$x_i^{[\nu]} - x(t_i) = O(\mathbf{h}^{\nu+1}), \quad \nu = 0, \ldots, m-1. \tag{2.14}$$

One of the most attractive features of the IDeC procedure is, that its fixed point is a certain superconvergent collocation solution of (2.1). In [6] and [7] a variety of modifications to this algorithm have been discussed. Some of these have been proposed only recently, and together they form a family of iterative techniques, each with its particular advantages.

Clearly, in each step of the classical IDeC procedure, we obtain not only an improved approximation $x_\Gamma^{[\nu]}$ for the exact solution values $R_\Gamma x$, but also an asymptotically correct estimate $p_\Gamma^{[\nu-1]} - R_\Gamma p^{[\nu-1]}$ for the global error of the basic method $x_\Gamma^{[0]} - R_\Gamma x$:

$$\begin{aligned}&(p_i^{[\nu]} - p^{[\nu]}(t_i)) - (x_i^{[0]} - x(t_i)) = \\ &\underbrace{(p_i^{[\nu]} - p^{[\nu]}(t_i)) - (x_i^{[0]} - x_i^{[\nu+1]})}_{=0} - (x_i^{[\nu+1]} - x(t_i)) = O(\mathbf{h}^{\nu+1}),\end{aligned}$$

$\nu = 0, \ldots, m-1$. Similarly, in each step of the iteration the difference $x_\Gamma^{[\nu]} - x_\Gamma^{[\nu+1]}$ can serve to estimate the global error $x_\Gamma^{[\nu]} - R_\Gamma x$ of the current approximation $x_\Gamma^{[\nu]}$,

$$(x_i^{[\nu]} - x_i^{[\nu+1]}) - \underbrace{(x_i^{[\nu]} - x(t_i))}_{=O(\mathbf{h}^{\nu+1})} = -(x_i^{[\nu+1]} - x(t_i)) = O(\mathbf{h}^{\nu+2}),\ \nu = 0, \ldots, m-1.$$

The DeC principle can also be used to estimate the global error of higher order schemes. It was originally proposed by Zadunaisky in order to estimate the global error of Runge-Kutta schemes. In this original version discussed in [9], [33], the error estimate for the high-order method is obtained by applying the given scheme twice, to the analytical problem (2.1) first, and to a properly defined neighboring problem next. In [12] and [31], this procedure was modified in order to reduce the amount of computational work. Here, the high-order method is carried out only to solve the original problem. Additionally, a computationally cheap low-order method is used twice to solve the original and the neighboring problem. We refer to [4] and [5] for further variants of the above error estimation strategies.

2.2 Estimate for the local truncation error in the IDeC context.

In [10] and [11] another variant of the IDeC procedure based on the estimation of the local truncation error was introduced. Let us again consider problem (2.1) and its numerical solution x_Γ obtained by the backward Euler scheme (2.10). If we knew the exact values of the local truncation error per unit step,

$$l_i^{us} := \frac{x(t_i) - x(t_{i-1})}{h_i} - f(t_i, x(t_i)), \ \ i = 1, \ldots, N, \tag{2.15}$$

and if we solved the perturbed BEUL scheme

$$\frac{y_i - y_{i-1}}{h_i} = f(t_i, y_i) + l_i^{us}, \;\; i = 1, \ldots, N, \tag{2.16}$$

then we would recover the correct values of the solution, $y_i = x(t_i)$, $i = 1, \ldots, N$. In practice we need to estimate the values of l_i^{us}. For this purpose consider m adjacent points to t_i, say $t_{i-m}, t_{i-m+1}, \ldots, t_{i-1}$, and define a polynomial $q_i(t)$[1] of degree $\leq m$ by requiring $q_i(t_i) = x_i$, $i = i - m, \ldots, i$. Using this polynomial we now construct the problem,

$$x_i'(t) = f(t, x_i(t)) + d_i^{us}(t), \;\; t \in [t_{i-m}, t_i], \quad x_i(t_0) = q_i(t_0), \tag{2.17}$$

where

$$d_i^{us}(t) := q_i'(t) - f(t, q_i(t)), \;\; i = 1, \ldots, N. \tag{2.18}$$

We again can expect that $q_i(t)$ is a good approximation for $x(t)$ in the interval $[t_{i-m}, t_i]$. Outside of $[t_{i-m}, t_i]$ the polynomial $q_i(t)$ may differ significantly from $x(t)$. Therefore, we could view (2.17) as a *local* neighboring problem for (2.1). Since $q_i(t)$ is the exact solution of (2.17), we know the associated local truncation error at t_i,

$$\begin{aligned} \ell_i^{us} :=& \frac{q_i(t_i) - q_i(t_{i-1})}{h_i} - f(t_i, q_i(t_i)) - q_i'(t_i) + f(t_i, q_i(t_i)) \\ =& \frac{x_i - x_{i-1}}{h_i} - q_i'(t_i), \;\; i = 1, \ldots, N, \end{aligned} \tag{2.19}$$

and thus we can use ℓ_i^{us} to estimate l_i^{us} in (2.16). Obviously, this process can be iteratively continued. However, in this paper, we are not interested in applying the related acceleration procedure, but in using the above idea to reliably estimate local errors of numerical methods for IVPs and consequently, to provide a basis for a step adaptation strategy.

We stress that we do not necessarily need to evaluate the defect in a way described in (2.18). In fact, it turns out that a modified defect definition will be more suitable in the case of very moderate smoothness in x. All we need is the property that ℓ_i^{us} is an asymptotically correct error estimate for l_i^{us},

$$\ell_i^{us} = l_i^{us} + O(h_i^2), \;\; i = 1, \ldots, N. \tag{2.20}$$

This requirement is motivated by the fact that for the backward Euler scheme both, ℓ_i^{us} and l_i^{us} are $O(h_i)$. Depending on the choice of d_i^{us} condition (2.20) holds under different smoothness requirements on x. It has been shown in [3] in context of equidistant grids that $x \in C^5[t_0, t_{end}]$ is sufficient for d_i^{us} specified via (2.18) to guarantee that (2.20) is satisfied.

[1] We denote this polynomial by q_i and not by p_i, as in the previous section, in order to underline that it is a local approximation for $x(t), t \in [t_{i-m}, t_i]$.

The following form of d_i^{us} suits both, less smooth solutions of (2.1) and arbitrary grids, see [4]:

$$\begin{aligned} d_i^{us} &:= \frac{q_i(t_i) - q_i(t_{i-1})}{h_i} - \frac{1}{2}\big(f(t_{i-1}, q_i(t_{i-1})) + f(t_i, q_i(t_i))\big) \\ &= \frac{x_i - x_{i-1}}{h_i} - \frac{1}{2}\big(f(t_{i-1}, x_{i-1}) + f(t_i, x_i)\big) = \frac{1}{2}\big(f(t_i, x_i) - f(t_{i-1}, x_{i-1})\big). \end{aligned} \tag{2.21}$$

For $x \in C^2[t_0, t_{end}]$ we have

$$\begin{aligned} \ell_i^{us} &= \frac{x_i - x_{i-1}}{h_i} - f(t_i, x_i) - \frac{x_i - x_{i-1}}{h_i} + \frac{1}{2}\big(f(t_{i-1}, x_{i-1}) + f(t_i, x_i)\big) \\ &= \frac{1}{2}\big(f(t_{i-1}, x_{i-1}) - f(t_i, x_i)\big). \end{aligned} \tag{2.22}$$

Recall that

$$l_i^{us} := \frac{x(t_i) - x(t_{i-1})}{h_i} - f(t_i, x(t_i)) = -\frac{1}{2} h_i x''(t_i) + o(h_i), \tag{2.23}$$

and thus

$$\ell_i^{us} - l_i^{us} = \frac{h_i}{2}\left(x''(t_i) - \frac{1}{h_i}\Big(\underbrace{f(t_i, x_i) - f(t_{i-1}, x_{i-1})}_{:=\Delta f_i} \Big) \right) + o(h_i) \tag{2.24}$$

with

$$\begin{aligned} \Delta f_i = &\underbrace{(f(t_i, x_i) - f(t_i, x(t_i)))}_{=J_i(x_i - x(t_i))} - \underbrace{(f(t_{i-1}, x_{i-1}) - f(t_{i-1}, x(t_{i-1})))}_{=J_{i-1}(x_{i-1} - x(t_{i-1}))} \\ &-(f(t_{i-1}, x(t_{i-1})) - f(t_i, x(t_i))), \end{aligned}$$

where $J_i = \int_0^1 f_x(t_i, sx_i + (1-s)x(t_i))\, ds$, $J_{i-1} = \int_0^1 f_x(t_{i-1}, sx_{i-1} + (1-s)x(t_{i-1}))\, ds$. Here, we assume that the right-hand side f is continuously differentiable with respect to x. From

$$\begin{aligned} x_i - x(t_i) &= x_{i-1} - x(t_{i-1}) + h_i f(t_i, x_i) - h_i x'(t_{i-1}) + O(h_i^2) \\ &= x_{i-1} - x(t_{i-1}) + h_i f(t_i, x_i) - h_i f(t_i, x(t_i)) + h_i f(t_i, x(t_i)) \\ &\quad - h_i(x'(t_i) + O(h_i)) + O(h_i^2) = x_{i-1} - x(t_{i-1}) + O(h_i^2) \end{aligned}$$

we have

$$\begin{aligned} \Delta f_i &= J_i(x_i - x(t_i)) - J_{i-1}(x_{i-1} - x(t_{i-1})) - (x'(t_{i-1}) - x'(t_i)) \\ &= J_i(x_{i-1} - x(t_{i-1}) + O(h_i^2)) - J_{i-1}(x_{i-1} - x(t_{i-1})) + h_i x''(t_i) + o(h_i) \\ &= h_i x''(t_i) + \underbrace{(J_i - J_{i-1})(x_{i-1} - x(t_{i-1}))}_{=O(h_i^2)} + o(h_i), \end{aligned}$$

and consequently,

$$\ell_i^{us} - l_i^{us} = \frac{h_i}{2}\left(x''(t_i) - \frac{1}{h_i}\big(h_i x''(t_i) + o(h_i)\big)\right) + o(h_i) = o(h_i).$$

Using similar arguments one could show that for $x \in C^3[t_0, t_{end}]$,

$$\ell_i^{us} - l_i^{us} = O(h_i^2)$$

holds.

Here, a remark is in order. In the above calculations we have taken advantage of the fact that in (2.24), x'' is approximated by differences of f-values. Note that the weighted sum of f-values in (2.21) can be related to a certain quadrature rule, cf. [5]. The defect evaluation in (2.21) can also be regarded as a substitution of the solution obtained from the numerical scheme of order $m - 1 = 1$ into another scheme of higher order, in this case of order $m = 2$. This idea is widely used in the design of error estimation procedures based on residual control. Its generalization in context of multi-step schemes will be discussed in the next section.

For the defect definition (2.18) we would have obtained

$$\ell_i^{us} - l_i^{us} = \frac{h_i}{2}\left(x''(t_i) - q_i''(t_i)\right) + O(h_i^2), \quad i = 1, \ldots, N, \tag{2.25}$$

which means that in this case x'' would be approximated by q_i'', where q_i is a polynomial interpolating the values of x_Γ. The different defect definitions (2.18) and (2.21) result in canceling of $f(t_i, q_i(t_i))$ terms in (2.19) and $(x_i - x_{i-1})/h_i$ terms in (2.22), respectively.

3 Estimates of the local error via defect evaluation.

We now exploit the advantages of the defect evaluation defined via (2.21) in context of a general multi-step scheme. We consider two different multi-step schemes, the basic one which we identify with the solver scheme, and the auxiliary one which is used to define the defect for the estimation of the local discretization error of the basic scheme. We first analyze the relation between the defect and the local error in context of two general schemes of order p and $\bar{p}$, respectively, and then apply the results to order one and order two schemes, to derive working formulas for the implementation.

3.1 Linear multi-step schemes.

Consider a linear multi-step scheme for the ODE (2.1) carried out on the grid Γ,

$$\sum_{j=0}^{k} \alpha_{j,i} x_{i-j} = h_i \sum_{j=0}^{k} \beta_{j,i} f(t_{i-j}, x_{i-j}), \quad i = k, \ldots N, \tag{3.1}$$

with given initial values $x_0, x_1, \ldots, x_{k-1} \in \mathbb{R}^n$. Let the coefficients of the scheme be normalized in such a way that $\alpha_{0,i} = 1$. The local truncation error[2] l_i of the scheme (3.1) is given by

$$l_i := \sum_{j=0}^{k} \alpha_{j,i} x(t_{i-j}) - h_i \sum_{j=0}^{k} \beta_{j,i} f(t_{i-j}, x(t_{i-j})), \quad i = k, \ldots N. \tag{3.2}$$

The linear multi-step method (3.1) is called consistent of order $p > 0$ if $|l_i| = O(\mathbf{h}^{p+1})$, where $|\cdot|$ denotes a vector norm in $\mathbb{R}^n$.

We identify (3.1) with a given solver routine providing an approximation for the solution of (2.1). Our aim is to design an error estimate for the local truncation error of this approximation which does not need more smoothness to work than the approximation procedure itself. For this purpose we use an auxiliary scheme,

$$\sum_{j=0}^{\bar{k}} \bar{\alpha}_{j,i} \bar{x}_{i-j} = h_i \sum_{j=0}^{\bar{k}} \bar{\beta}_{j,i} f(t_{i-j}, \bar{x}_{i-j}), \quad i = \bar{k}, \ldots N, \tag{3.3}$$

with given values $x_0, \bar{x}_1, \ldots, \bar{x}_{\bar{k}-1} \in \mathbb{R}^n$ and $\bar{\alpha}_{0,i} = 1$. As before, the local truncation error of (3.3) is given as

$$\bar{l}_i := \sum_{j=0}^{\bar{k}} \bar{\alpha}_{j,i} x(t_{i-j}) - h_i \sum_{j=0}^{\bar{k}} \bar{\beta}_{j,i} f(t_{i-j}, x(t_{i-j})), \tag{3.4}$$

and the scheme (3.3) is of order of consistency $\bar{p}$ if $|\bar{l}_i| = O(\mathbf{h}^{\bar{p}+1})$ holds. In this section, we are particularly interested in the case $p = \bar{p}$.

We first discuss the properties of the defect, defined by

$$d_i := \sum_{j=0}^{\bar{k}} \bar{\alpha}_{j,i} x_{i-j} - h_i \sum_{j=0}^{\bar{k}} \bar{\beta}_{j,i} f(t_{i-j}, x_{i-j}), \quad i = k, \ldots N, \tag{3.5}$$

obtained by substituting the approximations x_i computed from (3.1) into the scheme (3.3). Let us assume that the starting values for the schemes (3.1) and (3.3) are exact, and denote the solutions computed after the first step by $x_i^\star$ and $\bar{x}_i^\star$ respectively,

$$x_i^\star = -\sum_{j=1}^{k} \alpha_{j,i} x(t_{i-j}) + h_i \beta_{0,i} f(t_i, x_i^\star) + h_i \sum_{j=1}^{k} \beta_{j,i} f(t_{i-j}, x(t_{i-j})), \tag{3.6}$$

$$\bar{x}_i^\star = -\sum_{j=1}^{\bar{k}} \bar{\alpha}_{j,i} x(t_{i-j}) + h_i \bar{\beta}_{0,i} f(t_i, \bar{x}_i^\star) + h_i \sum_{j=1}^{\bar{k}} \bar{\beta}_{j,i} f(t_{i-j}, x(t_{i-j})), \tag{3.7}$$

[2]It is defined by substituting the values of the exact solution into the scheme. Note that now the scaling of the numerical scheme is different from (2.10). The local truncation error, l_i, is related to the local truncation error per unit step, l_i^{us}, by $l_i = h_i l_i^{us}$.

for $i = k, \dots N$. For explicit schemes ($\beta_{0,i} = 0$ and $\bar{\beta}_{0,i} = 0$) we immediately have

$$l_i = x(t_i) - x_i^\star \quad \text{and} \quad \bar{l}_i = x(t_i) - \bar{x}_i^\star,$$

but in general,

$$l_i = x(t_i) - x_i^\star - h_i\beta_{0,i}\big(f(t_i, x(t_i)) - f(t_i, x_i^\star)\big) = \big(I - h_i\beta_{0,i}J_i\big)(x(t_i) - x_i^\star), \tag{3.8}$$

and

$$\bar{l}_i = x(t_i) - \bar{x}_i^\star - h_i\bar{\beta}_{0,i}\big(f(t_i, x(t_i)) - f(t_i, \bar{x}_i^\star)\big) = \big(I - h_i\bar{\beta}_{0,i}\bar{J}_i\big)(x(t_i) - \bar{x}_i^\star). \tag{3.9}$$

Here, $J_i = \int_0^1 f'_x(t_i, sx(t_i) + (1-s)x_i^\star)\, ds$, $\bar{J}_i = \int_0^1 f'_x(t_i, sx(t_i) + (1-s)\bar{x}_i^\star)\, ds$, and f is supposed to be differentiable with respect to x. The properties of the defect d_i from (3.5) are formulated in the following lemma.

LEMMA 3.1. *Let $f(t,x)$ be continuous and continuously differentiable with respect to x. Let the step-size* **h** *be sufficiently small to guarantee that the matrix $\big(I - h_i\beta_{0,i}J_i\big)$ is nonsingular. Then the defect $d_i^\star$ defined*[3] *by*

$$d_i^\star := x_i^\star + \sum_{j=1}^{\bar{k}} \bar{\alpha}_{j,i}x(t_{i-j}) - h_i\bar{\beta}_{0,i}f(t_i, x_i^\star) - h_i\sum_{j=1}^{\bar{k}} \bar{\beta}_{j,i}f(t_{i-j}, x(t_{i-j})), \tag{3.10}$$

satisfies

$$d_i^\star = \bar{l}_i - l_i + h_i(\bar{\beta}_{0,i} - \beta_{0,i})J_i\big(I - h_i\beta_{0,i}J_i\big)^{-1}l_i. \tag{3.11}$$

The quantities x_i^*, $\bar{x}_i^*$ and the defects d_i and $d_i^\star$ in context of backward and forward Euler schemes are visualized in Figure 3.1. There we have $d_i = x_i^{\text{BEUL}} - x_{i-1} - h_i f(t_{i-1}, x_{i-1}) = x_i^{\text{BEUL}} - x_i^{\text{FEUL}} = x_i - \bar{x}_i$ and $d_i^* = x_i^* - \bar{x}_i^*$.

PROOF. Using the definitions (3.10), (3.4) and the relation (3.8) we obtain

$$\begin{aligned}
d_i^\star &= x_i^\star - h_i\bar{\beta}_{0,i}f(t_i, x_i^\star) + \sum_{j=1}^{\bar{k}} \bar{\alpha}_{j,i}x(t_{i-j}) - h_i\sum_{j=1}^{\bar{k}} \bar{\beta}_{j,i}f(t_{i-j}, x(t_{i-j})) \\
&= x_i^\star - h_i\bar{\beta}_{0,i}f(t_i, x_i^\star) + \bar{l}_i - x(t_i) + h_i\bar{\beta}_{0,i}f(t_i, x(t_i)) \\
&= x_i^\star - x(t_i) - h_i\bar{\beta}_{0,i}\big(f(t_i, x_i^\star) - f(t_i, x(t_i))\big) + \bar{l}_i \\
&= \big(I - h_i\bar{\beta}_{0,i}J_i\big)\big(x_i^\star - x(t_i)\big) + \bar{l}_i \\
&= -\big(I - h_i\bar{\beta}_{0,i}J_i\big)\big(I - h_i\beta_{0,i}J_i\big)^{-1}l_i + \bar{l}_i \\
&= \bar{l}_i - l_i + h_i\big(\bar{\beta}_{0,i} - \beta_{0,i}\big)J_i\big(I - h_i\beta_{0,i}J_i\big)^{-1}l_i\,.
\end{aligned}$$

□

COROLLARY 3.2. *Let the suppositions of Lemma 3.1 be satisfied. Moreover, let the schemes (3.1) and (3.3) be consistent of order p and $\bar{p}$, respectively. For the case $p = \bar{p}$, we additionally assume that*

$$l_i = c_i\, x^{(p+1)}(t_i)\, h_i^{p+1} + o(h_i^{p+1}), \quad \bar{l}_i = \bar{c}_i\, x^{(p+1)}(t_i)\, h_i^{p+1} + o(h_i^{p+1}), \tag{3.12}$$

[3] $d_i^\star$ is obtained by substituting $x_i^\star$ from (3.6) into (3.3)

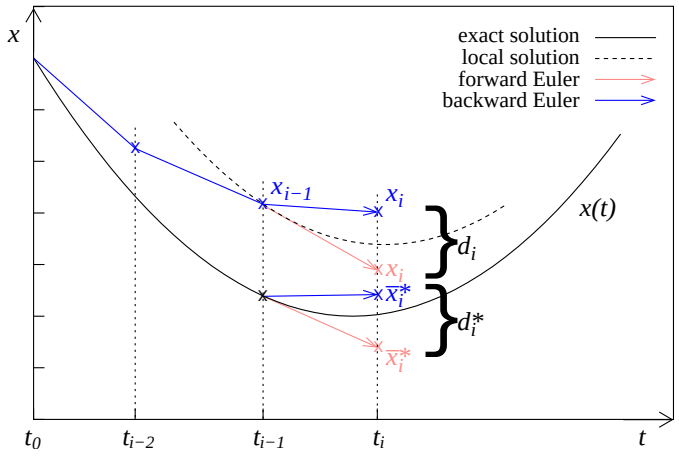

Figure 3.1: The defects d_i and $d_i^\star$ in context of backward and forward Euler schemes.

with constants $c_i \neq \bar{c}_i$, which depend only on the ratios of step-sizes. Then we have

(i) $\bar{l}_i = d_i^\star + O(h_i^{\bar{p}+2})$, *if* $p > \bar{p}$,

(ii) $l_i = -d_i^\star + O(h_i^{p+2})$, *if* $p < \bar{p}$,

(iii) $l_i = \dfrac{c_i}{\bar{c}_i - c_i} d_i^\star + o(h_i^{p+1}), \quad \bar{l}_i = \dfrac{\bar{c}_i}{\bar{c}_i - c_i} d_i^\star + o(h_i^{p+1}),$ *if* $p = \bar{p}$.

PROOF. Equation (3.11) immediately implies the properties (i), (ii), and (iii). □

Corollary 3.2 offers two options for designing an estimate ℓ_i for the local truncation error l_i of (3.1): According to *(ii)* we may choose a higher order scheme (3.3) to evaluate d_i given by (3.5) and set $\ell_i := -d_i$, cf. (2.8). We then have

$$\ell_i - l_i = -(d_i - d_i^\star) + O(h_i^{p+2}).$$

According to *(iii)* we may choose a scheme (3.3) with the same order $\bar{p} = p$ to evaluate d_i and set $\ell_i := \dfrac{c_i}{\bar{c}_i - c_i} d_i$. We then have

$$\ell_i - l_i = \frac{c_i}{\bar{c}_i - c_i}(d_i - d_i^\star) + o(h_i^{p+1}).$$

In both cases ℓ_i can be considered as an asymptotically correct estimate for l_i only if $d_i - d_i^\star = o(h_i^{p+1})$, i.e., if $|d_i - d_i^\star|$ is asymptotically smaller than the local truncation error itself.

In Section 2.2 the defect structured as a weighted sum of f-values, see (2.21), proved advantageous in the case when the solution x is only moderately smooth.

To obtain this structure for the defect (3.5) we choose an auxiliary scheme (3.3) with the same left-hand side as (3.1), i.e., $\bar{\alpha}_{j,i} = \alpha_{j,i}$, $j = 0, \ldots, k_{max}$, $\bar{\alpha}_{j,i} = 0$, $j = k_{max}+1, \ldots, \bar{k}$, where k_{max} is the maximal index j with $\alpha_{j,i} \neq 0$ in (3.1). Obviously this yields

(3.13)
$$d_i = \sum_{j=0}^{k} \alpha_{j,i} x_{i-j} - h_i \sum_{j=0}^{\bar{k}} \bar{\beta}_{j,i} f(t_{i-j}, x_{i-j}) = h_i \sum_{j=0}^{\max(k,\bar{k})} (\beta_{j,i} - \bar{\beta}_{j,i}) f(t_{i-j}, x_{i-j}) \,.$$

The structure of d_i displayed in (3.13) is crucial for the property that the difference $|d_i - d_i^\star|$ is asymptotically smaller than the local truncation error itself, because it provides an additional factor h_i. For the local exact solution $x(t; t_{i-k}, x_{i-k})$ in $d_i^\star$, it follows that $|d_i - d_i^\star| = O(h_i^{p+2})$. The freedom to choose the auxiliary scheme (3.3) now reduces to determine the coefficients $\bar{\beta}_{j,i}, j = 0, \ldots, \bar{k}$, which additionally have to fulfill consistency conditions to ensure that the scheme (3.3) has at least the order of convergence p. In Sections 3.2 and 3.3 we illustrate how this principle can be realized in the context of first and second order schemes. In Section 4 we generalize this technique to specially structured DAEs.

What we would like to control in praxis are the local errors $(x(t_i) - x_i^\star) = \big(I - h_i\beta_{0,i}J_i\big)^{-1} l_i$, see (3.8). As long as the problem is not stiff, the values of $h_i J_i$ are small compared to the identity matrix I. In this case l_i itself is a good approximation to $(x(t_i) - x_i^\star)$. However, for stiff problems the values of $h_i J_i$ can become considerably large and therefore l_i should be scaled by $\big(I - h_i\beta_{0,i}J_i\big)^{-1}$ or by an approximation to this matrix. Since $\big(I - h_i\beta_{0,i}J_i\big)$ is the Jacobian of the discrete scheme (3.1) this matrix (or its good approximation) and its factorization are usually available.

3.2 First order schemes.

For the first order methods we assume the analytical solution x to be in $C^2[t_0, t_{end}]$. Here we deal only with one-step schemes which simplifies matters, since the coefficients and the error constant are not step dependent. We first consider the forward Euler scheme.

EXAMPLE 3.1. We identify the forward Euler scheme (FEUL) with (3.1). The related local truncation error satisfies

$$l_i = l_i^{\text{FEUL}} = x(t_i) - x(t_{i-1}) - h_i x'(t_{i-1}) = \frac{h_i^2}{2} x''(t_i) + o(h_i^2).$$

Thus, the forward Euler scheme has the order of consistency $p = 1$ with the error constant $c = c_{\text{FEUL}} = \frac{1}{2}$. We choose the auxiliary scheme as a linear one-step method with the same left-hand side and obtain

$$\bar{x}_i - \bar{x}_{i-1} = h_i\big(\bar{\beta}_0 f(t_i, \bar{x}_i) + \bar{\beta}_1 f(t_{i-1}, \bar{x}_{i-1})\big),$$

where the coefficients $\bar{\beta}_0$ and $\bar{\beta}_1$ have to satisfy the consistency condition, $\bar{\beta}_0 + \bar{\beta}_1 = 1$, to ensure that this scheme is at least consistent of order 1. Consequently,

we arrive at the one-parameter family, where $\bar{\beta}_0 = \theta$, $\bar{\beta}_1 = (1-\theta)$, and the local truncation error has the form

$$\bar{l}_i = l_i^\theta = x(t_i) - x(t_{i-1}) - h_i\big(\theta x'(t_i) + (1-\theta)x'(t_{i-1})\big) = \left(\frac{1}{2} - \theta\right) h_i^2 x''(t_i) + o(h_i^2).$$

For $\theta \neq \frac{1}{2}$ the order is 1 and the error constant reads $\bar{c} = c^\theta = (\frac{1}{2} - \theta)$. The defect $d_i = d_i^\theta$ from (3.5) is given by

$$\begin{aligned} d_i^\theta &= x_i^{\text{FEUL}} - x_{i-1} - h_i\big(\theta f(t_i, x_i^{\text{FEUL}}) + (1-\theta) f(t_{i-1}, x_{i-1})\big) \\ &= -h_i\theta\big(f(t_i, x_i^{\text{FEUL}}) - f(t_{i-1}, x_{i-1})\big), \end{aligned} \tag{3.14}$$

and the error estimate ℓ_i^θ, the scaled defect, is

$$\ell_i^\theta = \frac{c}{c^\theta - c} d_i^\theta = \frac{\frac{1}{2}}{-\theta} d_i^\theta = \frac{1}{2} h_i\big(f(t_i, x_i^{\text{FEUL}}) - f(t_{i-1}, x_{i-1})\big).$$

While d_i^θ depends on the parameter θ, the error estimate $\ell_i = \ell_i^\theta$ does not. The same error estimate is obtained for $\theta = \frac{1}{2}$ which corresponds to the trapezoidal rule of order 2, see (*ii*) in Corollary 3.2.

It is important to note that the value of $f(t_i, x_i^{\text{FEUL}})$ necessary for the computation of ℓ_i will be used in the next step of the integration procedure which means that we do not face any additional evaluation of the right-hand side f.

EXAMPLE 3.2. We now identify the backward Euler scheme with (3.1). Here the error constant is $c = c_{\text{BEUL}} = -\frac{1}{2}$. For the auxiliary scheme (3.3) we have exactly the same choices as in the previous example. Again, this results in an error estimate ℓ_i independent of the free parameter θ. The most simple way to derive this error estimate is to set $\theta = 0$ which means that (3.3) is the forward Euler scheme with the error constant $\bar{c} = c_{\text{FEUL}} = \frac{1}{2}$. Now, the defect d_i from (3.5) is given by

$$d_i^{\text{FEUL}} = x_i^{\text{BEUL}} - x_{i-1} - h_i f(t_{i-1}, x_{i-1}) = h_i\big(f(t_i, x_i^{\text{BEUL}}) - f(t_{i-1}, x_{i-1})\big)$$

and the resulting error estimate is

$$\ell_i = \frac{c}{c^{\text{FEUL}} - c} d_i^{\text{FEUL}} = -\frac{1}{2} d_i^{\text{FEUL}} = -\frac{1}{2} h_i\big(f(t_i, x_i^{\text{BEUL}}) - f(t_{i-1}, x_{i-1})\big).$$

As in Example 3.1, the computation of ℓ_i does not cost any additional evaluations of the right-hand side f.

3.3 Second order schemes.

We now deal with two-step schemes of order 2 which, for $\alpha_2 = \beta_2 = 0$, include the trapezoidal rule. We assume the analytical solution x to be in $C^3[t_0, t_{end}]$. Except for the trapezoidal rule itself, we have to cope with variable coefficients and error constants that depend on the ratio of the step-sizes $\kappa_i = h_i/h_{i-1}$. We consider the linear two-step schemes of the form

$$x_i + \alpha_{1,i} x_{i-1} + \alpha_{2,i} x_{i-2} = h_i\Big(\beta_{0,i} f(t_i, x_i) + \beta_{1,i} f(t_{i-1}, x_{i-1}) + \beta_{2,i} f(t_{i-2}, x_{i-2})\Big), \tag{3.15}$$

normalized by the choice $\alpha_{0,i} = 1$. The remaining coefficients $\alpha_{1,i}, \alpha_{2,i}, \beta_{0,i}, \beta_{1,i}$ and $\beta_{2,i}$ have to satisfy the Dahlquist's root condition, i.e., $\alpha_{2,i} \in [-1, 1)$, to guarantee numerical stability, and three further conditions for the consistency of order 2, namely

$$1 + \alpha_{1,i} + \alpha_{2,i} = 0 \ , \tag{3.16}$$

$$-\alpha_{1,i} - \Big(1 + \frac{1}{\kappa_i}\Big)\alpha_{2,i} = \beta_{0,i} + \beta_{1,i} + \beta_{2,i} \ , \tag{3.17}$$

$$\frac{1}{2}\alpha_{1,i} + \frac{1}{2}\Big(1 + \frac{1}{\kappa_i}\Big)^2 \alpha_{2,i} = -\beta_{1,i} - \Big(1 + \frac{1}{\kappa_i}\Big)\beta_{2,i} \ . \tag{3.18}$$

What remains is a two-parameter family whose coefficients can be expressed using $\alpha_{2,i} \in [-1, 1)$, $\beta_{2,i} \in \mathbb{R}$,

$$\begin{aligned} \alpha_{1,i} &= -1 - \alpha_{2,i} \ , \\ \beta_{1,i} &= \frac{1}{2} - \frac{1}{\kappa_i}\Big(1 + \frac{1}{2\kappa_i}\Big)\alpha_{2,i} - \Big(1 + \frac{1}{\kappa_i}\Big)\beta_{2,i} \ , \\ \beta_{0,i} &= \frac{1}{2} + \frac{1}{2\kappa_i^2}\alpha_{2,i} + \frac{1}{\kappa_i}\beta_{2,i} \ . \end{aligned} \tag{3.19}$$

The error constant of the resulting method is

$$\begin{aligned} c_i &= -\frac{1}{6}\alpha_{1,i} - \frac{1}{6}\Big(1 + \frac{1}{\kappa_i}\Big)^3 \alpha_{2,i} - \frac{1}{2}\beta_{1,i} - \frac{1}{2}\Big(1 + \frac{1}{\kappa_i}\Big)^2 \beta_{2,i} \\ &= -\frac{1}{12} - \alpha_{2,i}\Big(\frac{1}{4\kappa_i^2} + \frac{1}{6\kappa_i^3}\Big) - \beta_{2,i}\Big(\frac{1}{2\kappa_i} + \frac{1}{2\kappa_i^2}\Big) \ . \end{aligned}$$

In the sequel, we discuss three important second order multi-step schemes, the implicit trapezoidal rule (ITR), the two-step backward differentiation formula (BDF_2), and the explicit two-step Adams Bashforth scheme (AB_2). The ITR,

$$x_i - x_{i-1} = h_i \cdot \frac{1}{2}\big(f(t_i, x_i) + f(t_{i-1}, x_{i-1})\big), \tag{3.20}$$

results for $\alpha_{2,i} = \beta_{2,i} = 0$ and its error constant reads $c_i^{\mathrm{ITR}} = -\frac{1}{12}$. The BDF_2,

$$x_i - \frac{(\kappa_i + 1)^2}{2\kappa_i + 1} x_{i-1} + \frac{\kappa_i^2}{2\kappa_i + 1} x_{i-2} = h_i \frac{\kappa_i + 1}{2\kappa_i + 1} f(t_i, x_i), \tag{3.21}$$

follows for $\beta_{2,i} = 0$, $\alpha_{2,i} = \frac{\kappa_i^2}{2\kappa_i + 1}$. The related error constant is given by $c_i^{\mathrm{BDF}_2} = -\frac{(\kappa_i+1)^2}{6\kappa_i(2\kappa_i+1)}$. We obtain AB_2,

$$x_i - x_{i-1} = h_i \cdot \Big(\Big(\frac{\kappa_i}{2} + 1\Big) f(t_{i-1}, x_{i-1}) - \frac{\kappa_i}{2} f(t_{i-2}, x_{i-2})\Big), \tag{3.22}$$

for $\alpha_{2,i} = 0$, $\beta_{2,i} = -\frac{\kappa_i}{2}$. The error constant is $c_i^{\mathrm{AB}_2} = \frac{1}{4\kappa_i} + \frac{1}{6}$. The size of $\alpha_{2,i} \in [-1, 1)$ may result in restrictions on the step-size ratio κ_i. In particular, in case of BDF_2 this means $\kappa_i \in (0\,,\, 1+\sqrt{2})$.

We will now apply techniques developed in Section 3.1 to estimate the local error

$$l_i = c_i \cdot h_i^3 \cdot x^{(3)}(t_i) + o(h_i^3)$$

of the linear two-step scheme (3.15), (3.19). The auxiliary scheme (3.3) is chosen to be a linear two-step scheme with the same left-hand side and coefficients $\bar\beta_{0,i}$, $\bar\beta_{1,i}$, $\bar\beta_{2,i}$ which satisfy consistency conditions (3.16) – (3.18). As in (3.19), these coefficients can be represented in terms of $\bar\alpha_{2,i} = \alpha_{2,i}$, and the remaining free parameter $\bar\beta_{2,i}$. The defect d_i resulting from the auxiliary scheme is now given by

$$\begin{aligned} d_i &= \sum_{j=0}^{2} \alpha_{j,i} x_{i-j} - h_i \cdot \sum_{j=0}^{2} \bar\beta_{j,i} f(t_{i-j}, x_{i-j}) \\ &= h_i \cdot \sum_{j=0}^{2} (\beta_{j,i} - \bar\beta_{j,i}) f(t_{i-j}, x_{i-j}) \\ &= h_i (\beta_{2,i} - \bar\beta_{2,i}) \Big(\frac{1}{\kappa_i} f(t_i, x_i) - (1 + \frac{1}{\kappa_i}) f(t_{i-1}, x_{i-1}) + f(t_{i-2}, x_{i-2}) \Big), \end{aligned}$$

and the parameter $\bar\beta_{2,i}$ is specified in such a way that the error constant of the resulting scheme satisfies

$$1 = \bar c_i - c_i = -(\beta_{2,i} - \bar\beta_{2,i}) \frac{\kappa_i + 1}{2\kappa_i^2}, \qquad \beta_{2,i} - \bar\beta_{2,i} = -\frac{2\kappa_i^2}{\kappa_i + 1}.$$

Thus, the following representation for the defect d_i and the error estimate ℓ_i follows:

$$\begin{aligned} d_i &= h_i \cdot \Big(\frac{2\kappa_i}{\kappa_i + 1} f(t_i, x_i) - 2\kappa_i f(t_{i-1}, x_{i-1}) + \frac{2\kappa_i^2}{\kappa_i + 1} f(t_{i-2}, x_{i-2}) \Big), \\ \ell_i &= c_i d_i \,. \end{aligned} \tag{3.23}$$

Note that d_i in (3.23) coincides with $h_i^3\, q_f''(t)$, where $q_f(t)$ is the quadratic polynomial that interpolates the values of $f(t_i, x_i)$, $f(t_{i-1}, x_{i-1})$, $f(t_{i-2}, x_{i-2})$. This is due to the fact, that $d_i^\star$ has to approximate the term

$$h_i^3\, x^{(3)}(t_i) = h_i^3\, \frac{d^2}{dt^2} f(t, x(t))(t_i)$$

with an accuracy of $o(h_i^3)$ by using only the three corresponding values of f.

REMARK 3.1. Throughout this paper we are interested in providing an asymptotically correct estimate for the local truncation error l_i of the basic scheme (3.1). The error estimates derived up to now, rely on the assumption that the leading term in

$$l_i = c_i\ h_i^{p+1}\ x^{(p+1)}(t_i) + o(h_i^{p+1})$$

does not vanish and therefore the asymptotic behavior of l_i does not change. Unfortunately, at least in case of oscillatory solutions, there always exist time

points $\hat{t}$ where the derivative $x^{(p+1)}(\hat{t})$ vanishes[4]. In the vicinity of such points our error estimates will tend to underestimate the true size of the error, often leading to incorrect step-size prediction and step rejections afterwards.

In order to remedy this situation we assume more smoothness and take the next higher derivative into consideration. Let us assume that $x \in C^{p+2}[t_0, t_{end}]$ and that we have the following representation of the local truncation error of a pth order method (3.1):

$$l_i = c_i^{[p+1]} h_i^{p+1} x^{(p+1)}(t_i) + c_i^{[p+2]} h_i^{p+2} x^{(p+2)}(t_i) + o(h_i^{p+2}).$$

¿From now on, we specify the auxiliary scheme (3.3) in such a way that the relation $\bar{c}_i^{[p+1]} - c_i^{[p+1]} = 1$ holds. Using Lemma 2.1 we conclude

$$\begin{aligned} d_i^\star &= \bar{l}_i - l_i + O(h_i\ l_i) \\ &= h_i^{p+1} x^{(p+1)}(t_i) + \underbrace{\left(\bar{c}_i^{[p+2]} - c_i^{[p+2]}\right)}_{\gamma_i} h_i^{p+2} x^{(p+2)}(t_i) + O(h_i\ l_i) + o(h_i^{p+2}), \end{aligned}$$

$$d_{i-1}^\star = \frac{h_i^{p+1}}{\kappa_i^{p+1}} x^{(p+1)}(t_{i-1}) + \left(\bar{c}_{i-1}^{[p+2]} - c_{i-1}^{[p+2]}\right) \frac{h_i^{p+2}}{\kappa_i^{p+2}} x^{(p+2)}(t_{i-1}) + O(h_i l_{i-1}) + o(h_i^{p+2}),$$

and therefore

$$\begin{aligned} d_i^\star - \kappa_i^{p+1} d_{i-1}^\star &= h_i^{p+1}(x^{(p+1)}(t_i) - x^{(p+1)}(t_{i-1})) \\ &\quad + h_i^{p+2}\left(\gamma_i x^{(p+2)}(t_i) - \frac{\gamma_{i-1}}{\kappa_i} x^{(p+2)}(t_{i-1})\right) + O(h_i\ l_i) + o(h_i^{p+2}), \end{aligned}$$

or equivalently

$$\begin{aligned} (3.24)\quad \frac{d_i^\star - \kappa_i^{p+1} d_{i-1}^\star}{h_i^{p+2}} &= \frac{(x^{(p+1)}(t_i) - x^{(p+1)}(t_{i-1}))}{h_i} \\ &\quad + \left(\gamma_i x^{(p+2)}(t_i) - \frac{\gamma_{i-1}}{\kappa_i} x^{(p+2)}(t_{i-1})\right) + O(\ l_i/h_i^{p+1}) + o(1). \end{aligned}$$

It is clear that $d_i^\star / h_i^{p+1}$ approximates $x^{(p+1)}(t_i)$ with order of accuracy $O(h_i)$. On the other hand the term $(d_i^\star - \kappa_i^{p+1} d_{i-1}^\star)/h_i^{p+2}$ is a reasonable approximation for $x^{(p+2)}(t_i)$ only under certain conditions. First of all l_i has to behave at least as $o(h_i^{p+1})$ which is true when the derivative $x^{(p+1)}(t_i)$ is nearly zero. Secondly, the term $\left(\gamma_i x^{(p+2)}(t_i) - \frac{\gamma_{i-1}}{\kappa_i} x^{(p+2)}(t_{i-1})\right)$ has to be appropriately small. This is guaranteed in case when the last p steps have been executed with constant step-size.

Motivated by the above arguments we propose to extend the error estimate ℓ_i to ℓ_i^{ext} by a heuristic term that only comes into play in the vicinity of time points $\hat{t}$ where componentwise (for $\nu = 1, \ldots, n$) $x_\nu^{(p+1)}(\hat{t}) = 0$,

$$(3.25)\qquad \ell_{i,\nu}^{ext} := \begin{cases} c_i^{[p+1]} d_{i,\nu} \ \text{ if } \ c_i^{[p+1]} d_{i,\nu} > c_i^{[p+2]}(d_{i,\nu} - \kappa_i^{p+1} d_{i-1,\nu}), \\ c_i^{[p+1]} d_{i,\nu} + c_i^{[p+2]}(d_{i,\nu} - \kappa_i^{p+1}\, d_{i-1,\nu}) \ \text{ else.} \end{cases}$$

[4]For example the third derivative vanishes at points where the curvature of the solution becomes extremal.

Even if the term $(d_i - \kappa_i^{p+1} d_{i-1})$ arising in (3.25) does not approximate the value of $h_i^{[p+2]} x^{(p+2)}(t_i)$, it prevents the error estimate from almost vanishing and consequently, stops the overgrowth of the predicted step-size. The above extension has been implemented in our code and it proved to work very dependably in practice.

4 Index 1 DAEs.

In this section we discuss how the ideas of the previous sections can be applied to DAEs of the form

$$A x'(t) - f(t, x(t)) = 0, \quad t \in \mathcal{J}, \tag{4.1}$$

where A is a constant singular $n \times n$ matrix. Due to the singularity of the matrix A, the system (4.1) involves constraints. The solution components lying in $\ker A$, called algebraic components, are never subject to differentiation and the inherent dynamics of the system live only in a lower dimensional subspace. DAEs are usually classified by their index. The literature on DAEs contains a number of different definitions of this term pointing to different properties of the considered DAEs. Fortunately, they widely coincide in characterizing the special type of DAEs (4.1) to be of index 1. We assume here that the DAE (4.1) has index 1 in the sense that the constraints are locally solvable for the algebraic variables. Then the DAE (4.1) involves a coupling of a nonlinear equation solving task and an integration task.

To be more precise we will distinguish the differential and algebraic solution components as well as the constraints by means of projectors[5]

$$Q \text{ onto } \ker A, \quad P := I - Q \text{ along } \ker A, \quad R \text{ along } \operatorname{im} A.$$

We split the solution into differential and algebraic components,

$$x = Px + Qx =: u + v, \; x \in \mathbb{R}^n, \; u \in \operatorname{im} P, \; v \in \operatorname{im} Q\,.$$

In a correct formulation of the problem the differential operator should be applied only to the components Px. This is realized by writing $A(Px)'$ instead of Ax' and searching for solutions in the space of functions $C^1_{\ker A} := \{x \in C(\mathcal{J}, \mathbb{R}^n) : Px \in C^1(\mathcal{J}, \mathbb{R}^n)\}$, which is independent of the special choice of the projector P (see e.g. [13]). In this setting, we deal with a DAE with properly stated leading term in the sense of [18], where the matrices A and P are well-matched.

By applying the projectors $(I-R)$ and R, we split the original system (4.1) into a set of differential equations and constraints:

$$A(Px)'(t) - (I-R) f(t, x(t)) = 0 \tag{4.2}$$

$$R f(t, x(t)) = 0. \tag{4.3}$$

[5] Any matrix Q is a projector iff $Q^2 = Q$. It projects onto its image and along its kernel.

Due to the index-1 assumption one can *theoretically* solve the constraints (4.3) for the algebraic components $Qx = v$,

$$Rf(t, u + v) = 0,\ Av = 0 \iff v = \hat{v}(t, u),$$

and insert v into the system (4.2). Finally, the system is scaled by a reflexive generalized inverse A^- with $AA^- = I - R$, $A^-A = P$, or equivalently, by some non-singular matrix H with $HA = P$. This yields a so-called *inherent regular ODE* for the differential components u,

$$u' - A^- f(t, u + \hat{v}(t, u)) = 0. \tag{4.4}$$

It can be shown that im P is an invariant subspace of (4.4), and that (4.4) together with $x(t) = u(t) + \hat{v}(t, u(t))$, is equivalent to (4.1). In contrast to this analytical decoupling, numerical schemes for DAEs should be directly applicable to the original problem (4.1). We refer to the monographs [2, 8, 13, 15, 17] or to the review papers [20, 21, 22] for a detailed analysis of DAEs and their numerics.

4.1 Linear multi-step schemes.

The straightforward generalization of linear multi-step schemes (3.1) to DAEs (4.1) results in

$$A \sum_{j=0}^{k} \alpha_{j,i} x_{i-j} - h_i \sum_{j=0}^{k} \beta_{j,i} f(t_{i-j}, x_{i-j}) = 0, \quad i = k, \dots N. \tag{4.5}$$

The above equations contain the following constraints

$$\sum_{j=0}^{k} \beta_{j,i} Rf(t_{i-j}, x_{i-j}) = 0, \quad i = k, \dots N, \tag{4.6}$$

that result by applying projector R. They represent a recursion in $Rf(t_i, x_i),\ i = k, \dots N$. For consistent initial values (i.e. $Rf(t_i, x_i) = 0,\ i = 0, \dots k-1$) and implicit methods , i.e. $\beta_{0,i} \neq 0$, it follows immediately that $Rf(t_i, x_i) = 0,\ i = k, \dots N$. This means that the exactly computed iterates x_i satisfy the constraints $Rf(t_i, x_i) = 0$. However, already small perturbations in the initial values or in the right-hand sides of (4.6) would cause a disastrous error amplification if the recursion (4.6) was not stable. The stability of (4.6) is therefore necessary for a well-posed discretized problem. Forcing the iterates to satisfy the constraints is the key issue that guarantees that a *theoretical* decoupling of the discrete scheme (4.8) leads to the same result as the corresponding discretization of the inherent regular ODE (4.4),

$$\sum_{j=0}^{k} \alpha_{j,i} u_{i-j} - h_i \sum_{j=0}^{k} \beta_{j,i} A^- f(t_{i-j}, u_{i-j} + \hat{v}(t_{i-j}, u_{i-j})) = 0, \quad i = k, \dots N. \tag{4.7}$$

One of the best known methods for the integration of DAEs is the BDF, which, applied to the DAE (4.1), takes the form

$$A\sum_{j=0}^{k}\alpha_{j,i}x_{i-j} - h_i\beta_{0,i}f(t_i,x_i) = 0, \quad i = k,\dots N. \tag{4.8}$$

This scheme involves the constraint $Rf(t_i,x_i)=0$ that replaces recursion (4.6). It guarantees consistent iterates x_i even if the initial values were inconsistent.

Other linear multi-step schemes may need to be realized in a modified way to guarantee a numerically stable formulation. To this end, more structural information has to be exploited. One option is to use different discretizations of the differential and constraint part of the DAE (4.1). For the case of explicitly given constraints, i.e. $A = \begin{pmatrix} A_1 \\ 0 \end{pmatrix}$ and $f = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$, where A_1 has the full row rank, this can be done via

$$\begin{aligned} A_1\sum_{j=0}^{k}\alpha_{j,i}x_{i-j} - h_i\sum_{j=0}^{k}\beta_{j,i}f_1(t_{i-j},x_{i-j}) &= 0\,, \\ f_2(t_i,x_i) &= 0\,. \end{aligned}$$

For general DAEs (4.1) a related stable scheme can be formulated using R,

$$A\sum_{j=0}^{k}\alpha_{j,i}x_{i-j} - h_i\sum_{j=0}^{k}\beta_{j,i}(I-R)f(t_{i-j},x_{i-j}) + Rf(t_i,x_i) = 0. \tag{4.9}$$

Note, that the solution of (4.9) also satisfies (4.5).

Another possibility is to use the projector P and to consider the scheme, see [13],

$$\begin{aligned} P\big(\sum_{j=0}^{k}\alpha_{j,i}x_{i-j} - h_i\sum_{j=0}^{k}\beta_{j,i}y_{i-j}\big) + Qy_i &= 0, \\ Ay_i - f(t_i,x_i) &= 0. \end{aligned} \tag{4.10}$$

For implicit methods this can be equivalently written as

$$\begin{aligned} A\frac{1}{\beta_{0,i}}\big(\sum_{j=0}^{k}\frac{1}{h_i}\alpha_{j,i}x_{i-j} - \sum_{j=1}^{k}\beta_{j,i}y_{i-j}\big) - f(t_i,x_i) &= 0, \\ P\frac{1}{\beta_{0,i}}\big(\sum_{j=0}^{k}\frac{1}{h_i}\alpha_{j,i}x_{i-j} - \sum_{j=1}^{k}\beta_{j,i}y_{i-j}\big) &= y_i\,. \end{aligned} \tag{4.11}$$

Again, note that the solution of (4.10) or(4.11) also satisfies (4.5).

The local truncation error l_i, defined as before by substituting the values of the exact solution into the scheme (4.5), is now given by

$$l_i := A\sum_{j=0}^{k}\alpha_{j,i}x(t_{i-j}) - h_i\sum_{j=0}^{k}\beta_{j,i}f(t_{i-j},x(t_{i-j})),\quad i=k,\dots N, \tag{4.12}$$

and satisfies the relation

$$\begin{aligned} l_i &= A(x(t_i)-x_i^\star) - h_i\beta_{0,i}\big(f(t_i,x(t_i)) - f(t_i,x_i^\star)\big)\\ &= \big(A - h_i\beta_{0,i}J_i\big)(x(t_i)-x_i^\star), \end{aligned} \tag{4.13}$$

where, as before, $x_i^\star$ is the result of a step with exact starting values $x(t_{i-j})$, $j=1,\dots,k$, and $J_i = \int_0^1 f_x(t_i, sx_i^\star + (1-s)x(t_i))\,ds$. Let us emphasize that the constraint part of l_i always vanishes, i.e., $Rl_i = 0$, and that l_i is related to the local truncation error l_i^{inh} of the discretized inherent ODE (4.7) by $l_i^{inh} = A^- l_i$ and $Al_i^{inh} = l_i$. The local truncation error l_i^{inh} of (4.7) is independent of the scaling of the given DAE and can be represented by an asymptotic expansion

$$l_i^{inh} = c_i\,(Px)^{(p+1)}(t_i)\,h_i^{p+1} + o(h_i^{p+1}), \tag{4.14}$$

provided that the applied linear multi-step scheme is of order p and that $Px \in C^{p+1}$. The local truncation error l_i defined by (4.12) depends on the scaling of the DAE (4.1). Instead of (3.12) or (4.14) we have

$$l_i = c_i\,(Ax)^{(p+1)}(t_i)\,h_i^{p+1} + o(h_i^{p+1}) = c_i\,A(Px)^{(p+1)}(t_i)\,h_i^{p+1} + o(h_i^{p+1}), \tag{4.15}$$

provided that $Ax \in C^{p+1}$, or equivalently, $Px \in C^{p+1}$. Approximations to l_i^{inh} will estimate the local error in the dynamic solution components $Px(t_i)$, while an approximation to l_i will approximate the local error of $Ax(t_i)$. An approximation to the local error in the complete solution vector $x(t_i)$ can be defined via the identity $x(t_i) - x_i^\star = \big(A - h_i\beta_{0,i}J_i\big)^{-1} l_i$, see also [27]. The matrix $\big(A - h_i\beta_{0,i}J_i\big)$ is the Jacobian of (4.5) and it is nonsingular for sufficiently small step-sizes h_i, cf. [13].

As before in context of ODEs, we use a second linear multi-step method with coefficients $\bar\alpha_{j,i}$, $\bar\beta_{j,i}$, and analyze the defect d_i of the computed iterates with respect to this second scheme.

LEMMA 4.1. *Let the DAE (4.1) be of index 1 and let $f(t,x)$ be continuous and continuously differentiable with respect to x. Let the step-size* $\mathbf{h}$ *be sufficiently small to guarantee that the matrix $\big(A - h_i\beta_{0,i}J_i\big)$ is nonsingular. Then the defect $d_i^\star$ defined by*

$$d_i^\star := A(x_i^\star + \sum_{j=1}^{\bar k}\bar\alpha_{j,i}x(t_{i-j})) - h_i\bar\beta_{0,i}f(t_i,x_i^\star) - h_i\sum_{j=1}^{\bar k}\bar\beta_{j,i}f(t_{i-j},x(t_{i-j})), \tag{4.16}$$

satisfies

$$d_i^\star = \bar l_i - l_i - h_i(\bar\beta_{0,i} - \beta_{0,i})J_i\big(A - h_i\beta_{0,i}J_i\big)^{-1}l_i. \tag{4.17}$$

The proof is fully analogous to that of Lemma 3.1.

Since the DAE (4.1) is of index 1, $(A - \beta h_i J_i)^{-1}(I - R) = O(1)$ holds and hence an analogue version of Corollary 3.2 applies.

COROLLARY 4.2. *Let the suppositions of Lemma 4.1 be satisfied. Let the schemes (3.1) and (3.3) be consistent of order p and $\bar{p}$, respectively. For the case $p = \bar{p}$, we additionally assume that*

$$l_i = c_i\,(Ax)^{(p+1)}(t_i)\,h_i^{p+1} + o(h_i^{p+1}), \quad \bar{l}_i = \bar{c}_i\,(Ax)^{(p+1)}(t_i)\,h_i^{p+1} + o(h_i^{p+1}),$$

with constants $c_i \neq \bar{c}_i$, which depend only on the ratios of step-sizes.

Then we have

(i) $\bar{l}_i = d_i^\star + O(h_i^{\bar{p}+2})$, *if* $p > \bar{p}$,

(ii) $l_i = -d_i^\star + O(h_i^{p+2})$, *if* $p < \bar{p}$,

(iii) $l_i = \dfrac{c_i}{\bar{c}_i - c_i} d_i^\star + o(h_i^{p+1}), \quad \bar{l}_i = \dfrac{\bar{c}_i}{\bar{c}_i - c_i} d_i^\star + o(h_i^{p+1})$, *if* $p = \bar{p}$.

The above corollary enables us to proceed as in the ODE case and use

$$d_i := A\Big(\sum_{j=0}^{\bar{k}} \bar{\alpha}_{j,i} x(t_{i-j})\Big) - h_i \sum_{j=0}^{\bar{k}} \bar{\beta}_{j,i} f(t_{i-j}, x(t_{i-j})) \tag{4.18}$$

to derive an estimate ℓ_i of the local error l_i of (4.5). Again, we choose an auxiliary scheme with $\bar{\alpha}_{j,i} = \alpha_{j,i}$ and $\bar{c}_i - c_i = 1$. With these settings the representations for the defects d_i and the estimates of the local error ℓ_i remain unchanged.

Recall that ℓ_i approximates the local error in Ax now. Depending on the available information we can monitor different quantities to satisfy accuracy requirements:

i) control $e_i := (A - \beta_{0,i} J_i)^{-1}\ell_i := (A - \beta_{0,i} J_i)^{-1} c_i d_i$ to match a given tolerance for x,

iia) control $e_i := \ell_i := c_i d_i$ to match a given tolerance for Ax, or

iib) control $e_i := A^- \ell_i := A^- c_i d_i$ to match a given tolerance for Px.

5 Step-size control.

Here we give the algorithmic details of a step-size control that is based on the results developed in the previous sections. We exemplify this for the important subclass of second order schemes, in particular for the implicit trapezoidal rule (ITR, $k := 1$) and the two-step backward differentiation formula (BDF_2, $k := 2$).

5.1 Initialization.

Since the initial value problem itself does not supply enough information to start a multi-step scheme any practical realization of such a scheme needs to address the problem of the necessary initialization. In the literature, see e.g. [8], [25], several more or less heuristic strategies to start the integration have been proposed. The first step always has to be computed by means of a one-step scheme. In the context of a variable step-size, variable order implementation of the BDF_2 the first step is carried out using the implicit Euler scheme, where the step-size has to be chosen in such a way that the estimated local error meets the prescribed tolerance.

For the numerical experiments in Section 6 we performed the first step by means of the trapezoidal rule, which is the only linear one-step scheme of order 2, but while estimating its local error we faced a problem: The formula (3.23) providing such estimate requires the knowledge of two preceding values of the right-hand side. However, with the iterate x_1 obtained from the ITR-step with the step-size h_1 only one preceding value of the right-hand side is available. Therefore, we used the estimated local error of the implicit (or explicit) Euler scheme,

$$\ell_1^{\mathrm{BEUL}} = -\frac{h_1}{2}\Big(f(t_1, x_1) - f(t_0, x_0)\Big),$$

to control the step-size. Since now an estimate which is correct for a first order scheme is used in context of a second order method, the first step may become unnecessarily small. Alternatively, one may perform two steps of the ITR with step-sizes h_1 and $h_2 = h_1$ and estimate the local error of the ITR step by

$$\ell_2^{\mathrm{ITR}} = \frac{h_1}{12}\Big(f(t_2, x_2) - 2f(t_1, x_1) + f(t_0, x_0)\Big).$$

Finally, both steps are accepted if this estimate satisfies the tolerance, or else they are rejected and repeated with a smaller step-size predicted from (5.6).

After the first step has been accepted and a step-size for the next step has been fixed the following algorithm can be carried out for any second order one-step or two-step method. The algorithm is described in terms of the DAE problem (4.1), but comprises also the ODE case by setting $A := I$.

5.2 Step-size control algorithm.

Let two initial values x_0, x_1 at time points t_0, $t_1 = t_0 + h_1$, an absolute and a relative tolerance ATOL, RTOL, and the step-size h_2 be given. Set $i := 2$.

1) Solve the system

$$A\sum_{j=0}^{2}\alpha_{j,i}x_{i-j} = h_i\sum_{j=0}^{2}\beta_{j,i}f(t_{i-j}, x_{i-j}), \tag{5.1}$$

or (4.9) or (4.11) for x_i, where the parameters $\alpha_{j,i}$ and $\beta_{j,i}$ are chosen to provide a second order two-step scheme (including the ITR with $\alpha_{2,i} = \beta_{2,i} = 0$).

2) Compute

$$d_i = h_i \cdot \Big(\frac{2\kappa_i}{\kappa_i+1} f(t_i, x_i) - 2\kappa_i f(t_{i-1}, x_{i-1}) + \frac{2\kappa_i^2}{\kappa_i+1} f(t_{i-2}, x_{i-2}) \Big),$$

and componentwise (for $\nu = 1, \dots, n$)

$$\ell_{i,\nu}^{ext} := \begin{cases} c_i^{[3]} d_{i,\nu} \text{ if } c_i^{[3]} d_{i,\nu} > c_i^{[4]} (d_{i,\nu} - \kappa_i^3 d_{i-1,\nu}), \\ c_i^{[3]} d_{i,\nu} + c_i^{[4]} (d_{i,\nu} - \kappa_i^3\, d_{i-1,\nu}) \text{ else.} \end{cases} \tag{5.2}$$

Depending on the problem setting and the available information define

$$\text{a) } e_i := (A - \beta_{0,i} J_i)^{-1} \ell_i^{ext} \text{ and } \hat{x} := x, \tag{5.3}$$

$$\text{b) } e_i := \ell_i^{ext} \text{ and } \hat{x} := Ax, \tag{5.4}$$

or, for DAEs only,

$$\text{c) } e_i := A^- \ell_i^{ext} \text{ and } \hat{x} := Px. \tag{5.5}$$

Compute componentwise (for $\nu = 1, \dots, n$)

$$\text{TOL}_\nu := \text{ATOL} + \text{RTOL} \cdot |\hat{x}_{i,\nu}|.$$

3) Apply a control strategy, c.f. [14], [28], [29], [32], predicting the new step-size h_{new} to match the tolerance multiplied by a safety factor θ, say $\theta = 0.7$. For example, apply the elementary control (eC),

$$\frac{h_{new}}{h_i} := \min_{\nu=1,\dots,n} \left(\frac{\theta \cdot \text{TOL}_\nu}{|e_{i,\nu}|} \right)^{\frac{1}{p+1}}, \tag{5.6}$$

the proportional integral control PI34 [14],

$$\begin{aligned} \frac{h_{new}}{h_i} &:= \min_{\nu=1,\dots,n} \left\{ \left(\frac{\theta \cdot \text{TOL}_\nu}{|e_{i,\nu}|} \right)^{\frac{0.3}{p+1}} \left(\frac{|e_{i-1,\nu}|}{|e_{i,\nu}|} \right)^{\frac{0.4}{p+1}} \right\} \\ &= \min_{\nu=1,\dots,n} \left\{ \left(\frac{\theta \cdot \text{TOL}_\nu}{|e_{i,\nu}|} \right)^{\frac{0.7}{p+1}} \left(\frac{\theta \cdot \text{TOL}_\nu}{|e_{i-1,\nu}|} \right)^{\frac{-0.4}{p+1}} \right\}. \end{aligned} \tag{5.7}$$

or the digital filter $H211b$ [28],

$$\frac{h_{new}}{h_i} := \min_{\nu=1,\dots,n} \left\{ \left(\frac{\theta \cdot \text{TOL}_\nu}{|e_{i,\nu}|} \right)^{\frac{1}{b(p+1)}} \left(\frac{\theta \cdot \text{TOL}_\nu}{|e_{i-1,\nu}|} \right)^{\frac{1}{b(p+1)}} \left(\frac{h_{i-1}}{h_i} \right)^{\frac{1}{b}} \right\}, \tag{5.8}$$

in each case with $p := 2$ (the order of the scheme). After step rejections the control should be restarted by using the elementary control.

4) If $|e_{i,\nu}| \le \text{TOL}_\nu$ for all $\nu = 1, \dots, n$, then accept the step. If $t_i > T$ then stop, else set $i := i + 1$, $h_i := h_{new}$ and go to 1.
If $|e_{i,\nu}| > \text{TOL}_\nu$ for at least one component $\nu \in \{1, \dots, n\}$, then reject the step and repeat it with the smaller step-size, i.e. set $h_i := h_{new}$ and go to 1.

6 Numerical experiments.

The strategies discussed in the previous sections have been implemented for the ITR and the BDF_2 and tested extensively on a set of ODEs and DAEs. By means of three model examples we now illustrate how the procedure performed. We start with a simple test problem where the exact solution is known and hence we can access the exact errors. We compare the results of the above algorithm with those where the extended local error estimate (5.2) has been replaced by the simpler formula $\ell_{i,\nu} := c_i^{[3]} d_{i,\nu}$, see Remark 3.1. In both cases the elementary control (5.6) with $\theta = 0.7$ is used. Moreover, we report on results from runs with two additional control strategies (5.7) and (5.8) which can be implemented by changing only a few lines of the code. Our second example is the so-called "Brusselator", a two dimensional nonlinear system exhibiting periodic solutions. The third test problem is a low-dimensional electronic circuit model. In all examples the scaling of the local error estimates is set according to (5.3).

EXAMPLE 6.1. Consider the scalar initial value problem

$$x'(t) = \lambda(x(t) - g(t)) + g'(t), \quad x(0) = g(0), \quad t \in [0, 10], \tag{6.1}$$

where $g(t) = \sin(t)$ and $\lambda = -100$. Its solution g is displayed in Figure 6.1.

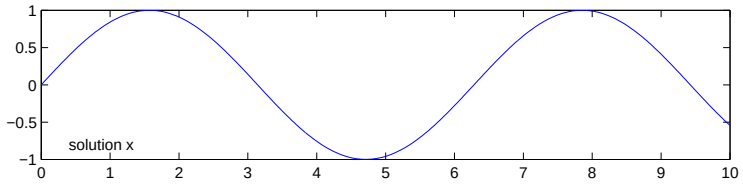

Figure 6.1: Solution $x(t) = g(t) = \sin(t)$ of (6.1).

Figure 6.2 gives simulation results for the BDF_2 computed without (top) and with (bottom) the extension (3.1). The tolerance parameters were set to ATOL = RTOL = 10^{-4}. In each case we display the accepted step-sizes (connected by a solid line), the rejected step-sizes (indicated by crosses ($\times$)), the tolerance (dotted line), the local truncation error estimates (solid line), and the true local error $x(t_i) - x_i^*$ (dashed line). The use of the extended formula is indicated by crosses ($\times$) in the dotted line for the tolerance. Similar figures of simulation results for the ITR can be found in [26].

We observe that the error estimate $(1 - \beta_{0,i}\lambda)^{-1}\ell_i = (1 - \beta_{0,i}\lambda)^{-1} c_i^{[3]} d_i$ decreases significantly when the third derivative of the solution tends to zero at $t = \pi/2 + k\pi, k = 0, 1, \ldots$. At these points the step-size becomes unreasonably small. Consequently, more rejected steps, and even twice rejected steps result for both schemes. The BDF_2 method requires generally smaller steps than the ITR due to its larger error constant. This behavior can also be observed for lower tolerances. By using the extension (5.2) the error estimate is prevented from vanishing and the predicted step-sizes are well related to the actual size

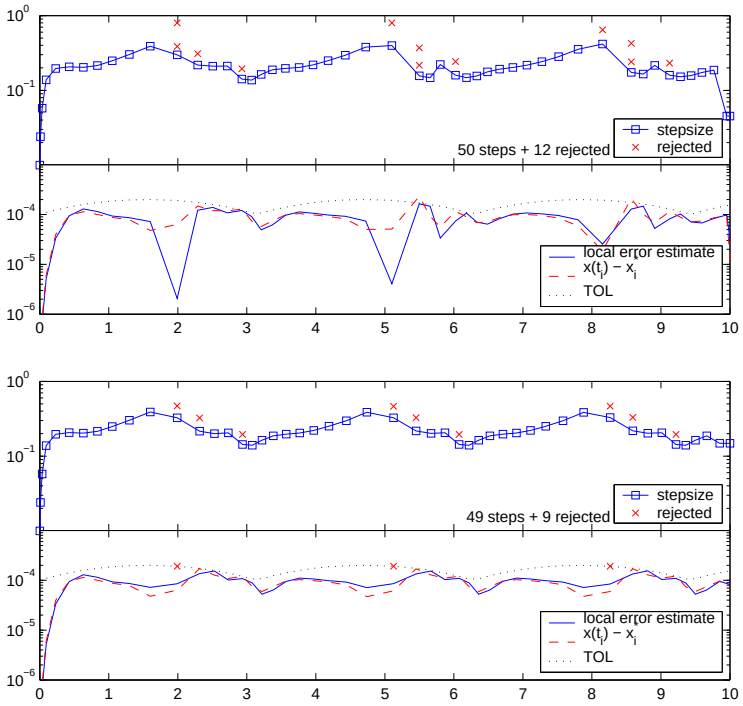

Figure 6.2: Step-size and local error estimate $(1-\beta_{0,i}\lambda)^{-1}\ell_i$ (top), and $(1-\beta_{0,i}\lambda)^{-1}\ell_i^{ext}$ (bottom), respectively, for the BDF_2 .

of the local error. The unnecessary step rejections are avoided. In all following experiments we use the extension (5.2).

Next, we compare the performance of the code for three different control strategies specified by (5.6), (5.7), and (5.8) with $b = 6$, which have been implemented for the ITR and the BDF_2. The results displayed in Table 6.1 have been computed for the tolerance parameters set to ATOL = RTOL = 10^{-5}. The numbers reflecting the performance of the control strategies are the sum of accepted and rejected steps, the ratio of rejected to accepted steps, and the maximal global error. For this simple problem whose solution is very smooth, we do not observe any significant differences. Both, for the ITR and for the BDF_2 the elementary control uses fewer steps at the price of larger global errors. The proportional integral control (5.7) and the digital filter (5.8) with $b = 6$, need nearly the same number of steps. In the following test runs we use the proportional integral control (5.7) with $\theta = 0.7$.

EXAMPLE 6.2. We now consider a two-dimensional system called Brusselator, cf. [16], a mathematical model for a certain chemical reaction,

$$x_1'(t) = 1 + x_1^2(t)x_2(t) - 4x_1(t),$$
$$x_2'(t) = 3x_1(t) - x_1^2(t)x_2(t),$$

Table 6.1: # steps: accepted + rejected and global errors

Controller	ITR		BDF_2	
	# steps	global error	# steps	global error
eC	96+13=109	$2{,}23\times10^{-5}$	132+10=142	$2{,}83\times10^{-5}$
PI34	106+12=118	$1{,}33\times10^{-5}$	139+9=148	$2{,}13\times10^{-5}$
$H211b, b=6$	101+15=116	$1{,}55\times10^{-5}$	136+11=147	$2{,}18\times10^{-5}$

with initial values $x_1(0) = 1.5,\ x_2(0) = 3$ and $t \in [0, 12]$. The solution components are plotted in Figure 6.3. We have executed the described algorithm with

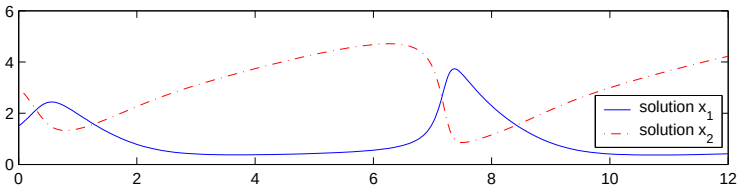

Figure 6.3: Solution components for the Brusselator.

the local error estimate (5.3), the control (5.7), and different values for the tolerance for both the ITR and the BDF_2. In Figure 6.4, we display the step-sizes, the error estimate, and the tolerance for the BDF_2. The results for the ITR can be found in [26].

As one would expect, the step-size decreases significantly in regions where the solution changes more rapidly. Many step rejections are observed when the step-size has to be significantly reduced. It is not easy to prevent this behavior, because the step size proposed by formula (5.6) is, apart from the safety factor θ, increased after an accepted step. A more pessimistic choice of the safety factor θ can help to prevent these step rejections, but it enhances the overall number of steps. The ratio of rejected to accepted steps becomes smaller with smaller tolerances.

Another option is to set $\theta = 1$, but allow errors in individual steps to be slightly larger than the tolerance, i.e., satisfy $|e_{i,\nu}| \le \hat{\theta} \cdot \mathrm{TOL}_\nu$ with a factor $\hat{\theta} > 1$, say $\hat{\theta} = 1.5$. This approach corresponds to the algorithm proposed in Section 5.2 with the modified data, $\mathrm{TOL} := \hat{\theta}^{-1}\mathrm{TOL}$ and $\theta := \hat{\theta}^{-1}$.

EXAMPLE 6.3. As an example for a system of DAEs we consider the model of a resistor-capacitor (RC) generator proposed in [34]. It can be used to trigger an electric oscillation by varying the capacities. The equivalent circuit diagram is given in Figure 6.5. The resonance frequency of the RC generator depends on the amplifier V, the resistances R_i $(i = 1, 2)$ and the capacities C_i $(i = 1, 2)$. By

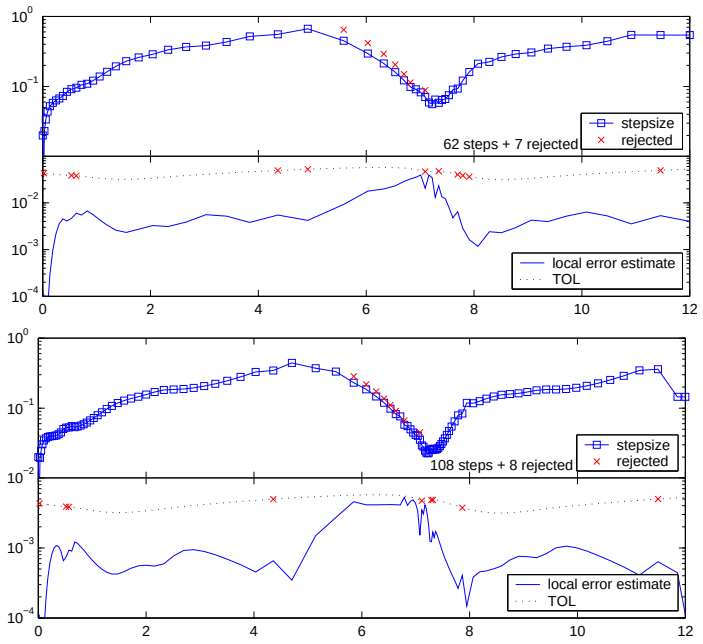

Figure 6.4: Brusselator: Step-size and local error estimate for the BDF_2, $\mathrm{RTOL} = \mathrm{ATOL} = 10^{-2}$ (top) and 10^{-3} (bottom).

Kirchhoff's Law we have

$$
\begin{array}{rrrrrcl}
C_2u_1' & & & + (G_1+G_2)u_1 & - G_1u_3 & = & 0, \\
C_1u_2' & - C_1u_3' & & + G_1u_1 & - G_1u_3 & = & 0, \\
 & & f(u_1) & - u_2 & & = & 0,
\end{array}
\tag{6.2}
$$

where u_1, u_2 and u_3 are the voltages at the corresponding nodes, see Figure 6.5, $G_i = R_i^{-1}$, $i = 1, 2$, and f is the characteristic of the amplifier V. We set

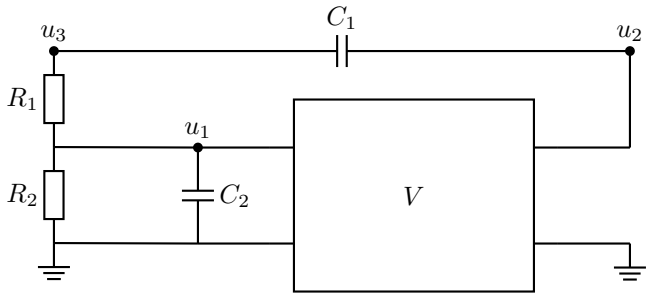

Figure 6.5: The RC generator circuit.

$f(u) = \arctan(5u)$, $C_i = 1$ [F] and $G_i = 1$ [$1/\Omega$], $(i = 1, 2)$ and obtain

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix}}_{:=A} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}' - \begin{pmatrix} -2u_1 + u_3 \\ -u_1 + u_3 \\ -\arctan(u_1) + u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{6.3}$$

Therefore

$$\ker A = \text{span}\{\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}\}, \qquad \text{im } A = \text{span}\{\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}\}.$$

In this example the matrix A is a projector itself. Hence we may choose the projectors as follows: $P = I - Q = A = I - R$. The generalized inverse A^- is given by $A^- := A$. Consistent initial values have to satisfy the constraint $u_2(0) = f(u_1(0))$. The solution for the consistent initial value $u_1(0) = 0.4$, $u_2(0) = f(u_1(0)) = \arctan(0.4)$, $u_3(0) = 0.6$ on the time-interval $\mathcal{J} = [0, 12]$ is given in Figure 6.6. Simulation results for two different values of the tolerance,

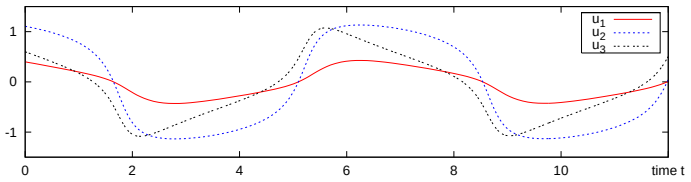

Figure 6.6: Solution components for the RC generator circuit.

ATOL = RTOL = $10^{-2}, 10^{-3}$, the local error estimate (5.3), and the control (5.7) for the BDF_2 are presented in Figure 6.7. Again, results for the ITR can be found in [26].

7 Conclusions.

In the present first part of the paper Defect Correction principle has been used to derive error estimation formulas for the discretization error of the numerical solution of ODEs and DAEs. The main focus of the work was to construct the error estimates in such a way that they require only very moderate smoothness of the analytical solution to work. We investigated how the local error of a multi-step scheme relates to a computable defect obtained by inserting the numerical solution provided by this scheme into a companion multi-step scheme of different (or the same) order. It turns out that those two quantities differ, up to the higher order terms, by a multiplicative constant. Consequently, a properly scaled defect may serve as an asymptotically correct estimate of the local error. The estimates derived here, have been implemented for second order schemes and their performance has been illustrated by means of numerical experiments. The present investigations were mainly motivated by stochastic differential equations

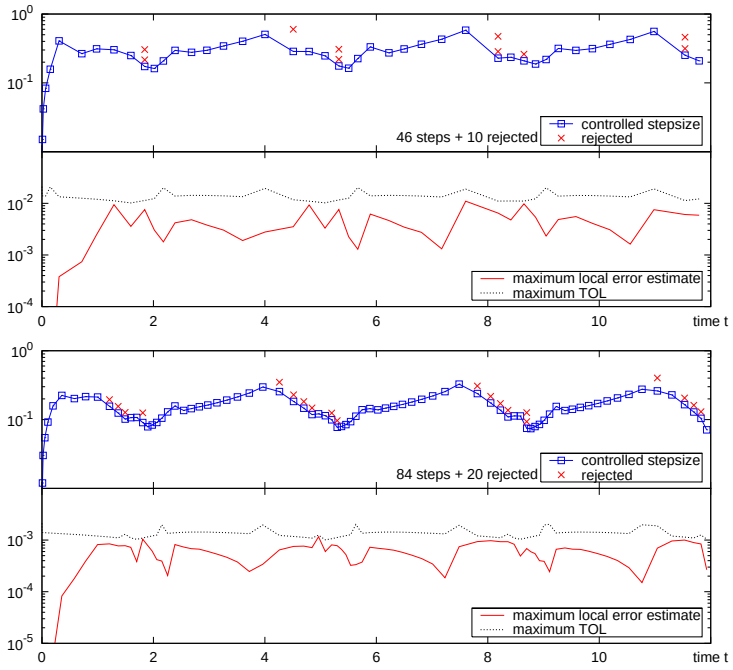

Figure 6.7: RC generator circuit: Step-size and local error estimate for the BDF_2, $\textsc{rtol} = \textsc{atol} = 10^{-2}$ (top) and 10^{-3} (bottom).

with small noise and constitute a theoretical basis for a forthcoming paper in which this class of problems will be discussed. However, the moderately smooth ODEs and DAEs are interesting in their own right and therefore the results of this paper may be also of interest in applications.

Acknowledgement.

We are indebted to Gustaf Söderlind for his valuable criticism during the revision process. His suggestions resulted in an essential improvement of the presentation of the paper.

REFERENCES

1. U. Ascher, R. M. M. Mattheij, and R.D. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1988.
2. U. Ascher and L. Petzold. *Computer methods for ordinary differential equations and differential-algebraic equations.* SIAM, Philadelphia, 1998.
3. W. Auzinger, R. Frank, F. Macsek, *Asymptotic error expansions for stiff equations: the implicit Euler scheme*, SIAM J. Numer. Anal., 27 (1990), pp. 67–104.

4. W. Auzinger, O. Koch, and E. Weinmüller, *Efficient collocation schemes for singular boundary value problems*, Numer. Algorithms 31 (2002), pp. 5–25.
5. W. Auzinger, O. Koch, and E. Weinmüller, *New variants of defect correction for boundary value problems in ordinary differential equations*, in Current Trends in Scientific Computing, Z. Chen, R. Glowinski, K. Li (eds), Publ. of AMS, Cont. Math. Series, 329 (2003), pp. 43–50.
6. W. Auzinger, O. Koch, W. Kreuzer, H. Hofstätter, and E. Weinmüller, *Superconvergent defect correction algorithms*, in WSEAS Transactions of Systems 4, Vol. 3(2004), pp. 1378-1383.
7. W. Auzinger, W. Kreuzer, H. Hofstätter, and E. Weinmüller, *Modified defect correction algorithms for ODEs. Part I: General Theory*, Numer. Algorithms 2, Vol. 36(2004), pp. 135-156.
8. K. Brenan, S. Campbell and L. Petzold, *Numerical solution of initial-value problems in differential-algebraic equations* , North-Holland, New York, 1989.
9. R. Frank, *Schätzungen des globalen Diskretisierungsfehlers bei Runge-Kutta-Methoden*, ISNM 27 (1975), pp. 45–70.
10. R. Frank, J. Hertling, and C. Überhuber, *Iterated Defect Correction Based on Estimates of the Local Discretization Error*, Technical Report No. 18 (1976), Department for Numerical Analysis, Vienna University of Technology, Austria.
11. R. Frank, J. Hertling, and C. Überhuber, *An extension of the applicabilty of iterated defect correction*, Math. of Comp. 31 (1977), pp. 907–915.
12. R. Frank, and C. Überhuber, *Iterated defect correction for differential equations, Part I: theoretical results*, Computing 20 (1978), pp. 207–228.
13. E. Griepentrog and R. März. *Differential-Algebraic Equations and Their Numerical Treatment.* Teubner-Texte Math. 88. Teubner, Leipzig, 1986.
14. K. Gustafsson, M. Lundh and G. Söderlind, *A PI stepsize control for the numerical solution of ordinary differential equations*, BIT, vol 28 (1988), pp. 270–287.
15. E. Hairer, C. Lubich, and M. Roche. *The numerical solution of differential-algebraic systems by Runge-Kutta methods.* Springer, Berlin, 1989.
16. E. Hairer, S. P. Nørsett, and G. Wanner, *Solving ordinary differential equations I, Second Edition*, Springer-Verlag, Berlin-Heidelberg-New York, 2000.
17. E. Hairer and G. Wanner. *Solving ordinary differential equations II, Stiff and differential-algebraic problems.* Springer, Berlin, second, rev. edition, 1996.
18. I. Higueras and R. März. *Differential Algebraic Equations with properly stated leading terms,* Computers and Mathematics with Applications 48 (2204), pp. 215–235.
19. H. Hofstätter and O. Koch, *Defect correction for geometric integrators*, in the Proceedings of APLIMAT 2004, pp. 465-470.
20. R. März. *Numerical methods for differential-algebraic equations,* Acta Numerica 1992, pp. 141–198.
21. R. März. *EXTRA-ordinary differential equations: Attempts to an analysis of differential-algebraic systems,* Progress in Mathematics 168 (1998), pp. 313–334.
22. L. Petzold. *Numerical solution of differential-algebraic equations,* in Theory and numerics of ordinary and partial differential equations, Oxford Univ. Press, New York, (1995), pp 123–142.

23. W. Römisch and R. Winkler. *Stepsize control for mean-square numerical methods for stochastic differential equations with small noise,* SIAM J. Sci. Comp. 28 (2006), pp. 604–625.

24. K. H. Schild, *Gaussian collocation via defect correction*, Numer. Math. 58 (1990), pp. 369–386.

25. L. F. Shampine, *Numerical solution of ordinary differential equations*, Chapman and Hall, London, 1994.

26. T. Sickenberger, E. Weinmüller, R. Winkler, *Local error estimates for moderately smooth ODEs and DAEs*, Preprint 06-1, Institut für Mathematik, Humboldt-Universität zu Berlin (2006).

27. J. Sieber, *Local error control for general index-1 and index-2 differential algebraic equations*, Preprint 97-21, Institut für Mathematik, Humboldt-Universität zu Berlin (1997).

28. G. Söderlind, *Digital Filters in Adaptive Time-Stepping*, ACM Trans. Math. Software 29 (2003), pp. 1–26.

29. G. Söderlind, *Time-step algorithms: Adaptivity, Control and Signal Processing*, Appl. Num. Math. 56 (2006), pp. 488–502.

30. H. J. Stetter, *Analysis of Discretization Methods for Ordinary Differential Equations*, Springer-Verlag, Berlin-Heidelberg-New York, 1973.

31. H. J. Stetter, *The defect correction principle and discretization methods*, Numer. Math., 29 (1978), pp. 425–443.

32. A. Verhoeven, T. G. J. Beelen, M. L. J. Hautus and E. J. W. ter Maten, *Digital linear control theory applied to automatic stepsize control in electrical circuit simulation*, in Progress in Industrial Mathematics at ECMI 2004, A. Di Bucchianico, R. M. M. Matheij, M. A. Peletier (eds), Springer (2006), pp. 198–203.

33. P. E. Zadunaisky, *On the estimation of errors propagated in the numerical integration of ODEs*, Numer. Math., 27 (1976), pp. 21–39.

34. Q. Zheng, *Ein Algorithmus zur Berechnung nichtlinearer Schwingungen bei DAEs*, Hamburger Beiträge zur Angewandten Mathematik, (1988).