

Analysis of a defect correction method for geometric integrators*

Harald Hofstätter** and Othmar Koch

*Institute for Analysis and Scientific Computing,
Vienna University of Technology, Vienna, Austria*

E-mail: hofi@aurora.anum.tuwien.ac.at; othmar@othmar-koch.org

We discuss a new variant of *Iterated Defect Correction (IDeC)*, which increases the range of applicability of the method. Splitting methods are utilized in conjunction with special integration methods for Hamiltonian systems, or other initial value problems for ordinary differential equations with a particular structure, to solve the neighboring problems occurring in the course of the IDeC iteration. We demonstrate that this acceleration technique serves to rapidly increase the convergence order of the resulting numerical approximations, up to the theoretical limit given by the order of certain superconvergent collocation methods.

Keywords: composition methods, geometric integration, Hamiltonian systems, iterated defect correction, splitting methods

AMS subject classification: 65L05

1. Introduction

In recent years, the importance of using special numerical integration schemes that reflect certain geometric properties or retain important conserved quantities of the flow of a differential equation has been widely recognized [4, 5]. Many of these methods are applicable to particular types of differential equations only. Examples of these are the Störmer/Verlet method for Hamiltonian systems and the exponential midpoint rule for homogeneous linear problems, but also higher order composition methods that we focus on in this paper. These are specified in section 3.

A cheap and efficient way to estimate the global error of a numerical method used to solve an ordinary differential equation is the defect correction principle [10, 11]. The idea can also be used to successively improve the accuracy of the numerical solution ([1, 3], and the references therein).¹ In this acceleration technique, a number

* This project was supported by the Special Research Program SFB F011 ‘AURORA’ of the Austrian Science Fund FWF.

** Corresponding author.

¹ The idea to use acceleration techniques in conjunction with composition methods in order to improve the order of accuracy of a numerical approximation is also discussed in [2].

of *neighboring problems* have to be solved, which are not necessarily of the same type as the original problem. Therefore it may happen that the neighboring problems cannot be solved by the same geometric integrator as the original problem. Thus, in [7] we proposed the method of *Splitting Defect Correction* to overcome this disadvantage. The resulting method, which we denote by *Iterated Splitting Defect Correction (ISDeC)*, is described in section 2, see also [1]. In section 4, we outline a proof of the convergence of the iteration, where we estimate the error of each respective ISDeC iterate as compared with the fixed point of the iteration in terms of the previous solution approximation, and show that the global error of the numerical approximations decreases rapidly. Crucial technical details of the proof are outlined in Appendix A, the complete proof is given in [8], since this would exceed the scope of this paper. In section 5 we give numerical examples illustrating our convergence results, where ISDeC is applied to solve the Kepler problem.

2. Iterated splitting defect correction

First, we describe the classical version of *Iterated Defect Correction (IDeC)* [3]. Consider an initial value problem in n dimensions

$$y'(t) = f(t, y(t)), \quad y(t_0) = y_0, \quad (1)$$

to be solved on the interval $[t_0, t_{\text{end}}]$. Subsequently, we assume that a sufficiently smooth solution y of the analytical problem exists on the whole interval. Moreover, we will require the existence of bounded Fréchet derivatives of f at various points throughout the convergence proof. The approximate solution $\eta^{[0]} := (\eta_0, \dots, \eta_N)$ is obtained by some discretization method Φ on a uniform grid² $\Gamma = (t_0, \dots, t_N)$, where $t_{i+1} - t_i = h$, $i = 0, \dots, N - 1$. Denote by $p^{[0]}(t)$ the polynomial of degree N interpolating the values of $\eta^{[0]}$. Using this interpolating function, called the *Zadunaisky polynomial*, we construct a neighboring problem associated with (1) whose exact solution is $p^{[0]}(t)$:

$$y'(t) = f(t, y(t)) + d^{[0]}(t), \quad y(t_0) = y_0, \quad (2)$$

where $d^{[0]}(t) := p^{[0]'}(t) - f(t, p^{[0]}(t))$. We now solve (2) using the same numerical method Φ and obtain an approximate solution $\pi^{[0]}$ for $p^{[0]}(t)$. This means that for the solution of the neighboring problem (2) we know the global error which is a good estimate for the unknown error of the original problem (1). This estimate can be used to improve the first solution,

$$\eta^{[1]} := \eta^{[0]} + \left(p^{[0]} - \pi^{[0]} \right). \quad (3)$$

Now, these values are used to define a new interpolating polynomial $p^{[1]}(t)$ by requiring $p^{[1]}(t_j) = \eta_j^{[1]}$. Again, $p^{[1]}(t)$ defines a neighboring problem in the same

²In fact, our arguments can easily be extended to piecewise equidistant grids; we will indicate the necessary changes at the appropriate places.

manner as in (2), where again the exact solution is known, and the numerical solution $\pi^{[1]}$ of this neighboring problem serves to obtain the second improved solution

$$\eta^{[2]} := \eta^{[0]} + \left(p^{[1]} - \pi^{[1]} \right).$$

This process can be continued iteratively. For obvious reasons one does not use one interpolating polynomial for the whole interval $[t_0, t_{\text{end}}]$ in practice. Instead, globally continuous piecewise functions composed of polynomials of (moderate) degree m are defined as the Zadunaisky polynomials for the specification of the neighboring problems.

In many situations, the defect correction principle yields an asymptotically correct error estimate and a successive improvement in the convergence orders of the respective iterates, up to a certain limit determined by the smoothness of the problem data and the value of m , see for example [1, 3].

If in the scheme described above the basic numerical solution method Φ is intended especially for ODEs with a particular structure, the neighboring problem (2) may have a form to which the integrator cannot be applied straightforwardly. For example, if the Störmer/Verlet method is applied to a Hamiltonian system, (2) is no longer an autonomous, separated system, cf. [7].

In order to be able to use IDeC even in such a case, we employ splitting methods, cf. [4, Sec. II.5]. To apply *Strang splitting* to (2), we split the time-dependent vector field into its components $f(t, y)$ and $d^{[0]}(t)$. We denote the numerical flow of $f(t, y)$ by $\Phi_{t,h}$, such that one step $(t, \eta_i) \mapsto (t + h, \eta_{i+1})$ with step size h of the basic scheme Φ applied to (1) can be written as $\eta_{i+1} = \Phi_{t,h}(\eta_i)$. Note that for autonomous problems (1), we can write the flow independently of t , $\eta_{i+1} = \Phi_h(\eta_i)$. The numerical flow $\Delta_{t,h}$ of the other component $d^{[0]}(t)$ is defined by the quadrature rule

$$\Delta_{t,h}(y) = y + \int_t^{t+h} D^{[0]}(\tau) d\tau, \tag{4}$$

where $D^{[0]}(t)$ is a piecewise polynomial interpolant of degree $\leq m - 1$ of $d^{[0]}(t)$.

To explain this more precisely, we require some additional notation. Choose the grid $\Gamma = (t_0, \dots, t_N)$ such that $N = mN_1$ for some integer N_1 , and denote $t_{i,j} := t_{im+j}$, $j = 0, \dots, m$, $\tau_i := t_{im}$, $i = 0, \dots, N_1 - 1$, $\tau_{N_1} := t_N$. We split the integration interval into subintervals $J_i := [\tau_i, \tau_{i+1}]$ of length $H = mh$. On the interval J_i , we define interpolation nodes

$$\sigma_{i,j} := \tau_i + H\rho_j, \quad j = 1, \dots, m, \quad 0 \leq \rho_1 < \rho_2 < \dots < \rho_m \leq 1. \tag{5}$$

The highest attainable convergence orders for the ISDeC iterates result if we use interpolation at Gaussian points in order to define $D^{[0]}(t)$. This implies that the maximal convergence order of IDeC iterates is $O(H^{2m})$, see [7] and section 4 of this paper.

Using $\Phi_{t,h}$ and $\Delta_{t,h}$ from above, the numerical solution of (2) is computed using the numerical flow

$$\Psi_{t,h} = \Delta_{t+h/2,h/2} \circ \Phi_{t,h} \circ \Delta_{t,h/2}, \quad (6)$$

where \circ denotes the *composition* of the numerical methods (which means that the result computed by one method is the starting value for the next method). We call the method where the solution of the neighboring problems is computed in this way *Iterated Splitting Defect Correction* (ISDeC).

3. High order geometric integration schemes

In the sequel, we describe high-order composition methods [4, Sec. II.4]. When these are based on low-order schemes with favorable geometric properties, as for example the Störmer/Verlet method for Hamiltonian systems, these properties are often inherited by the higher-order scheme, see for example [4, Sec. II.4] and the references therein. Let the basis for ISDeC be a composition method

$$\Phi = \Phi^{[s]} \circ \dots \circ \Phi^{[2]} \circ \Phi^{[1]}, \quad (7)$$

where $\Phi^{[j]}$ are any suitable low-order methods, see for example equation (9) below for the case of symmetric composition, or [7] and references therein for more general situations. Then, the numerical method Ψ for the solution of the neighboring problem (2) can be defined by

$$\Psi_{t,h} = \Delta_{t+\hat{\delta}_{s+1}h,\hat{\delta}_{s+1}h} \circ \Phi^{[s]} \circ \Delta_{t+\hat{\delta}_s h,\hat{\delta}_s h} \circ \dots \circ \Delta_{t+\hat{\delta}_2 h,\hat{\delta}_2 h} \circ \Phi^{[1]} \circ \Delta_{t+\hat{\delta}_1 h,\hat{\delta}_1 h}, \quad (8)$$

where $\delta_1, \dots, \delta_{s+1}$ are given real numbers which satisfy

$$\delta_1 + \dots + \delta_{s+1} = 1, \quad \hat{\delta}_j := \sum_{i=1}^{j-1} \delta_i,$$

and $\Delta_{t,h}$ is given by (4).

To illustrate the composition methods we consider (cf. [7]), we describe Yoshida's method as an example for symmetric composition of symmetric methods [4, Sec. V.3]. Our arguments also apply in more general situations, see the results for Suzuki's method or McLachlan's method (which is an example for symmetric composition of first order methods) given in [7]. The essential properties of the basic method Φ we use in our convergence proof are the basic convergence order q , existence of a modified differential equation and asymptotic error expansion, see section 4, and the suitable definition of the method Ψ to guarantee that this auxiliary method has the same order as Φ .

For symmetric composition we choose

$$\Phi^{[j]} = \phi_{t+\hat{\gamma}_j h, \gamma_j h}, \quad j = 1, \dots, s, \quad (9)$$

where ϕ is a symmetric second-order method, the coefficients $\gamma_s = \gamma_1, \gamma_{s-1} = \gamma_2, \dots$ are symmetric, and $\hat{\gamma}_j := \sum_{i=1}^{j-1} \gamma_i$. Examples of possible choices for ϕ are the Störmer/Verlet scheme, the implicit midpoint rule, the implicit trapezoidal rule, or the exponential midpoint rule [7]. For Yoshida's method, the coefficients γ_j are chosen as

$$s = 3, \quad \gamma_1 = \gamma_3 = 1/(2 - 2^{1/3}), \quad \gamma_2 = -2^{1/3}/(2 - 2^{1/3}). \quad (10)$$

This yields a method of order four, cf. [4, Sec. II.4]. A natural choice for the parameters δ_j in the splitting (8) is

$$\delta_1 = \gamma_1/2, \quad \delta_j = (\gamma_{j-1} + \gamma_j)/2, \quad j = 2, \dots, s, \quad \delta_{s+1} = \gamma_s/2, \quad (11)$$

since then Ψ can be written as

$$\Psi = \Psi^{[s]} \circ \dots \circ \Psi^{[1]}, \quad (12)$$

where

$$\Psi^{[j]} = \Delta_{t+\hat{\gamma}_j h+\gamma_j h/2, \gamma_j h/2} \circ \phi_{t+\hat{\gamma}_j h, \gamma_j h} \circ \Delta_{t+\hat{\gamma}_j h, \gamma_j h/2}. \quad (13)$$

Consequently, $\Psi^{[j]}$ is a symmetric second-order method, and Ψ is constructed from $\Psi^{[j]}$ by the same composition scheme as Φ . Hence, Ψ has the same order as Φ .

4. Fixed point convergence

We confine our analysis to autonomous problems

$$y'(t) = f(y(t)), \quad y(t_0) = y_0 \quad (14)$$

without restriction of generality. The neighboring problems

$$y'(t) = f(y(t)) + d(t), \quad y(t_0) = y_0 \quad (15)$$

are nonetheless non-autonomous. We use the standard procedure to rewrite this system as an autonomous differential equation by adding the trivial equation $s'(t) = 1$ with exact solution $s(t) = t$,

$$\tilde{y}'(t) = \tilde{f}(\tilde{y}(t)) + \tilde{d}(\tilde{y}(t)), \quad \tilde{y}(t_0) = \tilde{y}_0, \quad (16)$$

where

$$\tilde{y} = \begin{pmatrix} y \\ s \end{pmatrix}, \quad \tilde{f}(\tilde{y}) := \begin{pmatrix} f(y) \\ 0 \end{pmatrix}, \quad \tilde{d}(\tilde{y}) := \begin{pmatrix} d(s) \\ 1 \end{pmatrix}, \quad \tilde{y}_0 := \begin{pmatrix} y_0 \\ t_0 \end{pmatrix}. \quad (17)$$

To achieve formal correspondence between the original equation and the augmented neighboring problem, we also add the trivial equation to (14) to obtain

$$\tilde{y}'(t) = \hat{f}(\tilde{y}(t)), \quad \tilde{y}(t_0) = \tilde{y}_0, \quad (18)$$

where

$$\hat{f}(\tilde{y}) := \begin{pmatrix} f(y) \\ 1 \end{pmatrix}.$$

To analyze the convergence of ISDeC introduced in section 2, we consider one step of the iteration, starting from a grid vector $\eta = \eta_{i,j}$, $j = 0, \dots, m$, $i = 0, \dots, N_1 - 1$, and estimate the *iteration error* of the new approximation in terms of the iteration error of $\eta_{i,j}$. $\eta_{i,j}$ is either the solution of (14) by the basic scheme Φ , or an improved solution approximation computed in the course of the ISDeC iteration. For this type of analysis we use the fact that Iterated Defect Correction converges to a fixed point $p^* = (p_0^*, p_1^*, \dots, p_{N_1-1}^*)$ under fairly general assumptions [1, 9]. This fixed point is easily identified as the continuous collocating function consisting of polynomials of degree $\leq m$ which satisfies (14) at the points $\sigma_{i,j}$, $i = 0, \dots, N_1 - 1$, $j = 1, \dots, m$, cf. (5). The *iteration error* $\eta - p^*$ is the error of the respective grid vector as compared with the fixed point. The results we obtain for the iteration error directly translate into order results for the global error of the numerical solution as compared with the exact solution y of (14) by the triangle inequality. Note that in our estimates, we usually neglect the last, trivial component of (16), which is necessary however for technical reasons that will become clear presently.

Denote by $p = (p_0, p_1, \dots, p_{N_1-1})$ the piecewise polynomial function of maximal degree m interpolating $\eta_{i,j}$ at $t_{i,j}$, $j = 0, \dots, m$, $i = 0, \dots, N_1 - 1$. One step of the ISDeC procedure yields a new grid function $\eta_{i,j}^{\text{new}}$ and associated interpolant $p^{\text{new}} = (p_0^{\text{new}}, p_1^{\text{new}}, \dots, p_{N_1-1}^{\text{new}})$. Subsequently, we will derive estimates for the current iteration error $e^{\text{new}} := p^{\text{new}} - p^*$ in terms of estimates for $e := p - p^*$.

We now express all the quantities associated with one step of ISDeC using the calculus of Lie derivatives, cf. [4, Sec. III.5]. We start by rewriting the Zadunaisky polynomial p and the fixed point p^* in (22) and (26), and derive analogous expressions for the quantities computed numerically, the basic solution $\eta^{[0]}$ of (14) and the computational solution of the neighboring problem (15), see (36) and (37). Note that actually we use the flows for the autonomous formulations augmented by the trivial equations as in (16) or (18).

$p_i(t)$, $t \in J_i$, $i = 0, \dots, N_1 - 1$, is the exact solution of

$$y'(t) = f(y(t)) + d_i(t), \quad y(\tau_i) = p_i(\tau_i), \quad (19)$$

where the defect $d_i(t)$ is defined by

$$d_i(t) := p_i'(t) - f(p_i(t)). \quad (20)$$

Analogously as in (16), $\tilde{p}_i(t)$ is the exact solution of the augmented equations

$$\tilde{y}'(t) = \tilde{f}(\tilde{y}(t)) + \tilde{d}_i(\tilde{y}(t)), \quad \tilde{y}(\tau_i) = \tilde{\mathbf{p}}_i, \quad (21)$$

where $\tilde{\mathbf{p}}_i := \tilde{p}_i(\tau_i)$. $\tilde{p}_i(t)$ can thus be written as

$$\tilde{p}_i(\tau_i + t) = \exp(t(\mathcal{F} + \mathcal{D}_i))\text{Id}(\tilde{y})\Big|_{\tilde{y} = \tilde{\mathbf{p}}_i}, \quad (22)$$

where \mathcal{F} and \mathcal{D}_i are the differential operators (*Lie derivatives*)

$$\mathcal{F} = \sum_{j=1}^{n+1} \tilde{f}_j(\tilde{y}) \frac{\partial}{\partial \tilde{y}_j} = \sum_{j=1}^n f_j(y) \frac{\partial}{\partial y_j} \quad (23)$$

and

$$\mathcal{D}_i = \sum_{j=1}^{n+1} \tilde{d}_{i,j}(\tilde{y}) \frac{\partial}{\partial \tilde{y}_j} = \sum_{j=1}^n d_{i,j}(s) \frac{\partial}{\partial y_j} + \frac{\partial}{\partial s}, \quad (24)$$

see [4, Sec. III.5.1]. Here, f_j and $d_{i,j}$ denote the j -th component of f and d_i , respectively, and the operator

$$\hat{\mathcal{D}}^* := \frac{\partial}{\partial s} \quad (25)$$

only acts on the last component of \tilde{y} .

Now we derive a representation analogous to (22) for the fixed point $p^* = (p_0^*, p_1^*, \dots, p_{N-1}^*)$. With \tilde{p}^* , d^* , and $\hat{\mathcal{D}}^*$ defined analogously as before, we obtain for $\tilde{\mathbf{p}}_i^* := \tilde{p}_i^*(\tau_i)$

$$\tilde{p}_i^*(\tau_i + t) = \exp(t(\mathcal{F} + \mathcal{D}_i^*))\text{Id}(\tilde{y})\Big|_{\tilde{y} = \tilde{\mathbf{p}}_i^*} \quad (26)$$

with

$$\mathcal{D}_i^* = \sum_{j=1}^{n+1} \tilde{d}_{i,j}^*(\tilde{y}) \frac{\partial}{\partial \tilde{y}_j} = \sum_{j=1}^n d_{i,j}^*(s) \frac{\partial}{\partial y_j} + \frac{\partial}{\partial s}. \quad (27)$$

Choose as the basic method Φ for ISDeC a composition method according to section 3. Φ is composed of substeps using the method ϕ . For all the methods we consider, the numerical flow defined by ϕ satisfies a *modified differential equation*

$$y' = f(y) + hF_1(y) + h^2F_2(y) + \dots, \quad (28)$$

cf. [4, Ch. IX]. Thus, the numerical flow of ϕ can be written as

$$\phi_h(y) = \exp(h(\mathcal{F} + h\mathcal{F}_1 + h^2\mathcal{F}_2 + \dots))\text{Id}(y), \quad (29)$$

with the Lie derivatives

$$\mathcal{F}_\ell = \sum_{j=1}^n F_{\ell,j}(y) \frac{\partial}{\partial y_j}, \quad \ell = 1, 2, \dots \quad (30)$$

If we assume that Φ is a method of order q , then its numerical flow when applied to (14) can be written as

$$\Phi_h(y) = \exp(h(\mathcal{F} + h^q \mathcal{G}_q^* + h^{q+1} \mathcal{G}_{q+1}^* + \dots)) \text{Id}(y), \quad (31)$$

where the differential operators \mathcal{G}_ℓ^* are certain well-defined elements of the free Lie algebra generated by $\{\mathcal{F}, \mathcal{F}_1, \mathcal{F}_2, \dots\}$, i.e., they are certain linear combinations of iterated commutators of elements of $\{\mathcal{F}, \mathcal{F}_1, \mathcal{F}_2, \dots\}$, see [4, Sec. III.5.4].

The corresponding numerical flow for the augmented equation (18) is given accordingly by

$$\tilde{\Phi}_h(\tilde{y}) = \exp(h(\mathcal{F} + \hat{\mathcal{D}}^* + h^q \mathcal{G}_q^* + h^{q+1} \mathcal{G}_{q+1}^* + \dots)) \text{Id}(\tilde{y}). \quad (32)$$

The numerical flow $\tilde{\Delta}_{i,h}(\tilde{y})$ of the vector field $\tilde{d}_i(\tilde{y})$ of (21) is given by

$$\tilde{\Delta}_{i,h}(\tilde{y}) = \exp(h\hat{\mathcal{D}}_i) \text{Id}(\tilde{y}), \quad (33)$$

where $\hat{\mathcal{D}}_i$ is defined similarly as in (24),

$$\hat{\mathcal{D}}_i = \sum_{j=1}^n D_{i,j}(s) \frac{\partial}{\partial y_j} + \frac{\partial}{\partial s}. \quad (34)$$

Recall that $D_i(t)$ is the polynomial of degree $\leq m-1$ interpolating $d_i(t)$ at the collocation points $\sigma_{i,1}, \dots, \sigma_{i,m}$, cf. (4, 5).

The numerical solution method for the neighboring problem, Ψ , is constructed in such a way that a method of order q results and the composition of the submethods is parallel to the definition of Φ , see section 3. Thus, the numerical flow $\tilde{\Psi}_{i,h}(\tilde{y})$ for (21) is given by

$$\tilde{\Psi}_{i,h}(\tilde{y}) = \exp(h(\mathcal{F} + \hat{\mathcal{D}}_i + h^q \mathcal{G}_{i,q} + h^{q+1} \mathcal{G}_{i,q+1} + \dots)) \text{Id}(\tilde{y}), \quad (35)$$

where the differential operators $\mathcal{G}_{i,\ell}$ are well-defined elements of the free Lie algebra generated by $\{\mathcal{F}, \hat{\mathcal{D}}_i, \mathcal{F}_1, \mathcal{F}_2, \dots\}$. Note that if each occurrence of $\hat{\mathcal{D}}_i$ in the definition of $\mathcal{G}_{i,\ell}$ is replaced by $\hat{\mathcal{D}}^*$, the resulting differential operator coincides with \mathcal{G}_ℓ^* from (32).

Now we are in a position to write the numerical solutions of the original and the neighboring problem in terms of Lie derivatives as in (22) and (26).

The basic numerical solution $\tilde{\eta}_{i,j}^{[0]}$ of equation (18) is given by $\tilde{\eta}_{i,j}^{[0]} := \tilde{\pi}_i^*(\tau_i + jh)$, where the functions $\tilde{\pi}_i^*(t)$, $t \in J_i$, are recursively defined by

$$\begin{aligned} \tilde{\pi}_0^* &= \tilde{y}_0, \\ \tilde{\pi}_i^*(\tau_i + t) &= \exp(t(\mathcal{F} + \hat{\mathcal{D}}^* + h^q \mathcal{G}_q^* + h^{q+1} \mathcal{G}_{q+1}^* + \dots)) \text{Id}(\tilde{\pi}_i^*), \\ \tilde{\pi}_{i+1}^* &= \tilde{\pi}_i^*(\tau_i + H), \end{aligned} \tag{36}$$

cf. (32). Here and in the sequel, $\text{Id}(\tilde{\pi}_i^*)$ denotes $\text{Id}(\tilde{y})|_{\tilde{y}=\tilde{\pi}_i^*}$ and analogously for other arguments.

In the same way, the numerical solution $\tilde{\pi}_{i,j}$ of the neighboring problem (16) is given by $\tilde{\pi}_{i,j} := \tilde{\pi}_i(\tau_i + jh)$, where

$$\begin{aligned} \tilde{\pi}_0 &= \tilde{y}_0, \\ \tilde{\pi}_i(\tau_i + t) &= \exp(t(\mathcal{F} + \hat{\mathcal{D}}_i + h^q \mathcal{G}_{i,q} + h^{q+1} \mathcal{G}_{i,q+1} + \dots)) \text{Id}(\tilde{\pi}_i), \\ \tilde{\pi}_{i+1} &= \tilde{\pi}_i(\tau_i + H), \end{aligned} \tag{37}$$

cf. (35).

Now, one iteration step of the ISDeC method on the subinterval J_i can be written as

$$p_i^{\text{new}}(t_{i,j}) = \tilde{\pi}_i^*(t_{i,j}) + (\tilde{p}_i(t_{i,j}) - \tilde{\pi}_i(t_{i,j})), \tag{38}$$

see (3).

This yields a representation of the iteration error $e^{\text{new}} = p^{\text{new}} - p^*$, which is the piecewise interpolant of degree $\leq m$ at $t_{i,j}$ of the first n components³ of $\tilde{\varepsilon} = (\tilde{\varepsilon}_0, \tilde{\varepsilon}_1, \dots, \tilde{\varepsilon}_{N_1-1})$, where

$$\tilde{\varepsilon}_i(\tau_i + t) := (\tilde{\pi}_i^*(\tau_i + t) - \tilde{\pi}_i(\tau_i + t)) - (\tilde{p}_i^*(\tau_i + t) - \tilde{p}_i(\tau_i + t)). \tag{39}$$

Recall that the function $\tilde{\varepsilon}(t)$ is defined in a piecewise manner on the subintervals J_i , $i = 0, \dots, N_1 - 1$, cf. (36) and (37). Subsequently, we will derive estimates for the relevant quantities on each interval J_i , and use the following well-known result to obtain global estimates on $[t_0, t_{\text{end}}]$. We formulate the result for uniform grids, the modifications necessary for piecewise equidistant (nonuniform) grids are indicated in [6].

Lemma 1 (Discrete Gronwall Lemma). Let the sequence of nonnegative numbers ξ_i , $i = 0, 1 \dots$ satisfy

$$\xi_0 = \delta_0, \quad \xi_{i+1} \leq (1 + \omega)\xi_i + \delta, \quad i = 0, 1, \dots \tag{40}$$

³Note that the $(n + 1)$ -st component of $\tilde{\varepsilon}$ is zero.

with $\omega > 0$ and $\delta \geq 0$. Then the estimate

$$\xi_i \leq \frac{e^{i\omega} - 1}{\omega} \delta + e^{i\omega} \delta_0 \quad (41)$$

holds for all i .

Next, we introduce Sobolev-like norms by means of which we will estimate grid vectors or functions X_H occurring in the course of the ISDeC step $p \mapsto p^{\text{new}}$. The index H emphasizes that X_H depends on the underlying grid and thus on the step size H . Frequently, our estimates will be written as

$$X_H = O(H^\ell \|e\|_k), \quad (42)$$

where this short-hand notation implies that there are constants $H_0 > 0$ and C independent of H and p such that for all $0 < H \leq H_0$

$$\|X_H\| \leq CH^\ell \sum_{\kappa=0}^k \max_{i=0, \dots, N_1-1} \max_{t \in J_i} |e_i^{(\kappa)}(t)|. \quad (43)$$

Next, we formulate the main results of this paper. The proofs of these propositions are given in Appendix A, together with a number of technical lemmas required to derive the estimates. The numbering of the following lemma is chosen such as to be consistent with Appendix A.

Lemma 7. Let

$$\tilde{\epsilon}_i := \tilde{\epsilon}_i(\tau_i) = (\tilde{\pi}_i^* - \tilde{\pi}_i) - (\tilde{\mathbf{p}}_i^* - \tilde{\mathbf{p}}_i). \quad (44)$$

Then for each $u \geq 0$ the bound

$$|\tilde{\epsilon}_i| = O(H^{\min(q,m)} \|e\|_{\min(q,m)}) + O(H^m \|e\|_m) + O(H^u) \quad (45)$$

holds.

For the following result, we only consider the case where $q \leq m$. Although the arguments can be extended in principle to the case $q > m$, see [8], we omit this case here because it is not relevant in practice.

Using Lemma 7, the central convergence theorem for the iteration error of ISDeC can be proven:

Theorem 1. The iteration error $e^{\text{new}} = p^{\text{new}} - p^*$ satisfies estimates

$$\left| \frac{d^\kappa}{dt^\kappa} e^{\text{new}}(\tau_i + t) \right| = \begin{cases} O(H^m \|e\|_m) + O(H^q \|e\|_q) + O(H^u), & \kappa = 0, \\ O(H^{m+1-\kappa} \|e\|_m) + O(H^q \|e\|_{q-1+\kappa}) + O(H^{u+1-\kappa}), \\ \quad \kappa = 1, \dots, m-q, \\ O(H^{m+1-\kappa} \|e\|_m) + O(H^{u+1-\kappa}), & \kappa = m-q+1, \dots, m \end{cases} \quad (46)$$

for each $u \geq 0$. If the collocation abscissae ρ_j from (5) satisfy the condition

$$\sum_{j=1}^m \rho_j = \frac{m}{2}, \tag{47}$$

then the estimate (46) for $\kappa = m$ can be replaced by the sharper bound

$$\left| \frac{d^m}{dt^m} e^{\text{new}}(\tau_i + t) \right| = O(H^2 \|e\|_m) + O(H^{u+1-m}). \tag{48}$$

Remark. Note that condition (47) is satisfied if ρ_j are symmetric in $[0, 1]$, and consequently holds for Gaussian points for example.

5. Numerical examples

To illustrate the results derived in section 4, we now demonstrate the order sequences for the ISDeC iterates implied by Theorem 1 for the fourth order scheme resulting from Yoshida’s method based on the Störmer/Verlet method, see section 3. To this end, we first discuss the iteration error of the basic solution and its derivatives in the following theorem.

Theorem 2. The interpolant $e^{[0]}$ of the iteration error

$$\varepsilon^{[0]} = \eta^{[0]} - p^*$$

for the basic solution $\eta^{[0]}$ of (14) computed by a numerical method Φ of order q satisfies

$$\begin{aligned} \|e^{[0]}\|_k &= O(H^q), \quad k = 0, \dots, m - q + 1, \\ \|e^{[0]}\|_k &= O(H^{m+1-k}), \quad k = m - q + 2, \dots, m - 1, \\ \|e^{[0]}\|_m &= \begin{cases} O(H^2) & \text{if } \rho_j \text{ satisfy (47),} \\ O(H) & \text{otherwise,} \end{cases} \end{aligned} \tag{49}$$

if ISDeC is defined by polynomials of degree m .

Proof. The estimates above follow from

$$\|g\|_k = O(H^q), \quad k = 0, \dots, m,$$

where g is the piecewise polynomial interpolant of degree $\leq m$ of the global error $\eta^{[0]} - y$. This can be proven by means of an asymptotic expansion of the global error of Φ : If a sufficiently long error expansion exists, then there is a smooth function $E(t, H)$ such that

$$\eta_{i,j}^{[0]} - y(t_{i,j}) = E(t_{i,j}, H)H^q,$$

with $\frac{\partial^k}{\partial t^k} E(t, H) = O(1)$, $k = 0, \dots, m$. Consequently, we conclude

$$\begin{aligned} \left| \frac{d^k}{dt^k} g_i(t) \right| &\leq \left| \frac{d^k}{dt^k} g_i(t) - \frac{\partial^k}{\partial t^k} E(t, H) H^q \right| + \left| \frac{\partial^k}{\partial t^k} E(t, H) H^q \right| \\ &\leq O(H^{m+q+1-k}) + O(H^q) = O(H^q), \quad k = 0, \dots, m, \end{aligned}$$

which follows from Lemma 3. Moreover, we use the relations

$$\begin{aligned} \|Q\|_0 &= \begin{cases} O(H^{m+1}) & \text{if } \rho_j \text{ define collocation of order } \geq m+1, \\ O(H^m) & \text{otherwise,} \end{cases} \\ \|Q\|_k &= O(H^{m+1-k}), \quad k = 1, \dots, m-1, \\ \|Q\|_m &= \begin{cases} O(H^2) & \text{if } \rho_j \text{ satisfy (47),} \\ O(H) & \text{otherwise} \end{cases} \end{aligned}$$

for the piecewise polynomial interpolant Q of $p^* - y$ at $t_{i,j}$, $j = 0, \dots, m$. These are standard results for collocation methods, again taking into account Lemma 3. The improved estimate $\|Q\|_m = O(H^2)$ when (47) holds follows from

$$|Q^{(m)}(t)| = O(H^2), \quad \text{if } \rho_j \text{ satisfy (47),} \quad (50)$$

which is shown using Lemma 9, see [8].

Using this result for the basic approximation and Theorem 1, we can easily conclude the sequence of iteration errors for the respective ISDeC iterates. To illustrate the procedure, we give the results for the case where $m = 6, 7$, and 8 , respectively, and Φ is a fourth order composition method like Yoshida's method, cf. section 3.

If the condition (47) is satisfied, for the iteration errors we conclude the order sequences for $\|e^{[j]}\|_0$ for the polynomial degrees

$$\begin{aligned} m = 6 : & O(H^4), O(H^7), O(H^9), O(H^{11}), O(H^{13}), O(H^{15}), \dots \\ m = 7 : & O(H^4), O(H^8), O(H^{10}), O(H^{12}), O(H^{14}), O(H^{16}), \dots \\ m = 8 : & O(H^4), O(H^8), O(H^{10}), O(H^{12}), O(H^{14}), O(H^{16}), \dots \end{aligned}$$

These are also observed in the numerical experiments reported below, see also [7]. Note that the actually observed orders for $m = 6$ in this case are in fact higher in the numerical experiments reported in [7] than the orders concluded from the considerations above. These experimental results are no contradiction to the theory, however, since we only give sufficient conditions for our estimates to hold.

If conversely (47) is not satisfied, the resulting order sequences for $\|e^{[l]}\|_0$ are

$$m = 6 : O(H^4), O(H^7), O(H^8), O(H^9), O(H^{10}), O(H^{11}), \dots$$

$$m = 7 : O(H^4), O(H^8), O(H^9), O(H^{10}), O(H^{11}), O(H^{12}), \dots$$

$$m = 8 : O(H^4), O(H^8), O(H^{10}), O(H^{11}), O(H^{12}), O(H^{13}), \dots$$

To illustrate the convergence results discussed above, we consider a simple Hamiltonian test example, the *Kepler problem*. Let $x = (x_1, x_2)$, $y = (y_1, y_2)$, then the differential equations are defined by

$$x'(t) = -\nabla_y H(x, y), \quad y'(t) = \nabla_x H(x, y),$$

with

$$H(x, y) = \frac{1}{2}(x_1^2 + x_2^2) - \frac{1}{\sqrt{y_1^2 + y_2^2}}.$$

The exact solution of the Kepler problem is periodic with period 2π . Consequently, we choose the integration interval $[t_0, t_{\text{end}}] = [0, 2\pi]$. The initial values are given as

$$y_1(0) = 1 - e, \quad y_2(0) = 0, \quad y_1'(0) = 0, \quad y_2'(0) = \sqrt{\frac{1+e}{1-e}}, \quad (51)$$

where in our experiments we use $e = 0.6$. The test runs were implemented in C++ using ‘quad-double’ precision, see

<http://crd.lbl.gov/~dhbailey/mpdist/>

This extended precision (approximately 64 decimal digits) was necessary in order to observe unambiguously even very high convergence orders before reaching accuracies of the order of magnitude of round-off error.

In Table 1, we give the iteration errors of the basic solution computed by Yoshida’s method in conjunction with the Störmer/Verlet scheme, and of the first four steps of ISDeC based on this composition method and interpolation at $m = 7$ Gaussian points (thus, the fixed point of the iteration is a collocation solution of order 14, and the condition (47) is satisfied). Moreover, the empirical convergence orders computed for two successive step sizes H are listed. The order sequence corresponds to the theoretical results derived above.

Table 2 shows the corresponding results for the case where the interpolation is defined for $m = 7$ points ρ_j , see (5), chosen randomly in $[0, 1]$. The points do not satisfy the relation (47), and the order of the fixed point is seven. Again, the order sequence of the iteration errors reflects our theoretical considerations.

So far, we have considered the *iteration error* of ISDeC. The global error of the ISDeC iterates as compared with the exact solution of (14) is of course closely linked

Table 1
Iteration errors, $m = 7$ Gaussian points.

H	Yoshida	ISDeC 1	ISDeC 2	ISDeC 3	ISDeC 4
$2\pi/25$	1.06E-02	7.45E-05	1.58E-06	7.77E-09	5.20E-11
$\pi/25$	6.76E-04	3.05E-07	4.43E-10	5.71E-13	9.38E-16
$\pi/50$	4.25E-05	1.21E-09	1.42E-13	1.01E-16	7.19E-20
$\pi/100$	2.66E-06	4.72E-12	6.71E-17	2.38E-20	4.59E-24
$\pi/200$	1.66E-07	1.85E-14	4.82E-20	5.78E-24	2.83E-28
$\pi/400$	1.04E-08	7.22E-17	4.29E-23	1.41E-27	1.73E-32
$\pi/800$	6.50E-10	2.82E-19	4.09E-26	3.44E-31	1.06E-36
$2\pi/25$					
$\pi/25$	3.97	7.93	11.80	13.73	15.76
$\pi/50$	3.99	7.98	11.61	12.46	13.67
$\pi/100$	4.00	8.00	11.05	12.05	13.94
$\pi/200$	4.00	8.00	10.44	12.01	13.99
$\pi/400$	4.00	8.00	10.13	12.00	14.00
$\pi/800$	4.00	8.00	10.03	12.00	13.99

to this quantity. Namely, it follows from the triangle equality that the global error has the same order as the iteration error up to the order defined by the fixed point p^* . If we use Gaussian points $\sigma_{i,j}$ for the interpolation of the defect, see (4), (5), the maximally attainable order of the global error of ISDeC iterates is $2m$, cf. [7].

Further numerical experiments are reported in [7], where it is illustrated that the theory applies also to Suzuki’s composition method based on the Störmer/Verlet scheme for Hamiltonian systems or the exponential midpoint rule for linear, homogeneous ODEs. Moreover, McLachlan’s method based on either the symplectic

Table 2
Iteration errors, $m = 7$ random points.

H	Yoshida	ISDeC 1	ISDeC 2	ISDeC 3	ISDeC 4
$2\pi/25$	6.23E-02	2.98E-03	1.07E-04	1.10E-05	1.29E-06
$\pi/25$	5.45E-04	1.33E-05	1.55E-07	2.68E-08	6.52E-11
$\pi/50$	4.23E-05	4.26E-08	3.06E-11	2.06E-11	3.97E-14
$\pi/100$	2.66E-06	1.57E-10	1.43E-13	1.87E-14	1.70E-17
$\pi/200$	1.66E-07	6.01E-13	4.65E-16	1.77E-17	7.93E-21
$\pi/400$	1.04E-08	2.32E-15	1.09E-18	1.70E-20	3.80E-24
$\pi/800$	6.50E-10	9.03E-18	2.29E-21	1.65E-23	1.84E-27
$2\pi/25$					
$\pi/25$	6.84	7.81	9.43	8.68	14.27
$\pi/50$	3.69	8.29	12.31	10.35	10.68
$\pi/100$	3.99	8.08	7.74	10.11	11.19
$\pi/200$	4.00	8.03	8.26	10.05	11.07
$\pi/400$	4.00	8.02	8.74	10.02	11.03
$\pi/800$	4.00	8.01	8.89	10.01	11.01

or explicit/implicit Euler methods is covered by our treatment. The order results observed for Gaussian, Radau or random interpolation nodes illustrate our theoretical estimates, and it is also demonstrated that a choice of interpolation points such that (47) is satisfied, but $\sigma_{i,j}$ are not symmetric, still yields (48). Finally, as an interesting special case a method analyzed in [9] is also shown to fit into the framework of our discussion, where the trapezoidal rule is considered as a composition of the explicit and implicit Euler methods [7].

6. Discussion and outlook

In this paper, we have given a rigorous error analysis for a new defect correction method based on the ideas of splitting and composition, which we denote as Iterated Splitting Defect Correction (ISDeC). This method can be employed to make Iterated Defect Correction applicable in conjunction with geometric integrators for problems with a special structure, where a straightforward application of defect correction is not possible. Our analysis shows that high convergence orders can be achieved by ISDeC up to a theoretical limit defined in terms of certain (superconvergent) collocation methods. The main tool in our proof is the use of Lie derivatives [4, Sec. III.5.1] to show high-order fixed point convergence of the iterates. This technique naturally extends also to the analysis of classical versions of Iterated Defect Correction [1] and thus supplements standard techniques for this class of methods.

Our focus in this work has been on making the idea of Iterated Defect Correction applicable to a class of discretization methods for problems where a straightforward application of the method fails or is disadvantageous. Moreover, we were able to present a very general framework for the rigorous error analysis of defect correction methods, which allows to also explain the high convergence orders attainable by using defect interpolation at Gaussian points.

Unfortunately, the geometric properties of the basic methods are not preserved exactly in general for the high-order approximations obtained by our method. This has also been observed for other acceleration techniques in conjunction with geometric integrators [2].

We would like to point out nonetheless that our convergence results have simple implications on the approximation quality of the ISDeC iterates with respect to important conserved quantities of the exact flow, like the angular momentum or the Hamiltonian in the case of Hamiltonian systems, or the Euclidean norm of the solution for certain linear autonomous systems. If for instance we choose Gaussian interpolation nodes $\sigma_{i,j}$ for ISDeC applied to a Hamiltonian system, angular momentum and the Hamiltonian are conserved exactly by the fixed point of the ISDeC iteration. Consequently, the respective ISDeC iterates conserve these quantities up to terms of the order of the iteration error, which may be far better than the absolute error of the numerical solution. Unfortunately, no better approximation properties are observed in general [7].

We found that the application of ISDeC may also negatively affect the long-time approximation properties of the numerical solution. It appears that there is a trade-off between using lower-order methods which display favorable error growth for long time intervals, and the application of high-order schemes to achieve high accuracy on comparatively short intervals. The conclusion that the latter aim is also well justified in the context of geometric integration is supported by the claims of [2].

APPENDIX

A.1. Technical details of the proof

Here, we indicate the technical details necessary in order to prove the central convergence results formulated in section 4. For lack of space, not all the arguments are carried out exhaustively, for the complete proofs we refer the reader to [8]. The notation used here is introduced in section 4.

The following relations will be used in many of our arguments given in the course of the convergence proof for ISDeC. The proofs of the propositions are immediate, for details see [8].

Lemma 2. If X_H satisfies (42), then

$$X_H = O(H^{\ell-\kappa} \|e\|_{k-\kappa}), \quad \kappa = 0, \dots, k, \quad (52)$$

$$X_H = O(H^\ell \|e\|_m), \quad \text{if } k \geq m. \quad (53)$$

Lemma 3. Let p_i be a polynomial of degree $\leq m$ interpolating a sufficiently smooth function $y \in C^{m+1}[t_0, t_{\text{end}}]$ on the interval J_i . Then the estimates

$$\max_{t \in J_i} |p_i^{(\kappa)}(t) - y^{(\kappa)}(t)| \leq \text{const.} H^{m+1-\kappa} \max_{t \in J_i} |y^{(m+1)}(t)|, \quad (54)$$

$$\max_{t \in J_i} |p_i^{(\kappa)}(t) - y^{(\kappa)}(t)| \leq \text{const.} \max_{t \in J_i} |y^{(\kappa)}(t)| \quad (55)$$

hold for $\kappa = 0, \dots, m+1$.

Now, we derive bounds for quantities appearing in the convergence proof for ISDeC.

Lemma 4. The interpolant $D_{i,j}(t)$ of one component of the defect of p , $d_{i,j}(t) = p'_{i,j}(t) - f_j(p_i(t))$, satisfies

$$D_{i,j}^{(k)}(\tau_i) = O(\|e\|_{k+1}), \quad k = 0, \dots, m-1, \quad (56)$$

$$D_{i,j}^{(k)}(\tau_i) = 0, \quad k \geq m. \quad (57)$$

Proof. Let $F_{i,j}(t)$ and $F_{i,j}^*(t)$ be the polynomials of degree $\leq m - 1$ which interpolate the component functions $f_j(p_i(t))$ and $f_j(p_i^*(t))$, respectively, at the nodes $\sigma_{i,1}, \dots, \sigma_{i,m}$. From the definition of $p^*(t)$ we conclude $p_{i,j}'(t) = F_{i,j}'^*(t)$, and consequently

$$D_{i,j}^{(k)}(\tau_i) = (p_{i,j}^{(k+1)}(\tau_i) - p_{i,j}^{*(k+1)}(\tau_i)) - (F_{i,j}^{(k)}(\tau_i) - F_{i,j}^{*(k)}(\tau_i)). \quad (58)$$

Clearly,

$$p_{i,j}^{(k+1)}(\tau_i) - p_{i,j}^{*(k+1)}(\tau_i) = e_{i,j}^{(k+1)}(\tau_i) = O(\|e\|_{k+1}).$$

Using Lipschitz conditions for $f_{i,j}^{(\kappa)}(y)$, $\kappa = 0, \dots, k$, and Lemma 3 for the interpolant $F_{i,j}(t) - F_{i,j}^*(t)$ of $f_j(p_i(t)) - f_j(p_i^*(t))$,

$$F_{i,j}^{(k)}(\tau_i) - F_{i,j}^{*(k)}(\tau_i) = O(\|e\|_k) = O(\|e\|_{k+1})$$

is straightforward to show. □

Lemma 5. For the quantities defined in (22), (26), (36) and (37),

$$|\tilde{\boldsymbol{\pi}}_i - \tilde{\mathbf{p}}_i| = O(H^{\min(q,m)}) \quad (59)$$

and

$$|\tilde{\boldsymbol{\pi}}_i^* - \tilde{\mathbf{p}}_i^*| = O(H^{\min(q,m)}) \quad (60)$$

holds for $i = 0, \dots, N_1$.

Proof. To prove (59), we expand (22) and (37) up to terms of order $O(H^q)$ and obtain

$$\begin{aligned} \tilde{\boldsymbol{\pi}}_{i+1} - \tilde{\mathbf{p}}_{i+1} &= \tilde{\boldsymbol{\pi}}_i - \tilde{\mathbf{p}}_i + \sum_{k=1}^q \frac{H^k}{k!} \left((\mathcal{F} + \hat{\mathcal{D}}_i)^k \text{Id}(\tilde{\boldsymbol{\pi}}_i) - \right. \\ &\quad \left. - (\mathcal{F} + \mathcal{D}_i)^k \text{Id}(\tilde{\mathbf{p}}_i) \right) + O(H^{q+1}). \end{aligned} \quad (61)$$

Throughout the analysis given in this paper, it is important to realize the general form of terms occurring in expansions of expressions like $\frac{H^k}{k!} (\mathcal{F} + \mathcal{D}_i)^k \text{Id}(\tilde{\mathbf{p}}_i)$. We denote the generic form of the components of these terms by

$$\frac{H^k}{k!} \mathbf{f}(\mathbf{p}_i) \mathbf{d}_i(\tau_i), \quad (62)$$

where $\mathbf{f}(y)$ is either the constant function $\mathbf{f}(y) \equiv 1$ or a product of component functions $f_j(y)$ of $f(y)$ or derivatives thereof, all evaluated at $y = \mathbf{p}_i$ (the first n components of $\tilde{\mathbf{p}}_i$). $\mathbf{d}_i(\tau_i)$ denotes expressions of the form

$$\mathbf{d}_i(\tau_i) = d_{i,j_1}^{(\kappa_1)}(\tau_i) \cdots d_{i,j_r}^{(\kappa_r)}(\tau_i), \quad (63)$$

where the derivatives are of order $\leq k-1$ and $d_{i,j}$ denotes one component of the defect d_i . Clearly, the components of $\frac{H^k}{k!}(\mathcal{F} + \mathcal{D}_i)^k \text{Id}(\tilde{\mathbf{p}}_i)$ can be expanded in this way for all $k = 1, \dots, q$.

Likewise, the components of $\frac{H^k}{k!}(\mathcal{F} + \hat{\mathcal{D}}_i)^k \text{Id}(\tilde{\boldsymbol{\pi}}_i)$ have expansions which contain terms

$$\frac{H^k}{k!} \mathbf{f}(\boldsymbol{\pi}_i) \mathbf{D}_i(\tau_i), \quad (64)$$

where $\mathbf{f}(y)$ is evaluated at $y = \boldsymbol{\pi}_i$, and (63) is replaced by

$$\mathbf{D}_i(\tau_i) = D_{i,j_1}^{(\kappa_1)}(\tau_i) \cdots D_{i,j_r}^{(\kappa_r)}(\tau_i). \quad (65)$$

Combining corresponding terms in the above expansions we obtain

$$\mathbf{f}(\boldsymbol{\pi}_i) \mathbf{D}_i(\tau_i) - \mathbf{f}(\mathbf{p}_i) \mathbf{d}_i(\tau_i) = (\mathbf{f}(\boldsymbol{\pi}_i) - \mathbf{f}(\mathbf{p}_i)) \mathbf{D}_i(\tau_i) + \mathbf{f}(\mathbf{p}_i) (\mathbf{D}_i(\tau_i) - \mathbf{d}_i(\tau_i)).$$

Using Lemma 3 applied to the interpolating polynomials $D_{i,j}(t)$ of $d_{i,j}(t)$ which have degree $\leq m-1$, it is easy to show that

$$\mathbf{D}_i(\tau_i) - \mathbf{d}_i(\tau_i) = O(H^{m-k+1}), \quad (66)$$

if we recall that the highest derivatives occurring in this relation are of order $\leq k-1$. Using a Lipschitz condition for $\mathbf{f}(y)$, we now conclude

$$\begin{aligned} \frac{H^k}{k!} |\mathbf{f}(\boldsymbol{\pi}_i) \mathbf{D}_i(\tau_i) - \mathbf{f}(\mathbf{p}_i) \mathbf{d}_i(\tau_i)| &\leq \text{const.} H^k |\tilde{\boldsymbol{\pi}}_i - \tilde{\mathbf{p}}_i| + O(H^{m+1}) \\ &\leq \text{const.} H |\tilde{\boldsymbol{\pi}}_i - \tilde{\mathbf{p}}_i| + O(H^{m+1}). \end{aligned}$$

Summation over all $k = 1, \dots, q$ in (61) now yields

$$|\tilde{\boldsymbol{\pi}}_{i+1} - \tilde{\mathbf{p}}_{i+1}| \leq (1 + \text{const.} H) |\tilde{\boldsymbol{\pi}}_i - \tilde{\mathbf{p}}_i| + O(H^{\min(q,m)+1}). \quad (67)$$

The bound (59) now follows from Lemma 1, (60) is shown analogously. \square

Lemma 6. For each $u > 0$, the relation

$$|\tilde{\boldsymbol{\pi}}_i^* - \tilde{\boldsymbol{\pi}}_i| = O(\|e\|_1) + O(H^u) \quad (68)$$

holds.

Proof. Expansion of (36) and (37) up to terms of order $O(H^u)$ and combination of like powers of H yields

$$\begin{aligned} \tilde{\boldsymbol{\pi}}_{i+1}^* - \tilde{\boldsymbol{\pi}}_{i+1} &= \tilde{\boldsymbol{\pi}}_i^* - \tilde{\boldsymbol{\pi}}_i + \sum_{k=1}^u \frac{H^k}{k!} \left(\mathcal{F}^k \text{Id}(\tilde{\boldsymbol{\pi}}_i^*) - (\mathcal{F} + \hat{\mathcal{D}}_i)^k \text{Id}(\tilde{\boldsymbol{\pi}}_i) \right) + \\ &\quad + [\text{terms of order } \leq u \text{ involving } \mathcal{G}_\ell^* \text{ or } \mathcal{G}_{i,\ell}] + \\ &\quad + O(H^{u+1}) + (0, \dots, 0, H)^T. \end{aligned} \quad (69)$$

Here we have used

$$(\mathcal{F} + \hat{\mathcal{D}}^*)^k = \mathcal{F}^k + \hat{\mathcal{D}}^{*k},$$

$$\hat{\mathcal{D}}^* \text{Id}(\tilde{y}) = (0, \dots, 0, 1)^T \quad \text{and} \quad \hat{\mathcal{D}}^{*k} \text{Id}(\tilde{y}) = 0, \quad k \geq 2.$$

The last component of the term $(0, \dots, 0, H)^T$ cancels with the corresponding term in the expansion of $(\mathcal{F} + \hat{\mathcal{D}}_i)^k \text{Id}(\tilde{\pi}_i)$. Generally, the last component in (69) is zero and only the first n components are considered in our analysis.

The terms in the expansion of

$$\frac{H^k}{k!} (\mathcal{F}^k \text{Id}(\tilde{\pi}_i^*) - (\mathcal{F} + \hat{\mathcal{D}}_i)^k \text{Id}(\tilde{\pi}_i))$$

are estimated similarly as in Lemma 5.

To estimate the terms of order $\leq u$ involving \mathcal{G}_ℓ^* or $\mathcal{G}_{i,\ell}$ in (69), we have to understand the general form of the involved terms. This is explained in detail in [8]. We omit the precise, yet lengthy statement of these expressions here, the derivation can be reconstructed from results in [4], see in particular Sec. III.4.2 of this reference. The appearing terms take either of two possible forms:

$$\frac{H^{k_1} h^{k_2}}{k_1!} \mathbf{f} \mathbf{f}_1 \cdots \mathbf{f}_r, \tag{70}$$

where \mathbf{f} is as in (62), and likewise \mathbf{f}_ℓ denotes products of components $F_{\ell,j}$ of F_ℓ and derivatives thereof, all evaluated at either π_i^* or π_i . Additionally, there are terms of the form

$$-\frac{H^{k_1} h^{k_2}}{k_1!} \mathbf{f}(\pi_i) \mathbf{f}_1(\pi_i) \cdots \mathbf{f}_r(\pi_i) \mathbf{D}_i(\tau_i), \tag{71}$$

where $\mathbf{D}_i(\tau_i)$ is defined as in (65). Noting the precise number of occurrences of these terms and the orders of the involved derivatives, it is possible to derive the estimates

$$\begin{aligned} & \frac{H^{k_1} h^{k_2}}{k_1!} |\mathbf{f}(\pi_i^*) \mathbf{f}_1(\pi_i^*) \cdots \mathbf{f}_r(\pi_i^*) - \mathbf{f}(\pi_i) \mathbf{f}_1(\pi_i) \cdots \mathbf{f}_r(\pi_i)| \\ & \leq \text{const.} H |\tilde{\pi}_i^* - \tilde{\pi}_i|, \end{aligned} \tag{72}$$

and

$$\begin{aligned} & \frac{H^{k_1} h^{k_2}}{k_1!} |\mathbf{f}(\pi_i) \mathbf{f}_1(\pi_i) \cdots \mathbf{f}_r(\pi_i) \mathbf{D}_i(\tau_i)| \\ & = O(H^{q+1} \|e\|_q) = O(H \|e\|_1). \end{aligned} \tag{73}$$

Summing up the estimates for all the terms involved in equation (69), we obtain

$$\|\tilde{\boldsymbol{\pi}}_{i+1} - \tilde{\boldsymbol{\pi}}_{i+1}^*\| \leq (1 + \text{const.}H)\|\tilde{\boldsymbol{\pi}}_i - \tilde{\boldsymbol{\pi}}_i^*\| + O(H\|e\|_1) + O(H^{u+1}). \quad (74)$$

Now, (68) follows by an application of Lemma 1. \square

Lemma 7. Let

$$\tilde{\boldsymbol{\varepsilon}}_i := \tilde{\boldsymbol{\varepsilon}}_i(\tau_i) = (\tilde{\boldsymbol{\pi}}_i^* - \tilde{\boldsymbol{\pi}}_i) - (\tilde{\mathbf{p}}_i^* - \tilde{\mathbf{p}}_i). \quad (75)$$

Then for each $u \geq 0$ the bound

$$|\tilde{\boldsymbol{\varepsilon}}_i| = O(H^{\min(q,m)}\|e\|_{\min(q,m)}) + O(H^m\|e\|_m) + O(H^u) \quad (76)$$

holds.

Proof. Similarly as in the proof of Lemma 6 we obtain

$$\begin{aligned} \tilde{\boldsymbol{\varepsilon}}_{i+1} = & \tilde{\boldsymbol{\varepsilon}}_i + \sum_{k=1}^u \frac{H^k}{k!} \left(\mathcal{F}^k \text{Id}(\tilde{\boldsymbol{\pi}}_i^*) - (\mathcal{F} + \hat{\mathcal{D}}_i)^k \text{Id}(\tilde{\boldsymbol{\pi}}_i) - \right. \\ & \left. - (\mathcal{F} + \mathcal{D}_i^*)^k \text{Id}(\tilde{\mathbf{p}}_i^*) + (\mathcal{F} + \mathcal{D}_i)^k \text{Id}(\tilde{\mathbf{p}}_i) \right) + (0, \dots, 0, H)^T + \\ & + [\text{terms of order } \leq u \text{ involving } \mathcal{G}_\ell^* \text{ or } \mathcal{G}_{i,\ell}] + O(H^{u+1}), \end{aligned} \quad (77)$$

where the terms involving \mathcal{G}_ℓ^* or $\mathcal{G}_{i,\ell}$ are the same as in (69).

To treat the terms in the expansion of

$$\frac{H^k}{k!} (\mathcal{F}^k \text{Id}(\tilde{\boldsymbol{\pi}}_i^*) - (\mathcal{F} + \hat{\mathcal{D}}_i)^k \text{Id}(\tilde{\boldsymbol{\pi}}_i) - (\mathcal{F} + \mathcal{D}_i^*)^k \text{Id}(\tilde{\mathbf{p}}_i^*) + (\mathcal{F} + \mathcal{D}_i)^k \text{Id}(\tilde{\mathbf{p}}_i)),$$

we derive the following estimates:

$$\begin{aligned} & \frac{H^k}{k!} |\mathbf{f}(\boldsymbol{\pi}_i^*) - \mathbf{f}(\boldsymbol{\pi}_i) - \mathbf{f}(\mathbf{p}_i^*) + \mathbf{f}(\mathbf{p}_i)| \\ & \leq \text{const.}H|\tilde{\boldsymbol{\varepsilon}}_i| + O(H^{\min(q,m)+1}\|e\|_0), \end{aligned} \quad (78)$$

$$\begin{aligned} & \frac{H^k}{k!} |-\mathbf{f}(\boldsymbol{\pi}_i)\mathbf{D}_i(\tau_i) - \mathbf{f}(\mathbf{p}_i^*)\mathbf{d}_i^*(\tau_i) + \mathbf{f}(\mathbf{p}_i)\mathbf{d}_i(\tau_i)| \\ & = O(H^{m+1}\|e\|_m) + O(H^{\min(q,m)+k}\|e\|_k) \\ & = O(H^{m+1}\|e\|_m) + O(H^{\min(q,m)+1}\|e\|_1). \end{aligned} \quad (79)$$

To prove these assertions, we use the estimates

$$\mathbf{f}(\mathbf{p}_i) - \mathbf{f}(\mathbf{p}_i^*) = O(\|e\|_0), \quad (80)$$

$$\mathbf{f}(\boldsymbol{\pi}_i) - \mathbf{f}(\mathbf{p}_i) = O(H^{\min(q,m)}), \quad (81)$$

$$\mathbf{d}_i^*(\tau_i) - \mathbf{d}_i(\tau_i) = O(\|e\|_k), \tag{82}$$

$$\mathbf{d}_i^*(\tau_i) = O(H^{m-k+1}), \tag{83}$$

$$-\mathbf{D}_i(\tau_i) - \mathbf{d}_i^*(\tau_i) + \mathbf{d}_i(\tau_i) = O(H^{m-k+1}\|e\|_m). \tag{84}$$

To show these estimates, in [8] we used Lipschitz conditions for f and its derivatives and Lemmas 2, 3 and 5, taking into account the order of the derivatives occurring in the respective expressions. The remaining terms involving \mathcal{G}_ℓ^* or $\mathcal{G}_{i,\ell}$ can be treated analogously as in Lemma 6. The proof of the proposition of this lemma is thus completed using Lemma 1. \square

For the following results, we only consider the case $q \leq m$, see section 4.

Lemma 8. Let $\tilde{\varepsilon}_{i,u}(\tau_i + t)$ for $0 \leq t \leq H$ denote the part of the expansion of $\tilde{\varepsilon}_i(\tau_i + t)$ consisting of the terms of order up to $O(H^u)$. For the derivatives,

$$|\tilde{\varepsilon}_{i,u}^{(\kappa)}(\tau_i + t)| = \begin{cases} O(H^m\|e\|_m) + O(H^q\|e\|_q) + O(H^u), & \kappa = 0, \\ O(H^{m+1-\kappa}\|e\|_m) + O(H^q\|e\|_{q-1+\kappa}) + O(H^u), & \kappa = 1, \dots, m - q, \\ O(H^{m+1-\kappa}\|e\|_m) + O(H^u), & \kappa = m - q + 1, \dots, m \end{cases} \tag{85}$$

holds for $0 \leq t \leq H$.

Proof. To prove the assertion, $\tilde{\varepsilon}_{i,u}(\tau_i + t)$ is expanded similarly as in Lemma 7. Now, in the corresponding estimates terms of the form $\frac{H^\kappa}{k!}$ are replaced by $\frac{t^{k-\kappa}}{(k-\kappa)!}$ for the derivative of order κ , where $\kappa \leq k$. The estimates (85) are thus proven analogously as in Lemma 7. \square

The next lemma is used to obtain sharper bounds than those in Lemma 8 for special choices of the points $\sigma_{i,j}$ used for defect interpolation, see (4). The assumption of the lemma holds for instance if the points ρ_j from (5) are symmetric in the interval $[0, 1]$.

Lemma 9. Let $y(t)$ be an $(m + 2)$ times continuously differentiable function on $[t_0, t_0 + H]$. Let $p(t)$ be the interpolation polynomial of degree $\leq m$ which is defined by

$$p(t_0 + jh) = y(t_0 + jh), \quad j = 0, \dots, m, \tag{86}$$

and let $q(t)$ be any polynomial of degree $\leq m$ which satisfies

$$q'(t_0 + \rho_j H) = y'(t_0 + \rho_j H), \quad j = 1, \dots, m. \tag{87}$$

If ρ_j satisfy

$$\sum_{j=1}^m \rho_j = \frac{m}{2}, \quad (88)$$

then for the m -th derivatives of $p(t) - q(t)$ we have

$$p^{(m)}(t_0 + t) - q^{(m)}(t_0 + t) = O(h^2 \|y^{(m+2)}\|), \quad 0 \leq t \leq H, \quad (89)$$

where $\|y^{(\kappa)}\| := \max_{t \in [t_0, t_0+H]} |y^{(\kappa)}(t)|$.

Proof. The assertion of this lemma is an extension of results in [9], where a similar result is derived for *symmetric* points ρ_j . The details are given in [8].

Theorem 1. The iteration error $e^{\text{new}} = p^{\text{new}} - p^*$ satisfies estimates

$$\left| \frac{d^\kappa}{dt^\kappa} e^{\text{new}}(\tau_i + t) \right| = \begin{cases} O(H^m \|e\|_m) + O(H^q \|e\|_q) + O(H^u), & \kappa = 0, \\ O(H^{m+1-\kappa} \|e\|_m) + O(H^q \|e\|_{q-1+\kappa}) + O(H^{u+1-\kappa}), \\ \quad \kappa = 1, \dots, m-q, \\ O(H^{m+1-\kappa} \|e\|_m) + O(H^{u+1-\kappa}), & \kappa = m-q+1, \dots, m \end{cases} \quad (90)$$

for each $u \geq 0$. If the collocation abscissae ρ_j from (5) satisfy the condition (88), then the estimate (90) for $\kappa = m$ can be replaced by the sharper bound

$$\left| \frac{d^m}{dt^m} e^{\text{new}}(\tau_i + t) \right| = O(H^2 \|e\|_m) + O(H^{u+1-m}). \quad (91)$$

Proof. Define the new iteration error $e^{\text{new}} = (e_0^{\text{new}}, e_1^{\text{new}}, \dots, e_{N_1-1}^{\text{new}})$, where $e_i^{\text{new}}(t)$, $i = 1, \dots, N_1 - 1$, is the polynomial of degree $\leq m$ which interpolates $\varepsilon_i(t)$ at $t_{i,0}, \dots, t_{i,m}$. Here $\varepsilon_i(t)$ denotes the first n components of $\tilde{\varepsilon}_i(t)$, where the last, vanishing component is neglected. We rewrite $e_i^{\text{new}}(t)$ as

$$e_i^{\text{new}}(t) = e_{i;u}^{\text{new}}(t) + (e_i^{\text{new}}(t) - e_{i;u}^{\text{new}}(t)),$$

where $e_{i;u}^{\text{new}}(t)$ interpolates the first n components $\varepsilon_{i;u}(t)$ of $\tilde{\varepsilon}_{i;u}(t)$, and $e_i^{\text{new}}(t) - e_{i;u}^{\text{new}}(t)$ interpolates the remainder term $\varepsilon_i(t) - \varepsilon_{i;u}(t) = O(H^{u+1})$, cf. Lemma 8. From Lemma 3 we conclude

$$|e_{i;u}^{\text{new}(\kappa)}(t) - \varepsilon_{i;u}^{(\kappa)}(t)| \leq \text{const.} \max_{t \in J_i} |\varepsilon_{i;u}^{(\kappa)}(t)|, \quad t \in J_i,$$

whence

$$\frac{d^\kappa}{dt^\kappa} e_{i;u}^{\text{new}}(t) \leq \text{const.} \max_{t \in J_i} |\varepsilon_{i;u}^{(\kappa)}(t)|, \quad t \in J_i, \quad \kappa = 0, \dots, m, \quad (92)$$

follows by the triangle inequality.

For the interpolant of the remainder $\varepsilon_i(t) - \varepsilon_{i,u}(t) = O(H^{u+1})$ Lemma 3 implies

$$\frac{d^\kappa}{dt^\kappa} (e_i^{\text{new}}(t) - e_{i,u}^{\text{new}}(t)) = O(H^{u+1-\kappa}), \quad t \in J_i, \quad \kappa = 0, \dots, m. \quad (93)$$

From (92) and (93) together with Lemma 8, (90) now follows.

Finally, we prove (91). The bound $O(H\|e\|_m)$ in (90) for $\kappa = m$ is in fact a consequence of (84) when

$$\mathbf{D}_i(\tau_i) = D_{i,j}^{(k-1)}(\tau_i), \quad \mathbf{d}_i^*(\tau_i) = d_{i,j}^{*(k-1)}(\tau_i), \quad \mathbf{d}_i(\tau_i) = d_{i,j}^{(k-1)}(\tau_i). \quad (94)$$

All other contributions to the bounds (90) are actually $O(H^2\|e\|_m)$, see [8].

The terms given by (94) constitute components of

$$\begin{aligned} \tilde{\varepsilon}_{i;\mathcal{D}}(\tau_i + t) &:= \sum_{k=1}^{\infty} \frac{t^k}{k!} \left(-\hat{\mathcal{D}}_i^k \text{Id}(\tilde{\boldsymbol{\pi}}_i) - \mathcal{D}_i^{*k} \text{Id}(\tilde{\boldsymbol{\pi}}_i^*) + \mathcal{D}_i^k \text{Id}(\tilde{\boldsymbol{\pi}}_i) \right) \\ &= \sum_{k=1}^{\infty} \frac{t^k}{k!} \left(-\tilde{\mathcal{D}}_i^{(k-1)}(\tau_i) - \tilde{\mathbf{d}}_i^{*(k-1)}(\tau_i) + \tilde{\mathbf{d}}_i^{(k-1)}(\tau_i) \right) \\ &= \int_{\tau_i}^{\tau_i+t} \left(-\tilde{\mathcal{D}}_i(\sigma) - \tilde{\mathbf{d}}_i^*(\sigma) + \tilde{\mathbf{d}}_i(\sigma) \right) d\sigma. \end{aligned} \quad (95)$$

Now we rewrite the first n components of $\tilde{\varepsilon}_i(\tau_i + t)$ as

$$\varepsilon_i(\tau_i + t) = \varepsilon_{i;\mathcal{D}}(\tau_i + t) + (\varepsilon_i(\tau_i + t) - \varepsilon_{i;\mathcal{D}}(\tau_i + t)),$$

and treat the two terms separately in the interpolation process outlined above. Thus, let $e_{i;\mathcal{D}}^{\text{new}}(t)$ be the polynomial of degree $\leq m$ which interpolates $\varepsilon_{i;\mathcal{D}}(t)$ at $t_{i,0}, \dots, t_{i,m}$. Clearly, for $0 \leq t \leq H$,

$$\frac{d^m}{dt^m} (e_i^{\text{new}}(\tau_i + t) - e_{i;\mathcal{D}}^{\text{new}}(\tau_i + t)) = O(H^2\|e\|_m) + O(H^{u+1-m}). \quad (96)$$

Noting that $D_i(t)$ is the interpolation polynomial of $d_i(t) - d_i^*(t)$ at $\sigma_{i,1}, \dots, \sigma_{i,m}$, we now apply Lemma 9 component-wise with

$$\begin{aligned} t_0 &= \tau_i, \quad y(\tau_i + t) = \int_{\tau_i}^{\tau_i+t} (d_{i,j}(\sigma) - d_{i,j}^*(\sigma)) d\sigma, \\ q(\tau_i + t) &= \int_{\tau_i}^{\tau_i+t} D_{i,j}(\sigma) d\sigma, \quad p(\tau_i + t) - q(\tau_i + t) = e_{i,j;\mathcal{D}}^{\text{new}}(\tau_i + t). \end{aligned}$$

We conclude that

$$\frac{d^m}{dt^m} e_{i,j;\mathcal{D}}^{\text{new}}(\tau_i + t) = O(H^2\|e\|_m), \quad j = 1, \dots, n, \quad 0 \leq t \leq H, \quad (97)$$

if condition (88) is satisfied, since

$$y^{(m+2)}(t) = \frac{d^{m+1}}{dt^{m+1}}(f_j(p_i(t)) - f_j(p_i^*(t))) = O(\|e\|_m) \quad (98)$$

for $t \in J_i$, using Lipschitz conditions for the derivatives of $f_j(y)$, and Lemma 2. Together with (96) this completes the proof of (91).

References

- [1] W. Auzinger, H. Hofstätter, O. Koch, W. Kreuzer and E. Weinmüller, Superconvergent defect correction algorithms, *WSEAS Transactions on Systems* 4 (2004) 1378–1383.
- [2] S. Blanes and F. Casas, Raising the order of geometric numerical integrators by composition and extrapolation, *Numer. Algorithms* 38 (2005) 305–326.
- [3] R. Frank, The method of iterated defect correction and its application to two-point boundary value problems, Part I, *Numer. Math.* 25 (1976) 409–419.
- [4] E. Hairer, C. Lubich and G. Wanner, *Geometric Numerical Integration* (Springer, Berlin Heidelberg New York, 2002).
- [5] E. Hairer, C. Lubich, and G. Wanner, Geometric numerical integration illustrated by the Störmer/Verlet method, *Acta Numerica*, (2003) 1–51.
- [6] E. Hairer, S. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I* (Springer, Berlin Heidelberg New York, 1987).
- [7] H. Hofstätter and O. Koch, *Splitting Defect Correction*, AURORA TR-2003–23, Inst. for Appl. Math. and Numer. Anal., Vienna Univ. of Technology, Austria, 2003. Available at <http://www.vcpc.univie.ac.at/aurora/publications/>.
- [8] H. Hofstätter and O. Koch, *Convergence Proof for Iterated Splitting Defect Correction*, AURORA TR-2004–05, Inst. for Anal. and Sci. Comput., Vienna Univ. of Technology, Austria, 2004. Available at <http://www.vcpc.univie.ac.at/aurora/publications/>.
- [9] K. Schild, Gaussian collocation via defect correction, *Numer. Math.* 58 (1990) 369–386.
- [10] H.J. Stetter, The defect correction principle and discretization methods, *Numer. Math.* 29 (1978) 425–443.
- [11] P. Zadunaisky, On the estimation of errors propagated in the numerical integration of ODEs, *Numer. Math.* 27 (1976) 21–39.