BIT 27 (1987), 350–367

# IDeC – CONVERGENCE INDEPENDENT OF ERROR ASYMPTOTICS

W. AUZINGER and J. P. MONNET

*Institut für Angewandte und Numerische Mathematik, Technische Universität Wien, Wiedner Hauptstrasse 6-10, A-1040 Wien, Austria*

**Abstract.**

The present paper is concerned with the method of Iterated Defect Correction (IDeC) for two-point boundary value problems. We investigate the contractive behaviour of the IDeC iteration in a completely discrete setting. Our results (which are a generalization of "classical" results based on asymptotic expansions of the discretization error) imply the stability of the collocation method which defines the fixed point of the IDeC iteration.

*1980 AMS Subject Classification:* 65B05, 65L10.

## 1. Introduction.

In this paper we present an analysis of the so-called method of "Iterated Defect Correction" (IDeC method) for the efficient and highly accurate numerical solution of two-point boundary value problems (BVP) of the type

$$y''(t) = f(t, y(t)), \quad t \in (0, 1), \quad y(0) = \alpha, \quad y(1) = \beta. \tag{1.1}$$

The idea behind the IDeC method (and similar methods of defect/deferred correction type) is to combine a simple discretization (which will be called "basic discretization" in the following) with another, very accurate "target discretization" in a (semi-)iterative manner in order to obtain a high-order approximate solution to the given problem. Only the basic scheme is inverted; the target scheme is used to define defects. The particular version considered in the present paper uses defect definitions by collocating polynomials; it is essentially due to Frank ([4], [5]).

The "classical" analysis of the method, as, for instance, presented by Frank [5], makes essential use of asymptotic error expansions (for the basic discretization). In [5] it is shown that, if such an expansion exists up to a sufficient length, a high-order approximation is obtained after a certain ($O(1)$) number of steps. These classical results show what can be expected under optimal circumstances.

Received October 1986. Revised April 1987.



However, a further analysis of the IDeC method from a more general point of view is of interest for several reasons:

i) When the discrete equations are (approximately) solved by some numerical procedure, the result will not always obey the laws of error asymptotics.
ii) In some situations (i.e., for certain variants of deferred correction algorithms) the theoretical justification in the classical sense remains incomplete.
iii) There may be more general classes of problems (where asymptotic error expansions do not exist or exist only in some weaker sense) which can be tackled by defect correction techniques but require a more refined analysis.

For example, the questions i) and iii) arise if one aims at applying defect correction techniques to partial differential equations (cf. [1], [6]). [1] contains an analysis for a specific PDE problem, which does not require the existence of an asymptotic error expansion: The contractivity of the IDeC operator and its interrelation with error smoothness are discussed in an entirely discrete setting. For first order ODE's such a type of analysis has been done by Christiansen and Russell [2] and by Skeel [11]. (See also Fox [3] and Lees [9]).

A new situation arises in the analysis of the IDeC method with collocating polynomials, applied to (1.1). As our analysis will show, the principal part of the IDeC operator is *nilpotent* of low index. From this we shall be able to derive estimates for the contractive power of the iteration operator. These results can be used to conclude the stability of the target discretization. We hope that our work may also be helpful for more general situations.

Some preliminaries: We assume that $f(t, y)$ is Lipschitz continuous w.r.t. $y$ and satisfies the monotonicity condition

$$\frac{\partial}{\partial y} f(t, y) \geq 0. \tag{1.2}$$

The exact solution of the BVP (1.1) is denoted by $y^*(t)$. The "basic discretization" underlying the IDeC method is the usual stable, second order difference scheme

$$\frac{1}{h^2}[\eta_{h,\nu-1} - 2\eta_{h,\nu} + \eta_{h,\nu+1}] = f(t_\nu, \eta_{h,\nu}), \qquad \nu = 1, \ldots, N-1, \tag{1.3}$$

$$\eta_{h,0} = \alpha, \qquad \eta_{h,N} = \beta,$$

on an equidistant grid with $h = 1/N$ and $t_\nu = \nu h$, $\nu = 0, \ldots, N$. The target discretization is described in Section 2.

The present paper is divided into 5 sections. After a review of the IDeC method, presented in Section 2, we investigate the principal linear part of the IDeC operator in Section 3. Sections 4 and 5 deal with the general linear and nonlinear case, respectively. For some of the details we shall refer to Monnet [10].

## 2. The IDeC method.

Let us first introduce some operator notation. The given problem (1.1) is written as

$$Fy = 0 \tag{2.1}$$

with $F: B \to C[0,1] \times \mathbb{R}^2$, $B \subset C^2[0,1]$. The discrete analogue of a function $y(t)$ is a "grid function" $\eta_h$:

$$\eta_h(t_\nu) = \eta_{h,\nu}, \qquad \nu = 0, \ldots, N. \tag{2.2}$$

The space of all grid functions over the $h$-grid will be denoted by $\mathscr{E}_h$. Let $\mathscr{E}_h^0$ denote the subspace

$$\mathscr{E}_h^0 := \{\eta_h \in \mathscr{E}_h: \ \eta_{h,0} = \eta_{h,N} = 0\}. \tag{2.3}$$

The basic discretization (1.3) of (1.1) reads

$$\tilde{F}_h \eta_h = 0, \tag{2.4}$$

where $\tilde{F}_h: \mathscr{E}_h \to \mathscr{E}_h$ is defined by

$$\tilde{F}_h\eta_h(t_\nu) := \begin{cases} \eta_{h,0} - \alpha, & \nu = 0, \\ h^{-2}[\eta_{h,\nu-1} - 2\eta_{h,\nu} + \eta_{h,\nu+1}] - f(t_\nu, \eta_{h,\nu}), & \nu = 1, \ldots, N-1, \\ \eta_{h,N} - \beta, & \nu = N. \end{cases} \tag{2.5}$$

The principle of defect correction, applied to a discretization method, can be described as follows: Introduce a further discrete operator $F_h: \mathscr{E}_h \to \mathscr{E}_h$ which defines defects. This is only reasonable if $F_h$ is in some sense "closer to $F$" than $\tilde{F}_h$; but only $\tilde{F}_h$ needs to be inverted in the following. Denoting the solution operator of equation (2.4) by $\tilde{G}_h$, the so-called version B of the IDeC in the sense of Stetter [12] reads ($\eta_h^{(0)}$ is an initial guess):

$$\tilde{F}_h\eta_h^{(i+1)} := \Phi_h\eta_h^{(i)} = (\Phi_h\tilde{G}_h)\tilde{F}_h\eta_h^{(i)}, \qquad \text{where} \tag{2.6a}$$

$$\Phi_h := \tilde{F}_h - F_h. \tag{2.6b}$$

Our analysis of (2.6) is based on estimates for the contractive power of $\Phi_h\tilde{G}_h$ (the iteration operator for the $\tilde{F}_h\eta_h^{(i)}$).

We shall now define our defect-defining (or "target"-) operator $F_h$; the essential ideas are due to Frank [4]. Assume that the equidistant grid is such that



$N = nm$, where $m \geq 4$ is kept fixed and $n \to \infty$ as $h \to 0$; see Fig. 1. An intermediate approximation $\eta_h = \eta_h^{(i)}$ is interpolated by a piecewise polynomial function $p_h(t)$. I.e., $p_h(t)$ is the continuous concatenation of $n$ interpolating polynomials $p_{l,h}(t)$ of degree $\leq m$ over the subintervals

$$I_l := [t_{(l-1)m}, t_{lm}] = [(l-1)mh, lmh], \qquad l = 1, \ldots, n. \tag{2.7}$$

This interpolation can be represented by a linear operator $P_h: \mathscr{E}_h \to \hat{C}_h^2[0,1]$, where $\hat{C}_h^2[0,1] \subset C[0,1]$ denotes the ($h$-dependent) space of continuous functions which are $C^2$ in each subinterval $I_l$ but may have a jump in the first derivative at the endpoints $t = t_{lm}$, $l = 1, \ldots, n-1$, of these intervals (Fig. 1). (At these points the defect will have to be defined very carefully, but let us ignore this for the moment.)

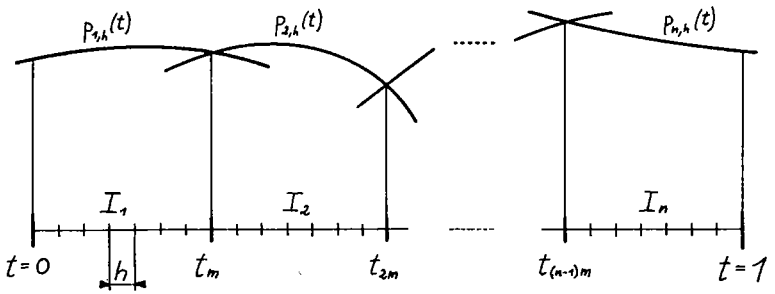


Fig. 1. $p_h(t) \in \hat{C}_h^2[0,1]$.

We can now apply (in a piecewise manner) the continuous operator $F$ to the interpolating function $p_h(t) = P_h\eta_h(t)$:

$$Fp_h(t) = \begin{cases} p_h(0) - \alpha, & t = 0, \\ p_h''(t) - f(t, p_h(t)), & t \in (0,1), \\ p_h(1) - \beta, & t = 1. \end{cases} \tag{2.8}$$

What remains to be done is to restrict the defect given by (2.8) to the discrete space $\mathscr{E}_h$. We shall use straightforward point evaluation at $t = t_\nu$. The corresponding restriction operator will be denoted by $R_h$.

As already mentioned, there remains a flaw in our definition of the defect: $p_h''(t)$ is not well-defined at the endpoints $t_{lm}$, $l = 1, \ldots, n-1$, of the subintervals $I_l$. Due to Frank [4], (2.8) is modified in the following way: Consider a function $p_h(t) \in \hat{C}_h^2[0,1]$ (represented by some $p_{l,h}(t)$ in the subinterval $I_l$) and define

(2.9) $\hat{F}_h p_h(t) :=$

$$\begin{cases} h^{-1}[p'_{i+1,h}(t)-p'_{i,h}(t)]+\frac{1}{2}[p''_{i,h}(t)+p''_{i+1,h}(t)]-f(t,p_h(t)), & \text{for } t=t_{im}, \\ Fp_h(t), & \text{otherwise.} \end{cases}$$

$\hat{F}_h$ is an extension of $F$ to the space $\hat{C}_h^2[0,1]$ which takes possible jumps in the derivatives into account. (For $y(t)\in C^2[0,1]$ it is obvious that $\hat{F}_h y = Fy$.) The introduction of the "jump defect" $h^{-1}[p'_{i+1,h}(t)-p'_{i,h}(t)]$ within (2.9) is essential; it can be motivated as follows. Since our goal is convergence to the smooth solution $y^*(t)$ of the BVP, it would certainly be unwise to ignore the jumps in the derivatives which inevitably occur. Moreover, for any pair $\tilde{F}_h$, $F_h$ to be used within an IDeC algorithm it is highly desirable that

(2.10) $$F_h = \tilde{F}_h + \textit{higher order corrections};$$

it will turn out in Section 3 that (2.10) is achieved by the particular form of (2.9) (cf. (3.5)–(3.8) below).

The target operator is now defined as

(2.11) $$F_h := R_h \hat{F}_h P_h.$$

The equation $F_h\eta_h = 0$ for the fixed point $\eta_h^*$ of the IDeC iteration defines a certain *collocation method* for the numerical solution of $Fy = 0$: $\eta_h^*$ is determined by the discrete values of a piecewise polynomial function for which the defect in the sense of (2.9), (2.11) vanishes. It can be shown by standard arguments that this collocation method has the usual consistency properties:

PROPOSITION 2.1:

(2.12) $$\|F_h R_h y^*\|_\infty = O(h^{m-1}),$$

*if the solution $y^*(t)$ of* (1.1) *is smooth enough, i.e., $y^* \in C^{m+1}[0,1]$.*

PROOF: See [10]. ■

In fact, $m-1$ is the optimal order which can be expected from our IDeC method (cf. [4], [5]). But to establish definitively the solvability of $F_h\eta_h^* = 0$ and the approximation order of the fixed point, i.e.,

(2.13) $$\|\eta_h^* - R_h y^*\|_\infty = O(h^{m-1}),$$

it remains to be shown that $F_h$ is stable. We shall demonstrate in Sections 4 and 5 how the local stability of $F_h$ can be concluded from the contractive behaviour of the linearized IDeC operator (cf. Theorems 4.3 and 5.1).



In the following sections we investigate the contractivity properties of $\Phi_h\tilde{G}_h$. As mentioned above, it will turn out that the incorporation of the "jump defect" within (2.9) is essential; we shall see in Section 3 that $\Phi_h$ is $h^{-2}$ times a difference operator of order 4. (Note that, by definition of $F_h$, $\Phi_h = \tilde{F}_h - F_h$ is always linear and does not depend on $f(t,y)$.)

## 3. Contractivity of the IDeC: the special case $f(t, y) = g(t)$.

In the simplest case $f(t,y) = g(t)$ the basic discretization (2.4), (2.5) reduces to the linear system

(3.1) $$\tilde{L}_h\eta_h = g_h,$$

where $\tilde{L}_h: \mathscr{E}_h \to \mathscr{E}_h$ is the principal, linear part of the general $\tilde{F}_h$:

(3.2) $$\tilde{L}_h\eta_h(t_\nu) := h^{-2}[\eta_{h,\nu-1} - 2\eta_{h,\nu} + \eta_{h,\nu+1}], \qquad \nu = 1,\dots,N-1.$$

It is well known that $\tilde{L}_h$ is stable with $\|\tilde{L}_h^{-1}\|_\infty \le \frac{1}{8}$. Further, $g_h(t_\nu)$ is given by $g(t_\nu)$ in the interior grid points and by $g_h(0) = \alpha, g_h(1) = \beta$ (boundary conditions). Now $\Phi_h\tilde{G}_h$ is affine with the linear part $\Phi_h\tilde{L}_h^{-1}$. Since the boundary conditions are always trivially satisfied, we can restrict our further considerations to the subspace $\mathscr{E}_h^0 \subset \mathscr{E}_h$ of zero boundary values.

Similarly to (3.1), $F_h\eta_h = 0$ can be written as $L_h\eta_h = g_h$. Thus, $\Phi_h = \tilde{L}_h - L_h$.

The following theorem is due to Monnet [10]. (Recall that $m$ is the degree of the interpolating polynomials defining $L_h$.)

THEOREM 3.1. *$\Phi_h\tilde{L}_h^{-1}: \mathscr{E}_h^0 \to \mathscr{E}_h^0$ is nilpotent of degree $\lfloor (m+1)/2 \rfloor$:*

(3.3) $$(\Phi_h\tilde{L}_h^{-1})^{\lfloor (m+1)/2 \rfloor} = 0, \qquad \varrho(\Phi_h\tilde{L}_h^{-1}) = 0.$$

By regularity of $\tilde{L}_h$ this is also true for $\tilde{L}_h^{-1}\Phi_h$. In order to prove Theorem 3.1 we first illustrate the important role of the "jump defect".

Define subspaces $\mathscr{U}_h^{(i)}, \mathscr{V}_h^{(i)} \subset \mathscr{E}_h^0$ $(i \ge 0)$ in the following way:

i) $\mathscr{U}_h^{(i)}$ := space of those grid functions $u_h \in \mathscr{E}_h^0$ where $u_h(t_\nu)$ is the projection of the value of a polynomial $p_{l,h}(t)$ of degree $\le i$ in the interior points of the subinteryals $I_l$ (that means, for $\nu = (l-1)m+1,\dots,lm-1$, $l = 1,\dots,n-1$,), but is arbitrary for $\nu = lm$, $l = 1,\dots,n-1$.

ii) $\mathscr{V}_h^{(i)}$ := space of those grid functions $v_h \in \mathscr{E}_h^0$ which are the projections of some continuous piecewise polynomial functions (degree $\le i$) from $\hat{C}_h^2[0,1]$ (jumps in the first derivatives allowed at $t_\nu = t_{lm}$).

Since there are $m-1$ grid points in the interior of each of the subintervals $I_l$,

it can easily be seen that

$$\mathcal{U}_h^{(m-2)} = \mathcal{V}_h^{(m)} = \mathcal{E}_h^0, \qquad m \geq 2. \tag{3.4}$$

In addition, let $\mathcal{U}_h^{(-3)} := \mathcal{U}_h^{(-2)} := \mathcal{O}_h$ (null space), and let $\mathcal{U}_h^{(-1)} \subset \mathcal{E}_h^0$ denote the space of grid functions which vanish except at $t_\nu = t_{lm}$, $l = 1, \ldots, n-1$. Fig. 2 and 3 illustrate $u_h \in \mathcal{U}_h^{(-1)}$, $v_h \in \mathcal{V}_h^{(1)}$.

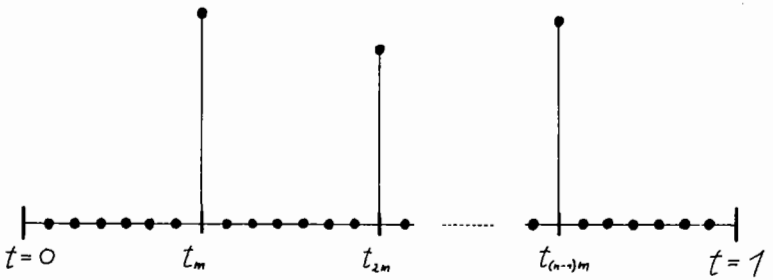


Fig. 2. $u_h \in \mathcal{U}_h^{(-1)}$.

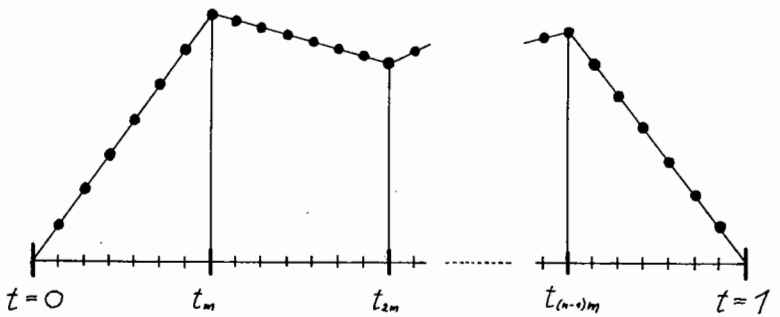


Fig. 3. $v_h \in \mathcal{V}_h^{(1)}$.

Consider a grid function $u_h \in \mathcal{U}_h^{(-1)}$ and let $v_h := \tilde{L}_h^{-1} u_h$. By definition, the second difference quotient $\tilde{L}_h v_h(t_\nu)$ vanishes at the grid points $t_\nu$ in the interior of the subintervals $I_l$. Hence $v_h \in \mathcal{V}_h^{(1)}$, i.e., $\tilde{L}_h^{-1}(\mathcal{U}_h^{(-1)}) \subset \mathcal{V}_h^{(1)}$. On the other hand, $\tilde{L}_h v_h \in \mathcal{U}_h^{(-1)}$ if $v_h \in \mathcal{V}_h^{(1)}$. Due to the definition of $F_h$ (cf. (2.9), (2.11)) the difference operator $L_h$ maps $\mathcal{V}_h^{(1)}$ into $\mathcal{U}_h^{(-1)}$, too. The crucial point is that actually

$$\tilde{L}_h v_h = L_h v_h, \qquad v_h \in \mathcal{V}_h^{(1)}, \tag{3.5}$$

holds. This would *not* be true if we had omitted the term involving the



difference of first derivatives within (2.9)! (In that case, $L_h v_h = 0$ for all $v_h \in \mathcal{V}_h^{(1)}$.) Thus we have shown $\Phi_h(\mathcal{V}_h^{(1)}) = \mathcal{O}_h$ and

$$(\Phi_h \tilde{L}_h^{-1})(\mathcal{U}_h^{(-1)}) \subset \mathcal{U}_h^{(-3)} = \mathcal{O}_h. \tag{3.6}$$

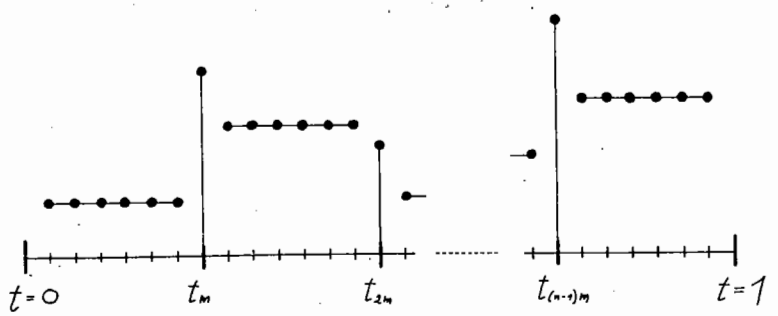


Fig. 4. $u_h \in \mathcal{U}_h^{(0)}$.

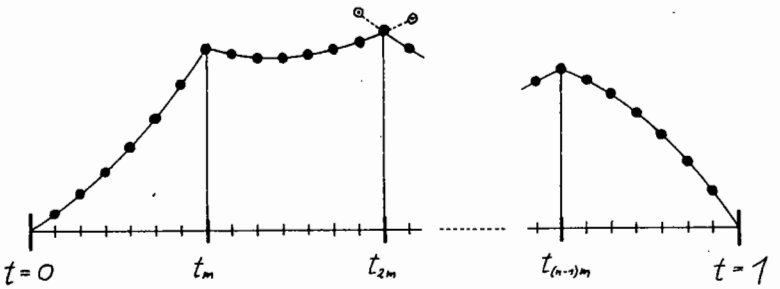


Fig. 5. $v_h \in \mathcal{V}_h^{(2)}$.

Now let $u_h \in \mathcal{U}_h^{(0)}$ and $v_h := \tilde{L}_h^{-1} u_h$ (see Fig. 4 and 5). $\tilde{L}_h v_h$ is piecewise constant in the interior of the subintervals $I_l$; thus, $v_h$ can be interpolated by a piecewise polynomial function of degree 2, i.e., $v_h \in \mathcal{V}_h^{(2)}$. Thus, $\tilde{L}_h^{-1}(\mathcal{U}_h^{(0)}) \subset \mathcal{V}_h^{(2)}$. For $v_h \in \mathcal{V}_h^{(2)}$, $\tilde{L}_h v_h \in \mathcal{U}_h^{(0)}$ is trivial. It is easy to see that $L_h v_h(t_\nu) = \tilde{L}_h v_h(t_\nu)$ in the interior of the subintervals. The situation at the points $t = t_{lm}$ is illustrated in Fig. 6. For the polynomials $p_{l,h}(t)$, $p_{l+1,h}(t)$ of degree 2 interpolating $v_h$, the first and second derivatives within (2.9) are exactly given by the first and second *central* difference quotients. Fig. 6 shows the resulting weights in

$$h^{-1}[p'_{l+1,h}(t) - p'_{l,h}(t)] + \tfrac{1}{2}[p''_{l,h}(t) + p''_{l+1,h}(t)] \tag{3.7}$$

(cf. (2.9)). We see that the quantities $p_{l,h}(t+h)$, $p_{l+1,h}(t-h)$ are eliminated by the

combination of terms in (3.7). We have again $L_h v_h(t_\nu) = \bar{L}_h v_h(t_\nu)$ due to the presence of the jump defect $h^{-1}[p'_{l+1,h}(t) - p'_{l,h}(t)]$. Thus, $\Phi_h(\mathcal{V}_h^{(2)}) = \mathcal{O}_h$, and together with $\bar{L}_h^{-1}(\mathcal{U}_h^{(0)}) \subset \mathcal{V}_h^{(2)}$ we obtain

$$(\Phi_h \bar{L}_h^{-1})(\mathcal{U}_h^{(0)}) \subset \mathcal{U}_h^{(-2)} = \mathcal{O}_h. \tag{3.8}$$

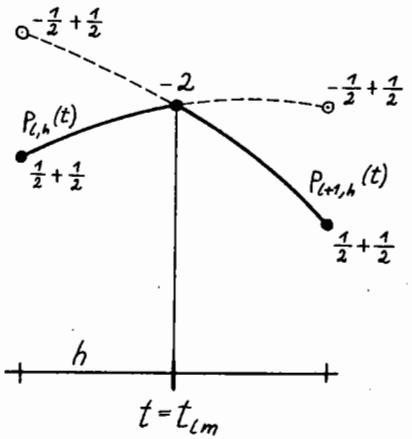


Fig. 6. $L_h v_h(t_{lm})$, $v_h \in \mathcal{V}_h^{(2)}$.

With (3.6) and (3.8), which are essential, it is simple to prove Theorem 3.1. To complete the proof, we shall at first derive some further properties of the operator $\Phi_h$ which will also be needed for the analysis of the general linear case in Section 4. To this end we adopt the following notation for operations with finite differences. Let $y(t)$ be any function and define

$$\begin{aligned} \sigma_h y(t) &:= y(t+h), \\ \partial_h y(t) &:= y(t+h) - y(t) = (\sigma_h - I)y(t), \\ \bar{\partial}_h y(t) &:= y(t) - y(t-h) = (I - \sigma_h^{-1})y(t), \\ \delta_h y(t) &:= y(t+h/2) - y(t-h/2). \end{aligned} \tag{3.9}$$

We think of the symbols $\sigma_h$, $\partial_h$, $\bar{\partial}_h$, $\delta_h$ as *operators*; in this sense, powers and general functions of these operators are well-defined, e.g., $\sigma_h^{-1} y(t) = y(t-h)$. (For an introduction to this operational concept see for instance Hildebrand [7].) Note that the operation $\delta_h$ is not well-defined for *grid* functions. But due to

$$\delta_h^2 y(t) = (\sigma_h^{-1} - 2I + \sigma_h)y(t) = (\partial_h - \bar{\partial}_h)y(t), \tag{3.10}$$

*even* powers of $\delta_h$ can be applied to $\eta_h \in \mathcal{E}_h$.



We shall need some properties of the operators (3.9) applied to polynomials $p(t)$ of degree $\leq m$. First of all, the differences $\partial_h^j p(t), \ldots$ vanish for $j > m$. The identities

$$\begin{aligned} \sigma_h p(t) &= [I - \bar{\partial}_h]^{-1} p(t) = [I + \bar{\partial}_h + \ldots + \bar{\partial}_h^m] p(t), \\ \sigma_h^{-1} p(t) &= [I + \partial_h]^{-1} p(t) = [I - \partial_h + \ldots + (-1)^m \partial_h^m] p(t), \end{aligned} \tag{3.11}$$

allow the representation of $p(t+ih)$, $i > 0$, by backward differences, and vice versa. Together with (3.10) we obtain analogous expansions for $\delta_h^2$:

$$\begin{aligned} &\text{(a)} \quad \delta_h^2 p(t) = [\bar{\partial}_h^2 + \ldots + \bar{\partial}_h^m] p(t), \\ &\text{(b)} \quad \delta_h^2 p(t) = [\partial_h^2 + \ldots + (-1)^m \partial_h^m] p(t). \end{aligned} \tag{3.12}$$

*Derivatives* of the polynomial $p(t)$ can be expanded into finite differences. In particular,

$$\begin{aligned} &\text{(a)} \quad p'(t) = \frac{1}{h}\left[\bar{\partial}_h + \frac{1}{2}\bar{\partial}_h^2 + \ldots + \frac{1}{m}\bar{\partial}_h^m\right] p(t), \\ &\text{(b)} \quad p'(t) = \frac{1}{h}\left[\partial_h - \frac{1}{2}\partial_h^2 + \ldots + (-1)^{m+1}\frac{1}{m}\partial_h^m\right] p(t). \end{aligned} \tag{3.13}$$

Furthermore, $p''(t)$ can be expanded into powers of $\delta_h^2$. The coefficients of this expansion are derived in [7]:

$$\begin{aligned} p''(t) &= \frac{1}{h^2}\left[\delta_h - \frac{1^2}{2^2 3!}\delta_h^3 + \frac{1^2 3^2}{2^4 5!}\delta_h^5 - \frac{1^2 3^2 5^2}{2^6 7!}\delta_h^7 + \ldots\right]^2 p(t) = \\ &= \frac{1}{h^2}\left[\delta_h^2 - \frac{1}{12}\delta_h^4 + \frac{1}{90}\delta_h^6 - \frac{1}{560}\delta_h^8 + \ldots\right] p(t). \end{aligned} \tag{3.14}$$

With these preliminaries we can show that $\Phi_h$ is a fourth order difference operator:

LEMMA 3.2. *Let $\eta_h \in \mathcal{E}_h$ and $p_{l,h}(t)$ (degree $\leq m$) interpolate $\eta_h(t_\nu)$ in $I_l$, $l = 1, .., n$. For $t_\nu$ in the interior of $I_l$,*

$$\Phi_h \eta_h(t_\nu) = (12h^2)^{-1} \delta_h^4 p_{l,h}(t_\nu) + \ldots. \tag{3.15a}$$

*For $t_\nu = t_{lm}$, $l = 1, \ldots, n-1$,*

$$\Phi_h \eta_h(t_\nu) = (24h^2)^{-1}[-5\sigma_h^{-2} + 4\sigma_h^{-1} + 4I + 4\sigma_h - 5\sigma_h^2]\delta_h^4 \eta_h(t_\nu) + \ldots. \tag{3.15b}$$

*The remainder terms in these expansions can be expressed by certain linear combinations of $h^{-2}$ times jth differences of $\eta_h$, $5 \le j \le m$.*

PROOF: (3.15a) is an immediate consequence of (3.14) and of

$$\tilde{L}_h \eta_h(t_\nu) = h^{-2}\delta_h^2 \eta_h(t_\nu).$$

The remainder terms originating from (3.14) involve values of $p_{l,h}(t)$ outside of $I_l$ which can be "shifted" to $I_l$ on the basis of (3.11); the resulting expressions consist of differences of order $\ge 5$ over $I_l$.

For the proof of (3.15b) we express (3.7) by finite differences, using the representation (3.14) for $t = t_{lm}$. Expanding $p'_{l,h}(t)$, $\delta_h^2 p_{l,h}(t)$ into backward differences (cf. (3.12a), (3.13a)) and $p'_{l+1,h}(t)$, $\delta_h^2 p_{l+1,h}(t)$ into forward differences (cf. (3.12b), (3.13b)), and using analogous expansions for $\delta_h^4 p_{l,h}(t)$, $\delta_h^4 p_{l+1,h}(t)$, we obtain

$$\begin{aligned}
&\frac{1}{h}\,[p'_{l+1,h}(t) - p'_{l,h}(t)] + \tfrac{1}{2}[p''_{l,h}(t) + p''_{l+1,h}(t)] = \\
&= \frac{1}{h^2}[\partial_h - \tfrac{1}{2}\partial_h^2 + \tfrac{1}{3}\partial_h^3 - \tfrac{1}{4}\partial_h^4 + \ldots] p_{l+1,h}(t) \\
&\quad - \frac{1}{h^2}[\bar{\partial}_h + \tfrac{1}{2}\bar{\partial}_h^2 + \tfrac{1}{3}\bar{\partial}_h^3 + \tfrac{1}{4}\bar{\partial}_h^4 + \ldots] p_{l,h}(t) \\
&\quad + \frac{1}{h^2}[\quad \tfrac{1}{2}\bar{\partial}_h^2 + \tfrac{1}{2}\bar{\partial}_h^3 + \tfrac{1}{2}\bar{\partial}_h^4 + \ldots] p_{l,h}(t) - \tfrac{1}{24}\bar{\partial}_h^4 p_{l,h}(t) + \ldots \\
&\quad + \frac{1}{h^2}[\quad \tfrac{1}{2}\partial_h^2 - \tfrac{1}{2}\partial_h^3 + \tfrac{1}{2}\partial_h^4 + \ldots] p_{l+1,h}(t) - \tfrac{1}{24}\partial_h^4 p_{l+1,h}(t) + \ldots = \\
&= \frac{1}{h^2}[(\partial_h - \bar{\partial}_h) - \tfrac{1}{6}(\partial_h^3 - \bar{\partial}_h^3) + \tfrac{5}{24}(\partial_h^4 + \bar{\partial}_h^4)]\eta_h(t) + \ldots;
\end{aligned}$$

the remainder terms are treated similarly as above. Together with

$$\partial_h - \bar{\partial}_h = \delta_h^2, \quad \partial_h^3 - \bar{\partial}_h^3 = (\sigma_h^{-1} + I + \sigma_h)\delta_h^4, \quad \bar{\partial}_h^4 = \sigma_h^{-2}\delta_h^4, \quad \partial_h^4 = \sigma_h^2 \delta_h^4,$$

(3.15b) follows easily. ■

Note that (3.6) and (3.8) can be considered as a consequence of (3.15a, b). A simple consequence of (3.15a) is

LEMMA 3.3.

$$(\Phi_h \tilde{L}_h^{-1})(\mathscr{U}_h^{(i)}) \subset \mathscr{U}_h^{(i-2)}, \qquad 1 \le i \le m-2. \tag{3.16}$$



PROOF: Consider $u_h \in \mathscr{U}_h^{(i)}$, $1 \le i \le m-2$. Let $v_h := \tilde{L}_h^{-1} u_h$ and let $p_{l,h}(t)$ be the polynomial (degree $\le m$) interpolating $v_h(t_\nu)$ in $I_l$; i.e., $p_{l,h}(t_\nu) = v_h(t_\nu)$, $\nu = (l-1)m, \ldots, lm$. Since $h^{-2}\delta_h^2 p_{l,h}(t_\nu) = \tilde{L}_h v_h(t_\nu) = u_h(t_\nu)$ for the grid points $t_\nu$ in the interior of $I_l$, it is easy to conclude that the degree of $p_{l,h}(t)$ is $\le i+2$. Thus,

$$\tilde{L}_h^{-1}(\mathscr{U}_h^{(i)}) \subset \mathscr{V}_h^{(i+2)}, \qquad 1 \le i \le m-2. \tag{3.17a}$$

Now let $v_h \in \mathscr{V}_h^{(i+2)}$ and $u_h := \Phi_h v_h$. Let $q_{l,h}(t)$ be the polynomial (degree $\le m-2$) interpolating $u_h(t_\nu)$ at the inner points of $I_l$: $q_{l,h}(t_\nu) = u_h(t_\nu)$, $\nu = (l-1)m+1, \ldots, lm-1$. It follows from (3.15a) that $q_{l,h}(t)$ is of degree $\le i-2$ because it is a linear combination of differences of order $\ge 4$ of a polynomial of degree $\le i+2$. This shows

$$\Phi_h(\mathscr{V}_h^{(i+2)}) \subset \mathscr{U}_h^{(i-2)}, \qquad 1 \le i \le m-2. \tag{3.17b}$$

(3.17a, b) imply (3.16). ■

Theorem 3.1 now follows as a direct consequence of (3.6) and (3.8) in conjunction with Lemma 3.3 and the fact that $\mathscr{U}_h^{(m-2)} = \mathscr{E}_h^0$ (cf. (3.4)).

The results above show that in the case $f(t, y) = g(t)$ the IDeC iteration attains its fixed point $\eta_h^*$ after a fixed number of $\lfloor (m+1)/2 \rfloor$ steps for *any* initial $\eta_h^{(0)}$. This is not true if one omits the jump term in the defect definition. (In [10] it is shown that $\varrho(\Phi_h \tilde{L}_h^{-1}) \ge 1$ in this case.)

In the sequel we shall use the notation

$$|\eta_h|_2 := \max_{\nu=1,\ldots,N-1} |h^{-2}\delta_h^2 \eta_h(t_\nu)|, \qquad |\eta_h|_4 := \max_{\nu=2,\ldots,N-2} |h^{-4}\delta_h^4 \eta_h(t_\nu)|. \tag{3.18}$$

$C_m$ will always denote a generic constant depending on $m$. A further conclusion from Lemma 3.2 is

COROLLARY 3.4. *For $\eta_h \in \mathscr{E}_h^0$,*

$$\|\Phi_h \tilde{L}_h^{-1} \eta_h\|_\infty \le C_m h^2 |\eta_h|_2 \le C_m \|\eta_h\|_\infty. \tag{3.19}$$

PROOF: The estimate $\|\Phi_h \eta_h\|_\infty \le C_m h^2 |\eta_h|_4$ follows directly from Lemma 3.2, whereas $|\eta_h|_4 \le |\tilde{L}_h \eta_h|_2$ is trivial. This proves (3.19) (the second inequality is obvious because $|\eta_h|_2 \le 4h^{-2}\|\eta_h\|_\infty$). ■

(3.19) shows that the actual *norm* contraction rate of a single IDeC step depends crucially on the smoothness of the error of the initial approximation. We are not discussing this dependence in detail here.*) For the proof of the

*) Note that *classical* estimates (utilizing asymptotic error expansions) are based on inequalities of the type $|\Phi_h \tilde{L}_h^{-1} \eta_h|_k \le Ch^2 |\eta_h|_{k+2}$, $k = 2, 4, \ldots$ (cf. [5]).

stability of the target operator $L_h$ we refer to Section 4 where the general linear case is discussed (cf. Theorem 4.3).

## 4. Contractivity of the IDeC: the general linear case.

Let us now look at the general linear problem where $f(t,y) = \gamma(t)y + g(t)$, $\gamma(t) \geq 0$. (Recall that $\Phi_h = \tilde{L}_h - L_h$ is the same as in the special case of Section 3.) The linear parts of the affine operators $\tilde{F}_h$, $F_h$ will be denoted by $\tilde{H}_h$ and $H_h$, resp.:

$$\tilde{H}_h := \tilde{L}_h - \Gamma_h, \qquad H_h := L_h - \Gamma_h, \tag{4.1a}$$

where

$$\Gamma_h := \operatorname{diag}(\gamma(t_\nu)). \tag{4.1b}$$

($\tilde{H}_h$ is stable because $\gamma(t) \geq 0$.) With

$$c_0 := \max_{0 \leq t \leq 1} |\gamma(t)| \tag{4.2a}$$

we have $\|\Gamma_h\|_\infty \leq c_0$. Let $\gamma(t)$ be sufficiently smooth, i.e., $\gamma(t) \in C^2[0,1]$, such that there exist constants $c_1$ and $c_2$ independent of $h$ with

$$\max_{\nu=0,\ldots,N-1} |\partial_h \gamma(t_\nu)| \leq hc_1, \qquad \max_{\nu=1,\ldots,N-1} |\delta_h^2 \gamma(t_\nu)| \leq h^2 c_2. \tag{4.2b}$$

Our smoothness assumptions enable the following estimate:

LEMMA 4.1. *For* $\eta_h \in \mathring{\mathcal{G}}_h^0$,

$$|\Gamma_h \eta_h|_2 \leq (c_0 + 2c_1 + c_2)|\eta_h|_2. \tag{4.3}$$

PROOF: The product formula

$$\delta_h^2(uv) = u\delta_h^2 v + \bar{\partial}_h u \bar{\partial}_h v + \partial_h u \partial_h v + v\delta_h^2 u$$

yields

$$|\delta_h^2 \Gamma_h \eta_h(t_\nu)| = |\delta_h^2(\gamma\eta_h)(t_\nu)| \leq c_0|\delta_h^2 \eta_h(t_\nu)| + hc_1[|\bar{\partial}_h \eta_h(t_\nu)| + |\partial_h \eta_h(t_\nu)|] + h^2 c_2 |\eta_h(t_\nu)|$$

for all $\nu = 1, \ldots, N-1$. (4.3) immediately follows due to the obvious inequalities

$$\max_{\nu=1,\ldots,N-1} |\eta_h(t_\nu)| \leq h^{-1} \max_{\nu=1,\ldots,N-1} |\bar{\partial}_h \eta_h(t_\nu)| \leq h^{-2} \max_{\nu=1,\ldots,N-1} |\delta_h^2 \eta_h(t_\nu)|. \quad \blacksquare$$



Now we derive norm estimates for $\Phi_h \tilde{H}_h^{-1}$ (which is not nilpotent unless $\gamma(t) \equiv 0$). We denote

$$\bar{m} := \lfloor (m+1)/2 \rfloor. \tag{4.4}$$

THEOREM 4.2.

$$\|\Phi_h \tilde{H}_h^{-1}\|_\infty \leq C_m(1 + c_0/8), \tag{4.5a}$$

*where* $C_m$ *is a bound for* $\|\Phi_h \tilde{L}_h^{-1}\|_\infty$ (*cf.* (3.19)). *Furthermore, there is a constant* $C_m$ *such that*

$$\|(\Phi_h \tilde{H}_h^{-1})^{\bar{m}}\|_\infty \leq C_m h^2. \tag{4.5b}$$

PROOF: (4.5a) follows from

$$\|\Phi_h \tilde{H}_h^{-1}\|_\infty \leq \|\Phi_h \tilde{L}_h^{-1}\|_\infty \|(\tilde{H}_h + \Gamma_h)\tilde{H}_h^{-1}\|_\infty \leq C_m(1 + c_0\|\tilde{H}_h^{-1}\|_\infty),$$

with $\|\tilde{H}_h^{-1}\|_\infty \leq \|\tilde{L}_h^{-1}\|_\infty \leq \frac{1}{8}$ (due to $\gamma(t) \geq 0$).

For the proof of (4.5b) we split $\Phi_h \tilde{H}_h^{-1}$ into

$$\Phi_h \tilde{H}_h^{-1} = \Phi_h \tilde{L}_h^{-1} + \Phi_h \tilde{L}_h^{-1} \Gamma_h \tilde{H}_h^{-1},$$

and estimate the second term separately. The inequality $\|\tilde{L}_h \tilde{H}_h^{-1}\|_\infty \leq 1 + c_0/8$ (cf. the proof of (4.5a) above) yields

$$|\eta_h|_2 = \|\tilde{L}_h \eta_h\|_\infty \leq (1 + c_0/8)\|\tilde{H}_h \eta_h\|_\infty.$$

Together with (3.19) and (4.3) this yields

$$\|\Phi_h \tilde{L}_h^{-1} \Gamma_h \tilde{H}_h^{-1}\|_\infty \leq C_m h^2. \tag{4.6}$$

Now let $A_h := \Phi_h \tilde{L}_h^{-1}$, $B_h := \Phi_h \tilde{L}_h^{-1} \Gamma_h \tilde{H}_h^{-1}$. From Theorem 3.1 we have $A_h^{\bar{m}} = 0$. Thus,

$$\|(\Phi_h \tilde{H}_h^{-1})^{\bar{m}}\|_\infty = \|(A_h + B_h)^{\bar{m}}\|_\infty \leq \|(A_h + B_h)^{\bar{m}} - A_h^{\bar{m}}\|_\infty$$

$$\leq \sum_{l=1}^{\bar{m}} \binom{\bar{m}}{l} \|A_h\|_\infty^{\bar{m}-l} \|B_h\|_\infty^l.$$

Together with $\|A_h\|_\infty \leq C_m$, $\|B_h\|_\infty \leq C_m h^2$ (cf. (4.5a), (4.6)), this implies (4.5b). $\blacksquare$

(4.5b) shows that, starting from *any* initial $\lambda_h^{(0)} = \tilde{F}_h \eta_h^{(0)}$, $\bar{m}$ IDeC steps

reduce the residual $\lambda_h^{(i)} = \tilde{F}_h \eta_h^{(i)}$ by a factor $O(h^2)$. Furthermore, we conclude from Theorem 4.2:

THEOREM 4.3. *$H_h$ is invertible and uniformly stable:*

$$\|H_h^{-1}\|_\infty \le C_m \quad \text{for} \quad h \le h_0. \tag{4.7}$$

PROOF: Due to (4.5b),

$$\varrho(I_h - H_h\tilde{H}_h^{-1}) = \varrho(\Phi_h\tilde{H}_h^{-1}) \le \|(\Phi_h\tilde{H}_h^{-1})^{\bar{m}}\|_\infty^{1/\bar{m}} \le (C_m h^2)^{1/\bar{m}} \le k < 1$$

for sufficiently small $h \le h_0$. Thus, $H_h$ is invertible. To prove stability, we write $(\Phi_h\tilde{H}_h^{-1})^{\bar{m}}$ in the form

$$(I_h - H_h\tilde{H}_h^{-1})^{\bar{m}} = \sum_{l=0}^{\bar{m}} (-1)^l \binom{\bar{m}}{l} (H_h\tilde{H}_h^{-1})^l =$$

$$= I_h - \sum_{l=1}^{\bar{m}} (-1)^{l-1} \binom{\bar{m}}{l} H_h\tilde{H}_h^{-1}(H_h\tilde{H}_h^{-1})^{l-1} =$$

$$= I_h - H_h\tilde{H}_h^{-1} \sum_{l=1}^{\bar{m}} (-1)^{l-1} \binom{\bar{m}}{l} (I_h - \Phi_h\tilde{H}_h^{-1})^{l-1}.$$

Thus, $(I_h - H_h\tilde{H}_h^{-1})^{\bar{m}} = I_h - H_h\hat{G}_h$, where

$$\hat{G}_h := \tilde{H}_h^{-1} \sum_{l=1}^{\bar{m}} (-1)^{l-1} \binom{\bar{m}}{l} (I_h - \Phi_h\tilde{H}_h^{-1})^{l-1}.$$

Due to (4.5a), $\hat{G}_h$ is uniformly bounded. Since $\|I_h - H_h\hat{G}_h\|_\infty \le k < 1$, we obtain

$$\|H_h^{-1}\|_\infty \le \|\hat{G}_h\|_\infty + \|H_h^{-1}(I_h - H_h\hat{G}_h)\|_\infty \le \|\hat{G}_h\|_\infty + k\|H_h^{-1}\|_\infty,$$

and therefore

$$\|H_h^{-1}\|_\infty \le (1-k)^{-1}\|\hat{G}_h\|_\infty, \qquad h \le h_0,$$

which proves (4.7). ∎

This completes our analysis for the linear problem.



## 5. The nonlinear case.

Consider at first the linearized discrete equations at $R_h y^*$ ($y^*$ is the exact solution of the BVP (1.1)). We denote

$$\tilde{F}_h'(R_h y^*) =: \tilde{H}_h^* = \tilde{L}_h - \Gamma_h^*, \qquad F_h'(R_h y^*) =: H_h^* = L_h - \Gamma_h^*, \tag{5.1a}$$

where

$$\Gamma_h^* := \operatorname{diag}(\gamma^*(t_\nu)), \qquad \gamma^*(t) := \frac{\partial}{\partial y} f(t, y^*(t)) \ge 0. \tag{5.1b}$$

Under the present smoothness requirements it is obvious that the results of Section 4 are valid for (5.1). I.e., the Fréchet derivative $H_h^* = F_h'(R_h y^*)$ is non-singular and stable. Let $S$ denote the stability threshold:

$$\|H_h^{*-1}\|_\infty \le S, \qquad h \le h_0. \tag{5.2}$$

We shall now show that the nonlinear target operator $F_h$ is stable in an $O(1)$-sphere around $R_h y^*$. (The argumentation below is essentially due to Keller [8].) Let

$$\mathcal{B}_h(r) := \{\eta_h \in \mathcal{E}_h : \|\eta_h - R_h y^*\|_\infty \le r\}, \tag{5.3}$$

$r$ independent of $h$. Assume that $M$ is a Lipschitz bound w.r.t. $y$ for $(\partial/\partial y) f(t, y)$ in the domain $\mathcal{B}_h(r)$:

$$\left|\frac{\partial}{\partial y} f(t_\nu, \eta_h(t_\nu)) - \frac{\partial}{\partial y} f(t_\nu, \zeta_h(t_\nu))\right| \le M\|\eta_h - \zeta_h\|_\infty, \qquad \eta_h, \zeta_h \in \mathcal{B}_h(r). \tag{5.4}$$

THEOREM 5.1. *For $r < 1/(SM)$, $F_h$ is stable in $\mathcal{B}_h(r)$:*

$$\|\eta_h - \zeta_h\|_\infty \le (S/(1 - SMr))\|F_h\eta_h - F_h\zeta_h\|_\infty, \qquad \eta_h, \zeta_h \in \mathcal{B}_h(r). \tag{5.5}$$

PROOF: By the Mean Value Theorem,

$$F_h\eta_h - F_h\zeta_h = \hat{H}_h(\eta_h, \zeta_h)(\eta_h - \zeta_h),$$

where

$$\hat{H}_h(\eta_h, \zeta_h) = L_h - \hat{\Gamma}_h(\eta_h, \zeta_h),$$

$$\hat{\Gamma}_h(\eta_h, \zeta_h) = \operatorname{diag}(\hat{\gamma}(t_\nu)), \qquad \hat{\gamma}(t_\nu) = \frac{\partial}{\partial y} f(t_\nu, \hat{\eta}_h(t_\nu)) \ge 0,$$

with $\hat{\eta}_h \in \mathcal{B}_h(r)$. We have

$$\hat{H}_h(\eta_h, \zeta_h) - H_h^* = \hat{\Gamma}_h(\eta_h, \zeta_h) - \Gamma_h^*.$$

By assumption (5.4),

$$\|\hat{\Gamma}_h(\eta_h, \zeta_h) - \Gamma_h^*\|_\infty \leq M\|\hat{\eta}_h - R_h y^*\|_\infty \leq Mr.$$

Thus,

$$\|I_h - H_h^{*-1}\hat{H}_h(\eta_h, \zeta_h)\|_\infty \leq \|H_h^{*-1}\|_\infty \|\hat{H}_h(\eta_h, \zeta_h) - H_h^*\|_\infty \leq SMr < 1$$

for $r < 1/(SM)$. Therefore, $\hat{H}_h(\eta_h, \zeta_h)$ is regular and

$$\|[\hat{H}_h(\eta_h, \zeta_h)]^{-1}\|_\infty \leq \|H_h^{*-1}\|_\infty \|H_h^*[\hat{H}_h(\eta_h, \zeta_h)]^{-1}\|_\infty \leq S/(1 - SMr),$$

which is equivalent to (5.5). ■

Together with consistency (2.12), the local unique solvability of $F_h\eta_h^* = 0$ and the error estimate $\|\eta_h^* - R_h y^*\|_\infty = O(h^{m-1})$ follow by standard argumentation (cf. [8]).

Note that the contractivity of the nonlinear IDeC operator was *not* required to prove the local stability of $F_h$ (in contrast to our argumentation in Section 4). In fact, the nonlinear IDeC iteration is not contractive in the general sense. The reason is that the effect of an unsmoothness in an initial approximation is *significantly stronger* than in the linear case: The Fréchet derivative $\Gamma_h'(\eta_h) = \mathrm{diag}[(\partial/\partial y)f(t_\nu, \eta_h(t_\nu))]$ at an *arbitrary* grid function $\eta_h$ does not satisfy the smoothness requirements of Section 4. (What can be shown is the uniform boundedness of the IDeC operator in $\mathcal{B}_h(r)$; this follows easily from (5.4).) To ensure contractivity without smoothness assumptions, one needs an initial approximation which is very close to $R_h y^*$ (i.e., in a $O(h^2)$-neighbourhood). See [10] for details.

## REFERENCES


1. W. Auzinger, *Defect corrections for multigrid solutions of the Dirichlet problem in general domains*, Math. Comp. (April 1987), 471–484.
2. J. Christiansen and R. D. Russell, *Deferred corrections using uncentered end formulas*, Numer. Math. 35 (1980), 21–33.
3. L. Fox, *The Numerical Solution of Two-point Boundary Value Problems*, Oxford University Press, Oxford (1957).
4. 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.
5. R. Frank, *The method of iterated defect correction and its application to two-point boundary value problems, part II*, Numer. Math. 27 (1977), 407–420.
6. R. Frank, J. Hertling and J. P. Monnet, *The application of iterated defect correction to variational methods for elliptic boundary value problems*, Computing 30 (1983), 121–135.





7. F. B. Hildebrand, *Introduction to Numerical Analysis*, 2nd ed., MacGraw-Hill, New York, Toronto, London (1974).
8. H. B. Keller, *Approximation methods for nonlinear problems with application to two-point boundary value problems*, Math. Comp. 29 (1975), 464–474.
9. M. Lees, in *Numerical Solution of Partial Differential Equations*, J. H. Bramble, Ed., 1966, 59–72.
10. J. P. Monnet, *Globale Kontraktionsaussagen und Konvergenzordnung der Defektkorrektur für 1-D Randwertprobleme (Einführung in die Problematik der 2-D Probleme)*, Ph.D. Thesis, Technical University of Vienna, 1986.
11. R. D. Skeel, *The order of accuracy for deferred corrections using uncentered end formulas*, SIAM J. Numer. Anal. 23 (1986), 393–402.
12. H. J. Stetter, *The defect correction principle and discretization methods*, Numer. Math. 29 (1978), 425–443.