Numer. Math. 48, 1-20 (1986)

Numerische Mathematik

Thirteen Ways to Estimate Global Error*

Robert D. Skeel

University of Illinois at Urbana-Champaign, Department of Computer Science,
(240 Digital Computer Laboratory, 1304 W. Springfield Avenue), Urbana, Illinois 61801, USA

Summary. Various techniques that have been proposed for estimating the accumulated discretization error in the numerical solution of differential equations, particularly ordinary differential equations, are classified, described, and compared. For most of the schemes either an outline of an error analysis is given which explains why the scheme works or a weakness of the scheme is illustrated.

Subject Classifications: AMS(MOS): 65B05, 65L99, 65J05; CR: G 1.7.

Introduction

In recent years there have been a number of surveys and comparison of techniques for estimating global discretization error in numerical solutions of differential equations [25, 29, 35, 36, 37, 39]. These have been an important stimulus to research in this area. In this paper we attempt to give an updated, organized, comprehensive survey of global error estimation techniques. For most of these techniques we outline an error analysis explaining why they work, and based on such theoretical considerations we discuss the relative merit of several of these techniques. To a large degree this paper is an extension of earlier work [30] by the author on deferred correction, which is one technique for global error estimation.

This survey is limited to error estimates, most of which are asymptotically correct in some sense, and excludes approximate or rigorous error bounds. Approximate error bounds, in the same spirit as matrix condition number estimators, would probably be more desirable. Some work on this has been done by Dahlquist [5] and Eriksson [9]. Rigorous error bounds are usually best achieved by interval analysis. Improvements in hardware and algorithms have made this approach feasible for problems of moderate size, and interval analy-

* This research is partially supported by NSF Grant No. MCS-8107046

sis now warrants greater recognition from the numerical mathematics community.

This paper is organized according to the type of error estimation technique:

1. Difference Correction
   - 1A. Stetter's Technique
   - 1B. Deferred Correction
   - 1L. Linearized Difference Correction
2. Differential Correction
   - 2A. Zadunaisky's Technique
   - 2B. Zadunaisky's Technique Reversed
   - 2L. Linearized Differential Correction
3. Integrating the Principal Error Equation
4. Solving for the Correction
5. Richardson Extrapolation
6. Error-Gradient Estimation
7. Using Two Different Tolerances
8. Using Two Different Methods
9. Using a Method with an Exact Principal Error Equation

Stetter [35] discusses techniques 1L, 2A, 2B, 2L, 3, and 9; Shampine and Watts [29] discuss 5, 7, and 8; Stetter [36] discusses 1B, 1L, 2L, 5, and 7; Stetter [37] discusses 1A, 1B, and 2A; Prothero [25] discusses 1B, 3, 5, 7, 8, and 9; and Stetter [39] discusses 2B, 2L, 5, and 7.

Zadunaisky's technique and Richardson extrapolation have been applied locally in order to estimate local truncation errors and/or improve the solutions. In such cases what we usually have is simply a systematic way of constructing higher order Runge-Kutta formulas. This topic is outside the scope of this paper.

Most of the thirteen listed techniques are more or less asymptotically correct, so that the error estimate ε^{est} can be subtracted from the computed solution to yield an improved solution $\bar{\eta} := \eta - \varepsilon^{\text{est}}$. In such a case ε^{est} is no longer really an error estimate, but it may still give a rough idea of the size of the error, and so we might call $|\varepsilon^{\text{est}}|$ a "fictitious error estimate"[1] or an "uncertainty estimate", a term which might be applied to any error estimate that is too crude for improving the solution.

The theory behind the various error estimation techniques is based on the assumption that the error is small, and so the accuracy of the error estimate increases with the accuracy of the numerical solution. This means that if the error estimate is large then the error must be large, but if the error estimate is small, then the error may or may not be small. The opposite situation would be preferable.

We believe that in order to model accurately the propagation of the error, it is necessary to do two integrations. All but technique 9 do this, and there are doubts about the validity of technique 9. Technique 6, in fact, uses three integrations.

---

[1] This term has been used by Stetter [36, p. 184] in connection with "local extrapolation"

Techniques 1 and 2 are variants of defect correction [37]. Difference correction involves the evaluation of a defect ($\equiv$residual) $\bar{\phi}(\eta)$ in the discrete realm where $\bar{\phi}$ is of higher accuracy than the discretization used for obtaining η. Differential correction involves the formation of a defect $F(\nabla\eta)$ in the continuous realm where $F(y)=0$ is the differential equation and $\nabla\eta$ is an interpolant of η. (The symbols ∇ and Δ introduced by Stetter [34] suggest, respectively, prolongation and restriction if one reads them bottom up.) Whether a defect is formed in the discrete or continuous realm there is a version A, a version B, and a linearized version of defect correction. In all cases one does a second integration using some cheap method ϕ of order p (usually $p=1$). In the case of linear methods (defined in subsection 1B), versions A and B of differential correction are special cases of difference correction in which $\bar{\phi}(\zeta) = \Delta_0^* F(\nabla\zeta)$, where Δ_0^* is a method-dependent restriction operator; for nonlinear methods this is not the case. Nonetheless, this observation suggests that difference correction offers more possibilities for error estimates. Efficiency considerations often rule against linearized versions of defect correction, and theoretical considerations slightly favor version B.

In the case of a conventional method ϕ for a linear problem, techniques 1A, 1B, and 1L become equivalent as do techniques 2A, 2B, 2L, and 4. Moreover, the latter become a special case of the former.

Of the thirteen techniques, we favor difference correction, Richardson extrapolation, and using two different tolerances, the third being included only because it is so simple. Of course, further research may alter this opinion.

Although applicable to PDEs, the ideas in this paper were primarily motivated by ODEs. In particular, the reader will find that the linear method

$$G(t_n, \eta_n, (\eta_n - \eta_{n-1})/h) = 0$$

for $G(t, y, y') = 0$ and the nonlinear method

$$(\eta_n - \eta_{n-1})/h + f(t_{n-1} + h/2, \eta_{n-1} + (h/2) f(t_{n-1}, \eta_{n-1})) = 0$$

for $y' - f(t, y) = 0$ are a rich source of examples for the theory.

1. Difference Correction

Assume that the given numerical solution η satisfies

$$\|\eta - \Delta y\|_q \leqq C h^r \tag{1.1}$$

where C, and every other such constant is independent of h and the subscript q denotes a discrete Sobolev norm involving divided differences up to order q. Difference correction methods depend on the evaluation of a discrete residual $\bar{\phi}(\eta)$ where the operator $\bar{\phi}$ is of higher order:

$$\|\bar{\phi}(\Delta y)\|_{*0} \leqq \bar{C} h^{r+p}.$$

The subscript $*0$ denotes a norm for the discrete residual space and may be the max norm or some weaker norm involving discrete integrals such as the Spijker norm [32]. For convenience we assume throughout that $p \leqq r$. A second solution $\bar{\eta}$ is obtained using a cheap method ϕ of order p. The exact details vary and are given in subsections 1A, 1B, and 1L. There are at least two reasons that this procedure may be preferable to obtaining $\bar{\eta}$ directly with $\bar{\phi}$. One is that $\bar{\phi}$ and ϕ^{-1} together may be cheaper to compute than $\bar{\phi}^{-1}$. The other is that $\bar{\phi}$ need not be stable if used to compute a difference correction. Of course ϕ must be stable, meaning that

$$\|\xi\|_0 \leqq S \|\phi(\zeta+\xi)-\phi(\zeta)\|_{*0}. \tag{1.2}$$

Error analysis in subsections 1A, 1B, and 1L shows that the accuracy of $\bar{\eta}$ depends on how close $\psi(\eta)$ is to the local truncation error $\phi(\Delta y)$ where $\psi := \phi - \bar{\phi}$. (The difference correction is $-\psi(\eta)$.) It follows immediately that

$$\|\psi(\eta)-\phi(\Delta y)\|_{*0} \leqq (KC+\bar{c})h^{r+p}$$

if ψ is sufficiently contractive, meaning that

$$\|\psi(\zeta+\xi)-\psi(\zeta)\|_{*0} \leqq Kh^p \|\xi\|_q.$$

The value of q used here determines the quality of the approximate solution needed (see (1.1)). Examples in Skeel [30] illustrate values of q from $p-k$ through $p+k$ where k is the order of the highest derivative in the differential operator F. The lower values of q are obtained by choosing $\bar{\phi}$ close to ϕ. In the differential correction technique of section 2 we must choose $q=p+k$ for a general error analysis. This, we believe, is a significant disadvantage of differential correction.

1A. Stetter's Technique

This technique, first described by Stetter [37, Sect. 8] under the name "defect correction, version A" is a synthesis of generalized difference correction (defect correction, version B), and Zadunaisky's technique. The basic idea is quite simple and appealing. We regard ϕ as an approximate method for solving $\bar{\phi}(\zeta)=0$ for ζ. Let η' be the solution obtained with ϕ applied to the original problem $\bar{\phi}(\zeta)=0$ and let η'' be the solution obtained with ϕ applied to the neighboring problem $\bar{\phi}(\zeta)-\bar{\phi}(\eta)=0$ where η is the given numerical solution. The solution of the neighboring problem is clearly η. Thus the global error of ϕ applied to the neighboring problem is $\eta''-\eta$ which should be close to the global error of ϕ applied to the original problem and hence can be used to improve the solution η':

$$\begin{aligned} \phi(\eta') &= 0, \\ \phi(\eta'')-\bar{\phi}(\eta) &= 0, \\ \bar{\eta} &= \eta' - (\eta''-\eta). \end{aligned}$$

The error analysis for this technique remains valid for the more general procedure

$$\phi(\eta'') - \bar{\phi}(\eta) = \phi(\eta'),$$
$$\bar{\eta} = \eta' - (\eta'' - \eta).$$

This procedure with $\eta' = \eta$ would be advantageous if $\phi(\eta)$ is cheaper to compute than $\phi^{-1}(0)$. And in any case the proof that follows suggests that it is more accurate.

Two extra assumptions are needed:

$$\|\eta' - \Delta y\|_k \leqq C' h^p \tag{1.3}$$

and

$$\|\phi'(\zeta + \xi) - \phi'(\zeta)\|_{k \to *0} \leqq L' \|\xi\|_k \tag{1.4}$$

for ζ and $\zeta + \xi$ in some neighborhood of Δy where the subscript $k \to *0$ refers to the operator norm induced by the k norm and the $*0$ norm. (In cases where the operator F is linear in its highest derivatives, one can use a norm weaker than the k norm.) The error estimate $\eta - \bar{\eta}$ differs from the actual error $\varepsilon = \eta - \Delta y$ by

$$\begin{aligned}
\|\bar{\eta} - \Delta y\|_0 &= \|\eta'' - (\eta' + \varepsilon)\|_0 \\
&\leqq S \|\phi(\eta') + \bar{\phi}(\eta) - \phi(\eta' + \varepsilon)\|_{*0} \\
&= S \|\phi(\Delta y) - \psi(\eta) + \int_0^1 \phi'(\Delta y + \theta\varepsilon) - \phi'(\eta' + \theta\varepsilon)\, d\theta \cdot \varepsilon\|_{*0} \\
&\leqq S[(KC + \bar{c}) h^{r+p} + L' \cdot C' h^p \cdot Ch^r]
\end{aligned}$$

where it has been assumed that the k norm is no stronger than the q norm.

1B. Deferred Correction

Invented by Fox [10], revived by Pereyra [24], and generalized by Stetter [37, p. 432], [36, Sect. II.B) II], and Lindberg [21], this technique can be expressed as

$$\phi(\bar{\eta}) = \psi(\eta)$$

where $\psi := \phi - \bar{\phi}$, which indicates that the local truncation error estimate is being removed in solving for $\bar{\eta}$. In contrast, if $\eta' = \eta$ in version A, then $\phi(\eta'') = -\psi(\eta)$, which indicates that the truncation error estimate is being added in, and so η'' has twice the global error of η. Because the second solution η'' is farther away from Δy instead of closer to it, one might expect version A to be inferior to B. Indeed, the following error bound for version B has one less term than that for version A:

$$\begin{aligned}
\|\bar{\eta} - \Delta y\|_0 &\leqq S \|\psi(\eta) - \phi(\Delta y)\|_{*0} \\
&\leqq S(KC + \bar{c}) h^{r+p}.
\end{aligned}$$

More details and examples are found in Skeel [30], and application to global error estimation for the stiff ODE integrator EPISODE [3] is in Skeel [31].

Stetter [37, p. 441] suggests the possible use of a ϕ on a coarser grid as the cheaper solution method. That is, we try to *coarsen* the problem

$$\phi^h(\bar{\eta}^h) - \phi^h(\eta^h) = -\bar{\phi}^h(\eta^h)$$

which using the nonlinear FAS multigrid idea of Brandt [1] gives

$$\phi^H(\bar{\eta}^H) - \phi^H(\Delta_h^H \eta^h) = -\Delta_h^{*H} \bar{\phi}^h(\eta^h) \tag{1.5}$$

where Δ_h^H and Δ_h^{*H} are restriction operators.

Under the assumption that ϕ is a linear method, we offer theoretical justification. If $\phi(\zeta; g)$ denotes the discretization of the more general operator $F(z) + g$ where g is simply a function, then a linear method is one for which

$$\phi(\zeta; g) = \phi(\zeta) + \Delta_0^* g \tag{1.6}$$

where Δ_0^* is a restriction operator. We say [30] that a method possesses *property* (E) if

$$\|\phi(\Delta(z+x); -F(z+x)) - \phi(\Delta z; -F(z))\|_{*k} \leqq c h^p \|x\|_{p+k} \tag{1.7}$$

where the subscript $p+k$ indicates a Sobolev norm involving derivatives up to order $p+k$. For linear methods this becomes

$$\|\phi(\Delta(z+x)) - \phi(\Delta z) - \Delta_0^* \{F(z+x) - F(z)\}\|_{*k} \leqq c h^p \|x\|_{p+k}, \tag{1.8}$$

which we assume is satisfied by both ϕ^h and ϕ^H.

First we show that ϕ^H is related to ϕ^h by an analog of property (E). We have

$$\begin{aligned} &\|\phi^H(\Delta^H(z+x)) - \phi^H(\Delta^H z) - \Delta_h^{*H}\{\phi^h(\Delta^h(z+x)) - \phi^h(\Delta^h z)\}\|_{*k} \\ &\quad \leqq \|(\Delta_0^{*H} - \Delta_h^{*H}\Delta_0^{*h})\{F(z+x) - F(z)\}\|_{*k} + c(H^p + h^p \|\Delta_h^{*H}\|_{*k \to *k}) \|x\|_{p+k} \\ &\quad \leqq c' H^p \|x\|_{p+k} \end{aligned} \tag{1.9}$$

under fairly obvious assumptions. Suppose

$$\Delta^H = \Delta_h^H \Delta^h$$

and there exists a prolongation operator ∇_h bounded in the $p+k \to p+k$ norm such that

$$\Delta^h \nabla_h = 1.$$

Letting $z = \nabla_h \zeta^h$ and $x = \nabla_h \xi^h$ in (1.9) we get property (E_h^H):

$$\|\phi^H(\Delta_h^H(\zeta^h + \xi^h)) - \phi^H(\Delta_h^H \zeta^h) - \Delta_h^{*H}\{\phi^h(\zeta^h + \xi^h) - \phi^h(\zeta^h)\}\|_{*k} \leqq c'' H^p \|\xi^h\|_{p+k}.$$

The error analysis justifying (1.5) is given by

$$\begin{aligned}\|\bar{\eta}^H - \Delta^H y\|_0 &\leqq S\,\|\phi^H(\Delta_h^H \eta^h) - \Delta_h^{*H}\bar{\phi}^h(\eta^h) - \phi^H(\Delta^H y)\|_{*0} \\ &= S\,\|\phi^H(\Delta_h^H \eta^h) - \phi^H(\Delta^H y) - \Delta_h^{*H}\{\phi^h(\eta^h) - \phi^h(\Delta^h y)\} \\ &\quad + \Delta_h^{*H}\{\psi^h(\eta^h) - \phi^h(\Delta^h y)\}\|_{*0} \\ &\leqq S\{c'' H^p \|\eta^h - \Delta^h y\|_{p+k} + \|\Delta_h^{*H}\|_{*k\to *0}(KC+\bar{c})\,h^{r+p}\},\end{aligned}$$

which is $O(h^{r+p})$ if (1.1) holds with $q = p+k$.

Remark. Generally $\|\Delta_h^{*H}\|_{*0\to *0}$ is not bounded independently of h and H. Such is the case when straight injection is used for Δ_h^{*H}. However, this norm is usually bounded if the restriction operator involves some kind of averaging.

1L. Linearized Deferred Correction

Recall that Fox's difference correction technique uses the iteration

$$\begin{aligned}\phi(\eta^1) &= 0, \\ \phi(\eta^{i+1}) &= \psi(\eta^i)\end{aligned}$$

until "convergence" where $-\psi$ consists of the higher order correction terms of ϕ. A linearized version of this was first proposed in a paper by Clenshaw and Oliver [4], which is discussed by Mayers [22, Sect. 29]. Instead of solving for η^{i+1}, one sets $\eta^{i+1} = \eta^i + \omega^{i+1}$, linearizes the difference equation in ω^i, and solves

$$\phi(\eta^i) + \phi'(\eta^i)\,\omega^{i+1} = \psi(\eta^i)$$

for the correction ω^{i+1}. This, of course, is simply nonlinear deferred correction with one Newton iteration, and for a linear operator ϕ it is equivalent to versions A and B of difference correction.

An error analysis is given by Pereyra [24, Sect. 3] for a single linearized deferred correction in a general setting with $\psi(\eta^1)$ regarded as a local error estimate. The error is analyzed also by Stetter [34, p. 46], who states that the linearized deferred correction procedure can be applied in an iterative updating fashion even in the nonlinear case although "the formulation becomes more complicated".

As an approach to global error estimation for ODEs, this is discussed by Stetter [35, Sect. 2], [36, Sect. II.B) I)], who calls it defect estimation by linearization in the discrete realm.

A single application of linearized deferred correction can be written as

$$\begin{aligned}\phi'(\eta)\,\omega &= -\bar{\phi}(\eta), \\ \bar{\eta} &= \eta + \omega.\end{aligned}$$

It is not difficult to show that stability implies

$$\|\phi'(\zeta)^{-1}\|_{*0\to 0} \leqq S,$$

and therefore

$$\begin{aligned}\|\bar{\eta}-\Delta y\|_0 &\leqq S\,\|\phi'(\eta)(\bar{\eta}-\Delta y)\|_{*0}\\ &= S\,\|\psi(\eta)-\phi(\Delta y)+\int_0^1(\phi'(\eta)-\phi'(\eta-\theta\varepsilon))\,d\theta\cdot\varepsilon\|_{*0}\\ &\leqq S((KC+\bar{c})h^{r+p}+\tfrac{1}{2}L'(Ch^r)^2)\end{aligned}$$

assuming that (1.4) holds.

2. Differential Correction

All versions of this approach are based on some cheap method of order p applied to a differential equation having as an inhomogeneous term the defect $F(\tilde{y})$ where $\tilde{y}$ is an approximation to y constructed from the given numerical solution of η. We assume that

$$\Delta\tilde{y}=\eta, \tag{2.1}$$

which can always be achieved by redefining η if necessary, that

$$\|\tilde{y}-y\|_{p+k}\leqq\tilde{C}h^r, \tag{2.2}$$

and that

$$\|\eta-\Delta y\|_k\leqq\tilde{C}h^r. \tag{2.3}$$

Assume also that

$$\|F(z+x)-F(z)\|_{*k}\leqq L\|x\|_k,$$

and it follows that

$$\|F(\tilde{y})\|_{*k}\leqq L\tilde{C}h^r.$$

Further assume that $\phi(\zeta;g)$ possesses property (E) given by (1.7), that $\phi(\zeta;g)$ is stable in some neighborhood of $(\Delta y;0)$ as given by (1.2), and that

$$\|\phi'(\zeta;g)-\phi'(\zeta)\|_{k\to *0}\leqq L_g\|g\|_{*k}. \tag{2.4}$$

Usually $\tilde{y}=V\eta$ for some linear operator V, typically an interpolation operator. It is straightforward to verify (2.1), (2.2), (2.3) if

$$\|\eta-\Delta y\|_{p+k}\leqq Ch^r$$

and the prolongation V satisfies

$$\begin{aligned}\Delta V&=1,\\ \|V\|_{p+k\to p+k}&\leqq M,\\ \|V\Delta-1\|_{r+p+k\to p+k}&\leqq\gamma h^r,\end{aligned}$$

for example, piecewise interpolation by polynomials of degree $r+p+k-1$.

Stetter [39, Sect. 1] suggests in the case of ODEs that rather than form $\tilde{y}$ explicitly one should define it implicitly by requiring $\tilde{y}(t_{n-1})=\eta_{n-1}$, $\tilde{y}(t_n)=\eta_n$, and $F(\tilde{y})(t)=\text{constant}=:\delta_n$ for $t_{n-1}\leqq t\leqq t_n$. Various estimates for δ_n are enumerated in [39, Sect. 2]. The analysis for this approach, which will not be

pursued here, would be quite different. It would take into account the discrepancy $\delta_n^{\text{est}} - \delta_n$ and it would probably require only that $\eta - \Delta y$ be $O(h^r)$ in the weaker k norm. The estimation of δ_n is far from trivial, much less straightforward than the evaluation of $F(V\eta)$. It is probably appropriate to classify this idea as differential correction because no special relationship is assumed between the construction of δ_n^{est} and the cheap method ϕ used to solve for the error.

In most cases $\tilde{y}$ and $F(\tilde{y})$ are only piecewise smooth, and the cheap method ϕ must take this into account. Also the definitions of norms for function in the continuous realm must be suitable modified. Appropriate details are found in Frank and Ueberhuber [16]. In fact, the complications involved in such an analysis is an argument in favor of difference correction.

Remark. Assumption (2.4), as well as assumption (2.7) of Sect. 2B, must be modified for methods like the Taylor methods which use analytical derivatives. Instead of $\|g\|_{*k}$ we need

$$\sum_{j=k}^{p} h^{j-k} \|g\|_{*j}$$

if $p > k$. The error analysis yields the same final result but the intermediate expressions become more complicated.

2A. Zadunaisky's Technique

This technique originates with Zadunaisky [40] and is discussed by Stetter [35, Sect. 5], [37, Eq. (4.9)]. It is applied iteratively and analyzed by Frank [12, 13] for ordinary boundary value problems. General theoretical results about the order of accuracy are proved by Frank and Ueberhuber [16] using asymptotic expansions in the meshsize parameter h, which are justified by Frank, Macsek, and Ueberhuber [15]. Their proof is complicated by a double use of the parameter h, a difficulty that can be avoided by the use of discrete Sobolev norms rather than asymptotic expansions.

The idea is to construct a neighboring problem $\tilde{F}(\tilde{y}) = 0$ which is close to the original problem $F(y) = 0$ and whose solution $\tilde{y}$ is known. Then we obtain a numerical solution η'' for this nearby problem using the same method that was used to obtain the original numerical solution η. The known global error $\eta'' - \Delta\tilde{y}$ for the neighboring problem is expected to be a reasonable estimate for $\eta - \Delta y$. More specifically, $\tilde{y}$ is constructed from η and

$$\tilde{F}(z) := F(z) - F(\tilde{y})$$

is used for the neighboring problem. As observed by Stetter [37], one can use a cheap method ϕ rather than the one that was used to obtain η. Then $\eta'' - \Delta\tilde{y}$ becomes an estimate of the error $\eta' - \Delta y$ where η' is the solution of the original problem obtained by ϕ. As we did for version A of difference correction, we

can generalize this still further to

$$\phi(\eta''; -F(\tilde{y})) = \phi(\eta'),$$
$$\bar{\eta} = \eta' - (\eta'' - \Delta\tilde{y}),$$

where we no longer assume $\phi(\eta') = 0$.

If we assume that (1.3) and (1.4) hold as we did for version A of difference correction, then

$$\begin{aligned}
\|\bar{\eta} - \Delta y\|_0 &= \|\eta'' - (\eta' + \varepsilon)\|_0 \\
&\leqq S\,\|\phi(\eta') - \phi(\eta' + \varepsilon; -F(\tilde{y}))\|_{*0} \\
&= S\Big\|\int_0^1 [\phi'(\eta + \theta\varepsilon') - \phi'(\eta + \theta\varepsilon'; -F(\tilde{y}))]\, d\theta \cdot \varepsilon' \\
&\quad + \int_0^1 [\phi'(\Delta y + \theta\varepsilon') - \phi'(\eta + \theta\varepsilon')]\, d\theta \cdot \varepsilon' \\
&\quad + \phi(\Delta y) - \phi(\eta; -F(\tilde{y}))\Big\|_{*0} \\
&\leqq S(L_g L \tilde{C} h^r C' h^p + L' \tilde{C} h^r C' h^p + c h^p \tilde{C} h^r).
\end{aligned}$$

For linear methods (1.6) holds and therefore if $\tilde{y} = V\eta$, Zadunaisky's technique becomes a special case of Stetter's technique with

$$\bar{\phi}(\zeta) = \Delta_0^* F(V\zeta).$$

In fact this gives us a defect estimate of the local error as proposed by Frank, Hertling, and Ueberhuber [14]. Under the assumptions of this section it can be shown [30, Sect. 2.4] that $\bar{\phi}$ satisfies the conditions of Sect. 1.

2B. Zadunaisky's Technique Reversed

Stetter [35, Sect. 7, process (B)] proposes the following iterated differential correction algorithm:

$$\begin{aligned}
&\phi\left(\eta^i; \sum_{j=1}^{i-1} F(\tilde{y}^j)\right) = 0, \\
&\qquad \tilde{y}^i = \text{interpolant of } \eta^i
\end{aligned} \tag{2.5}$$

for $i = 1, 2, 3, \ldots$. The first two iterations can be expressed

$$\begin{aligned}
&\phi(\eta) = 0, \\
&\quad \tilde{y} = \text{interpolant of } \eta \\
&\phi(\bar{\eta}; F(\tilde{y})) = 0,
\end{aligned}$$

this last equation being just the application of method ϕ to the problem given by

$$\bar{F}(z) := F(z) + F(\tilde{y}).$$

The sign of the inhomogeneous term is the reverse of what Zadunaisky uses and the effect of this is to produce an approximation to $\eta-\varepsilon$ rather than $\eta+\varepsilon$.

For global error estimation Stetter [39, Sect. 3D), third paragraph] suggests that the secondary solution method ϕ be cheaper than that used to obtain η, in which case one should solve the more general problem

$$\phi(\bar{\eta}; F(\tilde{y}))=\phi(\eta). \tag{2.6}$$

Assume that

$$\|\phi(\zeta; g)-\phi(\zeta)-\Delta_\zeta^* g\|_{*0} \leqq L_{gg}\|g\|_{*k}^2 \tag{2.7}$$

where Δ_ζ^* is the Frechet derivative of $\phi(\zeta; g)$ with respect to g evaluated at $(\zeta; 0)$. Then

$$\begin{aligned}
\|\bar{\eta}-\Delta y\|_0 &\leqq S\|\phi(\eta)-\phi(\Delta y; F(\tilde{y}))\|_{*0} \\
&= S\Bigg\|\int_0^1 [\phi'(\Delta y+\theta\varepsilon)-\phi'(\Delta y+\theta\varepsilon; -F(\tilde{y}))]\, d\theta\cdot\varepsilon \\
&\quad +2\phi(\Delta y)-\phi(\Delta y; -F(\tilde{y}))-\phi(\Delta y; F(\tilde{y})) \\
&\quad +\phi(\eta; -F(\tilde{y}))-\phi(\Delta y)\Bigg\|_{*0} \\
&\leqq S(L_g L\tilde{C}h^r \tilde{C}h^r+2L_{gg}(L\tilde{C}h^r)^2+ch^p \tilde{C}h^r).
\end{aligned}$$

If ϕ is a linear method and $\tilde{y}=\nabla\eta$, then (2.6) becomes a special case of deferred correction that uses a differential defect to estimate the local error. Recall the discussion at the end of Sect. 2A. Likewise, (2.5) becomes a special case of iterated deferred correction.

Remark. For a convergence proof of (2.5) we could define

$$\phi^i(\zeta; g) := \phi\left(\zeta; \sum_{j=1}^{i-1} F(\tilde{y}^j)+g\right)$$

and apply our analysis to ϕ^i after verifying that ϕ^i satisfies the various hypotheses.

2L. Linearized Differential Correction

This approach was used by Fox [11, Sect. 12] for two-point BVPs and is described by Stetter [35, Sect. 4; Sect. 7, process (A)], [36, Sect. II.B) I), continuous realm], [39, Sect. 3C)]. One solves the linearized problem

$$F(\tilde{y})+F'(\tilde{y})\, w=0$$

for a correction w. This is the continuous analog of linearized deferred correction and shares with it the disadvantage of having to form a Frechet derivative.

Conventional methods $\phi(\zeta; g)$ applied to a linear operator such as $F'(\tilde{y})z + g$ invariably have the form

$$\phi(\zeta; g) = \Phi \cdot \zeta + \Delta_0^* g$$

where both Φ and Δ_0^* may depend on $F'(\tilde{y})$. For such methods property (E), given by (1.8), simplifies to

$$\|\Phi \cdot \Delta x - \Delta_0^* F'(\tilde{y}) x\|_{*k} \leqq c h^p \|x\|_{p+k}.$$

With the additional assumptions

$$\|F'(z+x) - F'(z)\|_{k \to *k} \leqq L_F' \|x\|_k$$

and

$$\|\Delta_0^* g\|_{*0} \leqq M^* \|g\|_{*k}$$

we have that the solution ω of

$$\Phi \cdot \omega + \Delta_0^* F(\tilde{y}) = 0$$

satisfies

$$\begin{aligned} \|(\eta + \omega) - \Delta y\|_0 &\leqq S \|\Phi \cdot \Delta(\tilde{y} - y) - \Delta_0^* F(\tilde{y})\|_{*0} \\ &\leqq S\{\|\Phi \cdot \Delta(\tilde{y} - y) - \Delta_0^* F'(\tilde{y})(\tilde{y} - y)\|_{*0} + \tfrac{1}{2} M^* L_F' \|\tilde{y} - y\|_k^2\} \\ &\leqq S\{c h^p \tilde{C} h^r + \tfrac{1}{2} M^* L_F' (\tilde{C} h^r)^2\}. \end{aligned}$$

If the original problem is linear,

$$F y = g,$$

and if $\tilde{y} = V\eta$, then we solve

$$\Phi \cdot \omega + \Delta_0^* \{F \cdot V\eta - g\} = 0.$$

This is equivalent to linearized deferred correction, $\phi'(\eta) \cdot \omega = -\bar{\Phi}(\eta)$, with

$$\phi(\zeta) = \Phi \cdot \zeta - \Delta_0^* g$$

and

$$\bar{\phi}(\zeta) = \Delta_0^* F \cdot V\zeta - \Delta_0^* g.$$

This equivalence does not always hold for nonlinear problems. Let $\tilde{y}$ be constructed from a numerical solution η on an equidistant mesh for the ODE $y' - f(t, y) = 0$ by piecewise quadratic interpolation on successive pairs of subintervals. The correction w satisfies

$$w' - f_y(t, \tilde{y}(t)) w + \tilde{y}(t) - f(t, \tilde{y}(t)) = 0.$$

If we solve for w using the box scheme, then the linear part of the difference operator for even values of n is given by

$$(\Phi \cdot \omega)_n = \frac{\omega_n - \omega_{n-1}}{h} - f_y(t_{n-1/2}, \tfrac{3}{8}\eta_n + \tfrac{3}{4}\eta_{n-1} - \tfrac{1}{8}\eta_{n-2}) \frac{\omega_n + \omega_{n-1}}{2}.$$

If this is actually linearized deferred correction, then $(\Phi \cdot \omega)_n = (\phi'(\eta)\omega)_n$ for some ϕ, and so

$$\frac{\partial \phi_n}{\partial \eta_n} = \frac{1}{h} - \tfrac{1}{2} f_y(t_{n-1/2}, \tfrac{3}{8}\eta_n + \tfrac{3}{4}\eta_{n-1} - \tfrac{1}{8}\eta_{n-2})$$

and

$$\frac{\partial \phi_n}{\partial \eta_{n-2}} = 0.$$

However, this is impossible if $f(t, z)$ is nonlinear because

$$\frac{\partial}{\partial \eta_{n-2}} \frac{\partial \phi_n}{\partial \eta_n} = \frac{1}{16} f_{yy}(\ldots)$$

but

$$\frac{\partial}{\partial \eta_n} \frac{\partial \phi_n}{\partial \eta_{n-2}} = 0.$$

Shampine [27] suggests the use of an implicitly defined $\tilde{y}$ for ODEs, but instead of spreading out the local error uniformly on $[t_{n-1}, t_n]$ as Stetter [39] does, he would concentrate it all at t_n. That is, $\tilde{y}(t)$, $t_{n-1} \leqq t < t_n$, is the local ODE solution satisfying the initial condition $\tilde{y}(t_{n-1}) = \eta_{n-1}$, and so it has jumps before each meshpoint (which the correction integrator ϕ would take into account). This approach is advocated for stiff equations where an approximate Jacobian f_y is already available.

3. Integrating the Principal Error Equation

It is often the case that the global error has an asymptotic behavior in terms of the meshsize parameter h given by

$$\|\eta - \Delta(y + h^r e)\|_0 \leqq C' h^{r+p} \tag{3.1}$$

where the magnified error function e satisfies

$$F'(y)\, e = D^r(y)$$

for some nonlinear operator D^r. Invariably $p = 1$ or 2. With an approximation $\tilde{y}$ constructed from η we can solve the problem

$$F'(\tilde{y})\, \tilde{e} = D^r(\tilde{y})$$

instead. Applying a cheap method to this linear problem gives

$$\Phi \cdot \bar{\varepsilon} = \Delta_0^* D^r(\tilde{y}). \tag{3.2}$$

Such global error estimation algorithms for stiff ODEs have been tested by Robinson and Prothero [26] for DIFSUB [17] and by Prothero [25] for EPISODE [3]. This idea is very similar to linearized differential correction as Stetter [35, Sect. 2, 4] points out; however, it requires the existence of an asymptotic expansion for the global error, and furthermore the term $D^r(y)$ can be quite complicated.

In addition to the general assumptions of Sect. 2 and those of Sect. 2L, assume that

$$\|D^r(z+x)-D^r(z)\|_{*k} \leqq L_D \|x\|_{r+k}$$

and

$$\|\tilde{y}-y\|_{r+k} \leqq \tilde{C}h^p.$$

This last assumption is much like (2.2). Then

$$\begin{aligned}\|(\eta-h^r\bar{\varepsilon})-\Delta y\|_0 &\leqq h^r\|\Delta e-\bar{\varepsilon}\|_0 + C'h^{r+p}\\ &\leqq h^r S\|\Phi\cdot\Delta e-\Delta_0^* D^r(\tilde{y})\|_{*0}+C'h^{r+p}\\ &= h^r S\|\Phi\cdot\Delta e-\Delta_0^* F'(\tilde{y})e+\Delta_0^*\{(F'(\tilde{y})-F'(y))e\\ &\quad -(D^r(\tilde{y})-D^r(y))\}\|_{*0}+C'h^{r+p}\\ &\leqq h^r S(ch^p\|e\|_{p+k}+M^*(L_F'\tilde{C}h^r\|e\|_k+L_D\tilde{C}h^p))+C'h^{r+p}.\end{aligned}$$

The actual assumptions used are apparent from the constants in the bound.

The error bound just given does not exploit the apparent relationship between techniques 3 and 2L. If instead of (3.1) we had assumed that

$$\|\tilde{y}-(y+h^re)\|_k \leqq \tilde{C}'h^{r+p}$$

then it follows that

$$\begin{aligned}\|h^rD^r(\tilde{y})-F(\tilde{y})\|_{*k} &\leqq h^r\|D^r(\tilde{y})-D^r(y)\|_{*k}+\|F'(y)\{(y+h^re)-\tilde{y}\}\|_{*k}\\ &\quad +\|F(y)+F'(y)(\tilde{y}-y)-F(\tilde{y})\|_{*k}\\ &\leqq h^rL_D\tilde{C}h^p+L\tilde{C}'h^{r+p}+\tfrac{1}{2}L_F'(\tilde{C}h^r)^2,\end{aligned}$$

and this can be combined with the analysis for techniques 2L in order to yield an error bound for technique 3.

Prothero [25, p. 120] shows that the formation of the Frechet derivative F' can be avoided by solving

$$F(\tilde{y})-F(\tilde{y}-h^r\tilde{e})=h^rD^r(\tilde{y})$$

instead. This is just technique 4, to be discussed next, except that $h^rD^r(\tilde{y})$ replaces the defect $F(\tilde{y})$, which is the same relationship that (3.2) bears to technique 2L.

4. Solving for the Correction

Given a numerical solution η, one can construct an approximate solution $\tilde{y}$ and solve

$$F(\tilde{y}+w)=0$$

for the correction w using a cheap method ϕ. For example, if $y'=f(t,y)$ is the original problem, then w satisfies

$$w'=f(t,\tilde{y}(t)+w)-\tilde{y}'(t).$$

For linear problems this technique is clearly equivalent to differential correction. If we let $\phi(\zeta;;\tilde{y})$ denote the discretization of $F(\tilde{y}+z)$ then the discrete correction ω is obtained from

$$\phi(\omega;;\tilde{y})=0.$$

The validity of this approach is a consequence of the fact that most conventional methods satisfy property (F):

$$\|\phi(\Delta z;;y-z)\|_{*0}\leqq ch^p\|z\|_{p+k}$$

where c depends on only the original problem $F(y)=0$. If we also assume (2.2), we have

$$\|\omega-\Delta(y-\tilde{y})\|_0\leqq S\|\phi(\Delta(y-\tilde{y});;\tilde{y})\|_0\leqq Sch^p\tilde{C}h^r.$$

According to Johnson and Riess [18, p. 439], "this technique is widely used in the area of celestial mechanics (e.g., Encke's method)." They give an example in which one step of a fourth order collocation method is applied to a first order ODE and the built-in quadratic interpolant is used for $\tilde{y}$. Using a cheaper method with a smaller stepsize they determine ω and an improved solution $\bar{\eta}:=\Delta\tilde{y}+\omega$. The errors in $\Delta\tilde{y}$ and $\bar{\eta}$ are given below:

t	$\Delta\tilde{y}-\Delta y$	$\bar{\eta}-\Delta y$
$0.25h$	-0.01743	-0.01210
$0.5h$	-0.05580	-0.01035
$0.75h$	-0.03505	-0.00466
h	-0.01946	-0.02156

Note that the $O(h^3)$ interpolation error of the quadratic interpolant is reduced but the $O(h^5)$ discretization error at $t=h$ is not. This illustrates the importance of high accuracy interpolation, as assumed in (2.2).

5. Richardson Extrapolation

This very well known idea is mentioned in practically all surveys of global error estimation, and it is the technique chosen for implementation by Shampine and Watts [29] in their code GERK. An excellent survey of earlier work is given by Joyce [19]. In addition to the primary numerical solution η^h one obtains with the same method a solution η^{2h} on a mesh which is twice as coarse and then uses

$$\varepsilon^{\text{est}}=\frac{\eta^{2h}-\eta^h}{2^p-1}$$

where p is the order of the method. The cost of this technique may be greater than it appears because the coarse mesh must be chosen fine enough to ensure numerical stability and to ensure that the discrete equations (linear or nonlinear) can be readily solved. Also, unlike most other techniques, the validity of this approach really does depend on the existence of an asymptotic expansion for the global error in terms of h.

Stetter [37, p. 441] mentions a "Version B" of Richardson extrapolation that he will explain in a separate paper.

6. Error-Gradient Estimation

This idea of Epstein and Hicks [7, 8] is similar in spirit to Richardson extrapolation except that instead of extrapolating from meshsizes h and $2h$ one extrapolates from meshsize h alone using the discrete solution and its derivative with respect to h. This necessitates the construction of an approximate *continuum* solution for meshsizes θh, $0<\theta\leqq 1$.

As an example the forward Euler method applied to $y'=f(t,y)$, $y(0)=y^0$ with meshsize θh can be extended to the continuum by

$$y(\theta;t)=y^0+tf(0,y^0), \qquad 0\leqq t<\theta h,$$
$$y(\theta;t+\theta h)=y(\theta;t)+\theta h f(t,y(\theta;t)), \qquad t\geqq\theta h.$$

An estimate of the error is given by the error gradient $y_\theta(1;t)$, which can be determined by the difference equations

$$y^{n+1}=y^n+hf^n,$$
$$y_t^{n+1}=y_t^n+h(f_t^n+f_y^n\cdot y_t^n),$$
$$y_\theta^{n+1}=y_\theta^n+h(f_y^n\cdot y_\theta^n-(y_t^{n+1}-f^n))$$

where the superscript n denotes evaluation at either $(1;nh)$ or $(nh,y(1;nh))$, whichever is appropriate. An assymptotically correct error estimate is given by

$$e^{n+1}=e^n+h(f_y^n\cdot e^n-\tfrac{1}{2}(f^{n+1}-f^n)),$$

and we see that the error gradient y_θ^n gives double the asymptotically correct estimate. The problem is that $y(\theta;h)$ is not a differentiable function of θ.

Despite the lack of theoretical support, good results are reported in some shock calculations where the deferred correction error estimate blows up.

7. Using two Different Tolerances

This obvious approach is mentioned in most surveys. In addition to the original solution η computed with some requested tolerance τ, one computes a second solution η' with some cruder tolerance $R\tau$. If global error were proportional to the tolerance, then one could extrapolate and use

$$\frac{\eta'-\eta}{R-1}$$

as an error estimate for η; but this is not recommended [36]. It is safer to use $|\eta'-\eta|$ as an uncertainty estimate for η. Even this is often unreliable because there are many algorithms for which reductions in the tolerance fail to produce

roughly proportional changes in the errors, such as the example in Gear [17, p. 101] where reducing the tolerance by two orders of magnitude has no effect on the actual errors. Conditions under which tolerance proportionality can be achieved are examined by Stetter [38].

Shampine and Baca [28] suggest a greater coupling between the two integrations to achieve greater reliability. The same method (the extrapolated midpoint rule) is used on the same mesh but the more accurate integration is required to produce on each step an estimated local error that is at most half the size of the estimated local error for the less accurate integration.

8. Using two Different Methods

Shampine and Watts [29] state that "unless there is some special relation between the methods used, its performance will be unsatisfactory on a non-asymptotic basis".

Prothero [25, p. 114] suggests for ordinary IVPs using the same method on the same mesh with and without local extrapolation.

9. Using a Method With an Exact Principal Error Equation

Work by Butcher [2] suggested to Stetter [33] a technique for obtaining asymptotically correct global error estimates for ordinary IVPs without doing a second integration. It involves the use of special Runge-Kutta or predictor-corrector formulas whose principal error term can be evaluated from local information. Additional such formulas have been proposed by Dalle Rive and Pasciutti [6] and Merluzzi and Brosilaw [23]. However, as Krogh [20] has noted, such a technique is suspect because the error estimate does not, in the case of an autonomous ODE, depend on the history of the integration. For example, the global error estimate for an orbit problem would be roughly the same after ten revolutions as it is after one!

References

1. Brandt, A.: Guide to multigrid development. In: Multigrid methods; Proc. Köln-Porz 1981 (W. Hackbusch, U. Trottenberg, eds.), Lect. Notes Math., Vol. 960, pp. 220–312. Berlin-Heidelberg-New York: Springer 1982
2. Butcher, J.C.: The effective order of Runge-Kutta methods. In: Conference on the numerical solutions of differential equations; Proc. Dundee 1969. Lect. Notes Math., Vol. 109, pp. 133–139. Berlin-Heidelberg-New York: Springer 1969
3. Byrne, G.D., Hindmarsh, A.C.: A polyalgorithm for the numerical solution of ordinary differential equations. ACM Trans. Math. Software **1**, 71–96 (1975)
4. Clenshaw, C.W., Oliver, F.W.J.: Solution of differential equations by recurrence relations. Math. Tab. Wash. **5**, 34–39 (1951)
5. Dahlquist, G.: On the control of the global error in stiff initial value problems. In: Numerical Analysis; Proc. Dundee 1981. Lect. Notes Math., Vol. 912, pp. 38–49. Berlin: Springer 1981

6. Dalle Rive, L., Pasciutti, F.: Runge-Kutta methods with global error estimates. J. Inst. Maths. Applics. **16**, 381–388 (1975)
7. Epstein, B., Hicks, D.: Computational experiments on two error estimation procedures for ordinary differential equations. AFWL-TR-78-119, Air Force Weapons Lab, Kirtland AFB, NM, 1979
8. Epstein, B., Hicks, D.L.: Comparison between two error estimation procedures. In: Information linkage between applied mathematics and industry (P.C.C. Wang, ed.), pp. 293–298. New York: Academic Press 1979
9. Eriksson, L.O.: MOLCOL – an implementation of one-leg methods for partitioned stiff ODEs. TRITA-NA-8319, Numer. Anal., R. Inst. of Technology, Stockholm, 1983
10. Fox, L.: Some improvements in the use of relaxation methods for the solution of ordinary and partial differential equations. Proc. R. Soc. Lond., Ser. A **190**, 31–59 (1947)
11. Fox, L.: Ordinary differential equations: Boundary-value problems and methods. In: Numerical solution of ordinary and partial differential equations (L. Fox, ed.), pp. 58–72. Oxford: Pergamon Press 1962
12. Frank, R.: The method of iterated defect-correction and its application to two-point boundary value problems, I. Numer. Math. **25**, 409–419 (1976)
13. Frank, R.: The method of iterated defect-correction and its application to two-point boundary value problems, II. Numer. Math. **27**, 407–420 (1977)
14. Frank, R., Hertling, J., Ueberhuber, C.W.: Iterated defect correction based on estimates of the local discretization error. Report No. 18/76, Inst. Numer. Math., Technical University of Vienna, 1976
15. Frank, R., Macsek, F., Ueberhuber, C.W.: Asymptotic error expansions for defect correction iterates (manuscript) c1982
16. Frank, R., Ueberhuber, C.W.: Iterated defect corrections for differential equations, I: Theoretical results. Computing **20**, 207–228 (1978)
17. Gear, C.W.: Numerical initial value problems in ordinary differential equations. Englewood Cliffs, N.J.: Prentice-Hall 1971
18. Johnson, L.W., Riess, R.D.: Numerical analysis (2nd edition). Reading, MA: Addison-Wesley 1982
19. Joyce, D.C.: Survey of extrapolation processes in numerical analysis. SIAM Rev. **13**, 435–490 (1971)
20. Krogh, F.T.: Private communication (*c*1975)
21. Lindberg, B.: Error estimation and iterative improvement for discretization algorithms. BIT **20**, 486–500 (1980)
22. Mayers, D.F.: Ordinary differential equations: Prediction and correction; deferred correction. In: Numerical solution of ordinary and partial differential equations (L. Fox, ed.), pp. 28–45. Oxford: Pergamon Press 1962
23. Merluzzi, P., Brosilaw, C.: Runge-Kutta integration algorithms with built-in estimates of the accumulated truncation error. Computing **20**, 1–16 (1978)
24. Pereyra, V.: On improving an approximate solution of a functional equation by deferred corrections. Numer. Math. **8**, 376–391 (1966)
25. Prothero, A.: Estimating the accuracy of numerical solutions to ordinary differential equations. In: Computational techniques for ordinary differential equations (I. Gladwell, D.K. Sayers, eds.), pp. 103–128. London: Academic Press 1980
26. Robinson, A., Prothero, A.: Global error estimates for solutions to stiff systems of ordinary differential equations (contributed paper). Dundee Numer. Anal. Conf. 1977
27. Shampine, L.F.: Global error estimation for stiff ODEs (manuscript). Dundee Numer. Anal. Conf. 1983
28. Shampine, L.F., Baca, L.S.: Global error estimates for ODEs based on extrapolation methods. SIAM J. Sci. Stat. Comput. **6**, 1–14 (1985)
29. Shampine, L.F., Watts, H.A.: Global error estimation for ordinary differential equations. ACM Trans. Math. Software **2**, 172–186 (1976)
30. Skeel, R.D.: A theoretical framework for proving accuracy results for deferred corrections. SIAM J. Numer. Anal. **19**, 171–196 (1982)
31. Skeel, R.D.: Computational error estimates for stiff ODEs. In: Proceedings of the first international conference on computational mathematics; Benin City 1983 (S.O. Fatunla, ed.). Dublin: Boole Press (To appear)

32. Spijker, M.: On the structure of error estimates for finite difference methods. Numer. Math. **18**, 73–100 (1971)
33. Stetter, H.J.: Local estimation of the global discretization error. SIAM J. Numer. Anal. **8**, 512–523 (1971)
34. Stetter, H.J.: Analysis of discretization methods for ordinary differential equations. Berlin-Heidelberg-New York: Springer 1973
35. Stetter, H.J.: Economical global error estimation. In: Stiff differential systems (R.A. Willoughby, ed.), pp. 245–258. New York-London: Plenum Press 1974
36. Stetter, H.J.: Global error estimation in ODE-solvers. In: Numerical analysis Proc. Dundee 1977 (G.A. Watson, ed.), Lect. Notes Math., Vol. 630, pp. 179–189. Berlin-Heidelberg-New York: Springer 1978
37. Stetter, H.J.: The defect correction principle and discretization methods. Numer. Math. **29**, 425–443 (1978)
38. Stetter, H.J.: Tolerance proportionality in ODE-codes. In: Working papers for the 1979 SIGNUM meeting on numerical ordinary differential equations (R.D. Skeel, ed.), pp. 10.1–10.6. UIUCDCS-R-79-963, University of Illinois, Urbana 1979
39. Stetter, H.J.: Defect control and global error estimates. Manuscript, presented at a meeting at the University of Toronto, July 15–16 (1982)
40. Zadunaisky, P.E.: On the estimation of errors propagated in the numerical integration of ordinary differential equations. Numer. Math. **27**, 21–39 (1976)

Received October 18, 1984 / July 9, 1985

Note Added in Proof

Versions A, B and L of difference correction can all be expressed as special cases of the Newton-like process

$$\phi\left(\eta' - \frac{1}{\mu}(\bar{\eta} - \eta)\right) - \phi(\eta') = \frac{1}{\mu}\bar{\phi}(\eta), \qquad \mu \neq 0,$$

in which we are given η and η' and must solve for $\bar{\eta}$. The special case $\phi(\eta') = 0$ and $\mu > 1$ is suggested in [52, 45] as version C of defect correction. A similar generalization for Zadunaisky's technique yields

$$\phi\left(\eta' - \frac{1}{\mu}(\bar{\eta} - \eta);\ -\frac{1}{\mu}F(\nabla\eta)\right) - \phi(\eta') = 0.$$

The limiting case $\mu \to \infty$ does not yield technique 2L ("linearized differential correction") unless F is an affine operator. For this reason technique 2L ought to be classified separately. Moreover, the name "discrete Newton method" introduced by Böhmer [42] would be a better choice. Theoretical justification and practical details are provided in [41–44]. For such methods properties (E) and (F) are equivalent. In addition, the term "linear method" [42, 30] is easily confused with the different concept of "linear operator" and for this reason the term "additive method" [41, 43–45] is preferable. The approach of Robinson and Prothero [26] (see also [48, 53]) is better classified as linearized difference correction. Richardson extrapolation is chosen by Gladwell [49] for the NAG library routine DO2BDF, and a reliability indicator is constructed for GERK by Dekker and Verwer [47]. The first detailed discussion of tolerance proportionality is provided by Shampine and Gordon [54]. The approach of Shampine and Baca [28] is better classified as using two different methods. One integration is required to be at least $O(h^2)$ more accurate than the other. Other relevant work is in [46, 50, 51, 55, 56].

References Added in Proof

41. Allgower, E.L., Böhmer, K., McCormick, S.: Discrete correction methods for operator equations. In: Numerical solution of nonlinear equations, Proceedings, Bremen 1980 (E.L. Allgower, K. Glashoff, H.-O. Peitgen, eds.), pp. 30–97. Lecture Notes in Mathematics, Vol. 878. Berlin-Heidelberg-New York: Springer 1981
42. Böhmer, K.: A defect correction method for functional equations. In: Approximation theory, Proceedings, Bonn 1976 (R. Schaback, K. Scherer, eds.), pp. 16–29. Lecture Notes in Mathematics, Vol. 556. Berlin-Heidelberg-New York: Springer 1976
43. Böhmer, K.: Discrete Newton methods and iterated defect corrections. Numer. Math. **37**, 167–192 (1981)
44. Böhmer, K.: Asymptotic error expansions and discrete Newton methods. In: Numerical integration of differential equations and large linear systems, Proceedings, Bielefeld 1980 (J. Hinze, ed.), pp. 292–300. Lecture Notes in Mathematics, Vol. 968. Berlin-Heidelberg-New York: Springer 1982
45. Böhmer, K., Hemker, P., Stetter, H.J.: The defect correction approach. In: [46], pp. 1–32
46. Böhmer, K., Stetter, H.J. (eds.): Defect correction methods: theory and applications. Computing Supplementum; 5. Vienna: Springer 1984
47. Dekker, K., Verwer, J.G.: Estimating the global error of Runge-Kutta approximations for ordinary differential equations. In: Differential-difference equations (L. Collatz et al., eds.), pp. 55–71. ISNM Series, Vol. 62. Basel: Birkhäuser 1983
48. Dew, P.M., West, M.R.: Estimating and controlling the global error in Gear's method. BIT **19**, 135–137 (1979)
49. Gladwell, I.: Initial value routines in the NAG library. ACM Trans. Math. Software **5**, 386–400 (1979)
50. Hackbusch, W.: Bemerkungen zur iterierten Defektkorrektur und zu ihrer Kombination mit Mehrgitterverfahren. Revue Roumaine Math. Pures Appl. **26**, 1319–1329 (1981)
51. Hanson, P.M., Walsh, J.E.: Asymptotic theory of the global error and some techniques of error estimation. Numer. Math. **45**, 51–74 (1984)
52. Hemker, P.W.: Introduction to multigrid methods. Nw. Arch. Wisk. 190 (To appear)
53. Shampine, L.F.: Global error estimation for stiff ODEs. SAND79-1587, Sandia National Lab., Albuquerque, NM, 20 pp., 1979
54. Shampine, L.F., Gordon, M.K.: Computer solution of ordinary differential equations: the initial value problem. San Francisco: Freeman 1975
55. Shniad, H.: Global error estimation for the implicit trapezoidal rule. BIT **20**, 120–121 (1980)
56. Stetter, H.J.: Global error estimation in ordinary initial value problems. In: Numerical integration of differential equations and large linear systems, Proceedings, Bielefeld 1980 (J. Hinze, ed.), pp. 269–279. Lecture Notes in Mathematics, Vol. 968. Berlin-Heidelberg-New York: Springer 1982