Error estimation techniques based on defect computation and global reconstruction

Winfried Auzinger

Institute for Analysis und Scientific Computing, Vienna University of Technology, Wiedner Hauptstrasse 8-10/E101, 1040 Vienna, Austria

Abstract. The well-known technique of defect correction has been used in various moldings for numerical integration of differential or integral equations. Essentially, it can be traced back to [10] where the idea was presented in the context of initial value problems for ODEs, and to [9] where a general discussion of the principle is given. Here we focus on use of this principle as a tool in estimating local or global errors in a reliable and efficient way. Our general setting and guiding principle is presented in Section 1. In Section 2 we first consider boundary value problems for ODEs. Since the main purpose of an error estimate is mesh adaptation, an essential requirement is that it has to be robust with respect to arbitrarily distributed mesh points, and we will exemplify how this can be achieved. In particular, we consider explicit first and second order problems. We sketch the idea how to argue asymptotic correctness and present a numerical example. We also propose a related approach for estimating the error of a splitting scheme for evolution equations. Some remarks on elliptic PDEs are given in Section 3.

Keywords: error estimation, defect correction
PACS: 02.60.Lj, 02.70.Bf, 02.70.Jn

1. ABSTRACT SETTING

Let

\[ \mathcal{L} u = \mathcal{F}(u) \]  

(1)

represent a quasilinear differential equation with explicit, linear leading part \( \mathcal{L} u \). The exact solution for given initial or boundary data is denoted by \( u_* = u_*(x) \) and is assumed to be locally unique. We assume that a numerical approximation \( \tilde{u} \) for \( u_* \) has been computed, \(^1\) and we wish to estimate its error \( \tilde{e} := \tilde{u} - u_* \). If \( \tilde{u}(x) \) is a continuous function, e.g., a collocation polynomial, it is natural to consider the defect (or residual) of \( \tilde{u} \) with respect to (1),

\[ \tilde{d} := \mathcal{L} \tilde{u} - \mathcal{F}(\tilde{u}) \]  

(2)

as a first measure for the quality of \( \tilde{u} \). The classical idea due to [10] is to consider the ‘neighboring problem’ \( \mathcal{L} u = \mathcal{F}(u) + \tilde{d} \) with exact solution \( \tilde{u} \), compute its numerical approximation and use its error \( \tilde{e} \) as an estimate for \( \tilde{e} \). In other words: an improved approximation \( \tilde{u} - \tilde{e} \approx u_* \) is obtained via reconstruction, i.e., backsolving using the defect information. This procedure is often continued in an iterative fashion, cf., e.g. [1].

For the purpose of estimating \( \tilde{e} \) (without iterating) it was proposed in [9] to use any simple, low order scheme in the backsolving process. We are adopting this point of view but realize it in a modified way. To this end, let

\[ Lu = \mathcal{F}(u) \]  

(3)

represent your favorite low order, stable method, e.g., a simple finite difference scheme. Furthermore we assume that the given problem (1) can recast into a form where the leading term has the same form as in (3), i.e.,

\[ \mathcal{L} u = \mathcal{F}(u) \quad \rightarrow \quad Lu = G(u) \]  

(4)

with \( G \) appropriately defined, such that \( Lu_* - G(u_*) = 0 \). With this reformulation, we define the defect of \( \tilde{u} \) as \( \tilde{d} := L\tilde{u} - G(\tilde{u}) \) instead of (2). In practice, exact evaluation of \( G \) will not be possible and needs to be approximated, \( G \approx \hat{G} \), typically involving a sufficiently accurate quadrature scheme. With this numerically evaluated defect,

\[ \hat{d} := L\tilde{u} - \hat{G}(\tilde{u}) \]  

(5)

\(^1\) We are not very specific in denoting side conditions as well as mesh transfer and interpolation operators involved in the process.
we consider the solutions $u_0$, $u_d$ of
\[ Lu_0 = F(u_0), \quad Lu_d = F(u_d) + \hat{d}, \quad \text{satisfying} \quad L(u_d - u_0) = F(u_d) - F(u_0) + \hat{d} \] (6)

On the other hand, for $\hat{u}$ and $u_*$ we have
\[ L(\hat{u} - u_*) = \hat{G}(\hat{u}) - \hat{G}(u_*) + \Delta G(u_*) + \hat{d}, \quad \Delta G := \hat{G} - G \] (7)
with quadrature error $\Delta G(u_*)$. Since $F$ is an approximation for $G$, this suggests the a posteriori error estimate
\[ \hat{e} := u_d - u_0 \approx \hat{u} - u_* = \hat{e} \] (8)

From (6),(7) we see that the deviation of $\hat{e}$, i.e., the error $\hat{\delta} := \hat{e} - \hat{\delta}$ of the estimate satisfies \(^2\)
\[ L\hat{\delta} = (F(u_d) - F(u_0)) - (F(\hat{u}) - F(u_*)) + (\Delta F(\hat{u}) - \Delta F(u_*)) - \Delta G(u_*), \quad \Delta F := F - G \] (9)

The essential point is that $\hat{\delta}$ is not influenced by any approximation error concerning the leading part $\mathcal{L}u$. This is also essential for the robustness of the estimate over variable meshes. The quality of our estimate depends only on

- the size of $\Delta F(\hat{u}) - \Delta F(u_*)$,
- the numerical approximation error $\Delta G(u_*)$,
- and the stability of the auxiliary scheme (3).

2. APPLICATIONS

2.1. Boundary value problems for ODEs

The approach sketched in Section 1 has been realized and analyzed in [2, 4] in the following context:

- $\mathcal{L}u = \mathcal{F}(u)$ is a system of first ODEs, together with boundary conditions on an interval $[a, b]$,
- $\hat{u} \in C^0[a, b]$ is a collocation polynomial of degree $m$,
- $Lu = F(u)$ is a simple difference scheme over the collocation mesh $\{x_j\}$, e.g., an Euler or midpoint scheme.

Recasting the problem according to (4) is straightforward via integrating $u'$ between successive collocation points, such that $Lu = G(u)$ is a locally weighted version of the ODE. An appropriate approximation $\hat{G}(u)$ of $G(u)$ is obtained via quadrature over collocation subintervals, and the outcome is closely related to a higher order Runge-Kutta scheme defining the defect of $\hat{u}$. The analysis of the resulting error estimate $\hat{e}$ is based on the asymptotic properties of the collocation error, namely $\hat{e} = O(h^m)$, $\hat{\delta} = O(h^m)$.\(^3\) This permits an estimate of the critical quantity $\Delta F(\hat{u}) - \Delta F(u_*)$, leading to the conclusion that the estimate $\hat{e}$ is asymptotically correct, i.e., the deviation $\hat{\delta} = \hat{e} - \hat{\delta}$ satisfies $\hat{\delta} = O(h^{m+1})$. In [3, 6] this analysis was extended to the case of singular boundary value problems.

Let us discuss how this approach can be extended to the case of direct collocation approximations $\hat{u} \in C^4[a, b]$ for a second order two-point boundary value problem, for a quasilinear ODE with leading part $\mathcal{L}u = u''$,
\[ u''(x) = F(x, u(x), u'(x)), \quad x \in (a, b) \] (10)

For simplicity of presentation, we assume an equidistant mesh $\{x_j\}$ with meshwidth $h$; but this is on no way essential. $Lu$ is the standard second order difference operator. On an interval $[x_{j-1}, x_{j+1}] = [x_j - h, x_j + h]$, the identity
\[ Lu(x_j) = (\mathcal{X} u'')(x_j), \quad \text{with} \quad (\mathcal{X} f)(x) := \int_{-1}^{1} K(\xi) f(x + \xi h) d\xi, \quad K(\xi) = 1 - |\xi| \] (11)
is valid for each $u \in C^2[a, b]$. This extends to the case where a jump in $u''$ occurs at $x = x_j$ (this is the case for $u = \hat{u}$ at the end of a collocation subinterval). This leads to a reformulation in the spirit of (4) of the given ODE in the form
\[ Lu(x_j) = (\mathcal{X} f)(x_j), \quad f(x) = F(x, u(x), u'(x)) \] (12)

\(^2\) For linear problems, the estimate can be directly obtained by solving for $\hat{e}$ in (6), and (9) simplifies to $L\hat{\delta} = F(\hat{\delta}) + \Delta F(\hat{e}) - \Delta G(u_*)$.

\(^3\) Here, collocation at $m$ arbitrarily distributed inner collocation nodes is assumed, with $m$ even.
In this context, (15) plays the role of the auxiliary scheme, which is an approximation for the history shown as log-log-plot in Figure 1. The observed orders are using irregularly spaced collocation nodes. On coherent refinement of the collocation subintervals we obtain the error estimate (16). We consider a splitting step with increment \( t = h \):

\[
\begin{align*}
M &= A + B; & E(h) &:= \exp(hM) \approx \exp(hB) \exp(hA) =: \tilde{E}(h) \\
\end{align*}
\]

Splitting schemes are motivated by the fact that in many applications, \( \exp(hA) \) and \( \exp(hB) \) are much easier to realize or approximate than \( \exp(hM) \). See, e.g., [7], and references therein, for typical applications. Typically, the quality of a splitting approach relies on the size of the problem-dependent commutator \([A, B]\), and higher commutators.

To bring the ideas from Section 1 into play, we think of the Lie-Trotter approximation as a continuous object, \( \tilde{E} = \tilde{E}(t) \). It satisfies the Sylvester equation

\[
\tilde{E}' = \tilde{E}A + B\tilde{E}, \quad \tilde{E}(0) = I
\]

Thus, the defect \( \tilde{D} = \tilde{D}(t) \) of \( \tilde{E} \) is well-defined and takes the form of a commutator function, \( \tilde{D} := \tilde{E}' - M\tilde{E} = [\tilde{E}, A] \).

In this context, (15) plays the role of the auxiliary scheme, which is an approximation for \( E' = ME \), and \( \tilde{E}, E \) play the role of \( \tilde{u}, u \). Now we approximate the relation for the error \( \tilde{e} := \tilde{E} - E \),

\[
\tilde{e}' = M\tilde{e} + \tilde{D}, \quad \tilde{e}(0) = 0
\]

by the inhomogeneous Sylvester equation

\[
\tilde{e}' = \tilde{e}A + B\tilde{e} + \tilde{D}, \quad \tilde{e}(0) = 0
\]

where \( x_j \) runs over the complete collocation mesh including endpoints. Hence, the error estimation procedure is well-defined along the lines of Section 1. Sufficiently accurate quadrature approximations for the occurring integrals of the type (11) are readily constructible, and minor details like approximation of the first derivative in the basic difference approximation, or incorporating boundary conditions containing first derivatives are easy to fix. A complete analysis of this approach is given in the forthcoming thesis [8]. It turns out that for the second order case the estimate is always of a very high quality, namely with \( \hat{\delta} = \hat{e} - \tilde{e} = \Theta(M^{m+2}) \) for \( e = \Theta(M^m) \). Let us illustrate this for the nonlinear problem

\[
u'' = 1 - (u')^2, \quad u(0) = u(1) = 0
\]

with known analytic solution. \( \tilde{u} \) is computed as a piecewise polynomial collocating approximation of degree \( m = 3 \) using irregularly spaced collocation nodes. On coherent refinement of the collocation subintervals we obtain the error history shown as log-log-plot in Figure 1. The observed orders are \( m = 3 \) for \( ||\tilde{e}||_\infty \) and \( m + 2 = 5 \) for \( ||\hat{\delta}||_\infty = ||\hat{\delta}||_\infty \).

For related results concerning implicit first order systems, in particular index 1 DAEs, see [5].

FIGURE 1. Error and deviation of estimate for Example (13)
with exact solution (still to be practically approximated by quadrature)

\[ \tilde{e}(h) = \int_0^h \exp((h-t)B)\tilde{D}(t)\exp((h-t)A)\,dt \quad (18) \]

We may also think of recasting the original problem \( E' = ME \) into the form (in the spirit of (4))

\[ E' = EA + BE + D, \quad E(0) = I \quad (19) \]

Here, \( D := [A,E] \) is the truncation error of \( E \) with respect to the auxiliary scheme (15). The error \( \tilde{e} \) satisfies

\[ \tilde{e}' = \tilde{e}A + B\tilde{e} - D, \quad \tilde{e}(0) = 0 \quad (20) \]

Approximating \(-D = [E,A] \) by \( \tilde{D} = [\tilde{E},A] \) and backsolving again results exactly in the estimate \( \tilde{e} \) from (17),(18).

For an analysis of this approach, we observe that the defect \( \tilde{D} \) also satisfies a Sylvester equation,

\[ \tilde{D}' = \tilde{D}A + B\tilde{D} - [A,B]\tilde{E}, \quad \tilde{D}(0) = 0 \quad (21) \]

This shows \( \tilde{D}(t) = \mathcal{O}(t) \), where the \( \mathcal{O}() \) depends only on \( ||[A,B]|| \) and \( ||\tilde{E}|| \), and as a byproduct we, via (16), regain the well-known a priori estimate \( \tilde{e}(h) = \mathcal{O}(h^2) \). But our particular goal is a practical posteriori error estimate. To this end it can be shown that (18) is indeed an asymptotically correct approximation to \( \tilde{e}(h) \), and the same is true for the trapezoidal approximation to (18),

\[ \tilde{e}(h) := \frac{h}{2} (\tilde{D}(0) + \tilde{D}(h)) = \frac{h}{2} [\tilde{E}(h),A] \quad (22) \]

It satisfies \( \hat{e}(h) = \tilde{e}(h) - \tilde{e}(h) = \mathcal{O}(h^2) \), where the \( \mathcal{O}() \) depends also on \( ||[A,B]|| \) and \( ||B, [A,B]|| \).

Details, generalizations and applications will be discussed in a forthcoming paper by W. Auzinger and O. Koch.

### 3. DISCUSSION; ELLIPTIC PDE S

The successful applicability of our approach depends on the realization of the transformation (4). In the context of subsection 2.2 this was quite straightforward. In the ODE case (subsection 2.1) it could be established via a weighted integration with compactly supported kernel \( K(\xi) \). It can also be shown, e.g., \( K(\xi) \) from (11) an be represented as \( L_h^1 U(0) \), where \( L_h^1 \) is the local adjoint difference operator on a dimensionless mesh, and where \( U(x) \) is the fundamental solution associated with \( \Delta u = \psi \). For PDEs like the Poisson equation in 2D or 3D, such a transformation is much more complicated due the singularity of the fundamental solution associated with \( \Delta u \). Nevertheless, a reasonable local approximation may still be possible based on a Green’s identity over mesh elements. It is an open question worth investigating whether this enables us to extend the present approach to the PDE case, and how its performance compares with established error estimation techniques for elliptic problems.

### REFERENCES