Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

P.J. van der Houwen, B.P. Sommeijer

Improving the stability of predictor-corrector methods by residue smoothing

Department of Numerical Mathematics Report NM-R8707 February

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

Improving the Stability of Predictor-Corrector Methods by Residue Smoothing

P.J. van der Houwen, B.P. Sommeijer
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

Residue smoothing is usually applied in order to accelerate the convergence of iteration processes. Here, we show that residue smoothing can also be used in order to increase the stability region of predictor-corrector methods. We shall concentrate on increasing the real stability boundary. The iteration parameters and the smoothing operators are choosen such that the stability boundary becomes as large as $c(m,q)m^2 4^q$ where m is the number of right-hand side evaluations per step, q the number of smoothing operations applied to each right-hand side evaluation, and $c(m,q)$ a slowly varying function of m and q, of magnitude 1.3 in a typical case. Numerical results show that, for a variety of linear and nonlinear parabolic equations in one and two spatial dimensions, these smoothed predictor-corrector methods are at least competitive with conventional implicit methods.

1980 Mathematics Subject Classification: 65M20, 65L20.
Key Words & Phrases: numerical analysis, predictor-corrector methods, parabolic differential equations, stability, residue smoothing.

1. Introduction

Consider the initial value problem

$$\frac{dy}{dt} = f(t,y), \qquad y(t_0) \text{ prescribed}, \quad t_0 \leqslant t \leqslant T \tag{1.1}$$

and apply the implicit linear k-step method defined by the characteristic polynomials

$$\rho(\zeta) = \sum_{i=0}^{k} a_i \zeta^{k-i}, \qquad \sigma(\zeta) = \sum_{i=0}^{k} b_i \zeta^{k-i} \tag{1.2}$$

with $a_0 = 1$ and $b_0 \neq 0$. Then, in order to obtain the numerical approximation y_{n+1} to $y(t_{n+1})$, we have to solve the equation

$$y - b_0 \tau f(t_{n+1}, y) - \Sigma_n = 0, \tag{1.3a}$$

where $\tau := t_{n+1} - t_n$ and Σ_n denotes the sum of already computed back values, i.e.,

$$\Sigma_n := \sum_{i=1}^{k} [-a_i y_{n+1-i} + b_i \tau f(t_{n+1-i}, y_{n+1-i})]. \tag{1.3b}$$

We shall be particularly interested in the case where (1.1) originates from the semidiscretization of parabolic initial-boundary-value problems in two or three spatial dimensions. In such cases, the solution of (1.3) is usually rather time consuming. If functional iteration is used (e.g., predictor-corrector iteration), then rather small values of τ are required, not to obtain a sufficiently accurate solution of (1.3), but rather to keep the integration process stable. Therefore, functional iteration may cost a large amount of computational effort to reach the end point.

In van der Houwen & Sommeijer (1983) generalizations of predictor-corrector iteration, which allow for much larger values of τ and yet preserving stability, have been proposed. In the special case of semi-discrete partial differential equations, the efficiency of these *generalized predictor-corrector methods* (*GPC* methods) can be further improved by employing *residue smoothing*; that is, instead of

Report NM-R8707
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

(1.3), we solve the "preconditioned" equation

$$\mathbb{S}(y - b_0\tau f(t_{n+1},y) - \Sigma_n) = 0, \qquad (1.4)$$

where $\mathbb{S}$ is a nonsingular smoothing operator which removes the high frequencies from the vector to which it is applied. Residue smoothing has been used in several papers in order to accelerate the *convergence* of iteration processes (cf. e.g. LERAT (1979), JAMESON (1983), TURKEL (1985), JAMESON & MAVRIPLIS (1985), VAN DER HOUWEN et al. (1987)).

In this paper we show that residue smoothing can also be used to improve the *stability* of predictor-corrector methods. The smoothing operators employed are of *explicit* type (contrary to the smoothing operators employed by LERAT, JAMESON and TURKEL which are of implicit type), and are related to the smoothing techniques used in WUBS (1986) and VAN DER HOUWEN et al. (1986).

In Section 2 an expression for the local error of smoothed *GPC* methods is derived. From this expression the order conditions of the method easily follow. Section 3 presents the main part of the paper. It provides expressions for the iteration parameters which generate "almost" maximal *real stability boundaries* for a class of predictor-corrector pairs. The magnitude of the stability boundary β is of the form

$$\beta = c(m,q)m^2 4^q, \qquad (1.5)$$

where $c(m,q)$ is a slowly varying function of (m,q), m is the number of iterations performed by the *GPC* method, and q is the number of basic matrix-vector multiplications needed to apply the smoothing operator $\mathbb{S}$ (here, a basic matrix-vector multiplication is a tridiagonal matrix-vector multiplication in one-dimensional problems and a block-tridiagonal matrix-vector multiplication in two-dimensional problems).

The smoothed *GPC* method has been applied to a variety of parabolic Dirichlet-boundary-value problems, both of linear and nonlinear type and both in one and two spatial dimensions; its efficiency has been compared with the efficiency of more conventional implicit methods. On the basis of computational effort versus accuracy, the conventional methods are competitive for one-dimensional problems, but considerably less efficient in two-dimensional problems (see the tables of results in Section 5). However, in our opinion, the main advantage of the smoothed *GPC* method, is its extremely simple implementation (cf. Section 4).

2. THE *SGPC* METHOD

If a *GPC* method is applied to equation (1.4) we obtain the computional scheme:

$$y_{n+1}^{(0)} = \text{some initial approximation to the solution of (1.3)},$$

$$y_{n+1}^{(j)} = \sum_{l=1}^{j}[(\mu_{jl} + \frac{\bar{\mu}_{jl}}{b_0})y_{n+1}^{(l-1)} - \frac{\bar{\mu}_{jl}}{b_0}\mathbb{S}(y_{n+1}^{(l-1)} - b_0\tau f_{n+1}^{(l-1)} - \Sigma_n)], \qquad j = 1,2,...,m, \qquad (2.1a)$$

$$y_{n+1} = y_{n+1}^{(m)},$$

where $f_{n+1}^{(l)} := f(t_{n+1}, y_{n+1}^{(l)})$ and where the parameters $\bar{\mu}_{jl}$ and μ_{jl} satisfy the condition

$$\sum_{l=1}^{j}(\mu_{jl} + \frac{\bar{\mu}_{jl}}{b_0}) = 1, \qquad j = 1,2,...,m. \qquad (2.1b)$$

By virtue of this condition the solution of (1.3) satisfies the scheme (2.1). The smoothed *GPC* method (*SGPC* method) defined by (2.1) reduces to the *GPC* method analysed in VAN DER HOUWEN & SOMMEIJER (1983) if we set $\mathbb{S}=I$, I denoting the identity matrix. Following the proof of Theorem 3.1 given in this reference, we arrive at the theorem:

THEOREM 2.1 *Let f be sufficiently differentiable, and define the polynomials*

$$P_0(x)=1, \qquad P_j(x)=\sum_{l=1}^{j}[\mu_{jl}+\bar{\mu}_{jl}x]P_{l-1}(x), \qquad j=1,2,\ldots,m, \tag{2.2}$$

and the matrices

$$Z:=\tau\frac{\partial f}{\partial y}(t_{n+1},\eta), \qquad \hat{Z}:=\mathbb{S}Z+\frac{I-\mathbb{S}}{b_0}, \tag{2.3}$$

where η is the solution of (1.3). Then, the local error of $y_{n+1}^{(j)}$ in (2.1) is given by

$$y_{n+1}^{(j)}-y(t_{n+1})=[I-P_j(\hat{Z})](\eta-y(t_{n+1}))+P_j(\hat{Z})(y_{n+1}^{(0)}-y(t_{n+1}))$$
$$+\,O(\tau^{3+2\min(p,\tilde{p})}), \qquad j=0,\ldots,m,$$

where p and $\tilde{p}$ respectively denote the orders of accuracy of the corrector (1.3) and the predictor used to obtain $y_{n+1}^{(0)}$. □

COROLLARY 2.1 *Let the iteration polynomial $P_m(x)$ have a zero at $z=0$ of multiplicity r and let the smoothing operator $\mathbb{S}$ satisfy the condition*

$$\mathbb{S}=I+O(\tau^s) \text{ as } \tau\to 0, \qquad s>0.$$

Then the SGPC method has order

$$p^*:=\min\{p,\tilde{p}+r,\tilde{p}+r+s-1,2(1+\min(p,\tilde{p}))\}. \quad \square$$

Thus, if low order predictors are used, we have to choose $P_m(x)$ such that r is sufficiently high in order to compensate the low value of $\tilde{p}$. For example, for given p and $\tilde{p}$, we can only achieve $p^*=p$ if $\tilde{p}\geqslant(p-2)/2$ and $r\geqslant p-\tilde{p}-\min\{0,s-1\}$ (we observe that s is not necessarily an integer). Furthermore, we remark that $P_j(x)$ should always satisfy the condition $P_j(1/b_0)=1$ in order to fulfil condition (2.1b).

3. STABILITY

3.1 The characteristic equation

For the stability analysis we employ the linear test equation

$$\frac{dy}{dt}=Jy, \tag{3.1}$$

where J is a constant matrix. Let $y_{n+1}^{(0)}$ be computed by an explicit linear $\tilde{k}$-step method defined by polynomials $\tilde{\rho}(\zeta)$ and $\tilde{\sigma}(\zeta)$, and assume $\tilde{a}_0=1$. Then, on substitution of (3.1) into (2.1) we are led to the recursion

$$(P_m(\hat{Z})-I)\mathbb{S}(\rho(E)-Z\sigma(E))y_n=(I-b_0\hat{Z})P_m(\hat{Z})(\tilde{\rho}(E)-Z\tilde{\sigma}(E))y_{n+k-\tilde{k}}, \tag{3.2}$$

where we used the notation introduced in (2.2) and (2.3). From this recursion we easily deduce the following theorem:

THEOREM 3.1. *Let $\mathbb{S}$ and Z share the same eigensystem and let z and $\hat{z}$ denote the eigenvalues of Z and $\hat{Z}$ corresponding to the same eigenvector. Then the characteristic equation of the SGPC method in $P(E\mathbb{S}C)^mE$ mode is given by*

$$\rho(\zeta)-z\sigma(\zeta)=\frac{(1-b_0z)P_m(\hat{z})}{P_m(\hat{z})-1}[\tilde{\rho}(\zeta)-z\tilde{\sigma}(\zeta)]\zeta^{k-\tilde{k}}.\square \tag{3.3}$$

Let $z^*:=P_m(\hat{z})$, then we define the *stability domain* $\mathfrak{D}$ by the set of points (z,z^*) in the (z,z^*)-plane where (3.3) has its roots on the unit disk. Under the assumption of Theorem 3.1 we have that $\hat{z}$ is a function of z. This leads us to the stability criterion

$$(z, P_m(\hat{z}(z))) \in \mathfrak{D} \quad \text{for all eigenvalues } z \text{ of } Z = \tau J. \tag{3.4}$$

In Figure 3.1 a few stability domains are plotted in the real (z,z^*)-plane for the case where $\{\rho,\sigma\}$ is defined by the *p-th* order *backward differentiation* formula and $\{\tilde{\rho},\tilde{\sigma}\}$ is defined by the $\tilde{p}$-th order *extrapolation* formula, i.e., $\tilde{\rho}(\zeta)=(\zeta-1)^{\tilde{p}+1}$ and $\tilde{\sigma}(\zeta)\equiv 0$ (the generated *SGPC* method will be called a *smoothed* $EP_{\tilde{p}}-BD_p$ *method).*

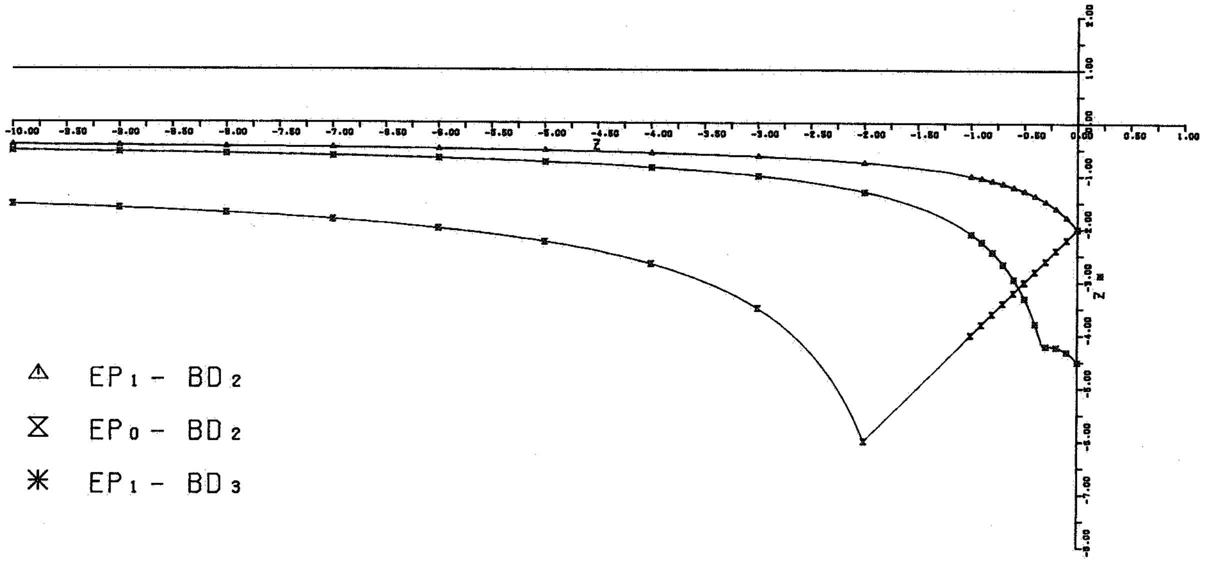

FIGURE 3.1. Stability domains of some $EP_{\tilde{p}}-BD_p$ methods

In order to apply the stability condition (3.4) we need to know the function $\hat{z}=\hat{z}(z)$. This will be discussed in the next subsection in the case where the smoothing operator $\mathbb{S}$ is suitable for use in *parabolic* problems. The general effect of these smoothing operators is the reduction of the length of the (real) eigenvalue interval of the matrix $\hat{Z}$ in (2.3) in comparison with the length of the eigenvalue interval of Z. It will be shown that such a reduction leads to increased real stability boundaries of the *SGPC* method.

3.2 Smoothing operators for parabolic problems

For elliptic difference equations with Dirichlet-type boundary conditions suitable smoothing operators for residue smoothing were derived in VAN DER HOUWEN et al. (1987). If (1.1) orginates from a parabolic problem with Dirichlet boundary conditions, i.e., $\partial f/\partial y$ possesses a negative spectrum, then (1.3) can be considered as an elliptic system of difference equations, so that these smoothing operators are expected to be suitable in the case (1.3) too. However, the boundary equations in (1.3) need some attention as we will see below.

3.2.1 One-dimensional problems

Let M be the number of internal grid points used to semi-discretize the parabolic problem. Then, the system (1.1) contains M equations approximating the parabolic equation at these internal grid points. In addition, we assume that the system (1.1) contains two equations representing the Dirichlet boundary conditions. If the boundary conditions are of the form $u(0,t)=a(t)$, $u(1,t)=b(t)$, where $u(x,t)$ denotes the solution of the parabolic problem, then the first and last equation of (1.1) are given by

$$\frac{dy_0}{dt}=\frac{da(t)}{dt}\ ,\quad \frac{dy_{M+1}}{dt}=\frac{db(t)}{dt}; \tag{3.5}$$

here, the subscripts refer to the components of the vector y and not to the time level. Thus, although the components y_0 and y_{M+1} are explicitly given by, respectively, $a(t)$ and $b(t)$, we assume that they are obtained numerically by integrating the equations (3.5) as part of the system (1.1). As a consequence, the sytem (1.3) also contains $M+2$ equations, whereas the useful approach defines a system of M equations by eliminating y_0 and y_{M+1} by means of the boundary conditions, i.e., the usual approach prescribes zero residues at the boundary points.

The reason for this unconventional approach can be traced back to the fact that we are not actually *solving* the system (1.3), but we stop the iteration process as soon as the last iterate satisfies the stability condition (3.4). Hence, the residue occurring in (2.1a) is not necessarily decreasing during the iteration process. If we ignore this feature of *GPC* methods, and just introduce zero-residues at boundary points, then we create a residue vector which may have jumps in the magnitude of its components near the boundary points. Obviously, when smoothing operators are applied to such "unsmooth" residue vectors, we introduce large errors into the scheme.

Using the equations (3.5) leads to the following boundary equations of the system (1.3):

$$\begin{aligned} &y_0 - b_0\tau\frac{da}{dt}(t_{n+1}) - (\Sigma_n)_0 = 0, \\ &y_{M+1} - b_0\tau\frac{db}{dt}(t_{n+1}) - (\Sigma_n)_{M+1} = 0. \end{aligned} \tag{3.6}$$

We are now in a position to apply the smoothing operators $\mathbb{S}$. For convenience, we reproduce the definition of the operator below.

Let the grid be uniform, let D be the difference operator

$$D=\frac{1}{4}\begin{bmatrix} 0 & & \cdots & & 0 \\ 1 & -2 & 1 & & \\ 0 & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & 1 & -2 & 1 \\ 0 & & \cdots & & 0 \end{bmatrix}, \tag{3.7}$$

and define the matrices F_j by the recursion

$$F_1 = I + D, \qquad F_{j+1} = (I-2F_j)^2, \qquad j \geqslant 1. \tag{3.8}$$

Then, the smoothing operator $\mathbb{S}$ is defined by the matrix

$$\mathbb{S} = \prod_{j=1}^{q} F_j, \qquad q \geqslant 1. \tag{3.9}$$

The matrices F_j are easily precomputed, so that the application of $\mathbb{S}$ requires q matrix-vector multiplications. It can easily be shown that the matrices F_j are essentially tridiagonal matrices. Hence, the application of $\mathbb{S}$ does not require much computational effort. In fact, this is due to the special form of the matrix D. In this connection, we remark that D is allowed to be any difference matrix, provided that it has its eigenvalues in the interval $[-1,0]$ and is such that for any smooth grid function y we have $Dy \rightarrow 0$ as the grid is refined.

An important property of the smoothing matrix $\mathbb{S}$ is the fact that, once the difference matrix D has been chosen, it does not depend on the particular problem to be integrated.

Furthermore, we note that $\mathbb{S}=I+O(D)$ as the grid is refined. Hence, if Δ is the mesh size, then

$$\mathbb{S} = I + O(\Delta^2) \ \text{ as } \ \Delta \rightarrow 0. \tag{3.10}$$

Finally, we actually computed the matrices F_j based on the difference matrix D as defined in (3.7) and determined the factorized operator $\mathbb{S}$ (cf. (3.9)) for arbitrary values of q. For convenience of the reader, we give a FORTRAN 77 subroutine which performs this operator $\mathbb{S}$:

```
      SUBROUTINE SMOOTH(M,Q,QAPPL,U,V)
      DIMENSION U(0:M+1),V(1:M)
      INTEGER Q,QAPPL,QMAX
C--------------------------------------------------------------------
C   THIS ROUTINE PERFORMS THE SMOOTHING OPERATOR S IN ITS FACTORIZED
C   FORM (CF. (3.7),(3.8) AND (3.9)).
C--------------------------------------------------------------------
C   M - THE NUMBER OF INTERNAL GRIDPOINTS.
C   U - VECTOR OF LENGTH (0:M+1), I.E. INCLUDING THE BOUNDARY POINTS.
C       ON INPUT, U SHOULD CONTAIN THE VECTOR TO BE SMOOTHED;
C       ON RETURN, U CONTAINS THE SMOOTHED RESULTVECTOR.
C   V - AN AUXILIARY VECTOR OF LENGTH M.
C   Q - THE NUMBER OF FACTORS IN THE OPERATOR S, I.E. THE REQUIRED
C       NUMBER OF SMOOTHING FACTORS.
C   QAPPL - OUTPUT PARAMETER; QAPPL IS THE NUMBER OF SMOOTHING
C       FACTORS ACTUALLY APPLIED. NOTE, THAT QAPPL MAY BE SMALLER THAN
C       Q, BECAUSE, IF 2**Q IS LARGER THAN M+1, THEN WE APPLY THE
C       MAXIMUM NUMBER OF SMOOTHING FACTORS ALLOWED BY THIS GRID.
C
C   RDOFF IS SET TO THE MACHINE ROUNDOFF AND MAY REQUIRE AMENDMENT ON
C   DIFFERENT MACHINES.
C--------------------------------------------------------------------
      DATA RDOFF/0.71E-14/
      QMAX=LOG(M+1.0)/LOG(2.0)+1.0E3*RDOFF
      QAPPL=MIN(Q,QMAX)

      DO 70 J=1,QAPPL
      L=2**(J-1)-1
      DO 10 I=1,M
   10 V(I)=2.0*U(I)
      DO 20 I=L+1,M
   20 V(I)=V(I)+U(I-L-1)
      DO 30 I=1,M-L
   30 V(I)=V(I)+U(I+L+1)
      DO 40 I=1,L
   40 V(I)=V(I)+2.0*U(0)-U(L+1-I)
      DO 50 I=M+1-L,M
   50 V(I)=V(I)+2.0*U(M+1)-U(2*M+1-L-I)
      DO 60 I=1,M
   60 U(I)=V(I)/4.0
   70 CONTINUE
      RETURN
      END
```

3.2.2 Two-dimensional problems

In the case of two-dimensional problems we proceed as in the preceding subsection and we define the system (1.3) in such a way that the boundary equations are analogous to (3.6), just by using the time derivatives of the Dirichlet-boundary conditions.

The smoothing matrix $\mathbb{S}$ can be defined by (3.8) and (3.9) if D is replaced by the two-dimensional analogue of (3.7). However, the precomputation of the matrices F_j is not as easy as in the one-dimensional case, and more important, it depends on the geometry of the domain on which the problem is defined.

A simple modification of the smoothing operator overcomes this difficulty: let the residue occurring in (2.1a) be arranged as a two-dimensional array in the natural way; then we first smooth all the rows of this array by applying the one-dimensional smoothing matrix defined in Section 3.2.1, and next we smooth all the columns of the resulting array again by this one-dimensional smoothing matrix. In this way, the two-dimensional smoothing process is broken down into a sequence of problem-independent, one-dimensional smoothing operators.

The analysis given in VAN DER HOUWEN et al. (1987) shows that this modified smoothing process reduces the spectral radius of the matrix $\hat{Z}$ by an extra factor ≈ 0.6 when compared with the unmodified smoothing process.

3.2.3. Non-Dirichlet conditions

In the preceding dicussions we have explicitly stipulated that the boundary conditions are of Dirichlet type. In the case of non-Dirichlet conditions the timederivatives of the boundary values are not explicitly available. However, if we are able to generate these time derivatives in a stable way, then the *SGPC* method so far described can be applied without any change. The generation of stable time-derivatives of boundary values in the case of Neuman-type boundary conditions is subject of future investigation of the present authors.

3.3 The real stability boundary of SGPC methods

We shall derive an approximation to the real stability boundary of *SGPC* methods for the model situation where the difference matrix D is not defined by (3.7), but, instead, by

$$D{:} = \frac{1}{R}\frac{\partial f}{\partial y}, \tag{3.7'}$$

where R denotes the spectral radius of $\partial f/\partial y$. Notice that this matrix has its eigenvalues in the interval $[-1,0]$ (recall that $\partial f/\partial y$ was assumed to have negative eigenvalues); furthermore, $Dy \rightarrow 0$ as the grid is refined for any smooth grid function y, while (3.10) is satisfied. In particular, if system (1.1) is the standard symmetric 3-point discretization of the diffusion equation $u_t = u_{xx}$, then (3.7) and (3.7′) coincide.

The main tool in deriving the real stability boundary is the following theorem (a proof can be given along the lines of the proof of Lemma 3.2 in VAN DER HOUWEN et al. (1986)):

THEOREM 3.2. *Let $k = 2^q - 1$, then the matrix $\mathbb{S}$ defined by (3.8) and (3.9) is given by*

$$\mathbb{S}{:} = \frac{T_{k+1}(I+2D)-I}{2(k+1)^2 D}, \qquad T_l(x){:} = \cos(l \arccos x). \square$$

We observe that the numerator of this expression contains the factor D so that $\mathbb{S}$ is actually a polynomial of degree k in D.

Consider the test equation (3.1), i.e. $\partial f/\partial y = J$ and $D = J/R$. Substitution of the resulting matrix $\mathbb{S}$ into (2.3) expresses the matrices Z and $\hat{Z}$ in terms of the fixed matrix J. From this the following relation between the eigenvalues z and $\hat{z}$ of Z and $\hat{Z}$ is immediate:

$$\hat{z} = \hat{z}(z){:} = \frac{1}{b_0}[1 + \frac{\tau R}{2(k+1)^2}(b_0 - \frac{1}{z})(T_{k+1}(1 + \frac{2z}{\tau R}) - 1)]. \tag{3.11}$$

By means of this relation we can proceed with the stability criterion (3.4). Suppose that we know the range of values assumed by the function $P_m(\hat{z}(z))$ for $-\tau R \leqslant z \leqslant 0$, i.e. the values of

$$z^*_{\min} := \min_{\mathcal{I}} P_m(\hat{z}), \qquad z^*_{\max} := \max_{\mathcal{I}} P_m(\hat{z}), \tag{3.12a}$$

where $\mathcal{I}$ is the interval of $\hat{z}$-values assumed by $\hat{z}(z)$ on the interval $[-\tau R,0]$. Then the stability condition (3.4) is certainly satisfied if

$$(z,z^*_{\min}),(z,z^*_{\max}) \in \mathcal{D} \text{ for } -\tau R \leqslant z \leqslant 0. \tag{3.12b}$$

In order to find the interval $\mathcal{I}$ we first observe that $\hat{z}(z) \leqslant 1/b_0$ for all $z \in [-\tau R,0]$. Furthermore, the minimal value of $\hat{z}(z)$ will be assumed at a point in the neighbourhood of $z=0$ where $T_{k+1}(1+2z/\tau R)$ is negative. It is easily shown that this point lies in the interval $[z_0,0]$, where

$$z_0 := \frac{1}{2}\tau R(\cos(\frac{\pi}{k+1})-1). \tag{3.12c}$$

Thus,

$$\mathcal{I} = [\min_{z_0 \leqslant z \leqslant 0} \hat{z}(z), \frac{1}{b_0}]. \tag{3.12d}$$

Before proceeding with the stability criterion (3.12) it should be observed that $P_m(1/b_0)=1$, so that $z^*_{\max} \geqslant 1$ and consequently, the stability domain should at least contain the line segment $\{(z,1): -\tau R \leqslant z \leqslant 0\}$ in order to let (3.12) be true. In the following we will concentrate on cases where $\mathcal{D}$ contains this line segment (we remark that the domains shown in Figure 3.1 satisfy this assumption).

We will now derive explicit expressions for the maximum value of τR for which the *SGPC* method is stable in the sense of (3.12). This value is called the *real stability boundary* of the method.

THEOREM 3.3. *Let the predictor-corrector pair be such that* $\mathcal{D}$ *contains the rectangle* $[-\tau R,0] \times [0,1]$, *and let* $P_m(x)$ *be given by*

$$P_m(x) = P_1(x) = 1 + a(x - \frac{1}{b_0}), \qquad 0 \leqslant a \leqslant b_0.$$

Then the SGPC method possesses a real stability boundary

$$\beta(k) \geqslant \frac{k(k+2)}{3b_0} = \frac{4^{q.}-1}{3b_0} \tag{3.13}$$

for all $a \in [0,b_0]$.

PROOF. First we observe that for small values of z the function $\hat{z}(z)$ behaves as

$$\hat{z}(z) \approx (1 - \frac{k(k+2)}{3b_0\tau R})z + O(z^2),$$

so that $\hat{z}(z)$ is positive in a left neighbourhood of $z=0$ if

$$\tau R < \frac{k(k+2)}{3b_0}. \tag{3.14}$$

It can be shown that $\hat{z}(z)$ is positive for all $z \in (-\tau R,0)$ if this inequality is satisfied. Hence, the interval $\mathcal{I}$ defined in (3.12d) is contained in $[0,1/b_0]$. From the definition of $P_1(z)$ and from (3.12a) it then follows that $z^*_{\min} \geqslant 0$ and $z^*_{\max} = 1$, so that the condition of the theorem on $\mathcal{D}$ implies that (3.12b) is satisfied. Thus, (3.13) is a sufficient condition for stability. This leads us to the given lower bound on β. □

There seems to be no advantage in choosing a other than b_0 which leads us to the smoothed version of the classical predictor-corrector method in *PECE* mode:

$$y_{n+1} = y^{(0)}_{n+1} - \mathcal{S}(y^{(0)}_{n+1} - b_0\tau f^{(0)}_{n+1} - \Sigma_n). \tag{3.15}$$

Its order follows from Corollary 2.1 with $r=1$ and, since $\tau=O(\Delta^2)$, $s=1$.

Although the stability boundary of (3.15) can be made arbitrarily large by increasing q (cf. (3.13)), we loose accuracy if q becomes too large with respect to the grid. In fact, one should never choose q greater than $\log_2 M$ where M is the number of internal grid points along a row or column of the grid (see the tables of results in Section 5).

In order to preserve stability and accuracy we have to perform more than a single iteration. Adopting the iteration polynomials derived in VAN DER HOUWEN and SOMMEIJER (1983), we arrive at the following theorem:

THEOREM 3.4. *Let the predictor-corrector pair be such that* $\mathfrak{D}$ *contains the rectangle* $[-\tau R,0]\times[-d,1]$, $d>0$ *and let* $P_m(x)$ *be given by*

$$P_m(x) = \tfrac{1}{2}[1-d + (1+d)T_m(w_0 + \frac{w_0+1}{\beta_m}x)], \tag{3.16}$$

where

$$w_0 := \cos(\frac{1}{m}\arccos\frac{d-1}{d+1}), \qquad \beta_m := \frac{1}{b_0}\frac{1+w_0}{1-w_0}.$$

Furthermore, let

$$\hat{z}_{\min}(k,\tau R) := \min_{z_0\leq z\leq 0}\hat{z}(z). \tag{3.17a}$$

Then the SGPC method possesses a real stability boundary $\beta_m(k)$ *which satisfies the inequality*

$$\hat{z}_{\min}(k,\beta_m(k)) > -\beta_m. \tag{3.17b}$$

PROOF. The polynomial $P_m(x)$ is chosen such that it is bounded by $-d$ and 1 in the interval $[-\beta_m,0]$ and by 0 and 1 in the interval $[0,1/b_0]$. Thus, $P_m(\hat{z})$ assumes values in the interval $[-d,1]$ if $\hat{z}\in[-\beta_m,1/b_0]$. Hence, condition (3.12b) is satisfied if $\mathcal{I}\subset[-\beta_m,1/b_0]$. From (3.12d) it then follows that (3.17b) should be satisfied. □

In Table 3.1 we have listed the stability boundaries $\beta_m(k)=\beta_m(2^q-1)$ for the EP_1-BD_2 method (cf. Figure 3.1).

It is possible to give a fairly accurate approximation to $\beta_m(k)$ directly in terms m and k. This approximation is based on the estimate

$$\hat{z}_{\min}\approx\hat{z}(z_0)$$

instead of (3.17a). Requiring $\hat{z}(z_0)>-\beta_m$, we find

$$\tau R<\tilde{\beta}_m(k) := (\beta_m + \frac{1}{b_0})(k+1)^2 - \frac{2}{b_0(1-\cos\frac{\pi}{k+1})}. \tag{3.18}$$

Thus, the true stability boundary $\beta_m(k)$ is approximated by $\tilde{\beta}_m(k)$. Notice that $\beta_m(0)=\tilde{\beta}_m(0)=\beta_m$ which is precisely the stability boundary of the *GPC* method without smoothing. For $k>0$ the approximation is rather accurate, especially for large values of m. In the case of the EP_1-BD_2 method, this can be easily verified from the true stability boundaries $\beta_m(k)$ listed in Table 3.1. For small m-values, however, (3.18) slightly overestimates the true boundaries. By taking the integer part of β_m, instead of β_m itself, we found that (3.18) yields a safe value, for all k and m.

We conclude our discussion of stability boundaries of *SGPC* methods with the observation that for large m and k we have

$$\beta_m(k) \approx \beta_m 4^q \approx \frac{4m^2 4^q}{b_0(\arccos\frac{d-1}{d+1})^2},$$

where $k+1=2^q$. In the case of the EP_1-BD_2 method ($d=1/3$, $b_0=2/3$) this yields

$$\beta_m(k) \approx 1.37m^2 4^q, \qquad k+1=2^q.$$

TABLE 3.1 Stability boundaries $\beta_m(2^q-1)$ of the EP_1-BD_2 method, characterized by m and q

m	$q=0$	$q=1$	$q=2$	$q=3$	$q=4$	$q=5$	$q=6$
1	.5	4.5	19.7	80.1	322.1	1289.7	5160.5
2	4.5	20.9	85.3	342.8	1372.5	5491.7	21968.3
3	11.3	48.2	194.7	780.5	3123.4	12495.0	49981.5
4	20.9	86.5	347.9	1393.3	5574.5	22299.6	89200.1
5	33.2	135.8	544.9	2181.1	8726.0	34905.6	139623.9
6	48.2	196.0	785.6	3144.1	12577.9	50312.9	201253.2
7	66.0	267.1	1070.1	4282.1	17130.0	68521.6	274088.1
8	86.0	349.2	1398.4	5595.3	22382.5	89531.6	358128.0
9	109.8	442.2	1770.5	7083.4	28335.3	113342.8	453372.1
10	135.8	546.1	2182.3	8746.7	34988.5	139955.4	559823.1
20	546.1	2187.6	8752.0	35009.4	140039.1	560157.9	2240633.2
50	3418.6	13677.6	54711.8	218848.4	875395.0	3501581.3	14006326.6
100	13677.4	54713.3	218853.9	875416.3	3501666.2	14006665.7	56026663.5
$m\to\infty$	$1.37m^2$	$5.47m^2$	$21.8m^2$	$87.5m^2$	$350m^2$	$1400m^2$	$5600m^2$

4. THE SMOOTHED EP_1-BD_2 METHOD

In this section we give a detailed specification of the $SGPC$ method based on the EP_1 predictor $\{\tilde{\rho},\tilde{\sigma}\}=\{(\zeta-1)^2,0\}$, the BD_2 corrector $\{\rho,\sigma\}=\{(3\zeta^2-4\zeta+1)/3,\ 2\zeta^2/3\}$ and the iteration polynomial (3.16), where $d=1/3$ and $b_0=2/3$. Following the implementational details given in VAN DER HOUWEN and SOMMEIJER (1983) we obtain the following scheme

$$\begin{aligned}
&y_{n+1}^{(0)} = 2y_n - y_{n-1};\\
&\text{if } m=1 \text{ then} \quad y_{n+1} = y_{n+1}^{(0)} - \mathbb{S}R_{n+1}^{(0)};\\
&\text{if } m\geqslant 2 \text{ then}\\
&y_{n+1}^{(1)} = y_{n+1}^{(0)} - (1-w_0)\mathbb{S}R_{n+1}^{(0)},\\
&y_{n+1}^{(j)} = 2y_{n+1}^{(j-1)} - y_{n+1}^{(j-2)} - 2(1-w_0)\mathbb{S}R_{n+1}^{(j-1)}, \qquad j=2,\dots,m-1,\\
&y_{n+1} = \tfrac{1}{3}y_{n+1}^{(0)} - \tfrac{2}{3}y_{n+1}^{(m-2)} + \tfrac{4}{3}y_{n+1}^{(m-1)} - \tfrac{4}{3}(1-w_0)\mathbb{S}R_{n+1}^{(m-1)}.
\end{aligned} \tag{4.1}$$

Here,

$$w_0 := \cos(\frac{1}{m}\arccos(-\tfrac{1}{2})), \qquad R_{n+1}^{(j)} := y_{n+1}^{(j)} - \tfrac{2}{3}\tau f(t_{n+1}, y_{n+1}^{(j)}) - \tfrac{4}{3}y_n + \tfrac{1}{3}y_{n-1}. \tag{4.2}$$

The matrix $\mathbb{S}$ is discussed in Section 3.2.

The smoothed EP_1-BD_2 method is, according to Corollary 2.1, second-order accurate in time, because $p=2$, $\tilde{p}=1$, $r=1$ and $s=1$. A sufficient condition for stability is given by

$$\tau R < (\tfrac{3}{2} + \lfloor\frac{3(1+w_0)}{2(1-w_0)}\rfloor)(k+1)^2 - \frac{3}{1-\cos(\frac{\pi}{k+1})}. \tag{4.3}$$

5. Numerical experiments

In this section we present a number of 1-D and 2-D initial-boundary value problems by which the smoothed predictor-corrector method will be compared with standard methods. A specification of these methods will be given in the next subsections.

The 1-D problems are defined on the unit interval and the 2-D problems on the unit square; both types are semidiscretized on a uniform spacegrid with meshes of size Δx, using symmetric second-order differences. The Dirichlet boundary conditions are treated as described in Section 3.2.1.

For the time-integration, we used a timestep equal to the meshwidth, i.e., $\tau = \Delta x$. The integration interval is $[0,1]$ in all experiments. The initial conditions, as well as the starting values needed in a multistep method, are taken from the exact solution.

To measure the accuracy obtained by the various schemes, we define

$$cd := \log_{10} \| \text{ global error in the endpoint } t = 1 \|_\infty , \tag{5.1}$$

where the global error is the difference between the numerical solution of the ODE (1.1) and the exact solution of the initial-boundary value problem restricted to the gridpoints. The value of cd can be considered as the number of correct digits in the numerical solution.

5.1. *One-dimensional problems*

To start with, we will test 4 one-dimensional problems. The specification of these problem, as well as the results of the various methods, can be found in the Tables 5.1-5.4. To these problems we applied the $EP_1 - BD_2(q)$ method for several values of q. In Section 4, this family of methods in completely defined.

The only free parameter is m, the number of stages. This parameter is chosen such that the stability condition (3.17) (or (4.3)) is satisfied. An approximation to the spectral radius R is obtained by using Gerschgorin's theorem.

In the tables of results, we only list the number of f-evaluations N (summed over all timesteps), as this is the major part of the computational work; this number is followed by the value of cd (cf. (5.1)).

To judge the merits of this $EP - BD$ method, we also implemented the fully implicit BDF_2, i.e., we directly solved the corrector of the above $SGPC$ method, using Newton's method. In the Tables 5.1-5.4 this method is denoted by BD_2. Again, we list the values of N/cd, where now N denotes the number of Newton iterations (performed in the whole integration process). For our test examples, it turned out that the accuracy furnished by the BD_2 method could not be improved by taking more than one Newton iteration; therefore, the given value of N corresponds to one Newton iteration per step.

Comparing both type of methods, we see that the BD_2 method is slightly more accurate than the $EP_1 - BD_2$ method for the same value of N. However, taking into account that one Newton iteration in the BD_2 method involves an f-evaluation, an evaluation and decomposition of the Jacobian matrix and the solution of a tridiagonal system, we think both type of methods are at least competitive for one-dimensional problems.

Finally, we observe that taking q as large as allowed by the grid causes a drop in the accuracy.

TABLE 5.1 N/cd-values for $u_t = e^u u_{xx} + u(9e^u - 1)$ with exact solution $u(t,x) = e^{-t}\sin(3x)$.

method	$\Delta x = 1/8$	$\Delta x = 1/16$	$\Delta x = 1/32$	$\Delta x = 1/64$
$EP_1 - BD_2(0)$	50/1.5	149/2.1	429/2.7	1218/3.3
(1)	27/1.5	79/2.1	222/2.7	625/3.3
(2)	14/1.6	45/2.1	120/2.7	332/3.3
(3)	8/1.7	30/2.2	63/2.7	189/3.3
(4)		15/1.7	33/3.2	126/3.4
(5)			31/1.9	63/3.1
(6)				63/2.1
BD_2	7/1.5	15/2.1	31/2.7	63/3.3

TABLE 5.2 N/cd-values for $u_t = u_{xx} + 3xt^2(x^2 - 2t)$ with exact solution $u(t,x) = 1 + x^3t^3$.

method	$\Delta x = 1/8$	$\Delta x = 1/16$	$\Delta x = 1/32$	$\Delta x = 1/64$
$EP_1 - BD_2(0)$	35/1.5	105/2.1	310/2.6	882/3.2
(1)	21/1.6	60/2.1	155/2.6	441/3.2
(2)	14/1.6	30/2.2	93/2.7	252/3.3
(3)	7/1.1	15/1.9	62/2.6	126/3.3
(4)		15/1.2	31/2.1	63/2.9
(5)			31/1.2	63/2.2
(6)				63/1.3
BD_2	7/1.6	15/2.2	31/2.7	63/3.3

TABLE 5.3 N/cd-values for $u_t = u^4 u_{xx} - u - 20x^3e^{-t}u^4$ with exact solution $u(t,x) = x^5e^{-t}$

method	$\Delta x = 1/8$	$\Delta x = 1/16$	$\Delta x = 1/32$	$\Delta x = 1/64$
$EP_1 - BD_2(0)$	22/2.6	55/3.1	147/3.7	409/4.3
(1)	12/2.3	30/3.1	81/3.7	223/4.3
(2)	8/1.6	20/2.5	49/3.2	125/4.0
(3)	7/1.1	15/1.7	34/2.6	81/3.4
(4)		15/1.2	31/1.8	63/2.7
(5)			31/1.2	63/2.0
(6)				63/1.3
BD_2	7/2.6	15/3.1	31/3.7	63/4.3

TABLE 5.4 N/cd-values for $u_t = e^u u_{xx} + u(x - t^2 e^u)$ with exact solution $u(t,x) = e^{tx}$.

method	$\Delta x = 1/8$	$\Delta x = 1/16$	$\Delta x = 1/32$	$\Delta x = 1/64$
$EP_1 - BD_2(0)$	87/1.9	256/1.9	744/2.5	2129/3.1
(1)	46/2.0	132/2.0	380/2.4	1084/3.1
(2)	25/1.5	70/2.2	199/2.4	556/3.2
(3)	15/1.6	38/2.5	110/3.0	296/3.2
(4)		23/1.6	66/2.5	161/3.4
(5)			36/1.6	96/2.5
(6)				63/1.6
BD_2	7/2.1	15/2.7	31/3.3	63/3.9

5.2 Two-dimensional problems

Next, we will test some two-dimensional problems. Increasing the dimension of the initial-boundary value problem has hardly consequences for the application of the *SGPC* method. Only the smoothing operator $\mathbb{S}$ has to be adapted, which can be performed in a straightforward way (see the discussion in Section 3.2.2).

If, on the other hand, a fully implicit scheme is applied (e.g. the BDF_2) we are faced with a huge algebraic problem, since now the Jacobian matrix has no longer a tridiagonal structure. Therefore, as an alternative to the BDF_2, we selected, as a reference method, the second-order *ADI* method which is defined by

$$\begin{aligned} y^* &= y_n + \frac{1}{2}\tau f_1(t_n + \frac{1}{2}\tau, y^*) + \frac{1}{2}\tau f_2(t_n, y_n), \\ y_{n+1} &= y^* + \frac{1}{2}\tau f_1(t_n + \frac{1}{2}\tau, y^*) + \frac{1}{2}\tau f_2(t_{n+1}, y_{n+1}). \end{aligned} \tag{5.2}$$

Here, f_1 and f_2 contain the dicretizations of the spatial derivatives in x_1- and x_2- direction, respectively. The inhomogeneous term in the initial-boundary value problem is equally distributed over both implicit relations in (5.2). However, due to the splitting, the tridiagonal structure of the systems to be solved, has been preserved.

In the tables of results, this method is abbreviated as $ADI(m)$, where m denotes the number of Newton iterations used in each of the implicit relations.

We applied the $EP_1 - BD_2(q)$ method and the $ADI(m)$ method to 3 two-dimensional problems. The Tables 5.5-5.7 contain the resulting N/cd-values. Note that for the *ADI* method, N means the total number of Newton interations summed over all steps and both stages in (5.2). For the nonlinear examples, it turned out that 2 Newton iterations are sufficient to solve the implicit relations in (5.2).

A comparison of both schemes reveals that the *ADI* method is superior to the $EP_1 - BD_2$ method in linear situations, but considerably less efficient for nonlinear problems.

TABLE 5.5. N/cd-values for $u_t=u_{x_1x_1}+u_{x_2x_2}+3t^2[x_1^3+x_2^3-2t(x_1+x_2)]$ with exact solution $u(t,x_1,x_2)=1+t^3(x_1^3+x_2^3)$.

method	$\Delta x=1/8$	$\Delta x=1/16$	$\Delta x=1/32$
$EP_1-BD_2(0)$	49/1.2	150/1.8	434/2.3
(1)	28/1.3	75/1.7	217/2.3
(2)	14/1.3	45/1.9	124/2.4
(3)	7/0.8	30/1.6	62/2.3
(4)		15/0.9	31/1.7
(5)			31/1.1
$ADI(1)$	14/1.9	30/2.3	62/2.8

As in the case of the one-dimensional problems, we see again an abrupt decrease of the accuracy if the largest possible value of q is used.

TABLE 5.6 N/cd-values for $u_t=e^u(u_{x_1x_1}+u_{x_2x_2})+u(9e^u-1)$ with exact solution $u(t,x_1,x_2)=e^{-t}(\sin(3x_1)+\sin(3x_2))$.

method	$\Delta x=1/8$	$\Delta x=1/16$	$\Delta x=1/32$
$EP_1-BD_2(0)$	95/2.4	286/2.9	826/3.7
(1)	50/2.4	147/3.0	420/3.7
(2)	26/2.5	76/3.1	220/3.7
(3)	15/1.8	42/2.8	116/3.6
(4)		27/1.9	67/2.9
(5)			37/2.0
$ADI(1)$	14/1.9	unstable	unstable
$ADI(2)$	28/1.9	60/2.5	124/3.1

TABLE 5.7 N/cd-values for $u_t=u^3_{x_1x_1}+u^3_{x_2x_2}+x_1x_2u-9t^2(x_1^2+x_2^2)$ with exact solution $u(t,x_1,x_2)=e^{tx_1x_2}$.

method	$\Delta x=1/8$	$\Delta x=1/16$	$\Delta x=1/32$
$EP_1-BD_2(0)$	144/1.1	436/1.6	1274/1.9
(1)	73/1.2	221/1.4	645/1.8
(2)	38/1.7	115/1.6	330/1.7
(3)	21/1.2	62/1.9	173/2.3
(4)		35/1.1	93/1.8
(5)			54/1.1
$ADI(1)$	14/1.2	unstable	unstable
$ADI(2)$	28/1.4	60/1.6	124/2.0

6. Conclusion

Explicit algorithms have been described for the efficient solution of parabolic initial-boundary value problems with Dirichlet boundary conditions. These methods are based on predictor-corrector type schemes and extended with residue smoothing. For a set of test problems, this technique turns out to be at least competitive with implicit methods.

A decisive advantage of the new methods is their extremely simple implementation.

References

1 P.J. van der Houwen and B.P. Sommeijer (1983) Predictor-corrector methods with improved absolute stability regions. *IMA J. Numer. Analysis* **3**, 417-437.

2 P.J. van der Houwen, B.P. Sommeijer, and F.W. Wubs (1986) Analysis of smoothing operators in the solution of partial differential equations by explicit difference schemes. *Report NM-R8617,* Centrum voor Wiskunde en Informatica, Amsterdam.

3 P.J. van der Houwen, C. Boon, and F.W. Wubs (1987) Analysis of smoothing matrices for the preconditioning of elliptic difference equations. *Report NM-R8705,* Centrum voor Wiskunde en Informatica, Amsterdam.

4 A. Jameson (1983) The evolution of computational methods in aerodynamics. *J. Appl. Mech.* **50**, 1052-1076.

5 A. Jameson and D. Mavriplis (1985) Finite volume solution of the two-dimensional Euler equations on a regular triangular mesh. *AIAA paper* 85-0435.

6 A. Lerat (1979) Une class de schémas aux differences implicites pour les systèmes hyperboliques de lois de conservation, *C.R. Acad. Sc. Paris, t. 288, Série A,* 1033-1036.

7 E. Turkel (1985) Acceleration to a steady state for the Euler equations, *Numerical methods for the Euler equations of fluid dynamics, SIAM,* Philadelphia, 218-311.

8 F.W. Wubs (1986) Stabilization of explicit methods for hyperbolic partial differential equations, *Int. J. Numer. Methods Fluids* **6**, 641-657.