Computers and Mathematics with Applications 63 (2012) 851–861

Contents lists available at SciVerse ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

# Perturbed MEBDF methods

Raffaele D'Ambrosio [a,*], Giuseppe Izzo [b], Zdzislaw Jackiewicz [c,d]

[a] Dipartimento di Matematica, Università di Salerno, Fisciano (Sa), 84084, Italy
[b] Dipartimento di Matematica e Applicazioni, Università di Napoli-"Federico II", 80126 Napoli, Italy
[c] Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287, United States
[d] AGH University of Science and Technology, Kraków, Poland

ARTICLE INFO

*Article history:*
Received 13 May 2011
Received in revised form 28 November 2011
Accepted 28 November 2011

*Keywords:*
MEBDF methods
General linear methods
Stability analysis
*A*-stability
$A(\alpha)$-stability

ABSTRACT

We investigate MEBDF methods of Cash from general linear methods point of view. Some perturbations of these methods are constructed which preserve the order of these formulas and improve their stability properties.



## 1. Introduction

To improve stability properties of backward differentiation formulas (BDFs) for ordinary differential equations (ODEs)

$$\begin{cases} y'(t) = f(t, y(t)), \quad t \in [t_0, T], \\ y(t_0) = y_0 \in \mathbb{R}^m, \end{cases} \tag{1.1}$$

Cash [1] proposed the extension of these methods which utilizes a future point at $t_{n+k+1}$. These extended BDF (EBDF) methods take the form

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h\beta_k f_{n+k} + h\beta_{k+1} f_{n+k+1}, \tag{1.2}$$

where $f_{n+k} = f(t_{n+k}, y_{n+k})$, $f_{n+k+1} = f(t_{n+k+1}, y_{n+k+1})$. The coefficients $\alpha_j$, $j = 0, 1, \ldots, k$, $\beta_k$, $\beta_{k+1}$, of these methods are computed by solving the appropriate order conditions for the order $p = k + 1$ and with the normalization $\alpha_k = 1$. These coefficients are listed in [1] for $k = 1, 2, \ldots, 8$. The resulting methods are *A*- and *L*-stable for $k = 1, 2$, and 3, and $A(\alpha)$-stable for $k = 4, 5, 6, 7$, and 8. The regions of absolute stability of these methods are plotted in [1,2].

Assuming that the approximations $y_n, y_{n+1}, \ldots, y_{n+k-1}$ to the solution $y$ of (1.1) at the points $t_n, t_{n+1}, \ldots, t_{n+k-1}$ are already computed, the algorithm based on EBDF methods is defined by the following three steps:

* Corresponding author.
*E-mail addresses:* rdambrosio@unisa.it (R. D'Ambrosio), giuseppe.izzo@unina.it (G. Izzo), jackiewi@math.la.asu.edu (Z. Jackiewicz).

0898-1221/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2011.11.050

(i) Compute $\bar{y}_{n+k}$ as the solution of the conventional BDF method

$$\bar{y}_{n+k} + \sum_{j=0}^{k-1} \widehat{\alpha}_j y_{n+j} = h\widehat{\beta}_k \bar{f}_{n+k}, \tag{1.3}$$

$\bar{f}_{n+k} = f(t_{n+k}, \bar{y}_{n+k})$.

(ii) Compute $\bar{y}_{n+k+1}$ as the solution of the same BDF advanced one step, that is,

$$\bar{y}_{n+k+1} + \widehat{\alpha}_{k-1}\bar{y}_{n+k} + \sum_{j=0}^{k-2} \widehat{\alpha}_j y_{n+j+1} = h\widehat{\beta}_k \bar{f}_{n+k+1}, \tag{1.4}$$

$\bar{f}_{n+k+1} = f(t_{n+k+1}, \bar{y}_{n+k+1})$.

(iii) Discard $\bar{y}_{n+k}$, insert $\bar{f}_{n+k+1}$ into EBDF method (1.2), and solve for $y_{n+k}$:

$$y_{n+k} + \sum_{j=0}^{k-1} \alpha_j y_{n+j} = h\beta_k f_{n+k} + h\beta_{k+1}\bar{f}_{n+k+1}. \tag{1.5}$$

The coefficients $\widehat{\alpha}_j, j = 0, 1, \ldots, k, \widehat{\alpha}_k = 1, \widehat{\beta}_k$, of BDF methods are listed in [3] for $k = 1, 2, \ldots, 6$ and in [1] for $k = 7$ and 8. If the EBDF method (1.2) is of order $k + 1$ and BDF methods (1.3) and (1.4) are of order $k$, then the overall algorithm (i)–(iii) based on (1.3)–(1.5) is of order $k + 1$ [1].

It was observed by Cash [4] and Hairer and Wanner [5] that the disadvantage of the algorithm given above is that stages (i) and (ii) represent nonlinear systems with the same Jacobian $I - h\widehat{\beta}_k J, J = \partial f/\partial y$, but stage (iii) has a different Jacobian, $I - h\beta_k J$, which requires extra LU decomposition. To remedy this situation, Cash [4] proposed an algorithm where the last stage (iii) was replaced by a modified EBDF (MEBDF) method of the form

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h\widehat{\beta}_k f_{n+k} + h(\beta_k - \widehat{\beta}_k)\bar{f}_{n+k} + h\beta_{k+1}\bar{f}_{n+k+1}. \tag{1.6}$$

These methods have order $k+1$ and are also $A$- and $L$-stable for $k = 1, 2$, and 3, and $A(\alpha)$-stable for $k = 4, 5, 6, 7$, and 8 with larger angles $\alpha$ than that of the corresponding EBDF methods. These angles for BDF, EBDF, and MEBDF methods are listed in [1,4,2,5] and reproduced also in [6]. The stability regions of MEBDF methods (1.6) have been plotted in [5].

In Section 2 the MEBDF methods will be reformulated as general linear methods (GLMs) for ODEs. In Section 3 we propose a perturbation of MEDBF methods which will preserve their order and improve their stability properties. In Section 4 we provide examples of perturbed MEBDF methods for $k = 4, 5, 6, 7$, and 8. In Section 5 we discuss local error estimation for small and large stepsizes. Finally, in Section 6 some concluding remarks are given and plans for future research are briefly discussed.

## 2. MEBDF methods as GLMs

GLMs for the numerical solution of ODEs (1.1) are defined by

$$\begin{cases} Y_i^{[n]} = h\displaystyle\sum_{j=1}^{s} a_{ij} f(Y_j^{[n]}) + \sum_{j=1}^{r} u_{ij} y_j^{[n-1]}, & i = 1, 2, \ldots, s, \\ y_i^{[n]} = h\displaystyle\sum_{j=1}^{s} b_{ij} f(Y_j^{[n]}) + \sum_{j=1}^{r} v_{ij} y_j^{[n-1]}, & i = 1, 2, \ldots, r, \end{cases} \tag{2.1}$$

$n = 1, 2, \ldots, N$, where $Nh = T - t_0$. Here, the internal stages $Y_i^{[n]}$ are approximations of stage order $q$ to $y(t_{n-1} + c_i h)$, and the external stages $y_i^{[n]}$ are approximations of order $p$ to the linear combinations of scaled derivatives of $y(t)$ in $t = t_n$, compare [6]. These methods are specified by the abscissa vector $\mathbf{c} = [c_1, \ldots, c_s]^T$ and the coefficient matrices

$$\mathbf{A} \in \mathbb{R}^{s\times s}, \qquad \mathbf{U} \in \mathbb{R}^{s\times r}, \qquad \mathbf{B} \in \mathbb{R}^{r\times s}, \qquad \mathbf{V} \in \mathbb{R}^{r\times r}.$$

Putting

$$Y^{[n]} = \begin{bmatrix} Y_1^{[n]} \\ \vdots \\ Y_s^{[n]} \end{bmatrix}, \qquad hf(Y^{[n]}) = \begin{bmatrix} hf(Y_1^{[n]}) \\ \vdots \\ hf(Y_s^{[n]}) \end{bmatrix}, \qquad y^{[n]} = \begin{bmatrix} y_1^{[n]} \\ \vdots \\ y_r^{[n]} \end{bmatrix},$$

the GLM (2.1) can be written in vector form as follows

$$\left[\begin{array}{c} Y^{[n]} \\ \hline y^{[n]} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A}\otimes\mathbf{I} & \mathbf{U}\otimes\mathbf{I} \\ \hline \mathbf{B}\otimes\mathbf{I} & \mathbf{V}\otimes\mathbf{I} \end{array}\right] = \left[\begin{array}{c} hf(Y^{[n]}) \\ \hline y^{[n-1]} \end{array}\right], \tag{2.2}$$

$n = 1, 2, \ldots, N$. Here, $\mathbf{I}$ is the identity matrix of dimension $m$ and '$\otimes$' stands for Kronecker product of matrices.

Substituting (1.3) into (1.4), we obtain

$$\bar{y}_{n+k+1} = \widehat{\alpha}_{k-1}\widehat{\alpha}_0 y_n + \sum_{j=1}^{k-1}(\widehat{\alpha}_{k-1}\widehat{\alpha}_j - \widehat{\alpha}_{j-1})y_{n+j} - h\widehat{\alpha}_{k-1}\widehat{\beta}_k\bar{f}_{n+k} + h\widehat{\beta}_k\bar{f}_{n+k+1}. \tag{2.3}$$

Then it can be verified that an algorithm based on formulas (1.3), (2.3) and (1.6) can be written as a GLM of the form (2.2) with $s = 3$, $r = k$, and with the vectors of internal approximations $Y^{[n]}, f(Y^{[n]})$, and the vector of external approximations $y^{[n]}$ defined by

$$Y^{[n]} = \begin{bmatrix} \bar{y}_{n+k} \\ \bar{y}_{n+k+1} \\ y_{n+k} \end{bmatrix}, \qquad f(Y^{[n]}) = \begin{bmatrix} \bar{f}_{n+k} \\ \bar{f}_{n+k+1} \\ f_{n+k} \end{bmatrix}, \qquad y^{[n]} = \begin{bmatrix} y_{n+k} \\ y_{n+k-1} \\ \vdots \\ y_{n+1} \end{bmatrix},$$

and with the coefficient matrices $\mathbf{A}$, $\mathbf{U}$, $\mathbf{B}$, and $\mathbf{V}$ given by

$$\mathbf{A} = \begin{bmatrix} \widehat{\beta}_k & 0 & 0 \\ -\widehat{\alpha}_{k-1}\widehat{\beta}_k & \widehat{\beta}_k & 0 \\ \beta_k - \widehat{\beta}_k & \beta_{k+1} & \widehat{\beta}_k \end{bmatrix},$$

$$\mathbf{U} = \begin{bmatrix} -\widehat{\alpha}_{k-1} & -\widehat{\alpha}_{k-2} & \cdots & -\widehat{\alpha}_1 & -\widehat{\alpha}_0 \\ \widehat{\alpha}_{k-1}\widehat{\alpha}_{k-1} - \widehat{\alpha}_{k-2} & \widehat{\alpha}_{k-1}\widehat{\alpha}_{k-2} - \widehat{\alpha}_{k-3} & \cdots & \widehat{\alpha}_{k-1}\widehat{\alpha}_1 - \widehat{\alpha}_0 & \widehat{\alpha}_{k-1}\widehat{\alpha}_0 \\ -\alpha_{k-1} & -\alpha_{k-2} & \cdots & -\alpha_1 & -\alpha_0 \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} \beta_k - \widehat{\beta}_k & \beta_{k+1} & \widehat{\beta}_k \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \mathbf{V} = \begin{bmatrix} -\alpha_{k-1} & -\alpha_{k-2} & \cdots & -\alpha_1 & -\alpha_0 \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

We have

$$\mathbf{A} \in \mathbb{R}^{3\times 3}, \qquad \mathbf{U} \in \mathbb{R}^{3\times k}, \qquad \mathbf{B} \in \mathbb{R}^{k\times 3}, \qquad \mathbf{V} \in \mathbb{R}^{k\times k}.$$

## 3. Perturbed and fully perturbed MEBDF methods

Observe that for the algorithm based on MEBDF methods written as GLMs we have

$$Y_3^{[n]} = y_1^{[n]},$$

and

$$y_i^{[n]} = y_{i-1}^{[n-1]}, \quad i = 2, 3, \ldots, k.$$

We now consider the perturbation of these methods, where the coefficient matrix $\mathbf{B}$, which will be denoted by the same symbol, takes the form

$$\mathbf{B} = \begin{bmatrix} \beta_k - \widehat{\beta}_k + b_{11} & \beta_{k+1} + b_{12} & \widehat{\beta}_k + b_{13} \\ b_{21} & b_{22} & b_{23} \\ \vdots & \vdots & \vdots \\ b_{k-1,1} & b_{k-1,2} & b_{k-1,3} \\ b_{k,1} & b_{k,2} & b_{k,3} \end{bmatrix},$$

and we choose the coefficients $b_{ij}$ in such a way that the order of the underlying MEBDF method is preserved, i.e.

$$Y_3^{[n]} = y_1^{[n]} + O(h^{p+1}), \tag{3.1}$$

and

$$y_i^{[n]} = y_{i-1}^{[n-1]} + O(h^{p+1}), \quad i = 2, 3, \ldots, k, \tag{3.2}$$

where $p = k + 1$ is the order of MEBDF methods. Since

$$Y_3^{[n]} = -\sum_{j=0}^{k-1} \alpha_j y_{k-j}^{[n-1]} + (\beta_k - \widehat{\beta}_k) h\bar{f}_{n+k} + \beta_{k+1} h\bar{f}_{n+k+1} + \widehat{\beta}_k h f_{n+k},$$

$$y_1^{[n]} = -\sum_{j=0}^{k-1} \alpha_j y_{k-j}^{[n-1]} + (\beta_k - \widehat{\beta}_k + b_{11}) h\bar{f}_{n+k} + (\beta_{k+1} + b_{12}) h\bar{f}_{n+k+1} + (\widehat{\beta}_k + b_{13}) h f_{n+k},$$

$$y_i^{[n]} = y_{i-1}^{[n-1]} + b_{i1} h\bar{f}_{n+k} + b_{i2} h\bar{f}_{n+k+1} + b_{i3} h f_{n+k}, \quad 2 \le i \le k,$$

and

$$h\bar{f}_{n+k} = hy'(t_{n+k}) + O(h^{p+1}), \qquad h\bar{f}_{n+k+1} = hy'(t_{n+k+1}) + O(h^{p+1}), \qquad h f_{n+k} = hy'(t_{n+k}) + O(h^{p+1}),$$

this leads to

$$b_{i1} = b_i, \qquad b_{i2} = 0, \qquad b_{i3} = -b_i, \quad i = 1, 2, \ldots, k,$$

where $b_i$ are arbitrary parameters. Hence, the resulting coefficient matrix $\mathbf{B}$ takes the form

$$\mathbf{B} = \begin{bmatrix} \beta_k - \widehat{\beta}_k + b_1 & \beta_{k+1} & \widehat{\beta}_k - b_1 \\ b_2 & 0 & -b_2 \\ \vdots & \vdots & \vdots \\ b_{k-1} & 0 & -b_{k-1} \\ b_k & 0 & -b_k \end{bmatrix}.$$

To preserve the FSAL property (first same as last, compare [7]) of the MEBDF, we will distinguish two different cases for which $b_1 = 0$ and $b_1 \neq 0$, respectively. These methods will be called perturbed MEBDF (PMEBDF) and fully perturbed MEBDF (FPMEBDF), respectively. It follows from (3.1) and (3.2) that PMEBDF and FPMEBDF methods have the same order as the underlying MEBDF methods for any $b_i$, $i = 1, 2, \ldots, k$. These free parameters will then be chosen to maximize the angle $\alpha$ of $A(\alpha)$-stability.

Let us recall that the stability polynomial of a GLM method (2.1) can be obtained as $p(w, z) = det(M - w\, I)$, where $M = \mathbf{V} + z\, \mathbf{B}(I - z\, \mathbf{A})^{-1}\mathbf{U}$ is the stability matrix of the method itself [6]. It can be easily shown that for PMEBDF and FPMEBDF the stability polynomial $p(w, z)$ has the form

$$p(w, z) = \frac{1}{(\widehat{\beta}_k\, z - 1)^3} \sum_{j=0}^{k} a_j(z) w^j, \tag{3.3}$$

where each $a_j(z), j = 0, \ldots, k$, is a polynomial of degree at most three in $z$, whose coefficients depend on the parameters $b_i, i = 1, \ldots, k$ ($b_1 = 0$ for PMEBDF). With the aim of maximizing the angle $\alpha$ of $A(\alpha)$-stability, we exploited the boundary locus technique [3] to define an objective function

$$f_n : (b_1, \ldots, b_k) \in \mathbb{R}^k \to \left[0, \frac{\pi}{2}\right],$$

which approximates the value of the angle $\alpha$ for specific choices of parameters $b_i$, $i = 1, 2, \ldots, k$. In order to understand the strategy we use to compute this objective function, we briefly remark how the boundary locus technique works for FPMEBDF methods. Taking into account (3.3), we consider the family of equations

$$\sum_{j=0}^{k} a_j(z) e^{ij\vartheta} = 0,$$

where $i$ is the imaginary unit and $\vartheta \in [0, 2\pi]$. Then, the solution

$$z = z(\vartheta)$$

defines the so-called boundary locus curve of the complex plane, that contains the boundary of the stability region of the corresponding FPMEBDF method. We next consider $n$ points on the unit circle which identify the angles

$$\vartheta_\nu = \frac{2\pi\nu}{n}, \quad \nu = 1, \ldots, n,$$

and, correspondingly, we introduce the sets

$$I_\nu = \left\{ \arctan \left| \frac{Im(z)}{Re(z)} \right| : Re(z) < 0, \sum_{j=0}^{k} a_j(z) e^{ij\vartheta_\nu} = 0 \right\}, \quad \nu = 1, \ldots, n,$$

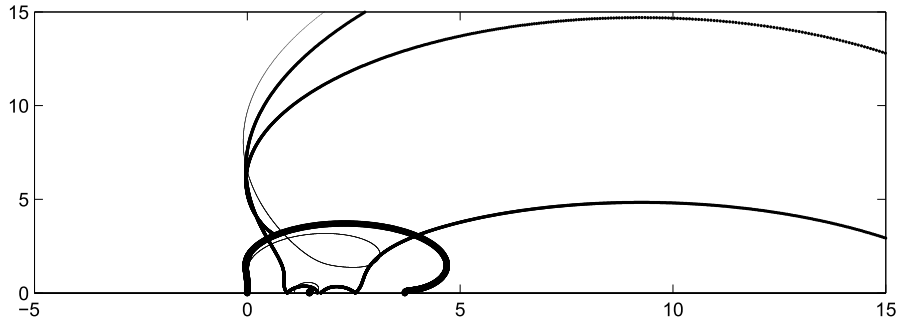


**Fig. 4.1.** Regions of stability of PEMBDF method (thin line), FPMBDF method (medium line) and MEBDF method (thick line) for $k = 4$.

of the angles formed by the negative real axis of the complex plane and the half-line from the origin to $z$. The objective function (and, thus, the approximation to the angle $\alpha$ of A($\alpha$)-stability) is then computed as $f_n(b_1, \ldots, b_k) := -\min(S)$, where $S$ is the following set

$$S = \bigcup_{\nu=1}^{n} I_\nu \cup \left\{\frac{\pi}{2}\right\}.$$

The objective function $f_n$ was then minimized (for increasing values of $n$) using the function `fminsearch` from Matlab. In this search we started with initial values $b_i = 0$, $i = 1, 2, \ldots, k$, which correspond to MEBDF methods, as well as with random initial values from the interval $[-3, 3]$. The examples of methods obtained in this way and their respective stability domains are presented in Section 4.

## 4. Examples of PMEBDF and FPMEBDF methods

Since the MEBDF methods corresponding to $k = 1$, 2, and 3 are already $A$- and $L$-stable we performed our search for $k = 4, 5, 6, 7$, and 8. The values of the coefficients are expressed in rational form; such rational approximations have been provided by using Matlab `rats` function, with default accuracy $10^{-6}$.

For $k = 4$ an example of the PMEBDF method with the large region of $A(\alpha)$-stability is given by

$$b_1 = 0, \qquad b_2 = -\frac{337}{374}, \qquad b_3 = -\frac{982}{207}, \qquad b_4 = -\frac{1365}{137}.$$

This method is $A(\alpha)$-stable for $\alpha = 89.32$. An example of the FPMEBDF method with $k = 4$ is given by

$$b_1 = -\frac{432}{199}, \qquad b_2 = -\frac{2181}{206}, \qquad b_3 = -\frac{1821}{71}, \qquad b_4 = -\frac{4099}{93},$$

whose angle of $A(\alpha)$-stability is for $\alpha = 89.71$. The region of stability of the PMEBDF method is plotted by a thin line in Fig. 4.1 and that of the FPMEBDF is plotted by a medium line, together with a region of stability of the corresponding MEBDF method, which is plotted by a thick line.

For $k = 5$ an example of the PMEBDF method with the large region of $A(\alpha)$-stability is given by

$$b_1 = 0, \qquad b_2 = -\frac{264}{281}, \qquad b_3 = -\frac{16329}{4082}, \qquad b_4 = -\frac{1399}{165}, \qquad b_5 = -\frac{3002}{187}.$$

This method is $A(\alpha)$-stable for $\alpha = 86.19$. An example of the FPMEBDF method with $k = 5$ is given by

$$b_1 = -\frac{96}{47}, \qquad b_2 = -\frac{1411}{135}, \qquad b_3 = -\frac{8367}{298}, \qquad b_4 = -\frac{7914}{137}, \qquad b_5 = -\frac{3817}{36},$$

whose angle of $A(\alpha)$-stability is for $\alpha = 88.01$. The corresponding stability regions are plotted in Fig. 4.2.

For $k = 6$ an example of the PMEBDF method with the large region of $A(\alpha)$-stability is given by

$$b_1 = 0, \qquad b_2 = -\frac{319}{305}, \qquad b_3 = -\frac{236}{71}, \qquad b_4 = -\frac{2220}{437}, \qquad b_5 = -\frac{570}{161}, \qquad b_6 = \frac{728}{75}.$$

This method is $A(\alpha)$-stable for $\alpha = 80.60$. An example of the FPMEBDF method with $k = 6$ is given by

$$b_1 = -\frac{92}{63}, \qquad b_2 = -\frac{652}{103}, \qquad b_3 = -\frac{707}{58}, \qquad b_4 = -\frac{389}{42}, \qquad b_5 = \frac{2029}{81}, \qquad b_6 = \frac{3155}{23},$$

whose angle of $A(\alpha)$-stability is for $\alpha = 84.67$. The corresponding stability regions are plotted in Fig. 4.3.

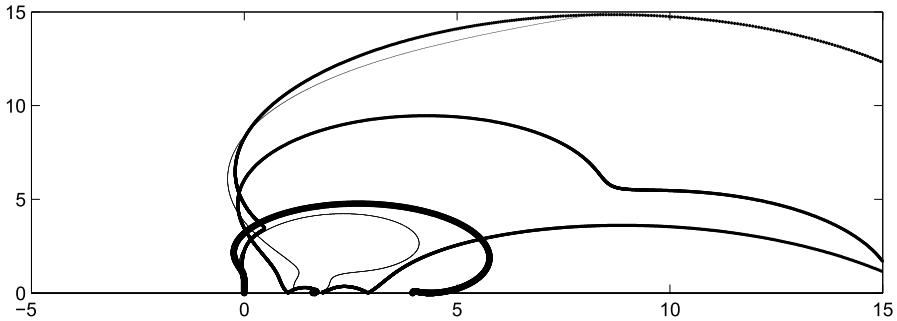


**Fig. 4.2.** Regions of stability of PEMBDF method (thin line), FPMBDF method (medium line) and MEBDF method (thick line) for $k = 5$.

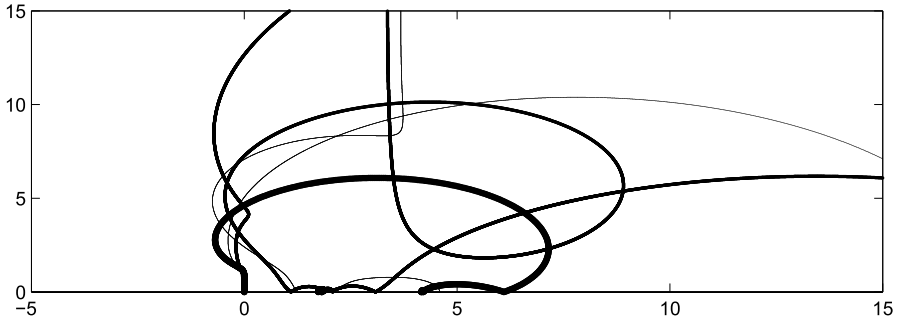


**Fig. 4.3.** Regions of stability of PEMBDF method (thin line), FPMBDF method (medium line) and MEBDF method (thick line) for $k = 6$.

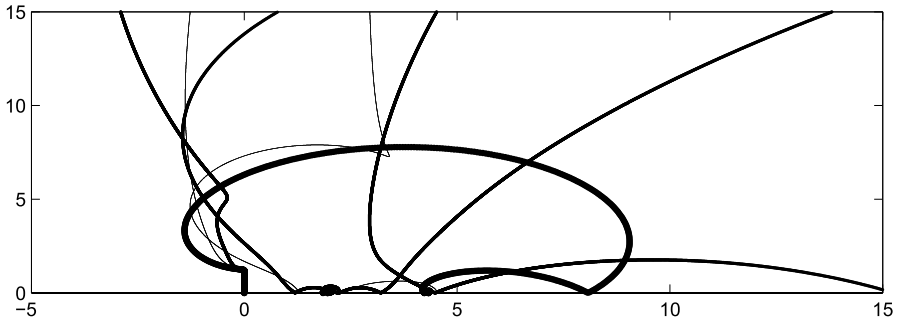


**Fig. 4.4.** Regions of stability of PEMBDF method (thin line), FPMBDF method (medium line) and MEBDF method (thick line) for $k = 7$.

For $k = 7$ an example of the PMEBDF method with the large region of $A(\alpha)$-stability is given by

$$b_1 = 0, \qquad b_2 = -\frac{199}{304}, \qquad b_3 = -\frac{30}{19},$$
$$b_4 = -\frac{690}{427}, \qquad b_5 = -\frac{259}{760}, \qquad b_6 = \frac{665}{383}, \qquad b_7 = -\frac{317}{153}.$$

This method is $A(\alpha)$-stable for $\alpha = 72.63$. An example of the FPMEBDF method with $k = 7$ is given by

$$b_1 = -\frac{50}{49}, \qquad b_2 = -\frac{1063}{259}, \qquad b_3 = -\frac{695}{92},$$
$$b_4 = -\frac{959}{130}, \qquad b_5 = -\frac{169}{214}, \qquad b_6 = \frac{472}{123}, \qquad b_7 = -\frac{3590}{101},$$

whose angle of $A(\alpha)$-stability is for $\alpha = 78.70$. The corresponding stability regions are plotted in Fig. 4.4.

For $k = 8$ an example of the PMEBDF method with the large region of $A(\alpha)$-stability is given by

$$b_1 = 0, \qquad b_2 = -\frac{25}{163}, \qquad b_3 = \frac{3}{763}, \qquad b_4 = \frac{447}{880},$$
$$b_5 = \frac{111}{166}, \qquad b_6 = \frac{371}{729}, \qquad b_7 = -\frac{5}{401}, \qquad b_8 = -\frac{17}{21}.$$

**Table 4.1**
Angles $\alpha$ of $A(\alpha)$-stability for BDF, EBDF, MEBDF, and PMEBDF formulas for $k = 1, 2, \ldots, 8$.

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ for BDF | 90° | 90° | 88° | 73° | 51° | 18° | * | * |
| $\alpha$ for EBDF | 90° | 90° | 90° | 87.61° | 80.21° | 67.73° | 48.82° | 19.98° |
| $\alpha$ for MEBDF | 90° | 90° | 90° | 88.36° | 83.07° | 74.48° | 61.98° | 42.87° |
| $\alpha$ for PMEBDF | 90° | 90° | 90° | 89.32° | 86.19° | 80.60° | 72.63° | 60.60° |
| $\alpha$ for FPMEBDF | 90° | 90° | 90° | 89.71° | 88.01° | 84.67° | 78.70° | 65.01° |

**Table 4.2**
Absolute error for methods MEBDF, PMEBDF, FPMEBDF applied to problem (4.1).

| $h$ | $a$ | $b$ | $k$ | MEBDF | PMEBDF | FPMEBDF |
|---|---|---|---|---|---|---|
| 0.1 | 5 | 25 | 6 | 9.1458e+67 | 1.0827e−10 | 6.4619e−10 |
| 0.05 | 5 | 25 | 6 | 9.8280e−46 | 4.2093e−42 | 3.1724e−51 |
| 0.1 | 10 | 25 | 7 | 3.7745e+60 | 2.8380e−08 | 1.8857e−10 |
| 0.05 | 10 | 25 | 7 | 4.2158e−24 | 8.6327e−43 | 1.0682e−41 |
| 0.1 | 10 | 15 | 8 | 3.2440e+19 | 2.2573e−10 | 4.7513e−13 |
| 0.05 | 10 | 15 | 8 | 2.1582e−21 | 5.9876e−31 | 6.2765e−38 |

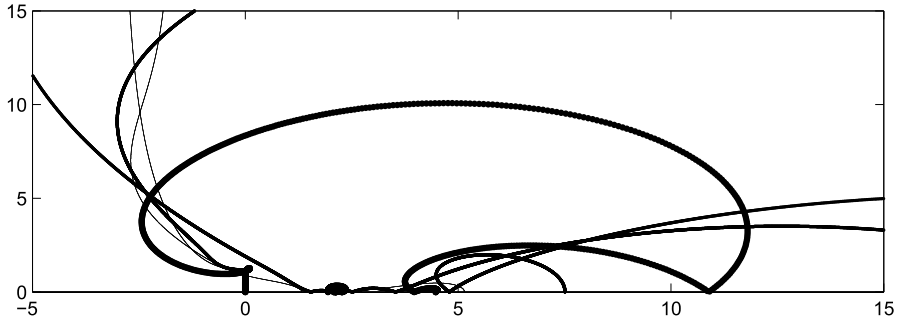


**Fig. 4.5.** Regions of stability of PEMBDF method (thin line), FPMBDF method (medium line) and MEBDF method (thick line) for $k = 8$.

This method is $A(\alpha)$-stable for $\alpha = 60.60$. An example of the FPMEBDF method with $k = 8$ is given by

$$b_1 = -\frac{337}{783}, \quad b_2 = -\frac{382}{225}, \quad b_3 = -\frac{921}{314}, \quad b_4 = -\frac{1013}{377},$$
$$b_5 = -\frac{35}{188}, \quad b_6 = \frac{1172}{349}, \quad b_7 = \frac{1099}{268}, \quad b_8 = -\frac{359}{672}$$

whose angle of $A(\alpha)$-stability is for $\alpha = 65.01$. The corresponding stability regions are plotted in Fig. 4.5.

These angles $\alpha$ of $A(\alpha)$-stability for PEMBDF methods are presented in Table 4.1 together with the corresponding angles for BDF, EBDF, and MBDF formulas. The asterisk in this table indicates that the method is not $A(\alpha)$-stable. As can be noticed from the results reported in Table 4.1, an increase in the angles of $A(\alpha)$-stability is visible for perturbed and fully perturbed formulas with respect to the MEBDF methods. In particular, as $k$ increases, the increase in angle is more remarkable.

In order to provide a numerical confirmation of the theoretical expectations regarding the increase achieved in the angles of $A(\alpha)$-stability, we consider the following test problem

$$y' = Ay, \quad t \in [0, 50], \tag{4.1}$$

with

$$A = \begin{bmatrix} -a & -b \\ b & -a \end{bmatrix}, \quad a > 0,\ b > 0,$$

for which $\sigma(A) = \{-a + bi, -a - bi\}$. We compared the perturbed and fully perturbed formulas with $k = 8$, with the MEBDF methods in a fixed stepsize environment with stepsize $h = h_1, h_2$, where $h_1 = 0.05$ and $h_2 = 0.1$. For $h = h_1$, the points $h(-10 \pm 15i)$ are inside of stability regions for all methods and the numerical approximations tend to 0 as the numerical solution advances. For $h = h_2$ the points $h(-10 \pm 15i)$ are outside of stability region for MEBDF method and inside of stability regions for PMEBDF and FPMEBDF methods. We have confirmed numerically that in this case the numerical approximation computed by the MEBDF method is divergent, where those computed by PMEBDF and FPEMBDF formulas tend to zero as the numerical solution advances. We repeated the numerical tests also in the cases $k = 6$ and $k = 7$, choosing respectively $a = 5$, $b = 25$ and $a = 10$, $b = 25$. The results of such a computation are listed in Table 4.2, where the

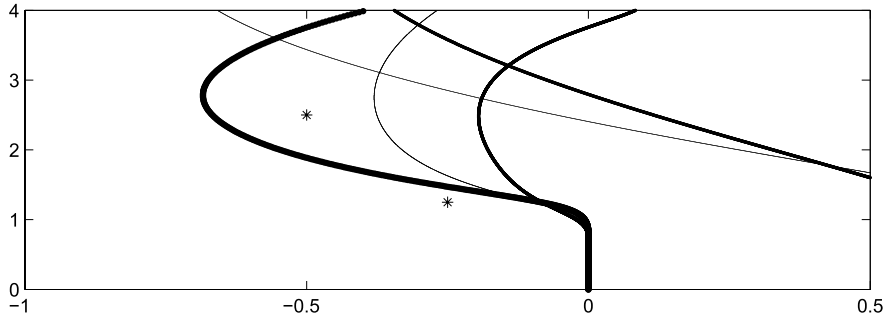


**Fig. 4.6.** Stability regions near the origin of MEBDF, PMEBDF, FPMEBDF methods corresponding to $k = 6$ and the points $h_1(-5+25i)$, $h_2(-5+25i)$, where $h_1 = 0.05$ and $h_2 = 0.1$.

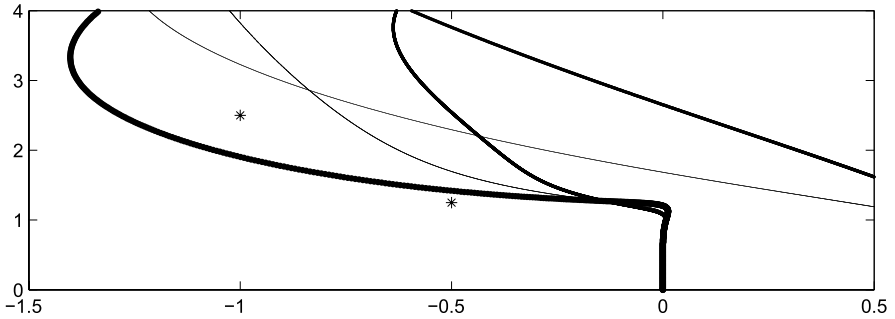


**Fig. 4.7.** Stability regions near the origin of MEBDF, PMEBDF, FPMEBDF methods corresponding to $k = 7$ and the points $h_1(-10+25i)$, $h_2(-10+25i)$, where $h_1 = 0.05$ and $h_2 = 0.1$.

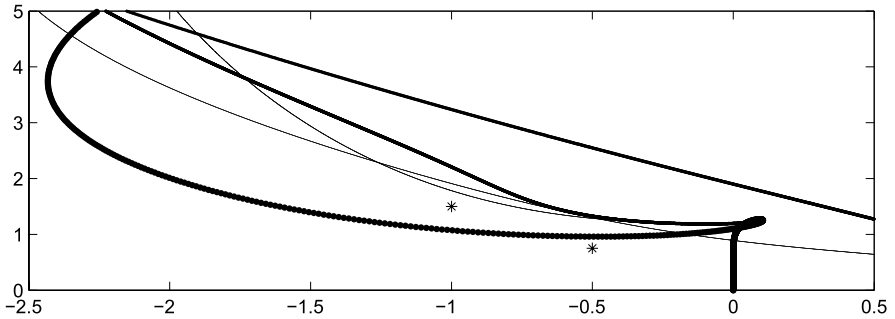


**Fig. 4.8.** Stability regions near the origin of MEBDF, PMEBDF, FPMEBDF methods corresponding to $k = 8$ and the points $h_1(-10+15i)$, $h_2(-10+15i)$, where $h_1 = 0.05$ and $h_2 = 0.1$.

absolute error $E_N = \|y(T) - y_N\|_1$, $N = hT$, is reported. The related points $h(-a + b\,i)$, for $h = h_1$ and $h = h_2$ are plotted in Figs. 4.6–4.8.

## 5. Local error estimation

This section is devoted to the derivation of a reliable estimation to principal term of the local truncation error $h^p y^{(p+1)}(t_n)$. In order to accomplish this purpose, some preliminary considerations are needed.

Let us consider the local solution $\widetilde{y}(t)$, i.e. the solution to the initial-value problem

$$\begin{cases} \widetilde{y}'(t) = f(\widetilde{y}(t)), \quad t \in [t_n, t_{n+1}], \\ \widetilde{y}(t_n) = y_n, \end{cases} \tag{5.1}$$

where the function $f(y)$ appearing in (1.1) and (5.1) satisfies the Lipschitz condition of the form

$$\|f(y) - f(z)\| \leq L\|y - z\|,$$

with a constant $L \geq 0$. Subtracting the integral forms of (1.1) and (5.1) we obtain

$$\|y(t) - \widetilde{y}(t)\| \leq \|y(t_n) - y_n\| + L \int_{t_n}^{t} \|y(s) - \widetilde{y}(s)\| ds,$$

$t \in [t_n, t_{n+1}]$. Using Gronwall's lemma (compare for example [7]) yields

$$\|y(t) - \widetilde{y}(t)\| \leq \|y(t_n) - y_n\| e^{L(t-t_n)}.$$

Hence,

$$\|y(t) - \widetilde{y}(t)\| = O(h^p), \quad t \in [t_n, t_{n+1}].$$

Assuming that the function $f(y)$ is sufficiently smooth we have a similar conclusion for the derivatives of $y(t)$ and $\widetilde{y}(t)$

$$\|y^{(i)}(t) - \widetilde{y}^{(i)}(t)\| = O(h^p), \quad t \in [t_n, t_{n+1}],\ i = 1, 2, \ldots,$$

compare [8].

In this section we aim to provide an estimation to the leading term of the local truncation error having the form

$$h^{p+1}\widetilde{y}^{(p+1)}(t_n) \approx \sum_{j=0}^{k} \sigma_j y_{n+j} + \sigma_{k+1} h\overline{f}_{n+k} + \sigma_{k+2} h\overline{f}_{n+k+1}. \tag{5.2}$$

The following result holds.

**Theorem 5.1.** *Assume that the solution $\widetilde{y}(t)$ to the problem (5.1) is sufficiently smooth. Then the constants $\sigma_0, \sigma_1, \ldots, \sigma_{k+2}$ appearing in (5.2) satisfy the following linear system of equations*

$$\begin{cases} \displaystyle\sum_{j=0}^{k} \sigma_j = 0, \\ \displaystyle\sum_{j=0}^{k} \sigma_j \frac{j^\ell}{\ell!} + \sigma_{k+1} \frac{k^{\ell-1}}{(\ell-1)!} + \sigma_{k+2} \frac{(k+1)^{\ell-1}}{(\ell-1)!} = 0, \quad \ell = 1, 2, \ldots, p, \\ \displaystyle\sum_{j=0}^{k} \sigma_j \frac{j^{p+1}}{(p+1)!} + \sigma_{k+1} \frac{k^p}{p!} + \sigma_{k+2} \frac{(k+1)^p}{p!} = 1. \end{cases} \tag{5.3}$$

**Proof.** Under the localizing assumption, it is $\widetilde{y}_{n+j} = \widetilde{y}(t_{n+j}), j = 0, 1, \ldots, k-1$. Expanding $\widetilde{y}(t_n + jh)$ and $\widetilde{y}'(t_n + jh)$ in Taylor series around $t_n$, leads to

$$\widetilde{y}(t_n + jh) = \sum_{\ell=0}^{p} \frac{(jh)^\ell}{\ell!} \widetilde{y}^{(\ell)}(t_n) + O(h^{p+1}), \qquad \widetilde{y}'(t_n + jh) = \sum_{\ell=1}^{p} \frac{(jh)^{\ell-1}}{(\ell-1)!} \widetilde{y}^{(\ell)}(t_n) + O(h^{p+1}).$$

Substituting these relations in (5.2) we obtain

$$h^{p+1}\widetilde{y}^{(p+1)}(t_n) = \sum_{j=0}^{k} \sigma_j \widetilde{y}(t_n) + \sum_{\ell=1}^{p+1} \left( \sum_{j=0}^{k} \sigma_j \frac{j^\ell}{\ell!} + \sigma_{k+1} \frac{k^{\ell-1}}{(\ell-1)!} + \sigma_{k+2} \frac{(k+1)^{\ell-1}}{(\ell-1)!} \right) h^\ell \widetilde{y}^{(\ell)}(t_n) + O(h^{p+2}).$$

Comparing the terms of order $O(h^k)$ for $k = 0, 1, \ldots, p+1$ yields the system (5.3). □

We observe that system (5.3) can be written in a more compact vector form, as follows:

$$K\boldsymbol{\sigma} = e_{k+3},$$

where the coefficient matrix $K$ assumes the form

$$\begin{bmatrix} 1 & 1 & \ldots & 1 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & j & \ldots & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & & \vdots & \vdots \\ 0 & 1 & \ldots & \dfrac{j^\ell}{\ell!} & \ldots & \dfrac{k^{\ell-1}}{(\ell-1)!} & \dfrac{(k+1)^{\ell-1}}{(\ell-1)!} \\ \vdots & \vdots & & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \ldots & \dfrac{j^p}{p!} & \ldots & \dfrac{k^{p-1}}{(p-1)!} & \dfrac{(k+1)^{p-1}}{(p-1)!} \\ 0 & 1 & \ldots & \dfrac{j^{p+1}}{(p+1)!} & \ldots & \dfrac{k^p}{p!} & \dfrac{(k+1)^p}{p!} \end{bmatrix} \in \mathbb{R}^{(p+2)\times(k+3)},$$

$\boldsymbol{\sigma} = [\sigma_0, \sigma_1, \ldots, \sigma_{k+2}]^T$ and $e_{k+3}$ is the $(k+3)$-rd vector of the standard basis of $\mathbb{R}^{k+3}$. We observe that, even if the matrix $K$ is not the Vandermonde matrix, its non-singularity can be proved with arguments analogous to the ones used for the Vandermonde matrix (similar analysis have been carried out, for instance, in [9]). In particular, it can be proved that, if $p = k+1$, the system (5.3) has a unique solution for any value of $k$. These unique solutions to the system (5.3) for different values of $k$ are listed below:

- $k = 1$:
$$\boldsymbol{\sigma} = \left[-\frac{12}{5} \quad \frac{12}{5} \quad -\frac{18}{5} \quad \frac{6}{5}\right]^T$$
- $k = 2$:
$$\boldsymbol{\sigma} = \left[\frac{30}{17} \quad -\frac{168}{17} \quad \frac{138}{17} \quad -\frac{132}{17} \quad \frac{24}{17}\right]^T$$
- $k = 3$:
$$\boldsymbol{\sigma} = \left[-\frac{170}{111} \quad \frac{330}{37} \quad -\frac{930}{37} \quad \frac{1970}{111} \quad -\frac{500}{37} \quad \frac{60}{37}\right]^T$$
- $k = 4$:
$$\boldsymbol{\sigma} = \left[\frac{555}{394} \quad -\frac{1820}{197} \quad \frac{5310}{197} \quad -\frac{10020}{197} \quad \frac{12505}{394} \quad -\frac{4110}{197} \quad \frac{360}{197}\right]^T$$
- $k = 5$:
$$\boldsymbol{\sigma} = \left[-\frac{1379}{1035} \quad \frac{455}{46} \quad -\frac{2240}{69} \quad \frac{13090}{207} \quad -\frac{2065}{23} \quad \frac{34811}{690} \quad -\frac{686}{23} \quad \frac{140}{69}\right]^T$$
- $k = 6$:
$$\boldsymbol{\sigma} = \left[\frac{644}{503} \quad -\frac{80584}{7545} \quad \frac{19950}{503} \quad -\frac{43680}{503} \quad \frac{191660}{1509} \quad -\frac{72744}{503} \quad \frac{186578}{2515} \quad -\frac{20328}{503} \quad \frac{1120}{503}\right]^T$$
- $k = 7$:
$$\boldsymbol{\sigma} = \left[-\frac{3018}{2429} \quad \frac{4004}{347} \quad -\frac{83524}{1735} \quad \frac{41370}{347} \quad -\frac{67970}{347} \quad \frac{79604}{347} \quad -\frac{75684}{347} \quad \frac{1253418}{12145} \quad -\frac{18264}{347} \quad \frac{840}{347}\right]^T$$
- $k = 8$:
$$\boldsymbol{\sigma} = \left[\frac{46845}{38596} \quad -\frac{840060}{67543} \quad \frac{557340}{9649} \quad -\frac{1550472}{9649} \quad \frac{2880675}{9649} \quad -\frac{3787980}{9649} \quad \frac{3699780}{9649} \quad -\frac{3020040}{9649} \quad \frac{37211841}{270172} \quad -\frac{641610}{9649} \quad \frac{25200}{9649}\right]^T.$$

We have provided the estimation (5.2) to the local truncation error, which is asymptotically correct for $h$ tending to 0. However, in order to approach stiff systems, this property of correctness is not sufficient, since their solution also requires the usage of large stepsizes with respect to certain features of the problem. Shampine and Baca in [10] focused their attention on the assessment of the quality of the error estimate for large values of the stepsize, by using similar arguments as in the classical theory of absolute stability. We now specialize the results obtained in [10] to our class of PMEBDF methods.

Following [10], we consider a restricted class of problems of the form $y' = Jy$, where $J$ is a constant matrix that can be diagonalized by a similarity transformation $M^{-1}JM = \mathrm{diag}(\xi_i)$. Then, it is sufficient to consider the scalar problem

$$\begin{cases} y'(t) = \xi y, \quad t \geq 0, \\ y(0) = 1, \end{cases} \tag{5.4}$$

where $\xi \in \mathbb{C}$ is one the eigenvalues of $J$, which is supposed to have negative real part. The solution of the problem (5.4) is $y(t) = e^{\xi t}$ and, therefore,

$$y_{n+j} = e^{\xi(t_n + jh)} + O(h^{p+1}).$$

As a consequence, we obtain

$$\mathrm{le}(t_n) = e^{\xi t_n}\left(\sum_{j=0}^{k} \alpha_j e^{jz} - z\beta_k e^{kz} - z\beta_{k+1} e^{(k+1)z}\right) + O(z^{p+1}),$$

where $z = \xi h$. Using the results contained in Theorem 5.1, we next provide the estimate (5.2) $\mathrm{est}(t_n)$, obtaining

$$\mathrm{est}(t_n) = C_p(t_n)e^{\xi t_n}\left(\sum_{j=0}^{k}\sigma_j e^{jz} + z\sigma_{k+1}e^{kz} - z\sigma_{k+2}e^{(k+1)z}\right) + O(z^{p+1}).$$

To investigate the behavior of error estimates for large values of $z$, we define the functions $R_{\mathrm{le}}(z)$ and $R_{\mathrm{est}}(z)$ by

$$R_{\mathrm{le}}(z) = \sum_{j=0}^{k}\alpha_j e^{jz} - z\beta_k e^{kz} - z\beta_{k+1}e^{(k+1)z},$$

$$R_{\mathrm{est}}(z) = \sum_{j=0}^{k}\sigma_j e^{jz} + z\sigma_{k+1}e^{kz} - z\sigma_{k+2}e^{(k+1)z},$$

corresponding to $\mathrm{le}(t_n)$ and $\mathrm{est}(t_n)$. To assess the quality of $\mathrm{est}(t_n)$ for large stepsizes, we examine the ratio

$$r(z) = \frac{R_{\mathrm{est}}(z)}{R_{\mathrm{le}}(z)}. \tag{5.5}$$

We observe that the ratio (5.5) behaves in the following way:

$$r(z) \sim \frac{\sigma_0}{\alpha_0}, \quad |z| \to \infty, \quad \mathrm{Re}(z) < 0,$$

and this behavior would suggest that the original estimate $\mathrm{est}(t_n)$ can be used for all the values of the stepsize. However, it is important to observe that the denominator appearing in the above expression could be quite small and, as a consequence, the ratio $r(z)$ results to be very large and, therefore, the error estimate $\mathrm{est}(t_n)$ would not be reliable at all. To compensate for this, Shampine and Baca proposed in [10], in the context of RK methods, premultiplying $\mathrm{est}(t_n)$ by the so-called *filter matrix*,

$$(I - h\mathbf{J}(t_n))^{-1},$$

where $\mathbf{J}(t_n)$ is an approximation to th Jacobian matrix of the problem (1.1) at the point $t_n$. This choice is suitable to damp the large, stiff error components. As observed in [10], the improved error estimator does not alter the behavior for small $h$ but it corrects the behavior of the estimate for large values of $h$.

## 6. Concluding remarks and future work

We have analyzed modified extended BDF of Cash [1,4] in the framework of GLMs for ODEs. This analysis leads to the new classes of perturbed MEBDF methods of the same order, which have better stability properties than the MEBDF formulas for $k = 4, 5, 6, 7$, and 8. The resulting methods are $A(\alpha)$-stable with larger angles $\alpha$ of stability. The improved stability properties were then confirmed by some numerical experiments. The future work will involve the incorporation of these methods into a variable stepsize variable order software for stiff systems of ODEs, by employing the error estimate provided in Section 5 and suitably extending the results obtained in [9,11,12].

## Acknowledgments

The results reported in this paper were obtained during the visit of the third author to the University of Naples and University of Salerno in May 2009, March 2010 and May 2011. This author wishes to express his gratitude to Giuseppe Izzo, Elvira Russo and Beatrice Paternoster for making these visits possible.

## References

[1] J.R. Cash, On the integration of stiff systems of O.D.E.s using extended backward differentiation formulae, Numer. Math. 34 (1980) 235–246.
[2] E. Hairer, G. Wanner, Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems, Springer-Verlag, New York, 1991.
[3] J.D. Lambert, Numerical Methods for Ordinary Differential Systems, Wiley, New York, 1991.
[4] J.R. Cash, The integration of stiff initial value problems in ODEs using modified extended backward differentiation formulae, Comput. Math. Appl. 9 (1983) 645–657.
[5] E. Hairer, G. Wanner, Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems, second rev. ed., Springer-Verlag, New York, 1996.
[6] Z. Jackiewicz, General Linear Methods for Ordinary Differential Equations, John Wiley & Sons, 2009.
[7] L.F. Shampine, Numerical Solution of Ordinary Differential Equations, Chapman & Hall, New York, London, 1994, pp. x+484.
[8] L.F. Shampine, Computer Solution of Ordinary Differential Equations, W.H. Freeman and Company, San Francisco, 1975, pp. x+318.
[9] R.D' Ambrosio, Highly stable multistage numerical methods for functional equations, Bi-Nationally supervised Ph.D. Thesis in Mathematics, University of Salerno-Arizona State University, 2010.
[10] L.F. Shampine, L.S. Baca, Error estimators for stiff differential equations, J. Comput. Appl. Math. 11 (2) (1984) 197–207.
[11] R.D' Ambrosio, Z. Jackiewicz, Construction and implementation of highly stable two-step continuous methods for stiff differential systems, Math. Comput. Simul. 81 (9) (2011) 1707–1728.
[12] Z. Jackiewicz, Implementation of DIMSIMs, Appl. Numer. Math. 42 (1–3) (2002) 251–267.