Numerical Approximation of Forward-Backward Differential Equations by a Finite Element Method

M.F. Teodoro\textsuperscript{1,2}, P.M. Lima\textsuperscript{2}, N.J. Ford\textsuperscript{3} and P.M. Lumb\textsuperscript{3}

\textsuperscript{1} Departamento de Matemática, EST, Instituto Politécnico de Setúbal, Portugal
\textsuperscript{2} CEMAT, Instituto Superior Técnico, UTL, Lisboa, Portugal
\textsuperscript{3} Departement of Mathematics, University of Chester, CH1 4BJ, Chester, UK
emails: mteodoro@est.ips.pt, plima@math.ist.utl.pt, njford@chester.ca.uk, p.lumb@chester.ca.uk

Abstract

This paper is concerned with the approximate solution of a linear first-order functional differential equation which involves forward and backward deviating arguments. We search for a solution $x$, defined for $t \in [-1,k], (k \in \mathbb{N})$, which takes given values on the intervals $[-1,0]$ and $(k-1,k]$. Continuing the work started in [13], [14] and [8], we introduce and analyse a new computational algorithm based on the finite element method for the solution of this problem which is applicable both in the case of constant and variable coefficients. Numerical results are presented and compared with the results obtained by other methods.

Key words: Mixed-type functional differential equation, method of steps, collocation method, least squares, finite element, splines
MSC 2000: 34K06; 34K10; 4K28; 65Q05

1 Introduction

This paper is devoted to the numerical solution of mixed-type functional differential equations (MTFDE). We consider equations of the type

$$x'(t) = \alpha(t)x(t) + \beta(t)x(t-1) + \gamma(t)x(t+1),$$  \hspace{1cm} (1)

where $x$ is the unknown function, $\alpha$, $\beta$ and $\gamma$ are known functions. The interest in this type of equation is motivated by its applications in optimal control [11], economic dynamics [12], nerve conduction [2] and travelling waves in a spatial lattice [1]. Important contributions to their analysis have appeared in the literature in the two last decades of the past century. The ill-posedness of MTFDEs has been discussed by Rustichini in...
1989 [11], where he considered linear autonomous equations. The same author extended his results to nonlinear equations [12]. J. Mallet-Paret applied the Fredholm theory to obtain new results for this class of equation [9] and introduced the idea of factorization of their solutions [10]. Independently, the authors of [6] have obtained results about factorization for the linear non-autonomous case. On the other hand, Krisztin [7] has analysed the roots of the characteristic equation of linear systems of MTFDE, which has led him to important results on the qualitative behaviour of their solutions. In particular, he has shown that a MTFDE may have a nonoscillatory solution in spite of the non-existence of real roots of its characteristic equation (unlike the case of delay-differential equations). Based on the existing insights on the qualitative behaviour of MTFDE, the authors of [3] and [4] have recently developed a new approach to the analysis of these equations in the autonomous case. More precisely, they have analysed MTFDE as boundary value problems, that is, for a MTFDE of the form (1), but with constant coefficients $\alpha$, $\beta$ and $\gamma$, they have considered the problem of finding a differentiable solution on a certain real interval $[-1, k]$, given its values at the intervals $[-1, 0]$ and $(k - 1, k]$. They have concluded that in general the specification of such boundary functions is not sufficient to ensure that a solution can be found. For the case where such a solution exists they have introduced a numerical algorithm to compute it. This approach was further developed in [13], where new numerical methods were proposed for the solution of such boundary value problems. In [14] and [8] these methods were extended to the non-autonomous case (when $\alpha$, $\beta$ and $\gamma$ are smooth functions of $t$).

In [5], we have decomposed the solution into a growing and a decaying component, following the ideas of J. Mallet-Paret and Verduyn-Lunel [10]. By approximating each of the components separately, we partially overcome the ill-conditioning of the problem, which enables us to improve the accuracy of the numerical results and to obtain numerical solutions on much larger intervals.

In [8] a theoretical basis was given to the computational methods, by relating them to the existing analytical results about MTFDE and with classical results of numerical analysis.

In the present paper, we give an overview of the numerical methods described in [8], [13], [14] and introduce a new algorithm, based on the finite element method. A comparative analysis of all these algorithms is provided.

2 Preliminaries

2.1 Method of Steps

The purpose is to compute a particular solution of equation (1) which satisfies

$$x(t) = \begin{cases} \varphi_1(t), & \text{if } t \in [-1, 0], \\ f(t), & \text{if } t \in (k - 1, k], \end{cases}$$

(2)

where $\varphi_1$ and $f$ are smooth real-valued functions, defined on $[-1, 0]$ and $(k - 1, k]$, respectively ($1 < k \in \mathbb{N}$). In order to analyze and solve this boundary value problem
(BVP) we consider an initial value problem (IVP), with the conditions:

\[ x(t) = \varphi(t), \quad t \in [-1, 1], \]  

(3)

where the function \( \varphi \) is defined by

\[ \varphi(t) = \begin{cases} 
\varphi_1(t), & \text{if } t \in [-1, 0], \\
\varphi_2(t), & \text{if } t \in (0, 1]. 
\end{cases} \]  

(4)

This reformulation provides the basis for both analytical and numerical construction of solutions using ideas based on Bellman’s method of steps for solving delay differential equations. One solves the equation over successive intervals of length unity. Assuming that \( \gamma(t) \neq 0, \forall t \geq 0 \), equation (1) can be rewritten in the form

\[ x(t + 1) = a(t)x'(t) + b(t)x(t) - c(t)x(t), \]  

(5)

where

\[ a(t) = \frac{1}{\gamma(t)}, \quad b(t) = -\frac{\beta(t)}{\gamma(t)}, \quad c(t) = -\frac{\alpha(t)}{\gamma(t)}. \]  

(6)

Using formula (5) repeatedly, we can construct a smooth solution of equation (1) on any interval \([1, k]\), starting from its definition on \([-1, 1]\) by formula (4). Continuing this process, we can extend the solution to any interval, provided that the initial function \( \varphi \) and the variable coefficients are smooth enough functions. The next theorem formulates this result in more precise terms.

**Theorem 2.1** Let \( x \) be the solution of problem (1), (4), where \( \alpha(t), \beta(t), \gamma(t) \in C^{2L}([-1, 2L + 1]), \gamma(t) \neq 0, \quad t \in [-1, 2L + 1], \quad \varphi_1(t) \in C^{2L}([-1, 0]), \quad \varphi_2(t) \in C^{2L}(0, 1]) \) (for some \( L \in \mathbb{N} \)). Then there exist functions \( \delta_{i,l}, \epsilon_{i,l}, \bar{\delta}_{i,l}, \bar{\epsilon}_{i,l} \in C([-1, 2L + 1]), \quad l = 1, \ldots, L, \) \( i = 0, 1, \ldots, 2l \), such that the following formulae are valid:

\[ x(t) = \sum_{i=0}^{2l-1} \delta_{i,l}(t)\varphi_1^{(i)}(t-2l) + \sum_{i=0}^{2l-1} \epsilon_{i,l}(t)\varphi_2^{(i)}(t-2l+1), \quad t \in [2l-1, 2l]; \]

\[ x(t) = \sum_{i=0}^{2l} \bar{\delta}_{i,l}(t)\varphi_2^{(i)}(t-2l) + \sum_{i=0}^{2l-1} \bar{\epsilon}_{i,l}(t)\varphi_1^{(i)}(t-2l-1), \quad t \in [2l, 2l+1]; \]  

(7)

A detailed proof can be found in [8]. For the autonomous case, similar results were obtained in [3].

### 2.2 Existence and Uniqueness of Solution

Results about existence and uniqueness of solution are found in [10], when analysing mixed-type functional differential equations on a real interval \([-\tau, \tau]\). Actually, this BVP can be reformulated (by a simple shift of the independent variable) as a BVP on an interval \([-\tau - 1, \tau + 1]\), with boundary conditions given at \([-\tau - 1, -\tau]\) and \([\tau, \tau + 1]\).

In [8] the results presented in [10] are applied to the analysis of the considered BVP.

To present the necessary conditions for the existence of at least one continuously differentiable solution to a given BVP of the form (1), (2), we begin by considering the ODE approach, introduced in two recent papers [13], [14].
Numerical Approximation of Forward-Backward Differential Equations

On the interval \([-1, 1]\) the solution of (1), (2) can be written in the form

\[
x(t) = x_0(t) + u(t), \quad t \in [-1, 1]
\]

where \(x_0\) is an initial approximation of the solution and \(u\) is a correction. More precisely, we require that \(x_0\) satisfies the following conditions:

- \(x_0(t) = \phi_1(t), \forall t \in [-1, 0];\)
- \(x_0\) is \(k\) times continuously differentiable on \([-1, 0]\), \(x_0(0) = \varphi_1(0)\) and \(x_0^{(j)}(0) = \varphi_1^{(j)}(0), j = 1, \ldots, k;\)
- \(x_0\) is \(k-1\) times continuously differentiable on \((0, 1]\), \(x_0(1) = x(1)\) and \(x_0^{(j)}(1) = x^{(j)}(1), j = 1, \ldots, k-1\) (where \(x\) is the required solution of the problem (1),(2)).

Once \(x_0\) is defined, our problem is reduced to the computation of the correction \(u\). First of all note that \(u(t) \equiv 0, \forall t \in [-1, 0]\) (otherwise, \(x\) does not satisfy the first boundary condition). Therefore, if we define \(u\) on \([0, 1]\), we can extend \(u\) to the whole interval \([-1, k]\) using the method of steps (as described in previous subsection). Let us denote as \(u[-1,k]\) the extension of \(u(t)\) to the interval \([-1, k]\). Then \(u[-1,k]\) is defined piece by piece in the following way:

\[
u[-1,k](t) = \begin{cases} 
0, & \text{if } t \in [-1, 0]; \\
u(t), & \text{if } t \in (0, 1]; \\
u[l-1,l](t), & \text{if } t \in (l - 1, l], \; l = 2, 3, \ldots, k.
\]

We shall now obtain an expression for \(u[l-1,l](t)\). Using (1) repeatedly, and taking into account that \(u(t) \equiv 0, \forall t \in [-1, 0]\), we can express \(u[l-1,l](t)\) in terms of \(u(t)\) and its derivatives, on any interval \((l - 1, l]\).

In particular, on the interval \((k - 1, k]\), we obtain

\[
u[k-1,k](t) = L^{k-1}u(t-k+1) := c_{k-1,k}(t)u^{(k-1)}(t-k+1) + \cdots + c_{0,k}(t)u(t-k+1),
\]

\(t \in (k - 1, k]\). Here, \(L^{k-1}\) denotes a linear differential operator of order \(k - 1\). The \(c_{ik}\) are coefficients that can be computed recursively, just as the \(\delta_{ik}\) and \(\epsilon_{ik}\) coefficients in the right-hand side of (7).

Notice that \(x_0\) can also be extended to the interval \([-1, k]\), using the same method. Now, since \(x\) must satisfy the second boundary condition in (2), we conclude that

\[
L^{k-1}u(t) = f(t + k - 1) - x_0[-1,k](t + k - 1), \quad t \in [0, 1].
\]

Imposing regularity conditions at \(t = 0\) and \(t = 1\), the following boundary conditions must be satisfied:

\[
u(0) = u'(0) = \cdots = u^{(k)}(0) = 0; \\
u(1) = u'(1) = \cdots = u^{(k-1)}(1) = 0.
\]

The number of boundary conditions in (12), \(2k + 1\), is higher than the order of the considered ODE (11). Therefore, there may not exist a solution of (11) which satisfies...
all the conditions (12). This is not surprising, since the existence of a solution to the original boundary value problem (1), (2) is also not guaranteed (as discussed in [3]).

When solving the problem (11), (12) we must keep only \( k - 1 \) conditions and ignore the remaining ones. The corresponding boundary value problem considering \( k - 1 \) boundary conditions is called the reduced BVP.

Finally, as it follows from [10] and [8], if a continuously differentiable solution \( x \) exists and \( b(t)c(t) > 0, \forall t \in [0, k - 1] \), then this solution is unique in the space \( W^{1,p}(0,k) \).

### 3 Numerical Methods

#### 3.1 Outline of the methods

We now give an outline of three numerical methods based on the ODE approach, described in the previous section.

According to these methods, we search for an approximate solution of (1) on \([-1, 1]\) in the form

\[
\hat{x}_N(t) = x_0(t) + \sum_{j=0}^{d-1} C_j x_j(t), \quad t \in [-1, 1]
\]  

where \( x_0 \) is an initial approximation of the solution; \( \{x_j\}_{0 \leq j \leq d-1} \) is a basis in the space of functions where the correction to the initial approximation is sought; \( d \) is the dimension of this space. The algorithms for computing \( \hat{x}_N \) consist of three steps:

1. constructing the initial approximation;
2. defining a set of basis functions;
3. computing the \( C_j \) coefficients.

1. **Constructing the initial approximation.**

From formula (5) it follows that if a solution of equation (1) belongs to \( C^n([-l+1, l]) \) (for certain \( l \geq 1, n \geq 1 \)), then it also belongs to \( C^{n-1}([-l, l+1]) \). Therefore, since we want \( \hat{x}_N \) to be at least continuous on \([-1, k]\) (for a certain \( k \geq 2 \)), we require that \( x_0 \) belongs to \( C^k([-l, l]) \). With this in mind, we define \( x_0 \) on \([-1, 1]\) in the following way:

\[
x_0(t) = \begin{cases} 
\varphi_1(t), & t \in [-1, 0]; \\
P_{2k}(t) = a_0 + a_1 t + \cdots + a_{2k} t^{2k}, & t \in [0, 1].
\end{cases}
\]  

Since \( x_0 \) must be \( k \) times continuously differentiable on \([-1, 0]\), \( P_{2k} \) must satisfy regularity conditions at \( t = 0 \) and at \( t = 1 \), we define a linear system of \( 2k + 1 \) equations with \( 2k + 1 \) unknowns \( a_0, a_1, \ldots, a_{2k} \). This system has a nonsingular matrix, for any \( k \geq 2 \).

Further, \( x_0 \) is extended from \([-1, 1]\) to \([-1, k]\) using the recurrence formulae (7). Let us denote by \( x_0^{[-1,k]} \) this extension.

2. **Definition of a basis set.**

With the purpose of computing a correction to the initial approximation on \([0, k]\),
we first consider this correction on [0, 1]. Let us define a grid of stepsize \( h \) on this interval. Let \( h = 1/N \) (where \( N \in \mathbb{N}, N \geq k + 1 \)) and \( t_i = ih, i = 0, \ldots, N \). The correction \( \tilde{x}_d(t) - x_0(t) \) on [0, 1] will be sought as a \( k \)-th degree spline, \( S_k(t) \), defined on this grid, which satisfies \( S_k(0) = S_k(1) = 0 \). As usual, we will use as basis functions \( x_j(t) \) the so-called \( B \)-splines of degree \( k \). From the definition it follows that the basis functions have the following properties: \( x_j \in C^{k-1}[0, 1] \); \( x_j(t) \) is different from zero only in \( (t_j, t_{j+k+1}) \) and \( x_j(t) \) is a polynomial of degree \( k \) on each interval \( [t_i, t_{i+1}] \), \( i = 0, \ldots, N - 1 \). Note that we have \( N - k \) functions \( x_j \) with these properties; therefore, we set \( d = N - k \).

Next the basis functions are extended to the interval \([0, k]\) using the formulae (7), where \( \varphi_1 \) is replaced by 0 and \( \hat{\varphi}_2 \) is replaced by \( x_j \). Let us denote the extended basis functions by \( x_j^{[0,k]} \). Each time we extend the basis function to the next interval, the smoothness of the splines decreases by 1 unit (though the polynomial degree remains constant). Therefore, on the interval \([k - 1, k]\) the basis functions are continuous but not continuously differentiable function. On the whole interval \([0, k]\), the approximate solution is given by

\[
\tilde{x}_N^{[0,k]}(t) = x_0^{[-1,k]}(t) + \sum_{j=0}^{N-k-1} C_j x_j^{[0,k]}(t), \quad t \in [0, k].
\]  

(15)

### 3. Computation of the Coefficients

Finally, we compute the coefficients \( C_j, j = 0, \ldots, N - k - 1 \) of the expansion (15) from the condition that \( \tilde{x}_N \) approximates \( f \) on the interval \([k - 1, k]\). Three alternative methods were used for this purpose. The collocation and least squares methods are described in [13], [14] and [8]. Here we will describe the finite element method. As in the previous two methods, the coefficients \( C_j, j = 0, \ldots, N - k - 1 \) are computed by solving a linear system with a \((N - k) \times (N - k)\) band matrix. In this last case, the coefficients \( C_j \) are obtained from the following orthogonality conditions:

\[
\int_{k-1}^{k} \left( f(t) - x_0^{[-1,k]}(t) - \sum_{j=0}^{N-k-1} C_j x_j^{[0,k]}(t) \right) x_j^{[0,k]}(t - k + 1) dt = 0, j = 0, \ldots, N - k - 1.
\]

The form of the basis functions leads us to the solution of a system of \( N - k \) linear equations with a band matrix \( A \) and independent term \( B \) whose generic element are respectively

\[
a_{i+1,j+1} = \begin{cases} 
\int_{k-1}^{k} x_i^{[0,k]}(t) x_j^{[0,k]}(t - k + 1) dt, & \text{if } i, j = 0, \ldots, N - k - 1, |i - j| \leq k, \\
0, & \text{if } i, j = 0, \ldots, N - k - 1, |i - j| > k;
\end{cases}
\]

and

\[
b_{j+1} = \int_{k-1}^{k} \left( f(t) - x_0^{[-1,k]}(t) \right) x_j^{[0,k]}(t - k + 1) dt, \quad \text{if } j = 0, \ldots, N - k - 1.
\]
Table 1: Numerical results for Example 4.1 with \( m = 2 \), by the finite element method. \( \epsilon = \|x - \tilde{x}_N\|_2 \) on each unit interval \([l-1, l]\), \( l = 1, \ldots, k-1 \), \( k = 3, 4, 5 \).

<table>
<thead>
<tr>
<th>N</th>
<th>( \epsilon )</th>
<th>( p )</th>
<th>( \epsilon )</th>
<th>( p )</th>
<th>( \epsilon )</th>
<th>( p )</th>
<th>( \epsilon )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>2.428e-11</td>
<td>4.60</td>
<td>1.907e-10</td>
<td>4.68</td>
<td>5.979e-9</td>
<td>4.35</td>
<td>2.602e-7</td>
<td>3.36</td>
</tr>
<tr>
<td>64</td>
<td>8.660e-13</td>
<td>4.81</td>
<td>6.638e-12</td>
<td>4.85</td>
<td>2.749e-10</td>
<td>4.44</td>
<td>2.495e-8</td>
<td>3.38</td>
</tr>
<tr>
<td>128</td>
<td>2.899e-14</td>
<td>4.90</td>
<td>2.194e-13</td>
<td>4.92</td>
<td>1.240e-11</td>
<td>4.47</td>
<td>2.311e-9</td>
<td>3.43</td>
</tr>
</tbody>
</table>

Once again, we can remark that the obtained system of linear equations is the same that we obtain if we apply the finite element method to the solution of the equation (11).

4 Numerical Results and Conclusions

We present an example of a BVP for MTFDE (4.1) also used in [8] to test some of the described numerical algorithms. The error norm at each interval \([l-1, l]\) is the discrete analog of the 2-norm:

\[
\|\tilde{x}_N^{[l,l+1]} - x\|_2 = \left(h \sum_{i=0}^{N} \left(\tilde{x}_N^{[l,l+1]}(t_{iN+i}) - x(t_{iN+i})\right)^2\right)^{1/2}.
\]

The estimate of the convergence order is obtained as usual by

\[
p = \log_2 \frac{\|x - \tilde{x}_N\|_2}{\|x - \tilde{x}_{2N}\|_2}.
\]

Example 4.1 The following example of an autonomous equation was first considered:

\[
x'(t) = (m - 0.5e^{-m} - 0.5e^m)x(t) + 0.5x(t-1) + 0.5x(t+1), \quad (16)
\]

with the boundary conditions \( \varphi_1(t) = e^{mt}, \quad t \in [-1, 0]; \quad f(t) = e^{mt}, \quad t \in (k-1, k), \) with \( m \in \mathbb{R}, m \neq 0 \). The exact solution is \( x(t) = e^{mt} \).
Numerical Approximation of Forward-Backward Differential Equations

Table 2: Numerical results for Example 4.1 using different methods.
\[
\epsilon = \|x - \hat{x}_N\|_2/\sqrt{(k-1)} \text{ on interval } [0,k-1], \ k = 3, 4, 5, \ m = 2.
\]

Table 1 contains numerical results for Example 4.1, including error norms and estimates of the convergence order. These values correspond to the finite element method and were computed separately on each interval \([l,l+1]\), in the case \(m = 2\), for different values of \(k\). In the case \(k = 3\) (resp. \(k = 4\)) the estimate of the convergence order is close to 4 in the first interval (resp. first two intervals) and decreases to approximately 3.5 in the interval \(([k-2,k-1])\). In the case \(k = 5\), the same estimate is higher than 4 in the first three intervals and less than four in \([3,4]\). This fall of the estimate is probably connected with the fact that the convergence order in \([0,1]\) is not the same for the solution and for its higher derivatives. As shown in [8], the convergence order of the solution in the subsequent intervals depends on the convergence order of its derivatives in the first one. In Table 2, the global error norm and the convergence order of the numerical solution are displayed. Here the error norms were computed globally on \([0,k-1]\), for the three methods: finite element, least squares and collocation method. As it was expected, the higher convergence order was obtained for the finite element method, \(p = 3.5\), and the lowest one corresponds to collocation method, \(p = 2\). For the least squares method, the estimated global order of convergence is close to 3. The finite element method seems to be the most efficient one, providing high (\(\approx 3.5\)) convergence order and giving results with absolute error less than \(6 \times 10^{-9}\), for \(h = 1/128\), when \(k = 3\) and \(m = 2\). In the future, we intend to carry out a more detailed numerical analysis of the presented methods, in particular the least squares and the finite element algorithms.
Acknowledgements

M.F. Teodoro acknowledges support from FCT, grant SFRH/BD/37528/2007. All the authors acknowledge support from the Treaty of Windsor Programme (project B-15/08).

References


