ELSEVIER

Applied Numerical Mathematics 24 (1997) 1–21

APPLIED NUMERICAL MATHEMATICS

A new mesh selection strategy for ODEs

Luigi Brugnano *, Donato Trigiante [1]

Dipartimento di Energetica, Universita di Firenze, Via C. Lombroso 6/17, 50134 Firenze, Italy

Received 21 December 1995; accepted 19 December 1996

Abstract

In this paper a new mesh selection strategy, based on the conditioning properties of continuous problems, is presented. It turns out to be particularly efficient when approximating solutions of BVPs. The numerical methods used to test the reliability of the strategy are symmetric Linear Multistep Formulae (LMF) used as Boundary Value Methods (BVMs) since they provide a wide choice of methods of arbitrary high order and have similar stability properties to each other. In particular, we shall consider a subclass of such methods, called *Top Order Methods* (TOMs) (Amodio, 1996; Brugnano and Trigiante, 1995, 1996), to carry out the numerical results on some singular perturbation test problems. © 1997 Elsevier Science B.V.

Keywords: Mesh selection; Conditioning; Boundary value problems; Boundary value methods

1. Introduction

The problem of variable stepsize selection is fundamental for the efficient numerical solution of ODEs. In fact, a uniform mesh is often not adequate to reach a given accuracy, since it would require a huge number of mesh points.

The strategies used so far are essentially based on the control of the local truncation errors. In particular, in the case of BVPs they use as the main tool the equidistribution of an appropriate *monitor function* defined through the estimated local errors [3,4,10–23]. This approach, however, is efficient under the basic assumption that the considered continuous BVP is well conditioned. Consequently, when this is not the case, numerical methods may not provide good approximations, since the selected mesh may be not an appropriate one.

In this paper we propose a new mesh selection strategy, which utilizes a monitor function based on the conditioning properties of the continuous problem. An algorithm is then presented, along with several numerical tests. For sake of brevity and clarity, we restrict ourselves to the linear case. However, the extension to nonlinear problems can be done by using standard arguments.

* Corresponding author. E-mail: na.brugnano@na-net.ornl.gov.

[1] E-mail: na.dtrigiante@na-net.ornl.gov.

PII S0168-9274(97)00007-X

The methods used in the numerical tests are LMF of symmetric type used as boundary value methods (BVMs) [5,7,9]. However this approach is sufficiently general that it can be used with the standard codes COLSYS [4] and TWPBVP [13] which appear in Netlib.

2. Classification of continuous problems

We shall consider the case of a two-point boundary value problem, although all the arguments can be extended to the more general case of multi-point conditions. Consider

$$\begin{aligned} &y' = L(t)y + f(t), \quad y, f, \eta \in \mathbb{R}^d, \\ &B_0 y(t_0) + B_1 y(T) = \eta. \end{aligned} \tag{1}$$

Here $L(t)$ is a square $d \times d$ matrix whose entries, as well as those of $f(t)$, belong to $C^{(1)}([t_0, T])$. The solution of this problem is

$$y(t) = \Phi(t, t_0) Q^{-1} \eta + \int_{t_0}^{T} G(t, s) f(s) \, \mathrm{d}s, \tag{2}$$

where

(a) $\Phi(t, t_0)$ is the fundamental matrix,

(b) $Q = B_0 + B_1 \Phi(T, t_0)$ is assumed to be nonsingular, and

(c)

$$G(t, s) = \begin{cases} \Phi(t, t_0) Q^{-1} B_0 \Phi(t_0, s), & \text{for } t \geqslant s, \\ -\Phi(t, t_0) Q^{-1} B_1 \Phi(T, s), & \text{for } t < s, \end{cases}$$

is the Green's function.

A perturbation $\delta\eta$ of the boundary condition will cause a perturbation δy to the solution which is bounded by

$$\|\delta y(t)\| \leqslant \|\Phi(t, t_0) Q^{-1}\| \, \|\delta\eta\|,$$

where $\|\cdot\|$ is any norm in $\mathbb{R}^d$. Let us define the function

$$\varphi(t) = \|\Phi(t, t_0) Q^{-1}\|, \tag{3}$$

and the norms in $C([t_0, T])$,

$$\|\delta y\|_\infty = \max_{t_0 \leqslant t \leqslant T} \|\delta y(t)\|, \qquad \|\delta y\|_1 = \frac{1}{T - t_0} \int_{t_0}^{T} \|\delta y(t)\| \, \mathrm{d}t. \tag{4}$$

One obtains

$$\|\delta y\|_\infty \leqslant \kappa_c \|\delta\eta\|, \qquad \|\delta y\|_1 \leqslant \gamma_c \|\delta\eta\|,$$

where

$$\kappa_c = \max_{t_0 \leqslant t \leqslant T} \varphi(t), \qquad \gamma_c = \frac{1}{T - t_0} \int_{t_0}^{T} \varphi(t) \, \mathrm{d}t.$$

The comparison between the two parameters κ_c and γ_c permits us to classify the problem as follows:

(1) Both κ_c and γ_c have moderate sizes. The continuous problem is well conditioned. The error is almost uniformly distributed over the integration interval and, therefore, a uniform mesh is appropriate.

(2) The parameter γ_c is of moderate size but $\kappa_c \gg \gamma_c$. In this case the error is concentrated in subintervals whose total measure is small with respect to $T - t_0$. The problem can be solved with a moderate number of mesh points only if they are appropriately chosen.

(3) Both parameters κ_c and γ_c are large. The problem is ill conditioned in both norms. In this case the numerical solution will need a large number of mesh points, even if a variable mesh is used.

We also mention that in [8] it was proposed to use the above parameters to define stiffness. In particular, problems having a large ratio $\sigma = \kappa_c/\gamma_c$ are stiff.

Let us now consider the effect on the solution due to a perturbation $\delta f(t)$, that is

$$\delta y(t) = \int_{t_0}^{T} G(t,s)\delta f(s)\,\mathrm{d}s.$$

Such a perturbation can be bounded by using again the above defined parameters. Suppose, for example, that the perturbation on the function $f(t)$ is impulsive, that is $\delta f(t) = c\delta(t-\bar{t})$, where c is a constant vector, $\bar{t} \in [t_0, T]$ and $\delta(t-\bar{t})$ is the Dirac function. One has,

$$\delta y(t) = \int_{t_0}^{T} G(t,s)c\,\delta(s-\bar{t})\,\mathrm{d}s = \begin{cases} \Phi(t,t_0)Q^{-1}B_0\Phi(t_0,\bar{t})c, & \text{for } t \geqslant \bar{t}, \\ -\Phi(t,t_0)Q^{-1}B_1\Phi(T,\bar{t})c, & \text{for } t < \bar{t}. \end{cases}$$

One then obtains,

$$\left\|\delta y(t)\right\|_\infty \leqslant \kappa_c\|c\| \max\left\{\left\|B_0\Phi(t_0,\bar{t})\right\|, \left\|B_1\Phi(T,\bar{t})\right\|\right\},$$

and

$$\left\|\delta y(t)\right\|_1 \leqslant \gamma_c\|c\| \max\left\{\left\|B_0\Phi(t_0,\bar{t})\right\|, \left\|B_1\Phi(T,\bar{t})\right\|\right\}.$$

3. Discrete problems

We choose to use numerical methods having the imaginary axis as the boundary of their absolute stability region. There are solid arguments to support such a choice, but we skip them for brevity (see [5,9]).

When applied to problem (1), a numerical method based on LMF generates a discrete problem such as

$$\boldsymbol{My} = \begin{pmatrix} \eta \\ h_1\hat{f}_1 \\ \vdots \\ h_N\hat{f}_N \end{pmatrix},$$

where $\boldsymbol{y} = (y_0, y_1, \dots, y_N)^{\mathrm{T}}$ is a block vector of dimension $(N+1)d \times 1$, whose ith block entry contains the approximation of the solution at t_i; $h_1, \dots, h_N$ are the stepsizes used, and, by setting $f_i = f(t_i)$,

$$\hat{f}_i = f_i + \mathrm{O}(h_i)$$

is a suitable combination of the values of the function $f(t)$ at the grid points near t_i.

The entries of the vector $\boldsymbol{y}$ are numbered starting from 0. Consequently, the (block) rows and columns of M will be numbered starting from the same value.

The discussion made in the previous section for continuous problems can be extended to discrete ones. In fact, let us define the matrices

$$M^{-1} = \begin{pmatrix} G_{00} & \dots & G_{0N} \\ \vdots & & \vdots \\ G_{N0} & \dots & G_{NN} \end{pmatrix}, \qquad \Omega = \begin{pmatrix} \|G_{00}\| & \dots & \|G_{0N}\| \\ \vdots & & \vdots \\ \|G_{N0}\| & \dots & \|G_{NN}\| \end{pmatrix}. \tag{5}$$

Then, a perturbation $\delta\eta$ of the boundary condition produces a perturbation $\delta\boldsymbol{y} = (\delta y_0, \dots, \delta y_N)^{\mathrm{T}}$ to the solution bounded by

$$|\delta\boldsymbol{y}| \leqslant \Omega_{*0}\|\delta\eta\|,$$

where $|\delta\boldsymbol{y}| = (\|\delta y_0\|, \dots, \|\delta y_N\|)^{\mathrm{T}}$, and Ω_{*j} denotes the jth column of Ω.

For brevity, let us now introduce the *mesh vector*

$$\boldsymbol{h} = (0, h_1, \dots, h_N)^{\mathrm{T}},$$

whose entries are the stepsizes used, and the vector $|\delta\tilde{\boldsymbol{y}}|$ defined as

$$|\delta\tilde{\boldsymbol{y}}|_0 = \|\delta y_0\|, \qquad |\delta\tilde{\boldsymbol{y}}|_i = \max\{\|\delta y_{i-1}\|, \|\delta y_i\|\}, \quad i = 1, \dots, N.$$

Then, the quantities

$$e_\infty(\boldsymbol{h}) = \max_i \|\delta y_i\|, \qquad e_1(\boldsymbol{h}) = \frac{1}{T - t_0}\boldsymbol{h}^{\mathrm{T}}|\delta\tilde{\boldsymbol{y}}|,$$

can be considered the discrete analogs of the corresponding continuous quantities (4). As before, we shall define the parameters

$$\kappa_{\mathrm{d}}(\boldsymbol{h}) = \max_i \Omega_{i0}, \qquad \gamma_{\mathrm{d}}(\boldsymbol{h}) = \frac{1}{T - t_0}\boldsymbol{h}^{\mathrm{T}}\tilde{\Omega}_{*0},$$

such that

$$e_\infty(\boldsymbol{h}) \leqslant \kappa_{\mathrm{d}}(\boldsymbol{h})\|\delta\eta\|, \qquad e_1(\boldsymbol{h}) \leqslant \gamma_{\mathrm{d}}(\boldsymbol{h})\|\delta\eta\|,$$

where the vector $\tilde{\Omega}_{*0}$ has components

$$\tilde{\Omega}_{00} = \Omega_{00}, \qquad \tilde{\Omega}_{i0} = \max\{\Omega_{i-1,0}, \Omega_{i0}\}, \quad i = 1, \dots, N.$$

Consequently, the discrete problem can be classified in the same way as the continuous one.

We observe that only the first (block) column of the matrix M^{-1} (see (5)) is needed to compute $\kappa_{\mathrm{d}}(\boldsymbol{h})$ and $\gamma_{\mathrm{d}}(\boldsymbol{h})$. This is not an expensive task, once a factorization of the matrix M has been computed.

In general the values $\gamma_{\mathrm{d}}(\boldsymbol{h})$ and γ_{c}, as well as the values $\kappa_{\mathrm{d}}(\boldsymbol{h})$ and κ_{c}, will differ. The differences between the continuous parameters and the corresponding discrete ones will be used as measure of the "closeness" of the two problems.

4. The new strategy

The new mesh selection strategy will be based on the requirement that both the discrete and the continuous problems belong to the same class of conditioning. Therefore, we shall impose the condition that the above definite discrete quantities $\kappa_{\mathrm{d}}(\boldsymbol{h})$ and $\gamma_{\mathrm{d}}(\boldsymbol{h})$ approximate as well as possible the corresponding continuous ones. The only possibility to achieve such result without increasing the number of the mesh points is to vary the mesh vector $\boldsymbol{h}$.

Let us now look for the vector $\boldsymbol{h}$ which makes $\gamma_{\mathrm{d}}(\boldsymbol{h})$, for a given N, a better approximation to γ_{c}. Consider the identity

$$\begin{aligned}\gamma_{\mathrm{d}}(\boldsymbol{h}) &= \left(\gamma_{\mathrm{c}} + \frac{1}{T-t_0}\left(\sum_{i=1}^{N} h_i\tilde{\varphi}(t_i) - \int_{t_0}^{T}\varphi(t)\,\mathrm{d}t\right) + \frac{1}{T-t_0}\sum_{i=1}^{N} h_i\left(\tilde{\Omega}_{i0} - \tilde{\varphi}(t_i)\right)\right)\\ &= \gamma_{\mathrm{c}} + E_1 + E_2,\end{aligned}$$

where

$$\tilde{\varphi}(t_i) = \max_{t_{i-1}\leqslant t\leqslant t_i} \varphi(t),$$

$\varphi(t)$ is the function defined in (3), and

$$E_1 = \frac{1}{T-t_0}\left(\sum_{i=1}^{N} h_i\tilde{\varphi}(t_i) - \int_{t_0}^{T}\varphi(t)\,\mathrm{d}t\right).$$

It follows that E_1 is the error in the quadrature formula for the function $\varphi(t)$. We observe that E_1 is positive by definition. It is not difficult to check that

$$E_1 \leqslant \frac{1}{T-t_0}\sum_{i=1}^{N} h_i\left(h_i|\varphi_i'|\right) \leqslant \frac{N}{T-t_0}\max_i h_i\left(h_i|\varphi_i'|\right),$$

where φ_i' is the value of the derivative of φ evaluated at a suitable point belonging to the interval $[t_{i-1}, t_i]$.

Concerning the term E_2, supposing that $\tilde{\Omega}_{i0}$ is not very small, it can be written as

$$|E_2| = \left|\frac{1}{T-t_0}\sum_{i=1}^{N} h_i\tilde{\Omega}_{i0}\frac{(\tilde{\Omega}_{i0} - \tilde{\varphi}(t_i))}{\tilde{\Omega}_{i0}}\right| \leqslant \frac{N}{T-t_0}\max_i\ h_i\tilde{\Omega}_{i0}\nu_i,$$

where

$$\nu_i = \frac{|\tilde{\Omega}_{i0} - \tilde{\varphi}(t_i)|}{\tilde{\Omega}_{i0}} \tag{6}$$

is the absolute value of the relative error. Each term in the sum giving $|E_2|$ is the product of two factors: $h_i\widetilde{\Omega}_{i0}$ and the factor ν_i representing the relative error in the ith interval. As a consequence, if the stepsizes are suitably small, $|E_2|$ is very small. Our strategy will then make this quantity negligible with respect to E_1. The problem of getting $|E_2|$ small is a difficult one and will be solved iteratively.

This is done by choosing the mesh vector $\boldsymbol{h}$ that minimizes $E_1+|E_2|$, that is by solving the minmax problem

$$\min_{\boldsymbol{h}} \max_i h_i\big(\nu_i\widetilde{\Omega}_{i0} + h_i|\varphi_i'|\big), \quad \sum_{i=1}^{N} h_i = T - t_0, \tag{7}$$

where the quantities $\{\nu_i\}$ are unknown. We assume that they are bounded by a quantity ν which will be taken, for example, equal to one. This will require us to solve the problem in different stages. In the first stage, all the ν_i are taken equal to one and the problem

$$\min_{\boldsymbol{h}} \max_i h_i\big(\widetilde{\Omega}_{i0} + h_i|\varphi_i'|\big), \quad \sum_{i=1}^{N} h_i = T - t_0,$$

is solved instead of (7). The unknown quantities $|\varphi_i'|$ are approximated by

$$|\varphi_i'| \approx \frac{|\Omega_{i0} - \Omega_{i-1,0}|}{h_i} \equiv \frac{|\Delta\Omega_{i-1,0}|}{h_i}, \quad i = 1,\dots,N,$$

so that the problem becomes

$$\min_{\boldsymbol{h}} \max_i h_i\big(\widetilde{\Omega}_{i0} + |\Delta\Omega_{i-1,0}|\big), \quad \sum_{i=1}^{N} h_i = T - t_0.$$

By introducing the *monitor function*

$$\psi(t) \equiv \widetilde{\Omega}_{i0} + |\Delta\Omega_{i-1,0}|, \quad \text{for } t \in (t_{i-1}, t_i), \tag{8}$$

the minmax problem is then solved by the process of equidistribution of the function ψ, which provides a new mesh vector $\boldsymbol{h}^{(1)}$. That is (see, for example, [3, p. 363]), the new mesh points $t_i^{(1)} = t_{i-1}^{(1)} + h_i^{(1)}$, $i = 1,\dots,N$, are chosen so that

$$\int_{t_{i-1}^{(1)}}^{t_i^{(1)}} \psi(t)\,\mathrm{d}t = \frac{1}{N}\int_{t_0}^{T} \psi(t)\,\mathrm{d}t, \quad i = 1,\dots,N.$$

The new mesh vector is then used to obtain a new matrix $M^{(1)}$ and, consequently, a new vector $\widetilde{\Omega}_{*0}^{(1)}$. As a result, new approximations $\kappa_\mathrm{d}(\boldsymbol{h}^{(1)})$ and $\gamma_\mathrm{d}(\boldsymbol{h}^{(1)})$ are obtained.

The process may be iterated. According to what has already been said, the new mesh vector will have small components where the monitor function $\psi(t)$ is large. Since the latter quantity is large at the points where Ω_{i0} is large, the process tends to concentrate the points in a neighborhood where κ_d occurs. This implies that at each successive iteration, better and better approximations to κ_c will be obtained. At the same time, the successive values of γ_d are decreasing, since smaller stepsizes are used where the entries of Ω_{*0} are larger.

A failure of the latter sequence to decrease, or a failure of the former sequence to converge, means that the number of mesh intervals N is not large enough.

Suppose now that a minimum value of γ_{d} has been reached in correspondence of the mesh vector $\boldsymbol{h}^*$. The first stage of the procedure terminates. At this point, one may check the reliability of the obtained approximations $\kappa_{\mathrm{d}}(\boldsymbol{h}^*)$ and $\gamma_{\mathrm{d}}(\boldsymbol{h}^*)$, that is to get an estimate for

$$\left|\kappa_{\mathrm{c}} - \kappa_{\mathrm{d}}(\boldsymbol{h}^*)\right|, \qquad \left|\gamma_{\mathrm{c}} - \gamma_{\mathrm{d}}(\boldsymbol{h}^*)\right|.$$

If the considered method has order p, this can be achieved either by mesh doubling, or by considering a more accurate method, with similar stability properties, over the same mesh $\boldsymbol{h}^*$. This allows us to obtain new approximations κ_{new} and γ_{new}, as well as a new approximated discrete solution $\boldsymbol{y}_{\mathrm{new}}$. If the values κ_{new} and γ_{new} are close to $\kappa_{\mathrm{d}}(\boldsymbol{h}^*)$ and $\gamma_{\mathrm{d}}(\boldsymbol{h}^*)$, respectively, one accepts the current mesh. If not, this means that N needs to be increased. If the mesh is accepted, then one also has an estimate of the global error.

Remark 1. Observe that it may seem a difficult task to find, in the class of LMF, methods of different order and having similar stability properties and higher order. This is certainly true when LMF are used as initial value methods, but the task becomes really trivial if one uses LMF as boundary value methods (BVMs). In fact, there are a lot of BVMs, namely the "symmetric schemes" [5,7,9], which essentially share the same stability properties of the trapezoidal rule, but having arbitrarily high order.

As a result, at the end of the first stage we have that the continuous function $\varphi(t)$ is well approximated in a set $\mathcal{I}_1 \subseteq [t_0, T]$, called the *precision set*. A criterion to estimate the precision set will be described in Section 4.1. For the moment, suppose that $\mathcal{I}_1$ is known.

The second stage then assumes $\nu_i = 0$ for the points belonging to $\mathcal{I}_1$ and $\nu_i = 1$ elsewhere. From (7) one then obtains a new monitor function. Some more mesh points, say N_1, are introduced in the mesh contained in $[t_0, T]\backslash\mathcal{I}_1$. Such new points, along with those already contained in $[t_0, T]\backslash\mathcal{I}_1$, are the only ones used to equidistribute the new monitor function. This will leave unchanged the mesh inside $\mathcal{I}_1$. One then obtains a new precision set $\mathcal{I}_2$ and so on. The process terminates when

$$\mathcal{I}_r \equiv [t_0, T].$$

In practice, in the above procedure it is preferable to use the perturbed monitor function (see (8))

$$\widehat{\psi}(t) = \psi(t) + \alpha,$$

where α is a suitable small positive parameter [2]. This is done in order to avoid the selection of too large stepsizes where $\psi(t)$ is small.

4.1. Estimate of the precision set

After the end of the first stage, we need to estimate the precision set, that is the set where the relative errors ν_i in (7) are suitably small. In principle, after the check of the parameters $\kappa_{\mathrm{d}}(\boldsymbol{h}^*)$ and $\gamma_{\mathrm{d}}(\boldsymbol{h}^*)$, we have the discrete functions

$$\widetilde{\Omega}_{i0}, \quad \widetilde{\Omega}_{i0}^{(\mathrm{new})}, \qquad i = 0, \ldots, N,$$

obtained with the less accurate and the more accurate method, respectively, in correspondence of the mesh vector $\boldsymbol{h}^*$. Consequently, one could use the estimates (see (6))

$$\nu_i \approx \frac{|\widetilde{\Omega}_{i0} - \widetilde{\Omega}_{i0}^{(\text{new})}|}{\widetilde{\Omega}_{i0}}.$$

We prefer, however, to use a different approach, which has been found to be more effective. The idea can be easily described by considering the scalar problem

$$y' = \lambda y, \qquad \lambda \in \mathbb{R}.$$

It follows that the function $\varphi(t)$ defined in (3) satisfies the relation

$$\varphi(t_i) = \varphi(t_{i-1} + h_i) = \varphi(t_{i-1})\,\mathrm{e}^{\lambda h_i}.$$

Then, by setting $q_i = \lambda h_i$, we have that

$$q_i = \log\big(\varphi(t_i)/\varphi(t_{i-1})\big).$$

Similarly, when a one step method is used, we have that the discrete approximation Ω_{i0} of $\varphi(t_i)$ satisfies

$$\Omega_{i0} = z_i \Omega_{i-1,0},$$

where z_i is the characteristic root of the method. If q_i is sufficiently small, then $z_i \approx \mathrm{e}^{q_i}$, so that

$$q_i \approx \log(\Omega_{i0}/\Omega_{i-1,0}).$$

Hence, we shall assume that t_i belongs to the precision set when the above quantity is suitably small. Conversely, it belongs to its complement.

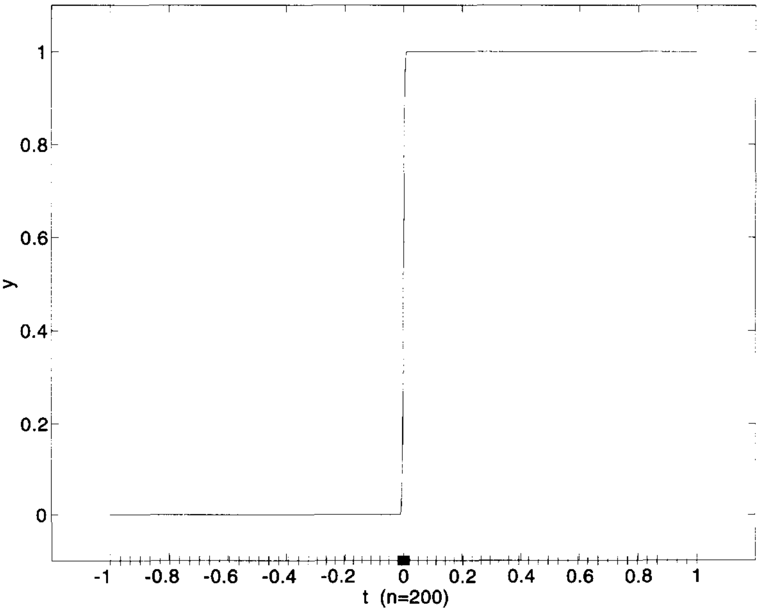

Fig. 1. Discrete approximation of the solution of problem (10), $\varepsilon = 10^{-5}$, at the end of the first stage.

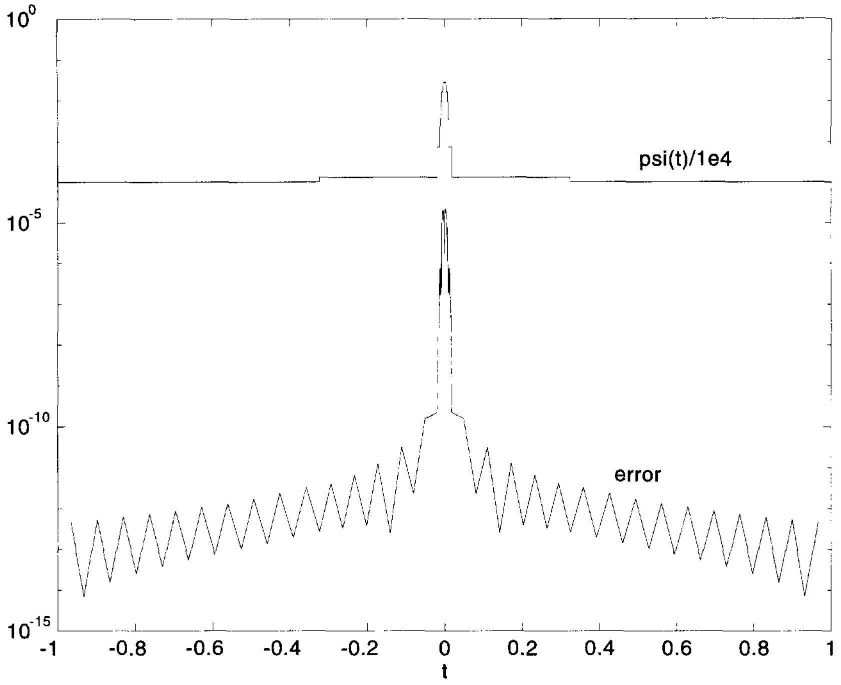

Fig. 2. Error on the computed solution of problem (10), $\varepsilon = 10^{-5}$, along with the monitor function (8).

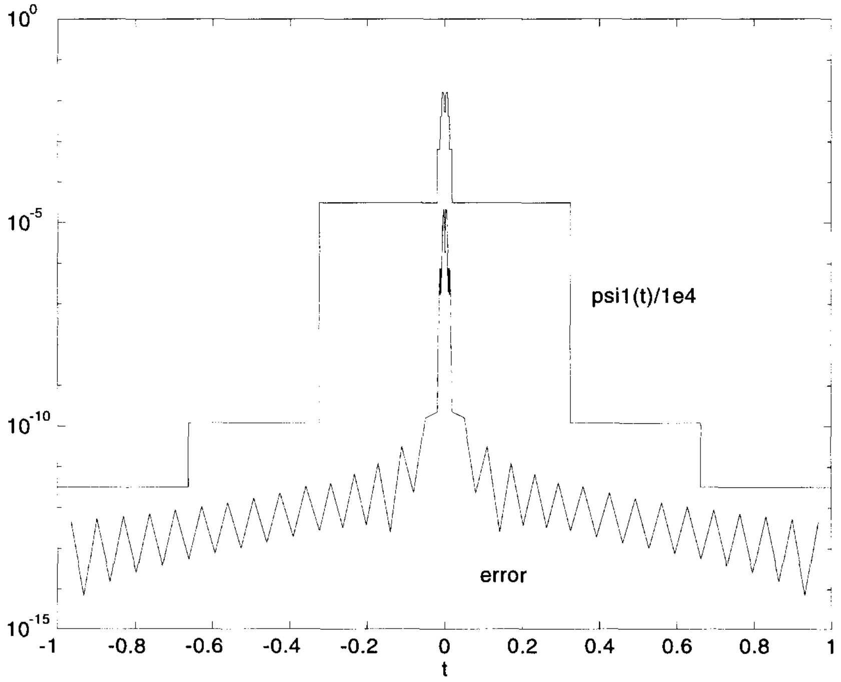

Fig. 3. Error on the computed solution of problem (10), $\varepsilon = 10^{-5}$, along with the modified monitor function (11).

The previous considerations are generalized to multistep methods, by considering that their solutions are essentially generated by only one of the characteristic roots [5,6,9].

In the case of systems of equations, the analogous quantity

$$\widehat{q}_i = \left| \log(\Omega_{i0}/\Omega_{i-1,0}) \right|, \tag{9}$$

is used. Observe that large values of $\widehat{q}_i$ correspond to large relative variations of the discrete function. In this case, it is reasonable to expect a large relative error. Conversely, a small value of $\widehat{q}_i$ means that the discrete function has small relative variations. Therefore, we consider the mesh points corresponding to small values of $\widehat{q}_i$ as belonging to the precision set.

The estimate (9) is very cheap to obtain. Nevertheless, it is quite reliable. As an example, let us consider the following singular perturbation problem,

$$\begin{aligned} &\varepsilon y'' + t y' = 0, \\ &y(-1) = 0, \qquad y(1) = 1, \end{aligned} \tag{10}$$

where $\varepsilon = 10^{-5}$. In Fig. 1 we report the computed discrete solution at the end of the first stage of the procedure, obtained by using the trapezoidal rule. The final mesh is also reported. It is easily seen that most of the 200 mesh points are around the layer at $t = 0$. The values κ_d and γ_d computed on the final mesh are very close to the corresponding continuous parameters. In fact, by considering the infinity norm, we have obtained $\kappa_\mathrm{d} \approx 252$, which is exactly the value of the continuous parameter, and $\gamma_\mathrm{d} \approx 2$, whereas $1 \leqslant \gamma_\mathrm{c} < 1.5$ [8].

In Fig. 2 the error on the discrete solution, along with the monitor function (8) (scaled by a factor 10^4), are reported. In Fig. 3 we plot the monitor function modified by taking $\nu_i = 0$ inside the estimated precision set, that is,

$$\psi_1(t) \equiv \nu_i \widetilde{\Omega}_{i0} + |\Delta\Omega_{i-1,0}|, \quad \text{for } t \in (t_{i-1}, t_i). \tag{11}$$

In particular, ν_i has been taken equal to zero if both t_{i-1} and t_i are inside the precision set, and equal to one otherwise. Moreover, a mesh point t_i has been considered inside the precision set if the corresponding value $\widehat{q}_i$ computed as in (9) was smaller than one.

By comparing Fig. 3 with Fig. 2, one realizes at once that the weights ν_i are equal to one only in a small neighborhood of $t = 0$, that is where the error is maximum.

5. The nonhomogeneous case

At the end of the above procedure, we have that both $|\kappa_\mathrm{c} - \kappa_\mathrm{d}(\boldsymbol{h}^*)|$ and $|\gamma_\mathrm{c} - \gamma_\mathrm{d}(\boldsymbol{h}^*)|$ are minimized. This means that the entries $\{G_{i0}\}$ in the first block column of M^{-1} (see (5)) are good approximations of the corresponding matrices $\{\Phi(t_i, t_0)Q^{-1}\}$. Concerning the remaining block entries of M^{-1}, the following result holds true.

Theorem 1. *If the matrix M is suitably scaled, and the first block column of M^{-1} is an accurate approximation of the function $\Phi(t, t_0)Q^{-1}$ at the grid points, then G_{ij} is an accurate approximation of $G(t_i, t_j)$, for all $i = 0, \ldots, N$, $j = 1, \ldots, N$.*

Proof. The proof will be carried out in the simpler case where the method used is the trapezoidal rule. However, the arguments can be extended to more general multistep methods, although the proof becomes longer.

Let us apply the trapezoidal rule to problem (1). Then, we scale the matrix M as follows,

$$M = \begin{pmatrix} B_0 & & & B_1 \\ -Z_1 & I & & \\ & \ddots & \ddots & \\ & & -Z_N & I \end{pmatrix},$$

$$Z_i = \left(I - \tfrac{1}{2}h_i L(t_i)\right)^{-1}\left(I + \tfrac{1}{2}h_i L(t_{i-1})\right), \quad i = 1, \dots, N.$$

Consequently, by setting $\widehat{Q} = B_0 + B_1 \prod_{j=1}^{N} Z_j$, we obtain

$$M^{-1} = \begin{pmatrix} \widehat{Q}^{-1} & -\widehat{Q}^{-1}B_1 \prod\limits_{r=2}^{N} Z_r & \dots & -\widehat{Q}^{-1}B_1 \\ Z_1\widehat{Q}^{-1} & Z_1\widehat{Q}^{-1}B_0 Z_1^{-1} & \dots & -Z_1\widehat{Q}^{-1}B_1 \\ \vdots & \vdots & & \vdots \\ \prod\limits_{r=1}^{N-1} Z_r\widehat{Q}^{-1} & \prod\limits_{r=1}^{N-1} Z_r\widehat{Q}^{-1}B_0 Z_1^{-1} & \dots & -\prod\limits_{r=1}^{N-1} Z_r\widehat{Q}^{-1}B_1 \\ \prod\limits_{r=1}^{N} Z_r\widehat{Q}^{-1} & \prod\limits_{r=1}^{N} Z_r\widehat{Q}^{-1}B_0 Z_1^{-1} & \dots & \prod\limits_{r=1}^{N} Z_r\widehat{Q}^{-1}B_0\left(\prod\limits_{r=1}^{N} Z_r\right)^{-1} \end{pmatrix}.$$

The proof is then almost completed, since if for all $i = 0, \dots, N$, we assume that $\prod_{r=1}^{i} Z_r\widehat{Q}^{-1}$ is a good approximation of $\Phi(t_i, t_0)Q^{-1}$, then for $j \geqslant 1$ and $i \geqslant j$ we get

$$G_{ij} = \prod_{r=1}^{i} Z_r\widehat{Q}^{-1}B_0\left(\prod_{r=1}^{j} Z_r\right)^{-1} = \prod_{r=1}^{i} Z_r\widehat{Q}^{-1}B_0\widehat{Q}^{-1}\left(\prod_{r=1}^{j} Z_r\widehat{Q}^{-1}\right)^{-1}$$

$$\approx \Phi(t_i, t_0)Q^{-1}B_0Q^{-1}\left(\Phi(t_j, t_0)Q^{-1}\right)^{-1} = \Phi(t_i, t_0)Q^{-1}B_0\Phi(t_0, t_j) = G(t_i, t_j).$$

The proof in the case $i < j$ is obtained by similar arguments. □

The above result justifies our strategy which tends to obtain good approximations on the first block column of M^{-1}. However, this may not be sufficient when the inhomogeneity $f(t)$ in (1) is not smooth enough. In fact, in this case the solution of the continuous problem can be written as (see (2))

$$y(t) = y_{\text{hom}}(t) + z(t), \tag{12}$$

where $y_{\text{hom}}(t)$ is the solution of the associated homogeneous problem, while $z(t)$ is the solution of the problem when $\eta = 0$. The first term $y_{\text{hom}}(t)$ has already been discussed in the previous sections and a good approximation of it has been obtained on the mesh $\boldsymbol{h}^*$. Moreover, from (2) we obtain

$$z(t) = \int_{t_0}^{T} G(t,s)f(s)\,\mathrm{d}s = \sum_{j=1}^{N} \int_{t_{j-1}}^{t_j} G(t,s)f(s)\,\mathrm{d}s = \sum_{j=1}^{N} G(t,t_j) \int_{t_{j-1}}^{t_j} \Phi(t_j,s)f(s)\,\mathrm{d}s$$

$$= \sum_{j=1}^{N} G(t,t_j) \int_{t_{j-1}}^{t_j} \bigl(I + \mathrm{O}(h_j)\bigr)f(s)\,\mathrm{d}s.$$

At the point t_i, the numerical method provides the value

$$z_i = \sum_{j=1}^{N} h_j G_{ij} \hat{f}_j = \sum_{j=1}^{N} h_j G_{ij} \bigl(f_j + \mathrm{O}(h_j)\bigr).$$

From Theorem 1, we have that $G_{ij} \approx G(t_i,t_j)$, so that from the previous expressions we obtain

$$z(t_i) - z_i \approx \sum_{j=1}^{N} G_{ij} \left(\int_{t_{j-1}}^{t_j} f(s)\,\mathrm{d}s - h_j f_j \right).$$

By proceeding as before, we may define the following global measure of these errors,

$$\frac{1}{T-t_0} \sum_{i=1}^{N} h_i \bigl\| z(t_i) - z_i \bigr\| \approx \frac{1}{T-t_0} \sum_{i=1}^{N} h_i \left\| \sum_{j=1}^{N} G_{ij} \left(\int_{t_{j-1}}^{t_j} f(s)\,\mathrm{d}s - h_j f_j \right) \right\| =: E_3.$$

It follows that, for suitable $\xi_j \in (t_{j-1}, t_j)$, $j = 1, \ldots, N$,

$$E_3 \leqslant \frac{d}{T-t_0} \sum_{i=1}^{N} h_i \sum_{j=1}^{N} h_j \Omega_{ij} \bigl\| f'(\xi_j) \bigr\|_\infty \leqslant d\,\Omega_{\max} \max_j \bigl(h_j \bigl\| f'(\xi_j) \bigr\|_\infty \bigr),$$

where d is the dimension of the continuous problem, and (see (5)),

$$\Omega_{\max} = \max_{ij} \Omega_{ij}.$$

Therefore, in the intervals where $h_j \| f'(\xi_j) \|_\infty$ is large (i.e., where $f(t)$ has large variations and a suitable small stepsize is not used), the error on the approximation of $z(t)$ may become large. To get such error small we proceed to a further equidistribution. This time we shall equidistribute the monitor function

$$\psi_2(t) \equiv \max\bigl(\bigl\| f'(t_{i-1}) \bigr\|_\infty, \bigl\| f'(t_i) \bigr\|_\infty \bigr), \quad \text{for } t \in (t_{i-1}, t_i).$$

The new equidistribution must maintain the points of the mesh $\boldsymbol{h}^*$, otherwise the errors in the approximation of $y_{\text{hom}}(t)$ (see (12)) could increase. This implies that the new equidistribution may only add new mesh points to the old ones. This will also have the effect to decrease the errors E_1 and E_2.

Remark 2. By observing that in the intervals where $f(t)$ varies rapidly the local errors are presumably large, one could then handle the inhomogeneous term also by equidistributing the principal term of the local errors. This is, in fact, the strategy used by most of the currently available codes.

The overall process is described by the following pseudocode.

```
0. $it = 0$, $\boldsymbol{h} = \langle$uniform mesh of $N$ intervals$\rangle$,
   $\kappa_d(\boldsymbol{h}_{\text{old}}) = 0$, $\gamma_d(\boldsymbol{h}_{\text{old}}) = \infty$, $sk = 0$
1. compute $\kappa_d(\boldsymbol{h})$, $\gamma_d(\boldsymbol{h})$, $\boldsymbol{y}(\boldsymbol{h})$ and $\boldsymbol{h}_{\text{new}}$
   if $\kappa_d(\boldsymbol{h}) \approx \kappa_d(\boldsymbol{h}_{\text{old}})$
     $sk = 1$
     if $\gamma_d(\boldsymbol{h}) \geqslant 0.95 * \gamma_d(\boldsymbol{h}_{\text{old}})$
       $\boldsymbol{h}^* = \boldsymbol{h}$, goto 2
     else
       $\boldsymbol{h}_{\text{old}} = \boldsymbol{h}$, $\boldsymbol{h} = \boldsymbol{h}_{\text{new}}$, goto 1
     end
   elseif $sk = 1$
     $\boldsymbol{h}^* = \boldsymbol{h}_{\text{old}}$, goto 2
   elseif $it > it_{\max}$
     if $N > N_{\max}$
       error(too many mesh points required)
     else
       increase $N$, goto 0
     end
   else
     $\boldsymbol{h}_{\text{old}} = \boldsymbol{h}$, $\boldsymbol{h} = \boldsymbol{h}_{\text{new}}$, $it = it + 1$, goto 1
   end
2. compute $\kappa_{\text{new}}$, $\gamma_{\text{new}}$ and $\boldsymbol{y}_{\text{new}}$
   if $\kappa_{\text{new}} \approx \kappa_d(\boldsymbol{h}^*)$ and $\gamma_{\text{new}} \approx \gamma_d(\boldsymbol{h}^*)$
     $err$ = estimate_error($\boldsymbol{y}_{\text{new}}$, $\boldsymbol{y}(\boldsymbol{h}^*)$, $\boldsymbol{h}^*$)
     if $err < tol$
       exit
     elseif $N < N_{\max}$
       call refine_mesh
       if $\langle$non void complement of the precision set$\rangle$
         goto 1
       else
         call new_equid_mesh, goto 2
       end
     else
```

```
        error(too many mesh points required)
      end
    elseif $N < N_{\max}$
      increase $N$, goto 0
    else
      error(too many mesh points required)
    end
```

In the above pseudocode, the routine refine_mesh computes the precision set. Finally, the routine new_equid_mesh handles the eventual inhomogeneity of the problem, as seen in this section.

We conclude this section by observing that, when the continuous problem is well conditioned (i.e., when both κ_c and γ_c have moderate size), the above described procedure does not work better than those based on the equidistribution of the local errors. Conversely, when κ_c is large (i.e., for stiff and ill conditioned problems), the new strategy seems to be superior, as shown in the numerical tests.

Regarding to the cost needed by the new strategy for each equidistribution step, it is essentially given by the computation of the first block column of M^{-1} (see (5)), and this is obtained by solving d (the size of the continuous problem) linear systems with the matrix M. On the other hand, strategies based on local errors would require only one or two linear systems with the matrix M to be solved. Nevertheless, since the main cost in the solution of the linear systems is due to the LU factorization of M (and this must be done anyway), the computational cost per step is comparable for both approaches.

6. Numerical examples

The solution of problem (10) has already been obtained by using the trapezoidal rule and the mesh selection strategy previously described. In order to show the effectiveness of this strategy on different kinds of problems, in this section we shall consider some more numerical examples. With the only exception of the third problem, all of them are chosen among singularly perturbed BVPs.

We continue to use the trapezoidal rule, but every *symmetric scheme* (see [7,9]) could be used. The check of the parameters $\kappa_d(\boldsymbol{h}^*)$ and $\gamma_d(\boldsymbol{h}^*)$ (and then, the estimate of the error), is carried out by considering the sixth order TOM over the same mesh. The details of the implementation will be presented in a forthcoming paper.

Example 1. Consider the singularly perturbed BVP,

$$\varepsilon y'' - ty' + y = 0, \qquad y(-1) = 1, \qquad y(1) = 2, \tag{13}$$

where $\varepsilon = 10^{-4}$, whose solution has two boundary layers. This problem is very difficult to solve, and most of the currently available BVP solvers fail to provide the correct solution, when started from a uniform mesh. For example, the current version of the popular code COLSYS, started from a uniform mesh with 50 subintervals and used with ispace = 150,000, fspace = 500,000, fails to provide a correct answer for $\varepsilon \leqslant 10^{-2}$. This is shown in Fig. 4, where we plot the discrete solution obtained for $\varepsilon = 10^{-2}$. It is easily realized that the right boundary layer is missed.

By using the new strategy, we obtain the approximated solution reported in Fig. 5. The final mesh contains 480 points, and the estimated error is 9×10^{-5}. Moreover, as by-product we obtain that

$$\kappa_d(\boldsymbol{h}^*) \approx 10^4, \qquad \gamma_d(\boldsymbol{h}^*) \approx 3,$$

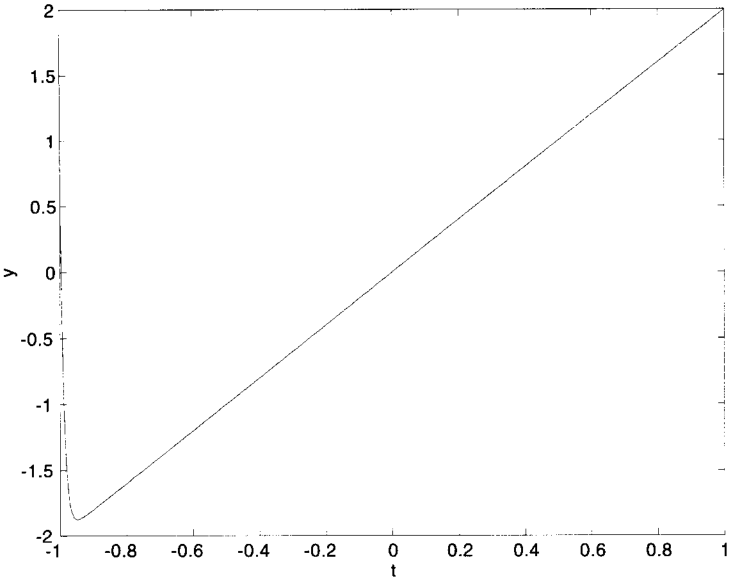

Fig. 4. Discrete solution for problem (13), $\varepsilon = 10^{-2}$, computed by COLSYS.

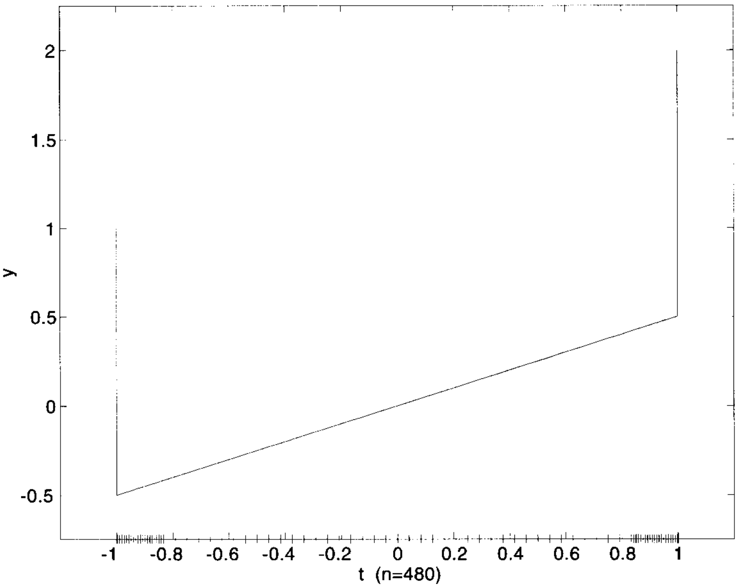

Fig. 5. Computed discrete solution for problem (13), $\varepsilon = 10^{-4}$.

thus confirming that this is a stiff problem. The same procedure, when $\varepsilon = 10^{-5}$, terminates with a mesh of 840 points, estimated error 10^{-4}, and estimated parameters

$$\kappa_{\mathrm{d}} \approx 10^5, \qquad \gamma_{\mathrm{d}} \approx 3,$$

that is, the problem becomes more stiff, as ε tends to zero.

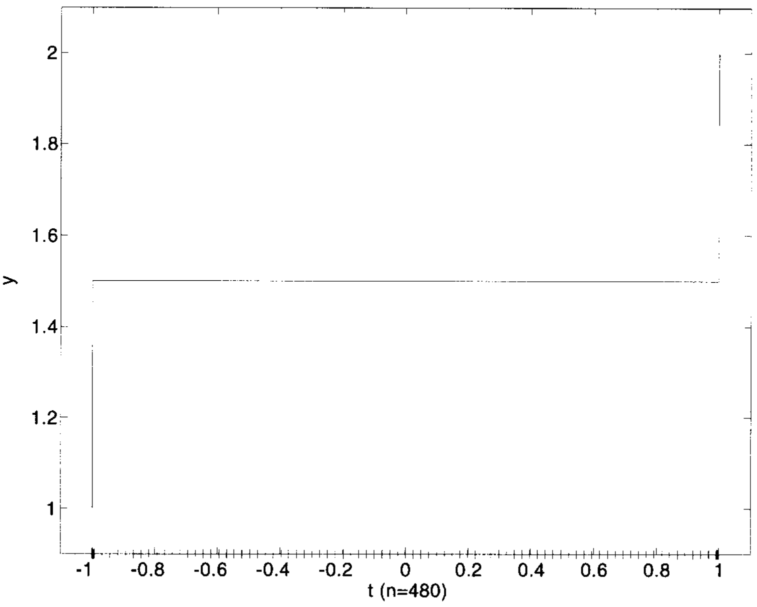

Fig. 6. Computed discrete solution for problem (14), $\varepsilon = 10^{-4}$.

Example 2. Consider now the problem,

$$\varepsilon y'' - 2ty' = 0, \qquad y(-1) = 1, \qquad y(1) = 2, \tag{14}$$

where $\varepsilon = 10^{-4}$. Even for this problem COLSYS, used with the same input parameters considered in the previous example, fails to provide the correct solution.

We obtain a final mesh of 480 points. The estimated error is 9×10^{-5} (see Fig. 6). Moreover, we have

$$\kappa_\mathrm{d}(\boldsymbol{h}^*) \approx 2 \times 10^4, \qquad \gamma_\mathrm{d}(\boldsymbol{h}^*) \approx 3,$$

showing that it is a stiff problem. In the case where $\varepsilon = 10^{-5}$, we obtain a final mesh of 760 points, an estimated error 2×10^{-5}, and estimated parameters

$$\kappa_\mathrm{d} \approx 2 \times 10^5, \qquad \gamma_\mathrm{d} \approx 3.$$

Example 3. Consider the following problem,

$$y'' = y + f_\varepsilon(t), \quad y(-1) = -y(1) = 1, \tag{15}$$

where the inhomogeneity is constructed such that the solution is given by (see Fig. 7)

$$y(t) = 1 + (t+1)\,\mathrm{erf}\big(-t\varepsilon^{-1/2}\big), \quad \varepsilon = 10^{-6}.$$

The function $f_\varepsilon(t)$ is quite smooth and has moderate size, except for a small neighborhood of $t = 0$, where it assumes values ranging from approximately -10^6 to 10^6. Such variation is responsible of the layer at $t = 0$ in Fig. 7.

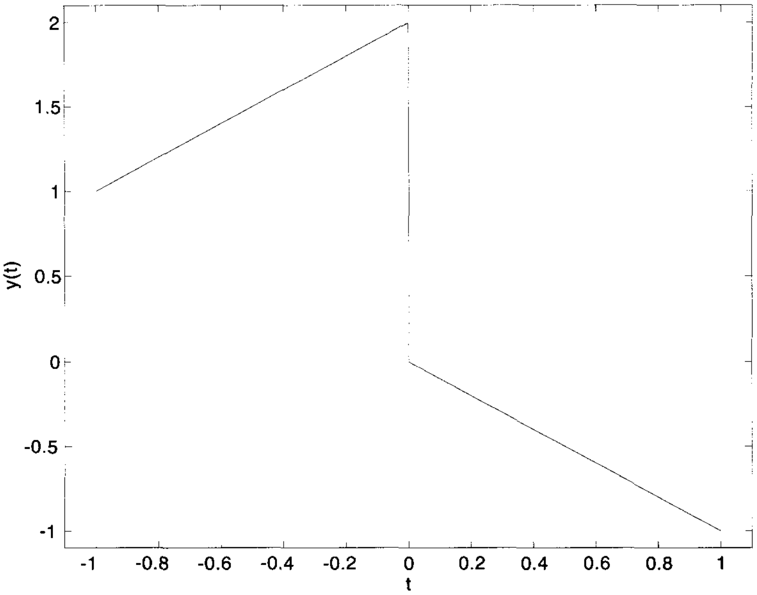

Fig. 7. Solution of problem (15), $\varepsilon = 10^{-6}$.

The first stage of the procedure ends producing an almost uniform mesh of 200 points, and estimated values

$$\kappa_{\mathrm{d}} \approx \gamma_{\mathrm{d}} \approx 2,$$

for the corresponding continuous parameters, thus confirming that the associated homogeneous problem is well conditioned. However, the computed solution (see Fig. 8) is far from the correct one. This drawback is recovered by the handling of the inhomogeneity, which produces a final mesh of 680 points and the discrete solution reported in Fig. 9.

Example 4. Consider the following problem,

$$\varepsilon y'' + t^2 y' + y = 0, \qquad y(-1) = 1, \qquad y(1) = 2, \tag{16}$$

where $\varepsilon = 10^{-4}$. This is a very hard to solve singular perturbation problem. In fact, its solution has a layer at $t = -1$, where, in a very short interval, it reaches a value $\approx 7.9 \times 10^9$ (see Fig. 10). Moreover, the solution heavily oscillates near $t = 0$ (see Fig. 11).

One obtains a final mesh of 3160 points, where the trapezoidal rule gives a maximum relative error of $\approx 2 \times 10^{-1}$. However, the estimated relative error on the solution of the sixth order TOM on the same mesh is $\approx 5 \times 10^{-7}$, with a maximum value 7.9151×10^9. The plots in Figs. 10 and 11 are relative to this solution. Moreover, we obtain the estimates

$$\kappa_{\mathrm{d}} \approx 4 \times 10^{13}, \qquad \gamma_{\mathrm{d}} \approx 5 \times 10^9.$$

One then concludes that the problem is both very ill conditioned and stiff.

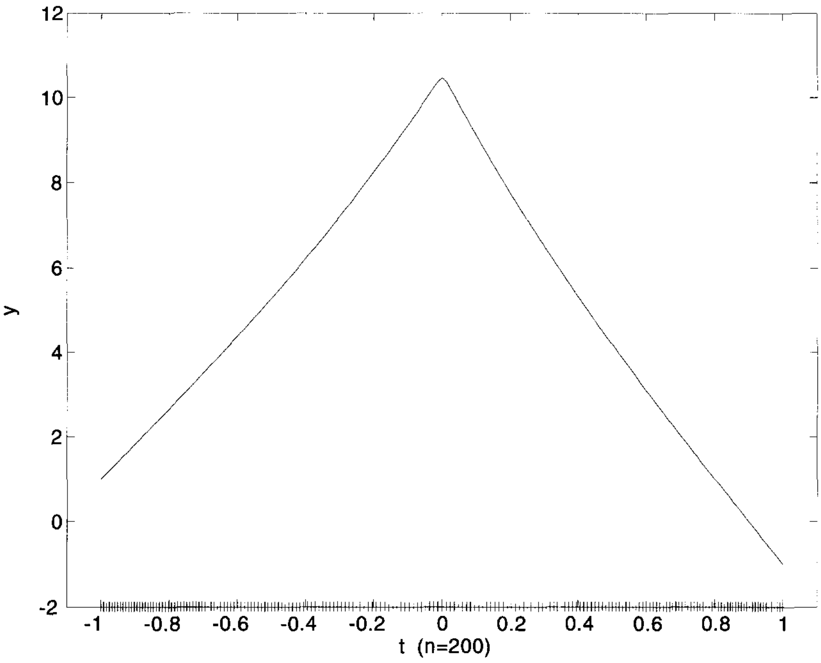

Fig. 8. Computed solution of problem (15), $\varepsilon = 10^{-6}$, before the handling of the inhomogeneity.

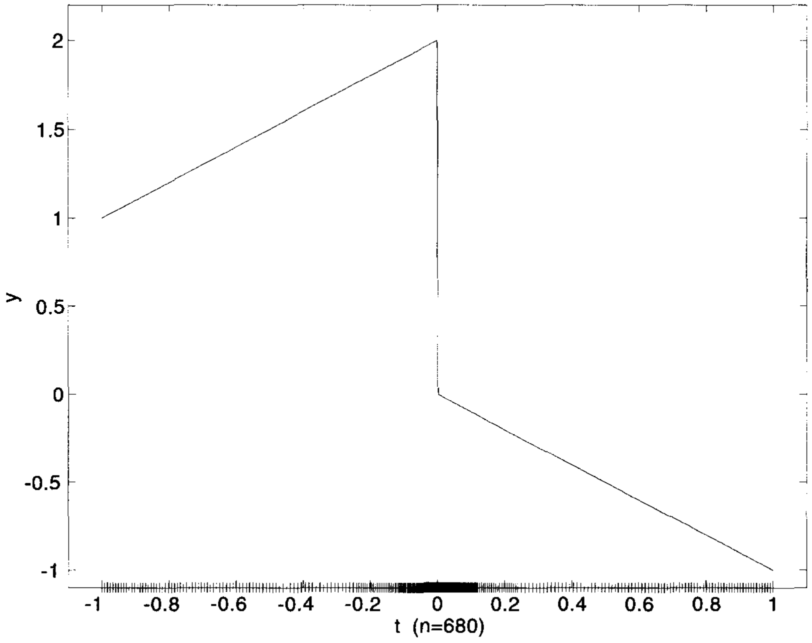

Fig. 9. Final computed solution of problem (15), $\varepsilon = 10^{-6}$.

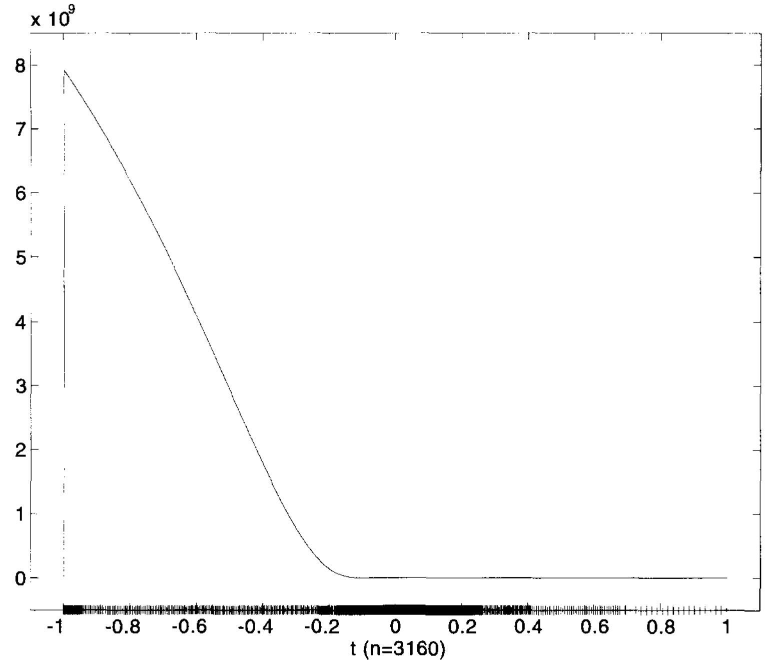

Fig. 10. Solution of problem (16).

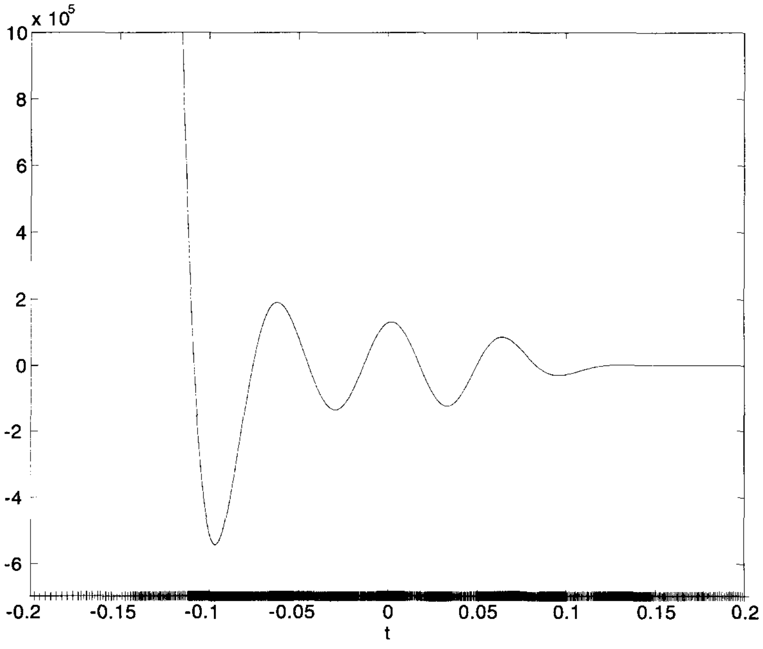

Fig. 11. Solution of problem (16), zoom at $t = 0$.

Example 5. Consider the following two-point BVP,

$$y'' + \frac{3\varepsilon}{(\varepsilon + t^2)^2} y = 0, \quad y(-0.1) = -y(0.1) = \frac{-0.1}{\sqrt{\varepsilon + 10^{-2}}}. \tag{17}$$

For $\varepsilon > 0$ and $\varepsilon \neq 10^{-2}$, this problem is well posed, and its solution is given by

$$y(t) = \frac{t}{\sqrt{\varepsilon + t^2}}.$$

However, for $\varepsilon = 10^{-2}$ the solution is not unique, so that the problem is ill posed. In fact, one verifies that for all $\alpha \in \mathbb{R}$

$$y_\alpha(t) = \frac{t}{\sqrt{\varepsilon + t^2}} + \alpha \frac{t^2 - \varepsilon}{\sqrt{\varepsilon + t^2}}$$

is a solution of problem (17).

If we apply the presented mesh selection with the trapezoidal rule to problem (17) with $\varepsilon = 10^{-2}$, the first stage of the procedure provides an almost uniform mesh of 560 points, and estimated discrete parameters

$$\kappa_{\mathrm{d}}(\boldsymbol{h}^*) \approx 1.5 \times 10^6, \qquad \gamma_{\mathrm{d}}(\boldsymbol{h}^*) \approx 1.2 \times 10^6. \tag{18}$$

One would then infer that this is an ill conditioned problem. In fact, the effect of the discretization is equivalent to consider a perturbed continuous problem. Since this perturbed problem is close to an ill posed one, an ill conditioned problem is then obtained.

However, the check of the approximations (18) by using the sixth order TOM on the same mesh provides

$$\kappa_{\mathrm{new}} \approx 8.5 \times 10^{13}, \qquad \gamma_{\mathrm{new}} \approx 6.7 \times 10^{13},$$

while the estimated maximum error on the discrete solution is approximately 10^{-2}. Since the new estimated parameters κ_{new} and γ_{new} are both much larger than $\kappa_{\mathrm{d}}(\boldsymbol{h}^*)$ and $\gamma_{\mathrm{d}}(\boldsymbol{h}^*)$, respectively, one may deduce that the continuous problem has κ_{c} and γ_{c} unbounded, that is it is ill posed.

Acknowledgements

The authors are pleased to thank the referees for their suggestions.

References

[1] P. Amodio, A-stable k-step linear multistep formulae of order $2k$ for the solution of stiff ODEs, Report 24/96 Dipartimento di Matematica, Università degli Studi di Bari, submitted.

[2] P. Amodio, W.L. Golik and F. Mazzia, Variable step boundary value methods based on reverse Adams schemes and their grid redistribution, *Appl. Numer. Math.* 18 (1995) 5–21.

[3] U.M. Ascher, R.M.M. Mattheij and R.D. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations* (Prentice-Hall, Englewood Cliffs, NJ, 1988).

[4] U. Ascher, J. Christiansen and R.D. Russell, A collocation solver for mixed order systems of boundary value problems, *Math. Comp.* 33 (1979) 659–679.

[5] L. Brugnano and D. Trigiante, High order multistep methods for boundary value problems, *Appl. Numer. Math.* 18 (1995) 79–94.

[6] L. Brugnano and D. Trigiante, Convergence and stability of boundary value methods for ordinary differential equations, *J. Comput. Appl. Math.* 66 (1996) 97–109.

[7] L. Brugnano and D. Trigiante, Block boundary value methods for linear Hamiltonian systems, *Appl. Math. Comput.* 81 (1997) 49–68.

[8] L. Brugnano and D. Trigiante, On the characterization of stiffness for ODEs, *Dynamics of Continuous, Discrete and Impulsive Systems* 2(3) (1996) 317–335.

[9] L. Brugnano and D. Trigiante, *Solving ODEs by Linear Multistep Formulae: Initial and Boundary Value Methods* (Gordon and Breach, to appear).

[10] J.R. Cash, A variable order deferred correction algorithm for the numerical solution of nonlinear two-point boundary value problems, *Comput. Math. Appl.* 9 (1983) 257–265.

[11] J.R. Cash, On the numerical integration of nonlinear two-point boundary value problems using iterated deferred corrections. Part 2: the development and analysis of highly stable deferred correction formulae, *SIAM J. Numer. Anal.* 25 (1988) 862–882.

[12] J.R. Cash and M.H. Wright, Implementation issues in solving nonlinear equations for two-point boundary value problems, *Computing* 45 (1990) 17–37.

[13] J.R. Cash and M.H. Wright, A deferred correction method for nonlinear two-point boundary value problems: Implementation and numerical evaluation, *SIAM J. Sci. Statist. Comput.* 12 (1991) 971–989.

[14] K. Chen, Error equidistribution and mesh selection, *SIAM J. Sci. Comput.* 15 (1994) 798–818.

[15] C. de Boor, Good approximation by splines with variable knots II, in: *Proceedings of Conference on the Numerical Solution of Differential Equations*, Lecture Notes in Mathematics (Springer, Berlin, 1973) 12–20.

[16] B. Kreiss and H.O. Kreiss, Numerical methods for singular perturbation problems, *SIAM J. Numer. Anal.* 18 (1981) 262–276.

[17] M. Lentini and V. Pereyra, A variable order finite difference method for nonlinear multipoint boundary value problems, *Math. Comp.* 28 (1974) 981–1003.

[18] M. Lentini and V. Pereyra, An adaptive finite difference solver for nonlinear two-point boundary value problems with mild boundary layers, *SIAM J. Numer. Anal.* 14 (1977) 91–111.

[19] V. Pereyra, Iterated deferred corrections for nonlinear operator equations, *Numer. Math.* 10 (1967) 316–323.

[20] V. Pereyra, Variable order variable step finite difference methods for nonlinear boundary value problems, in: *Proceedings of Conference on the Numerical Solution of Differential Equations*, Lecture Notes in Mathematics 363 (Springer, Berlin, 1973) 118–133.

[21] V. Pereyra, Difference solution of boundary value problems in ordinary differential equations, in: G.H. Golub, ed., *Studies in Numerical Analysis*, MAA Studies in Mathematics 24 (The Mathematical Association of America, 1984).

[22] V. Pereyra and E.G. Sewell, Mesh selection for discrete solution of boundary problems in ordinary differential equations, *Numer. Math.* 23 (1975) 261–268.

[23] R.D. Russell, Mesh selection methods, in: *Codes for Boundary Value Problems in Ordinary Differential Equations*, Lecture Notes in Computer Science 76 (Springer, New York, 1979).