Pergamon

Computers Math. Applic. Vol. 36, No. 1, pp. 21–32, 1998
© 1998 Elsevier Science Ltd. All rights reserved
Printed in Great Britain
0898-1221/98 $19.00 + 0.00

PII: S0898-1221(98)00105-9

Optimal Approximation of the Initial Value Problem

T. BENOUAZ* AND O. ARINO
Laboratoire de Mathématiques Appliquées
URA 1204 C.N.R.S, Avenue de l'Université
64000 Pau, France

(Received December 1997; accepted January 1998)

Abstract—We construct an optimal approximation procedure for the resolution of the initial value problem by using the optimal derivation. © 1998 Elsevier Science Ltd. All rights reserved.

Keywords—Initial value problem, Optimal derivative, Optimal approximation, Computational procedure.

1. INTRODUCTION

In [1–6], we presented a computational procedure defined as the optimal derivative of a nonlinear ODE. This procedure is a global approximation conceived initially to associate a linear equation to a nonlinear ODE in the neighborhood of a steady state, in particular when the nonlinearity is not smooth enough, near the steady state.

The aim of this paper is to construct an approximation of the solution of a nonlinear ODE based on the optimal derivative. This approximation will be called an "optimal approximation".

In fact, we are interested in the numerical resolution of the following initial value problem:

$$\begin{aligned} \frac{dx}{dt} &= F(x), \\ x(0) &= x_0, \end{aligned} \tag{1}$$

for $t \in [0,T]$, $x \in \mathbb{R}^n$. F is defined on an open subset Ω of $\mathbb{R}^n$ with values in $\mathbb{R}^n$.

We consider a subdivision $0 < t_1 < \cdots < t_i < t_{i+1} < \cdots < t_n = T$ of the interval $[0,T]$, and set $\tau_{i+1} = t_{i+1} - t_i$.

Many methods for solving this problem already exist (Euler, Runge-Kutta,...). The idea is to approach the nonlinear function with its successive derivatives.

What we propose here is to replace F by a linear map in the sense of the optimal derivative on each of the intervals of the subdivision. This permits the calculation of an approximation $\tilde{x}_i$ of the solution at each point t_i of the subdivision. An approximation $\tilde{x}$ of the solution is deduced by linear interpolation between t_i and t_{i+1}. This approximation does not use the value of the solution of the nonlinear equation, and hence, it is based on the computational value of the approximation of the solution using the optimal linear equation.

Now, we give a brief overview of the contents. The next two sections are devoted to preliminaries and a quick reminder of the optimal procedure. Then, the approximation procedure is

*Permanent address: Institut de Sciences Exactes, Bp. 119, Université de Tlemcen, 13000 Algérie.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

presented and an error estimate is obtained, first in the case when F is dissipative, then, in a general situation. Finally, we illustrate the applicability of the procedure through a simple example.

2. PRELIMINARIES

Starting with $\bar{x}(0) = \bar{x}_0$ an approximation of $x(0) = x_0$, we are in reality going to solve the following problem:

$$\begin{aligned} \frac{d\bar{x}}{dt} &= F(\bar{x}), \\ \bar{x}(0) &= \bar{x}_0. \end{aligned} \tag{2}$$

For this, we shall assume:

(H1) $\|F(x)\| \leq M$, $M > 0$,
(H2) F γ-Lipschitz continuous, and
(H3) there exists $M_2 < +\infty$ such that $\|F(y+x) - F(x) - DF(x)y\| \leq M_2\|y\|^2$, for all $x, y \in \mathbb{R}^n$.

Next, we intend to obtain the error introduced by this formulation. We start by changing the variable and function by centering F around $\bar{x}_0$. We set

$$\begin{aligned} \bar{x} &= \bar{y} + \bar{x}_0, \\ G(\bar{y}) &= F(\bar{x}_0 + \bar{y}) - F(\bar{x}_0), \\ b &= F(\bar{x}_0). \end{aligned} \tag{3}$$

Equation (2) yields

$$\begin{aligned} \frac{d\bar{y}}{dt} &= G(\bar{y}) + b, \\ \bar{y}(0) &= 0. \end{aligned} \tag{4}$$

Note that the solution of system (4) verifies the relation

$$\bar{x}(t) = \bar{y}(t) + \bar{x}_0, \tag{5}$$

where $\bar{x}(t)$ represents the solution of problem (2) and $\bar{y}(t)$ the solution of system (4). Applying the optimal derivative to system (4), G is replaced by a linear map $\tilde{A}$. We obtain the equation which defines the optimal problem

$$\begin{aligned} \frac{du}{dt} &= \tilde{A}u + b, \\ u(0) &= 0, \end{aligned} \tag{6}$$

and $u(t)$ is its solution.

3. OPTIMAL DERIVATIVE PROCEDURE IN AN INTERVAL $[\alpha, \beta]$

Let $[\alpha, \beta]$ be any interval of the real time, $x \in \mathbb{R}^n$, the function $G(\bar{y})$ is written as

$$G(\bar{y}) = F(\bar{y} + x) - F(x). \tag{7}$$

We will now briefly recall the procedure followed in the optimal derivative of G. We refer to [1,2] for more details.

One minimizes the functional

$$J(A) = \int_{\alpha}^{\beta} \|G(\bar{y}) - A\bar{y}(t)\|^2 \, dt \tag{8}$$

along a given solution. This gives

$$A = \left(\int_{\alpha}^{\beta} [G(\bar{y}(t))][\bar{y}(t)]^{\top} \, dt \right) \left(\int_{\alpha}^{\beta} [\bar{y}(t)][\bar{y}(t)]^{\top} \, dt \right)^{-1}. \tag{9}$$

This computation is then used iteratively on the interval $[\alpha, \beta]$ as follows.

FIRST STEP. **Compute the initial matrix A^0.** A^0 is the initial matrix: generally, one chooses for A^0 the Jacobian matrix of G at a point.

SECOND STEP. **Compute A^1** from the solution of the equation

$$\begin{aligned} \frac{d\bar{y}}{dt} &= (A^0)\bar{y} + b, \\ \bar{y}(0) &= 0, \end{aligned} \tag{10}$$

by minimizing the functional

$$J(A) = \int_{\alpha}^{\beta} \|G(\bar{y}(t)) - A\bar{y}(t)\|^2 \, dt, \tag{11}$$

$\bar{y}(t)$ being the solution of equation (10). A^1 is uniquely determined by formula (9).

THIRD STEP. Assuming that $A^1, \ldots, A^{(j-1)}$ have been computed, to compute $A^{(j)}$ from $A^{(j-1)}$, we first solve

$$\begin{aligned} \frac{d\bar{y}}{dt} &= \left(A^{(j-1)}\right)\bar{y} + b, \\ \bar{y}(0) &= 0. \end{aligned} \tag{12}$$

Let $\bar{y}_j(t)$ be the solution of equation (12). The minimization of the functional

$$J_j(A) = \int_{\alpha}^{\beta} \|G(\bar{y}_j(t)) - A\bar{y}_j(t)\|^2 \, dt \tag{13}$$

yields A^j.

In fact, we have the following relationships between A^{j-1} and A^j:

$$A^{(j)} = \left(\int_{\alpha}^{\beta} [G(\bar{y}_j(t))][\bar{y}_j(t)]^{\top} \, dt \right) \left(\int_{\alpha}^{\beta} [\bar{y}_j(t)][\bar{y}_j(t)]^{\top} \, dt \right)^{-1}, \tag{14}$$

where

$$G(\bar{y}_j) = F(\bar{y}_j + x) - F(x), \tag{15}$$

and

$$\begin{aligned} \bar{y}_j(\beta) &= \int_{\alpha}^{\beta} \exp\left[(\beta - s)A^{(j-1)}\right](b)\, ds, \\ b &= F(x). \end{aligned} \tag{16}$$

The limit of this sequence represents the optimal matrix $\tilde{A}(x, \theta)$, relative to the problem starting at x_0 and integrated on the interval $[\alpha, \beta]$ of length θ.

Solving the obtained linear system

$$\begin{aligned} \frac{du}{dt} &= \left(\tilde{A}(x, \theta)\right)u + b, \\ u(0) &= 0, \end{aligned} \tag{17}$$

with respect to u, gives

$$\begin{aligned} u(\beta) &= \int_{\alpha}^{\beta} \exp\left[(\beta - s)\tilde{A}(x, \theta)\right](b)\, ds, \\ b &= F(x). \end{aligned} \tag{18}$$

In this way, an approximation $u(\beta)$ of $\bar{y}(t)$ at point β is calculated, and

$$\tilde{y}_{\beta} = u(\beta).$$

4. OPTIMAL APPROXIMATION PROCEDURE IN THE INTERVAL $[0,T]$

We will now construct an optimal approximation on a whole interval $[0,T]$, using the optimal derivative presented above in each subinterval $[\alpha,\beta]$ of a suitable subdivision of $[0,T]$. For the time being, we consider an arbitrary subdivision $[t_i,t_{i+1}]$, $i=0,\dots,n$, $t_0=0$, $t_n=T$. We denote $\tau_{i+1}=t_{i+1}-t_i$.

In this subdivision, the nonlinear function G can be written

$$G(\bar{y})=F(\bar{y}+\bar{x}(t_i))-F(\bar{x}(t_i)). \tag{19}$$

Note that there does not exist a unique function G, but the functions obtained by centering F around the points $\bar{x}(t_i)$ on each of the subdivision $[t_i,t_{i+1}]$ of the interval $[0,T]$. We start with the initial value $\bar{x}(t_0)=\bar{x}_0=\tilde{x}_0$.

Algorithm

FIRST STEP. The solution of the optimal derivative problem on the interval $[t_0,t_1]$ from $\tilde{x}_0=\bar{x}_0$, departing from $A_0=DF(x_0)$, allows us to compute $\tilde{A}_1(\tilde{x}_0,\tau_1)$.

The solution of the corresponding linear system on the interval $[t_0,t_1]$

$$\begin{aligned}\frac{du}{dt}&=\left(\tilde{A}_1(\tilde{x}_0,\tau_1)\right)u+b_0,\\ u(0)&=0,\\ b_0&=F(\tilde{x}_0),\end{aligned} \tag{20}$$

gives an approximation $\tilde{y}_1=u(t_1)$ of $\bar{y}_1(t_1)$ and

$$\tilde{x}_1=\tilde{y}_1+\tilde{x}_0. \tag{21}$$

SECOND STEP. The solution of the optimal derivative problem on the interval $[t_1,t_2]$ from $\tilde{x}(t_1)=\tilde{x}_1$, departing from $\tilde{A}_1(\tilde{x}_0,\tau_1)$, allows us to compute $\tilde{A}_2(\tilde{x}_1,\tau_2)$.

The corresponding linear system can be written

$$\begin{aligned}\frac{du}{dt}&=\tilde{A}_2(\tilde{x}_1,\tau_2)u+b_1,\\ u(0)&=0,\\ b_1&=F(\tilde{x}_1).\end{aligned} \tag{22}$$

The solution of this system gives the value of the approximation $\tilde{y}_2=u(t_2)$, of $\bar{y}(t_2)$ and consequently

$$\tilde{x}_2=\tilde{y}_2+\tilde{x}_1. \tag{23}$$

THIRD STEP. Assuming that $\tilde{x}_1,\dots,\tilde{x}_i$ have been computed, to compute $\tilde{x}_{i+1}$ from $\tilde{x}_i$, we first solve the optimal derivative problem in the interval $[t_i,t_{i+1}]$.

We obtain the corresponding optimal matrix $\tilde{A}_{i+1}(\tilde{x}_i,\tau_{i+1})$ which defines a linear equation of the form

$$\begin{aligned}\frac{du}{dt}&=\tilde{A}_{i+1}(\tilde{x}_i,\tau_{i+1})u+b_i,\\ u(0)&=0,\\ b_i&=F(\tilde{x}_i),\end{aligned} \tag{24}$$

whose solution on the considered interval is

$$u(t_{i+1})=\int_{t_i}^{t_{i+1}}\exp\left[(t_{i+1}-s)\left(\tilde{A}_{i+1}\left(\tilde{x}_i,\tau_{i+1}\right)\right)\right](b_i)\,ds. \tag{25}$$

In this way, an approximate value $\tilde{y}_{i+1} = u(t_{i+1})$, of $\bar{y}(t_{i+1})$ is calculated and consequently, the general scheme of the optimal approximation solution of system (2) can be written

$$\begin{aligned} \tilde{x}_{i+1} &= \tilde{y}_{i+1} + \tilde{x}_i, \\ \tilde{x}_0 &= \bar{x}_0, \end{aligned} \qquad 0 \le i \le n. \tag{26}$$

Finally, the optimal approximation procedure permits us to construct a function $\tilde{x}(t)$ by recursive application of approximations found on each of the intervals $[t_i, t_{i+1}]$. $\tilde{x}(t)$ is defined by

$$\tilde{x}(t) = u_{i+1}(t) + \tilde{x}_i, \qquad \text{for } t_i \le t \le t_{i+1}, \quad 0 \le i \le n, \tag{27}$$

where u is the solution of equation (24), and $\tilde{x}$ is the optimal approximation of the solution x on $[0, T]$.

REMARK 1. Note that $\bar{x}$ and $\tilde{x}$ are in fact functions of the subdivision $\tau = (\tau_{i+1})$. This dependence may be emphasized by the notation $\bar{x}^{(\tau)}$, $\tilde{x}^{(\tau)}$, $|\tau| = \sup_i \tau_{i+1}$.

5. ERROR ESTIMATE

In this section, we are going to estimate the error introduced by the transformation of problem (1) into problem (2) and the error between the solution of problem (2) and the optimal approximation. We compute this error on the interval $[t_i, t_{i+1}]$, where $\tilde{A}_{i+1}(\tilde{x}_i, \tau_{i+1}) = \tilde{A}_{i+1}$ represents the optimal matrix calculated in Section 4. We will prove that under some conditions, the solution given by the optimal approximation converges to the solution of the nonlinear system in $L^1(0, T)$.

First, we consider the case when the function is dissipative on the open set containing the trajectory of the desired solution. That is to say, we assume that by selecting the canonical Euclidean norm in $\mathbb{R}^n$ and by denoting $\langle ., . \rangle$, the corresponding scalar product, there exists $R > 0$, $R > \|x_0\| + MT$, $\alpha > 0$, such that

$$\langle F(x) - F(y), x - y \rangle \le -\alpha \|x - y\|^2, \tag{28}$$

for all $x, y \in B_R = B(o, R)$.

The basic assumptions on F allow us to confirm that the desired solution remains inside a sphere with center o, and radius $\|x_0\| + MT$, such that $t \le T$.

LEMMA 2. *The matrix given by the relation (9) is bounded in the interval* $[t_i, t_{i+1}]$.

PROOF. In fact,

$$\begin{aligned} \|A\| &= \left\| \left(\int_{t_i}^{t_{i+1}} [G(\bar{y}(t))][\bar{y}(t)]^\top \, dt \right) \left(\int_{t_i}^{t_{i+1}} [\bar{y}(t)][\bar{y}(t)]^\top \, dt \right)^{-1} \right\| \\ &\le \left\| \int_{t_i}^{t_{i+1}} [G(\bar{y}(t))][\bar{y}(t)]^\top \, dt \right\| \left\| \left(\int_{t_i}^{t_{i+1}} [\bar{y}(t)][\bar{y}(t)]^\top \, dt \right)^{-1} \right\| \end{aligned} \tag{29}$$

in view of (H1) and (H2) in Section 2, and we have

$$\begin{aligned} &\int_{t_i}^{t_{i+1}} \|G(\bar{y}(t))\| \, \|[\bar{y}(t)]^\top\| \, dt \left(\int_{t_i}^{t_{i+1}} \|\bar{y}(t)\| \, \|[\bar{y}(t)]^\top\| \, dt \right)^{-1} \\ &\qquad \le \left(\int_{t_i}^{t_{i+1}} M \|\bar{y}(t)\|^2 \, dt \right) \left(\int_{t_i}^{t_{i+1}} \|\bar{y}(t)\|^2 \, dt \right)^{-1}. \end{aligned} \tag{30}$$

Finally

$$\|A\| \le M \frac{\int_{t_i}^{t_{i+1}} \|\bar{y}(t)\|^2 \, dt}{\int_{t_i}^{t_{i+1}} \|\bar{y}(t)\|^2 \, dt} \le M. \tag{31}$$ ∎

PROPOSITION 3. *With Lemma 2, under the assumptions (H1), (H2), and (H3) on F, and if F is α-dissipative for some $\alpha > 0$, then $\tilde{x}^{(\tau)}$ converges to $\bar{x}^{(\tau)}$ in $L^1(0,T)$, as the step size of the subdivision goes to zero.*

PROOF. We will first evaluate the error introduced by the optimal approximation on the interval $[t_i, t_{i+1}]$. We have

$$\begin{aligned}\left\|\bar{x}^{(\tau)}(t_{i+1}) - \tilde{x}^{(\tau)}(t_{i+1})\right\|^2 &= \int_{t_i}^{t_{i+1}} \frac{d}{ds}\left(\left\|\bar{x}^{(\tau)}(t) - \tilde{x}^{(\tau)}(t)\right\|^2\right) ds \\ &= 2\int_{t_i}^{t_{i+1}} \left\langle \bar{x}^{(\tau)}(s) - \tilde{x}^{(\tau)}(s), \dot{\bar{x}}(s) - \dot{\tilde{x}}(s)\right\rangle ds\end{aligned} \tag{32}$$

by denoting between $t_i \leq s \leq t_{i+1}$

$$\begin{aligned}\bar{x}^{(\tau)}(s) &= v_{i+1}(s) + \tilde{x}_i^{(\tau)}, \\ \tilde{x}^{(\tau)}(s) &= u_{i+1}(s) + \tilde{x}_i^{(\tau)},\end{aligned} \tag{33}$$

and

$$\begin{aligned}\dot{v}_{i+1}(s) &= G(v_{i+1}) + b_i, \\ \dot{v}_{i+1}(s) &= F\left(v_{i+1} + \tilde{x}_i^{(\tau)}\right), \\ \dot{u}_{i+1}(s) &= \left(\tilde{A}_{i+1}\left(\tilde{x_i}^{(\tau)}, \tau_{i+1}\right)\right) u_{i+1} + b_i.\end{aligned} \tag{34}$$

We obtain

$$\left\|\bar{x}^{(\tau)}(t_{i+1}) - \tilde{x}^{(\tau)}(t_{i+1})\right\| = 2\int_{t_i}^{t_{i+1}} \langle v_{i+1}(s) - u_{i+1}(s), \dot{v}_{i+1}(s) - \dot{u}_{i+1}(s)\rangle \, ds. \tag{35}$$

Now

$$\begin{aligned}\left\|\bar{x}^{(\tau)}(t_{i+1}) - \tilde{x}^{(\tau)}(t_{i+1})\right\| &= 2\int_{t_i}^{t_{i+1}} \left\langle v_{i+1} - u_{i+1}, G(v_{i+1}(s)) - \left(\tilde{A}_{i+1}\right) u_{i+1}(s)\right\rangle ds \\ &= 2\int_{t_i}^{t_{i+1}} \langle v_{i+1} - u_{i+1}, G(v_{i+1}(s)) - G(u_{i+1}(s))\rangle \, ds \\ &\quad + 2\int_{t_i}^{t_{i+1}} \left\langle v_{i+1} - u_{i+1}, G(u_{i+1}(s)) - \left(\tilde{A}_{i+1}\right) u_{i+1}(s)\right\rangle ds.\end{aligned} \tag{36}$$

The first integral is ≤ 0, since G is dissipative, so one can discard it:

$$\left\|\bar{x}^{(\tau)}(t_{i+1}) - \tilde{x}^{(\tau)}(t_{i+1})\right\| \leq 2\int_{t_i}^{t_{i+1}} \left\langle v_{i+1} - u_{i+1}, G(u_{i+1}(s)) - \left(\tilde{A}_{i+1}\right) u_{i+1}(s)\right\rangle ds. \tag{37}$$

Using the Cauchy-Schwartz inequality for the right-hand side,

$$\begin{aligned}&\left\|\bar{x}^{(\tau)}(t_{i+1}) - \tilde{x}^{(\tau)}(t_{i+1})\right\| \\ &\quad \leq 2 \sup_{t_i \leq s \leq t_{i+1}} \|v_{i+1} - u_{i+1}\| \left(\sqrt{\tau_{i+1}}\right) \left(\int_{t_i}^{t_{i+1}} \left\|G(u_{i+1}(s)) - \left(\tilde{A}_{i+1}\right) u_{i+1}(s)\right\|^2 ds\right)^{1/2},\end{aligned} \tag{38}$$

and

$$\sup_{t_i \leq s \leq t_{i+1}} \|v_{i+1} - u_{i+1}\| \leq 2\left(\sqrt{\tau_{i+1}}\right) \left(\int_{t_i}^{t_{i+1}} \left\|G(u_{i+1}(s)) - \left(\tilde{A}_{i+1}\right) u_{i+1}(s)\right\|^2 ds\right)^{1/2}, \tag{39}$$

we obtain

$$\left\|\bar{x}^{(\tau)}(t_{i+1}) - \tilde{x}^{(\tau)}(t_{i+1})\right\| \leq 2\sqrt{\tau_{i+1}} \left(\int_{t_i}^{t_{i+1}} \left\| G(u_{i+1}(s)) - \left(\tilde{A}_{i+1}\right) u_{i+1}(s) \right\|^2 ds \right)^{1/2}. \tag{40}$$

In the interval $[t_i, t_{i+1}]$, due to the minimizing property $\tilde{A}_{i+1}$ with respect to G, we have [2],

$$\begin{aligned} &\int_{t_i}^{t_{i+1}} \left\| G(u_{i+1}(s)) - \left(\tilde{A}_{i+1}\right) u_{i+1}(s) \right\|^2 ds \\ &\quad = \inf_{\forall A \in M_n(R), \mathrm{Re}(\sigma) \subset]-\infty,0[} \int_{t_i}^{t_{i+1}} \|G(u_{i+1}(s)) - A u_{i+1}(s)\|^2 \, ds. \end{aligned} \tag{41}$$

In particular, using $A = DG(0)$, we have

$$\int_{t_i}^{t_{i+1}} \left\| G(u_{i+1}(s)) - \left(\tilde{A}_{i+1}\right) u_{i+1}(s) \right\|^2 ds \leq \int_{t_i}^{t_{i+1}} \|G(u_{i+1}(s)) - DG(0) u_{i+1}(s)\|^2 \, ds, \tag{42}$$

which yields

$$\begin{aligned} \left\|\bar{x}^{(\tau)}(t_{i+1}) - \tilde{x}^{(\tau)}(t_{i+1})\right\| &\leq 2\sqrt{\tau_{i+1}} \left(\int_{t_i}^{t_{i+1}} \|G(u_{i+1}(s)) - DG(0) u_{i+1}(s)\|^2 \, ds \right)^{1/2} \\ &\leq 2\sqrt{\tau_{i+1}}\sqrt{\tau_{i+1}} \sup_{t_i \leq s \leq t_{i+1}} \|G(u_{i+1}(s)) - DG(0) u_{i+1}(s)\|. \end{aligned} \tag{43}$$

In view of (H3), we have

$$\|G(u_{i+1}(s)) - DG(0) u_{i+1}(s)\| \leq M_2 \|u_{i+1}(s)\|^2, \tag{44}$$

and

$$\left\|\bar{x}^{(\tau)}(t_{i+1}) - \tilde{x}^{(\tau)}(t_{i+1})\right\| \leq 2M_2(\tau_{i+1}) \sup_{t_i \leq s \leq t_{i+1}} \left[\|u_{i+1}(s)\|^2\right]. \tag{45}$$

Application of Gronwall's Lemma to the linear equation defining the optimal approximation in the interval $[t_i, t_{i+1}]$ yields

$$\begin{aligned} \frac{d}{dt}\|u_{i+1}(s)\| &\leq \left\|\tilde{A}_{i+1}\right\| \|u_{i+1}(s)\| + \|b_i\|, \\ \|u_{i+1}(0)\| &= 0, \\ b_i &= F\left(\tilde{x}_i^{(\tau)}\right), \end{aligned} \tag{46}$$

and

$$\sup_{t_i \leq s \leq t_{i+1}} \|u_{i+1}(s)\| \leq \|b_i\| \left(\frac{e^{\|\tilde{A}_{i+1}\|(t_{i+1}-t_i)} - 1}{\left\|\tilde{A}_{i+1}\right\|} \right). \tag{47}$$

In view of Lemma 2, and with τ_{i+1} sufficiently small, we obtain

$$\sup_{t_i \leq s \leq t_{i+1}} \|u_{i+1}(s)\| \leq 2\left(\|b_i\|\right) \tau_{i+1}. \tag{48}$$

This gives that

$$\sup_{t_i \leq s \leq t_{i+1}} \|u_{i+1}(s)\|^2 \leq 4\left(\|b_i\|\right)^2 (\tau_{i+1})^2. \tag{49}$$

Finally, denoting

$$k = 8 \sup_{0 \leq i \leq n} \|b_i\|^2 M_2 = 8 \sup_{\|x\| \leq \|x_0\| + MT} \|F(x)\|^2 M_2, \tag{50}$$

we have

$$\left\|\bar{x}^{(\tau)}(t_{i+1}) - \tilde{x}_{i+1}^{(\tau)}\right\| \leq k(\tau_{i+1})^3. \tag{51}$$

Multiplying by τ_{i+1}, we obtain

$$\tau_{i+1}\left\|\bar{x}^{(\tau)}(t_{i+1}) - \tilde{x}_{i+1}^{(\tau)}\right\| \leq k(\tau_{i+1})^4. \tag{52}$$

In the interval $[0,T]$, the error can be written as

$$\sum_{i=0}^{n} \tau_{i+1}\left\|\bar{x}^{(\tau)}(t_{i+1}) - \tilde{x}_{i+1}^{(\tau)}\right\| \leq k\sum_{i=0}^{n}(\tau_{i+1})^4. \tag{53}$$

With $\sum_{i=0}^{N}(\tau_{i+1})^4 \leq T(\sup_i \tau_{i+1})^3 = T|\tau|^3$, it holds

$$\sum_{i=0}^{n}\left\|\bar{x}^{(\tau)}(t_{i+1}) - \tilde{x}_{i+1}^{(\tau)}\right\| \leq k|\tau|^3 T. \tag{54}$$

Passing to the limit, we obtain

$$\lim_{|\tau|\longrightarrow 0}\int_0^T \left\|\bar{x}^{(\tau)}(s) - \tilde{x}^{(\tau)}(s)\right\| ds = 0. \tag{55}$$ ∎

We conclude that the solution $\tilde{x}$ computed from the optimal approximation converges to the solution of problem (2) in $L^1(0,T)$, as $\tau \longrightarrow 0$.

Now, we will evaluate the error between the problems (1) and (2).

LEMMA 4. *Under the assumptions (H1) and (H2) on the function F, and if F is α-dissipative for some $\alpha > 0$, then,*

$$\left\|x_{i+1} - \bar{x}^{(\tau)}(t_{i+1})\right\| \leq e^{-\alpha\tau_{i+1}}\left\|x_i - \tilde{x}_i^{(\tau)}\right\|. \tag{56}$$

PROOF. We have

$$\begin{aligned} 2\langle x - \bar{x}, \dot{x} - \dot{\bar{x}}\rangle &= 2\langle x(t) - \bar{x}(t), \dot{x}(t) - \dot{\bar{x}}(t)\rangle \\ &= 2\langle x(t) - \bar{x}(t), F(x(t)) - F(\bar{x}(t))\rangle. \end{aligned} \tag{57}$$

The dissipative character of F allows us to write

$$2\langle x - \bar{x}, \dot{x} - \dot{\bar{x}}\rangle \leq -\alpha\|x(t) - \bar{x}(t)\|^2, \tag{58}$$

and

$$\frac{d}{dt}\|x(t) - \bar{x}(t)\| \leq -\alpha\|x(t) - \bar{x}(t)\|. \tag{59}$$

By integrating between t_i and t, for $t_i \leq t \leq t_{i+1}$,

$$\|x(t) - \bar{x}(t)\| \leq e^{-\alpha(t-t_i)}\|x(t_i) - \bar{x}^{(\tau)}(t_i)\| \tag{60}$$

and setting $x(t_i) = x_i$ and $\bar{x}^{(\tau)}(t_i) = \tilde{x}_i^{(\tau)}$, we obtain

$$\|x(t) - \bar{x}(t)\| \leq e^{-\alpha(t-t_i)}\left\|x_i - \tilde{x}_i^{(\tau)}\right\|. \tag{61}$$

The relation (54) can be written as

$$\left\|x(t_{i+1}) - \bar{x}^{(\tau)}(t_{i+1})\right\| \leq e^{-\alpha(t_{i+1}-t_i)}\left\|x_i - \tilde{x}_i^{(\tau)}\right\| \tag{62}$$

and

$$\left\|x_{i+1} - \bar{x}^{(\tau)}(t_{i+1})\right\| \leq e^{-\alpha\tau_{i+1}} \left\|x_i - \tilde{x}_i^{(\tau)}\right\|. \tag{63}$$ ∎

PROPOSITION 5. *Under the same assumptions on the function F and under the assumptions of Lemma 4 and Proposition 3, then x converges to $\tilde{x}^{(\tau)}$ in $L^1(0,T)$, as the step size of the subdivision goes to zero.*

PROOF. The global error estimate can be written

$$\left\|x(t_{i+1}) - \tilde{x}^{(\tau)}(t_{i+1})\right\| \leq \left\|x(t_{i+1}) - \bar{x}^{(\tau)}(t_{i+1})\right\| + \left\|\bar{x}^{(\tau)}(t_{i+1}) - \tilde{x}^{(\tau)}(t_{i+1})\right\| \tag{64}$$

and

$$\left\|x_{i+1} - \tilde{x}_{i+1}^{(\tau)}\right\| \leq \left\|x_{i+1} - \bar{x}^{(\tau)}(t_{i+1})\right\| + \left\|\bar{x}^{(\tau)}(t_{i+1}) - \tilde{x}_{i+1}^{(\tau)}\right\|, \tag{65}$$

in view of the relations (51) and (63),

$$\begin{aligned}
\left\|x_{i+1} - \tilde{x}_{i+1}^{(\tau)}\right\| &\leq e^{-\alpha\tau_{i+1}} \left\|x_i - \tilde{x}_i^{(\tau)}\right\| + k(\tau_{i+1})^3 \\
&\leq e^{-\alpha\tau_{i+1}} \left(e^{-\alpha\tau_i} \left\|x_{i-1} - \tilde{x}_{i-1}^{(\tau)}\right\| + k(\tau_i)^3\right) + k(\tau_{i+1})^3 \\
&\leq e^{-\alpha(\tau_{i+1}+\tau_i)} \left\|x_{i-1} - \tilde{x}_{i-1}^{(\tau)}\right\| + ke^{-\alpha\tau_{i+1}}(\tau_i)^3 + k(\tau_{i+1})^3 \\
&\leq e^{-\alpha(\tau_{i+1}+\tau_i+\tau_{i-1})} \left\|x_{i-2} - \tilde{x}_{i-2}^{(\tau)}\right\| + ke^{-\alpha(\tau_{i+1}+\tau_i)}(\tau_{i-1})^3 + k(\tau_{i+1})^3.
\end{aligned} \tag{66}$$

Finally, for all steps we obtain

$$\begin{aligned}
\left\|x_{i+1} - \tilde{x}_{i+1}^{(\tau)}\right\| &\leq e^{-\alpha(\tau_{i+1}+\tau_i+\tau_{i-1}+\cdots+\tau_1)}\|x_0 - \tilde{x}_0\| \\
&\quad + k\sum_{j=1}^{i} e^{-\alpha(\tau_{i+1}+\tau_i+\ldots+\tau_{i-j})}(\tau_{i-1-j})^3 + k(\tau_{i+1})^3.
\end{aligned} \tag{67}$$

With $(\tau_{i+1})^3 \leq |\tau|^2\tau_{i+1}$ and

$$\sum_{j=1}^{i} e^{-\alpha(\tau_{i+1}+\tau_i+\cdots+\tau_{i-j})}(\tau_{i-1-j})^3 \leq \sum_{j=1}^{i}(\tau_{i-1-j})^3 \leq |\tau|^2 \sum_{j=1}^{i} \tau_{i-1-j}, \tag{68}$$

it holds

$$\begin{aligned}
\left\|x_{i+1} - \tilde{x}_{i+1}^{(\tau)}\right\| &\leq e^{-\alpha(\tau_{i+1}+\tau_i+\tau_{i-1}+\ldots+\tau_1)}\|x_0 - \tilde{x}_0\| + k|\tau|^2 \left(\tau_{i+1} + \sum_{j=1}^{i} \tau_{i-1-j}\right) \\
&\leq e^{-\alpha(\tau_{i+1}+\tau_i+\tau_{i-1}+\ldots+\tau_1)}\|x_0 - \tilde{x}_0\| + k|\tau|^2 T.
\end{aligned} \tag{69}$$

Multiplying by τ_i and in the interval $[0,T]$, the error can be written

$$\begin{aligned}
\sum_{i=0}^{n} \tau_{i+1} \left\|x_{i+1} - \tilde{x}_{i+1}^{(\tau)}\right\| &\leq \sum_{i=0}^{n}(\tau_{i+1})e^{-\alpha(\tau_{i+1}+\tau_i+\tau_{i-1}+\cdots+\tau_1)}\|x_0 - \tilde{x}_0\| + k\sum_{i=0}^{n}(\tau_{i+1})|\tau|^2 T \\
&\leq Te^{-\alpha T}\|x_0 - \tilde{x}_0\| + k|\tau|^3 T.
\end{aligned} \tag{70}$$

Passing to the limit, we obtain

$$\lim_{|\tau|\longrightarrow 0} \int_0^T \left\|x(s) - \tilde{x}^{(\tau)}(s)\right\| ds \leq Te^{-\alpha T}\|x_0 - \tilde{x}_0\|. \tag{71}$$

The global error is overestimated by the starting error on the initial conditions when we consider that $\bar{x}_0$ is an approximation of $x(0) = x_0$. Then, if we suppose this error is negligible, which is the case in general, we have

$$\lim_{|\tau|\longrightarrow 0} \int_0^T \left\|x(s) - \tilde{x}^{(\tau)}(s)\right\| ds = 0. \tag{72}$$ ∎

In this case, the solution $\tilde{x}^{(\tau)}$ of the optimal approximation converges to the solution x of the theoretical problem in $L^1(0,T)$, as $\tau \longrightarrow 0$.

Case When F is Not Dissipative

The above calculation is made under the assumption that the function F is dissipative. In the Lipschitz continuous case, one can always reduce to the dissipative case by the following change of variables:

$$x(t) = e^{\lambda t} z(t). \tag{73}$$

The initial equation can be written

$$\frac{dx}{dt} = \lambda e^{\lambda t} z(t) + e^{\lambda t} \frac{dz(t)}{dt} = F\left(e^{\lambda t} z(t)\right), \tag{74}$$

and

$$\begin{aligned} \frac{dz}{dt} &= e^{-\lambda t} F\left(e^{\lambda t} z(t)\right) - \lambda z(t) \\ &= H(t, z(t)). \end{aligned} \tag{75}$$

We obtain a new function $H(t, z)$ depending on time. For λ sufficiently large, $H(t, z)$ is dissipative on y uniformly with respect to time.

In the context in which we work, we subdivide the interval $[0, T]$ into a union of intervals $[t_i, t_{i+1}]$ in which we approximate H by a function independent of t. In what follows, we will evaluate the error of this approximation. It holds that

$$\begin{aligned} \|H(t, z) - H(t_i, z)\| &\leq \left|e^{-\lambda t} - e^{-\lambda t_i}\right| \left\|F\left(e^{\lambda t} z\right)\right\| + e^{-\lambda t} \left\|F\left(e^{\lambda t} z\right) - F\left(e^{\lambda t_i} z\right)\right\| \\ &\leq \lambda |t - t_i| M_0 + M_1 \lambda |t - t_i| \|z\| \end{aligned} \tag{76}$$

with $\lambda > M_1$. By denoting $\delta = |t - t_i|$, we obtain

$$\|H(t, z) - H(t_i, z)\| \leq \lambda \delta (M_0 + M_1 \|z\|). \tag{77}$$

$H(t_i, z)$ will be used as an approximation of $H(t, z)$ for $t \in [t_i, t_{i+1}]$.

This means that when we consider the function H not depending on the time in the interval $[t_i, t_{i+1}]$, we make an error characterized by the relation (77).

6. APPLICATION

In this section, we present numerical computations undertaken on an example, for comparison purposes. We consider the example introduced in [2, Example 11] as an illustration of the global least square approximation. We add two computations: first, a standard RK4 procedure, then the optimal approximation procedure presented in Section 4 are applied to the example. Comparisons of the two methods, on the one hand, and of the global and the local optimal methods, on the other hand, have been formulated in terms of the relative errors (Table 3).

EXAMPLE 6. Consider the following system:

$$\begin{aligned} \frac{dx}{dt} &= -x - \frac{2y}{\ln(x^2 + y^2)}, \\ \frac{dy}{dt} &= -y + \frac{2x}{\ln(x^2 + y^2)}, \end{aligned} \qquad (x_0, y_0) = (0, .5), \tag{78}$$

in the open unit disk $\{(x, y) \in \mathbb{R}^2;\ x^2 + y^2 < 1\}$.

The linearization of F at $(x_0, y_0) = (0, .5)$ gives

$$DF(x_0, y_0) = \begin{bmatrix} -1 & 3.524 \\ -1.4426 & -1 \end{bmatrix}, \qquad (x_0, y_0) = (0, .5). \tag{79}$$

After ten iterations, the least square approximation gives, at the level $\varepsilon = 10^{-6}$,

$$\tilde{A} = \begin{bmatrix} -1.4934 & 1.2489 \\ -0.5213 & -1.1254 \end{bmatrix}, \qquad (x_0, y_0) = (0, .5). \tag{80}$$

Results in Tables 1 and 2 represent the components of the solution of the nonlinear system (78) $(Xnl(t), Ynl(t))$, the components of the solution obtained by least square approximation (80) $(Xlin1(t), Ylin1(t))$, and the components of the solution obtained by optimal approximation procedure $(Xlin2(t), Ylin2(t))$.

The numerical data have been set at the following values: $t_0 = 0$, $T = 10$, step $= 0.1$, $\varepsilon = 10^{-4}$. We obtain the results shown in Tables 1 and 2.

Table 1.

t	$Xnl(t)$	$Xlin1(t)$	$Xlin2(t)$
0	0.0000000	$0.0000000E+00$	$0.0000000E+00$
1	$0.1432933E+00$	$0.1517739E+00$	$0.1432945E+00$
2	$0.6613008E-01$	$0.5793614E-01$	$0.6613086E-01$
3	$0.2476355E-01$	$0.1105308E-01$	$0.2476381E-01$
4	$0.8627967E-02$	$-0.3666260E-05$	$0.8628033E-02$
5	$0.2898245E-02$	$-0.8070500E-03$	$0.2898258E-02$
6	$0.9504736E-03$	$-0.3078052E-03$	$0.9504751E-02$
7	$0.3056629E-03$	$-0.5867217E-04$	$0.3056622E-03$
8	$0.9643728E-04$	$0.3895649E-07$	$0.9643653E-04$
9	$0.2978096E-04$	$0.4291440E-05$	$0.2978051E-04$
10	$0.8954184E-05$	$0.1635317E-05$	$0.8953958E-05$

Table 2.

t	$Ynl(t)$	$Ylin1(t)$	$Ylin2(t)$
0	$.5000000E+00$	$.5000000E+00$	$.5000000E+00$
1	$0.1153286E+00$	$0.1177924E+00$	$0.1153296E+00$
2	$0.1434192E-01$	$0.8519892E-02$	$0.1434179E-01$
3	$-0.2539680E-02$	$-0.5333531E-02$	$-0.2539879E-02$
4	$-0.3069739E-02$	$-0.2656959E-02$	$-0.3069846E-02$
5	$-0.1717583E-02$	$-0.6254748E-03$	$-0.1717630E-02$
6	$-0.7953907E-03$	$-0.4509668E-04$	$-0.7954108E-03$
7	$-0.3383072E-03$	$0.2837578E-04$	$-0.3383153E-03$
8	$-0.1372356E-03$	$0.1411885E-04$	$-0.1372386E-03$
9	$-0.5404247E-04$	$0.3321252E-05$	$-0.5404361E-04$
10	$-0.2085933E-04$	$0.2386980E-06$	$-0.2085973E-04$

In Table 3, we first give the relative error Er1 between the solution obtained by solving the nonlinear system (78), using RK4 procedure, and the solution calculated by the least square approximation. The column marked as Er2 gives the difference between the same RK4 solution, and the solution calculated using the procedure presented in Section 4.

7. COMMENTS

As a continuation of earlier work [1,2], we have presented here developments regarding the optimal derivative procedure. The emphasis is on the use of the optimal derivative as an optimal approximation method. This method allows us to solve numerically the initial value problem (1). Example 6 shows satisfactory adequacy of approximate results with respect, first, to the solution

Table 3.

t	Er1	Er2
0	0.	0.
1	0.048	$0.8E-05$
2	0.148	$1.1E-05$
3	0.562	$1.3E-05$
4	0.946	$1.3E-05$
5	1.14	$1.4E-05$
6	1.18	$1.6E-05$
7	1.06	$1.7E-05$
8	0.932	$1.8E-05$
9	1.01	$1.9E-05$
10	0.983	$2.E-05$

obtained by solving the nonlinear system (78), using the RK4 procedure, and second, with respect to the global optimal derivative presented in [2]. This is confirmed by the computation of the relative error, which permits us to see that the optimal approximation procedure presented in Section 4 is better than the global optimal derivative.

The proposed approach is fundamentally different from the existing methods, in the sense that we compute an approximation of the solution on each length by replacing the nonlinear equation with the corresponding optimal derivative in the considered interval. The error introduced by the optimal approximation is of order three with respect to the discretization length τ_i.

REFERENCES

1. T. Benouaz and O. Arino, Determination of the stability of a nonlinear ordinary differential equation by least square approximation. Computational procedure, *Appl. Math. and Comp. Sci.* **5** (1), 33–48 (1995).
2. T. Benouaz and O. Arino, Least square approximation of a nonlinear ordinary differential equation, *Computers Math. Applic.* **31** (8), 69–84 (1996).
3. T. Benouaz and O. Arino, Existence, unicité et convergence de l'approximation au sens des moindres carrés d'une équation différentielle ordinaire non-linéaire, Publications de l' U.A, CNRS 1204, No. 94/14, (1994).
4. T. Benouaz and O. Arino, Relation entre l'approximation optimale et la stabilité asymptotique, Publications de l' U.A, CNRS 1204, No. 95/10, (1995).
5. T. Benouaz, Least square approximation of a nonlinear ordinary differential equation: The scalar case, *Proceeding of the Fourth International Colloquium on Numerical Analysis*, Plovdiv, Bulgaria, August 13–17, 1995, (to appear).
6. T. Benouaz, Approximation of a nonlinear differential equation by an optimal procedure, *Proceeding of the 2^{nd} International Conference on Differential Equations*, Marrakech, June 16–20, 1995, (to appear).
7. M. Crouzeix and A.L. Mignot, *Analyse Numérique des Equations Différentielles*, Masson, (1984).
8. J.P. Demailly, *Analyse Numérique et Equations Différentielles*, Presses Universitaires de Grenoble, (1991).