Iranian Journal of Numerical Analysis and Optimization

Volume 6 , Number 2

Summer 2016

Ferdowsi University of Mashhad, Iran

In the Name of God

Iranian Journal of Numerical Analysis and Optimization (IJNAO)

This journal is authorized under the registration No. 174/853 dated 1386/2/26, by the Ministry of Culture and Islamic Guidance.

Volume 6, Number 2, Summer 2016

ISSN: 2423-6977

Publisher: Faculty of Mathematical Sciences, Ferdowsi University of Mashhad

Published by: Ferdowsi University of Mashhad Press

Circulation: 100

Address: Iranian Journal of Numerical Analysis and Optimization

Faculty of Mathematical Sciences, Ferdowsi University of Mashhad

P.O. Box 1159, Mashhad 91775, Iran.

Tel. : +98-51-38806222 , **Fax:** +98-51-38807358

E-mail: mjms@um.ac.ir

Website: http://ijnao.um.ac.ir

This journal is indexed by:

• Zentralblatt

• ISC

• SID

به اطلاع کلیه محققان، پژوهشگران، اساتید ارجمند، دانشجویان تحصیلات تکمیلی و نویسندگان محترم می رساند که نشریه ایرانی آنالیز عددی و بهینه سازی - IJNAO - طبق مجوز شماره ۳/۱۸/۵۴۸۹۱۳/ مورخه ۱۳۹۲/۱۰/۲۵ مدیر کل محترم سیاست گذاری و برنامه ریزی امور پژوهشی وزارت علوم، تحقیقات و فناوری، علمی – پژوهشی میباشد.
بر اساس نامه شماره ۹۲/۱۵۳۶/پ مورخه ۱۳۹۲/۱۱/۲۱ سرپرست محترم معاونت پژوهشی وفناوری پایگاه استنادی علوم جهان اسلام، نشریه ایرانی آنالیز عددی و بهینه سازی در پایگاه ISC نیز نمایه می شود.

Iranian Journal of Numerical Analysis and Optimization

Volume 6, Number 2, Summer 2016

Ferdowsi University of Mashhad - Iran

Iranian Journal of Numerical Analysis and Optimization

This journal is published under the auspices of Ferdowsi University of Mashhad

---

* Full Professor

** Associate Professor

We would like to acknowledge the help of Narjes khatoon Zohorian in the preparation of this issue.

Letter from the Editor in Chief

I would like to welcome you to the Iranian Journal of Numerical Analysis and Optimization (IJNAO). This journal is published biannually and supported by the Faculty of Mathematical Sciences at the Ferdowsi University of Mashhad. Faculty of Mathematical Sciences with three centers of excellence and three research centers is well-known in mathematical communities in Iran.

The main aim of the journal is to facilitate discussions and collaborations between specialists in applied mathematics, especially in the fields of numerical analysis and optimization, in the region and worldwide.

Our vision is that scholars from different applied mathematical research disciplines, pool their insight, knowledge and efforts by communicating via this international journal.

In order to assure high quality of the journal, each article is reviewed by subject-qualified referees.

Our expectations for IJNAO are as high as any well-known applied mathematical journal in the world. We trust that by publishing quality research and creative work, the possibility of more collaborations between researchers would be provided. We invite all applied mathematicians especially in the fields of numerical analysis and optimization to join us by submitting their original work to the Iranian Journal of Numerical Analysis and Optimization.

Mohammad Hadi Farahi

Contents

Iranian Journal of Numerical Analysis and Optimization
Vol. 6, No. 2, (2016), pp 1-15

Transcritical bifurcation of an immunosuppressive infection model

E. Shamsara, R. Mostolizadeh and Z. Afsharnezhad*

Abstract

In this paper, the dynamic behavior of an immunosuppressive infection model, specifically AIDS, is analyzed. We show through a simple mathematical model that a sigmoidal CTL response can lead to the occurrence of transcritical bifurcation. This condition usually occurs in immunodeficiency virus infections (such as AIDS infection) in which viruses attack immune cells $CD4^+T$. Our results imply that the dynamic interactions between the CTL immune response and HIV infection are very complex and in the CTL response, dynamics can exist the stable regions and unstable regions. At the end of the paper, numerical simulations are presented to illustrate the main results.

Keywords: CTL response; HAM/TSP; Transcritical bifurcation.

1 Introduction

One of the most complicated organs of higher organisms is the immune system. The function of the immune system is to fight off pathogenic organisms that enter and grow within the host (for example, viruses, bacteria, unicellular eukaryotic parasites such as malaria, and multicellular parasites such as worms). Immune responses can be subdivided broadly into two categories: (i) innate or nonspecific responses, and (ii) specific, adaptive responses. Innate immune mechanisms provide a first line of defense against an invading pathogen. They include physical barriers like the skin, changes in the environment of the body, such as fever, and immune cells that can fight pathogens

*Corresponding author
Received 23 January 2015; revised 7 November 2015; accepted 21 November 2015
E. Shamsara
Department of Applied Mathematics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Iran. email: elham.shamsara@stu.um.ac.ir

R. Mostolizadeh
Department of Applied Mathematics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Iran. email: re.mostolizadeh@stu.um.ac.ir

Z. Afsharnezhad
Department of Applied Mathematics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Iran. email: afsharnezhad@math.um.ac.ir

in a nonspecific way. Nonspecific is the key word here and means that these responses cannot specifically recognize the physical structure of the pathogen. Instead, these nonspecifics sense that an invader is present and react. While such responses slow down the initial growth of a pathogen, they are usually insufficient to clear an infection. For removing an infection, a specific and adaptive immune response is required [14]. The adaptive immune response consists of three main branches. 1. The B cells secret antibodies that neutralize free virus particles. 2. The CTL (also known as $CD4^+$ T cells) attack infected cells. 3. The $CD4^+$ T helper cells are very important regulators that ensure that CTL and B cell responses are developed efficiently. In immunosuppressive infection models, infected cells attack to $CD4^+$ T cells and infect them; subsequently, they cannot help CTL and $CD4^+$ T cells to act efficiently. Mathematical models have been of central importance for understanding the dynamics between viral infections and immune responses, particularly in the context of a human immunodeficiency virus (HIV) infection [7]. Significant emphasis has been placed on the viral side of these dynamics, including the estimation of basic viral parameters. Subsequent work has focused on the immune side of these interactions in trying to explain a variety of experimental observations about the dynamics of immune cells in various infections. One particular part of the immune system that is very important in the fight against viral infections is the killer T cells or cytotoxic T lymphocytes (CTL). They basically fight intracellular pathogens [15]. Clinical data have shown that for some human pathogens, such as HIV, hepatitis B virus (HBV) and hepatitis C virus (HCV), drug therapy sometimes is not completely effective [7,13]. Recently, in 2015 [2,12] and 2014 [9], impaired immune responses in immunosuppressive infection models have attracted more and more attention. Mathematical models have been developed to capture the interaction *in vivo* among HIV [2,4,5,9,12,14,15].

The following model is general and satisfied the clinical data, so it was pursued by scientists; see the above references. In 2003, this model was developed and considered [5].

$$\begin{cases} \dot{y} = yg_r(y) - yz \\ \dot{z} = zf(y). \end{cases} \tag{1}$$

In this system, y is the virus population and z is the population of the immune cells. The function $g_r(y)$ should be satisfied in:

$$\begin{cases} 1. g_r(0) > 0,\ \frac{\partial g_r}{\partial y} < 0\ \forall y \\ 2. \exists\, y^* > 0\ , g_r(y^*) > 0,\ \frac{\partial g_r(y)}{\partial y} > 0 \forall r, y. \end{cases} \tag{2}$$

Also the following conditions were assumed for $f(y)$ in [5]

$$\begin{cases} 3. \exists\, y_1,\ y_2 > 0\ such\ that\ \ f(y_1) = f(y_2) = 0 \\ 4.\ \frac{\partial f}{\partial y} > 0\ for\ \ y = y_1\ and\ \ \frac{\partial f}{\partial y} < 0\ for\ \ y = y_2 \end{cases} \tag{3}$$

$g_r(y)$ is the virus growth function that depends on the viral replication rate, r, and $f(y)$ is the immune expansion function that does not depend on r. In the above case, when viral replication is high and the virus load is between y_1 and y_2, immune expansion is increasing and levels of antigen are sufficient to trigger sustained immunity [5]. In [5], a special function for $g_r(y)$ and $f(y)$ is introduced (see Model (12)). Model (12), in 2015 [2,12], was considered to investigate the stability of the CTL immune response. Shu et al [9] in 2014 obtained saddle point for system (12) which shows stable and unstable. Note that all the above investigations were on the eigenvalues with *the non-zero real part* and they didn't consider the zero eigenvalue (bifurcation theory).
In this paper, we are interested in only one zero eigenvalue of the system (12) at the fixed point, which can lead to the occurance of transcritical bifurcation [1, 6, 8]. Since our concentration is on AIDS, we change the condition (3), in order to consider a weak immune system. The difference between HIV and AIDS is: HIV is the beginning of the AIDS disease, in AIDS; virus load rises more sharply, and the $CD4^+$ T cell (which defend against ADIS cells) drops sharply [14]. From a mathematical point of view, $\frac{\partial f}{\partial y} > 0$ means that the function f is a strictly increasing function with respect to the variable y. From a biological perspective, it means that the function of immune system responses to the disease increase. In this study the conditions for $f(y)$ are as follows:

$$\begin{cases} 1.\ \exists\, y^* > 0\,;\, f(y^*) = 0 \\ 2.\ \frac{\partial f}{\partial y} = 0\ for\ y^* > 0. \end{cases} \tag{4}$$

The new conditions cause a critical situation for the function f. For this case, we try to find a zero eigenvalue to apply transcritical bifurcation. Bifurcation theory helps us to obtain conditions for the parameters to keep the disease stable. In other words, by finding a region for parameter r with respect to parameter k, we tried to keep the immune system in proper condition as long as possible. Our work is organized as follows:

In Section 2, we give some preliminary definitions of bifurcation and theorems, which are going to be used in other Sections. Section 3 is devoted to bifurcation of system (12). Section 4 illustrates our numerical results. Section 5 is the conclusion.

2 Preliminaries

Bifurcation theory is fundamental for the qualitative study of dynamical systems, and can be used to reveal complex dynamical behaviors of the biological systems under study, such as bistability, recurrence, and regular oscillation. Characterized by a controllable parameter, called the bifurcation parameter,

bifurcation occurs at a critical value of this parameter where the properties of equilibria change significantly.

We consider bifurcations of equilibria of autonomous systems which depend on one single parameter μ:

$$\dot{x} = f(x, \mu) \ , x \in \mathbb{R}^n \ , \mu \in \mathbb{R}. \tag{5}$$

The system (5) is called smooth if $f(x, \mu)$ is differentiable up to any order in both x and μ. Equilibria of (5) are solutions of the algebraic equations

$$f(x, \mu) = 0. \tag{6}$$

In order to graphically illustrate the dependence of an equilibrium x on μ, we require a scalar measure of the n-vector x. We shall use the notation $[x]$ for such a measure of x. A diagram depicting $[x]$ versus μ, where (x, μ) solves equation (1), will be called a bifurcation diagram. The continuous curves of solutions of (1) under variation of μ are called branches. The branches of smooth systems are continuous and smooth but can split into more branches. On a regular point of a branch, that is, on a point where the branch does not split or turn around, we can define the slope of the branch. We will use the following abbreviations:

$$J(x, \mu) := \frac{\partial f(x, \mu)}{\partial x}, \quad f_\mu := \frac{\partial f(x, \mu)}{\partial \mu}. \tag{7}$$

Both derivatives exist for a smooth system. Using the implicit function theorem it follows that, provided that the Jacobian matrix $J(x, \mu)$ is non-singular, locally (1) is equivalent to writing x as a function of μ, i.e, $0 = f(x(\mu), \mu)$. Then it follows from differentiating (1) with respect to μ that

$$J(x, \mu)\frac{dx}{d\mu} + f_\mu(x, \mu) = 0. \tag{8}$$

As $J(x, \mu)$ is non-singular, we can solve for $\frac{dx}{d\mu}$. A point (x, μ) is called regular if $det(J(x, \mu)) \neq 0$.

Definition 1 (Bifurcation). The appearance of a topologically nonequivalent phase portrait under a variation of parameters is called a bifurcation [1, 6, 8, 13, 25].

Definition 2. Transcritical bifurcation is a particular kind of local bifurcation, meaning that it is characterized by an equilibrium having an eigenvalue whose real part passes through zero.

A transcritical bifurcation is one in which a fixed point exists for all values of a parameter and is never destroyed. However, such a fixed point interchanges its stability region with instability region as the parameter is

varied. In other words, both before and after the bifurcation, there is one unstable and one stable fixed point [1, 6, 8, 13].

Theorem 1. *(Sotomayor Theorem) Suppose that* $f_{\mu_0}(x_0) = 0$ *and that* $n \times n$ *matrix* $A = Df(x_0, \mu_0)$ *has a simple eigenvalue* $\lambda = 0$ *with eigenvector* ν *and that* A^T *has an eigenvector* ω *corresponding to the eigenvalue* $\lambda = 0$. *Furthermore, suppose that* A *has* k *eigenvalues with a negative real part and* $(n - k - 1)$ *eigenvalues with a positive real part and that the following conditions are satisfied:*

$$\omega^T f_\mu(x_0, \mu_0) \neq 0 \ , \ \omega^T [D^2 f(x_0, \mu_0(\nu, \nu))] \neq 0 \tag{9}$$

then there is a smooth curve of equilibrium points for $\dot{x} = f(x, \mu)$ *in* $\mathbb{R}^n \times \mathbb{R}$*passing through* (x_0, μ_0) *and tangent to the hyperplane* $\mathbb{R}^n \times \mu_0$. *Depending on the signs of the expressions in (6). In this case the system experiences a saddle node bifurcation. If the conditions (9) are changed to :*

$$\begin{array}{l} \omega^T f_\mu(x_0, \mu_0) = 0, \\ \omega^T [D f_\mu(x_0, \mu_0)] \neq 0, \\ \omega^T [D^2 f(x_0, \mu_0)(\nu, \nu)] \neq 0, \end{array} \tag{10}$$

then the system (5) experiences a Pitchfork bifurcation. And if the condition (9) changed to:

$$\begin{array}{ll} \omega^T f_\mu(x_0, \mu_0) = 0, & \omega^T [D f_\mu(x_0, \mu_0)\nu] \neq 0, \\ \omega^T [D^2 f(x_0, \mu_0)(\nu, \nu)] = 0, & \omega^T [D^3 f(x_0, \mu_0)] \neq 0, \end{array} \tag{11}$$

then the system (5) experiences a Transcritical bifurcation.

Proof. For the proof, one can see [8]. □

3 Bifurcation of the system (12)

Consider the following system of differential equations:

$$\begin{cases} \dot{y} = ry(1 - \frac{y}{k}) - ay - pyz \\ \dot{z} = \frac{czy}{1+dy} - qyz - bz \end{cases} \tag{12}$$

where y and z are as before. The virus population is assumed to grow logistically: r is the viral replication rate at low viral loads, and we assume that this rate is decreased linearly with increased viral load to reach zero at a viral load k. Immune cells are assumed to be inhibited by the virus at a rate qyz and die at a rate b.
Clearly $E_0 = (0, 0)$ is a trivial equilibrium of the system. There exist an equilibrium $E_1 = (\bar{y}, 0) = (\frac{k}{r}(r - a), 0)$ provided $r > a > 0$.

The equilibrium E_1 is called the virus dominant equilibrium (VDE). Moreover, we can find another equilibrium $E^* = (y^*, z^*)$, where $y^* > 0$ and $z^* > 0$, satisfying the following equations:

$$\begin{cases} r(1 - \frac{y^*}{k}) - a - pz^* = 0 \\ \frac{cy^*}{1+dy^*} - qy^* - b = 0 \end{cases} \tag{13}$$

$E^* > 0$ means that while the virus population is growing, immune cells start to increase; therefore, our main attention will be on equilibrium E^*. It follows from the first equation of (13)

$$z^* = \frac{r(k - y^*) - ak}{pk} > 0 \tag{14}$$

By $z^* > 0$, one can find $\bar{y}$ such that

$$y^* < \bar{y} \tag{15}$$

In order to find y^* for E^*, we should solve the quadratic equation

$$h(y) = qdy^2 + (-c + q + bd)y + b = 0, \quad y^* < \bar{y} \tag{16}$$

to obtain a double root for (16), one should have

$$\Delta = 0 \Rightarrow (c - q - bd)^2 = 4bqd \Rightarrow c - q - bd = \pm 2\sqrt{bqd} \tag{17}$$

The minus sign for the root is not applicable, so

$$c - q - bd = 2\sqrt{bqd} \tag{18}$$

or equivalently $c = (\sqrt{q} + \sqrt{bd})^2$.

Conditions (17) and (18) on polynomial (16) lead to

$$g(y) = (y - \frac{c - q - bd}{2qd})^2 = (y - \frac{2\sqrt{bqd}}{2qd})^2 = (y - \sqrt{\frac{b}{qd}})^2 \tag{19}$$

Consequently,

$$y^* = \sqrt{\frac{b}{qd}} \tag{20}$$

and

$$E^* = (\sqrt{\frac{b}{qd}}, \frac{r(k - \sqrt{\frac{b}{qd}}) - ak}{pk}) \tag{21}$$

Because $y^* < \bar{y}$, we can define a threshold (see the following definition) as follow:

$$rk - ry^* > ak \Rightarrow r(k - y^*) > ak \Rightarrow r > \frac{ak}{k - y^*} \tag{22}$$

$$\Rightarrow r_t = \begin{cases} \frac{ak}{k-y^*} & \text{if } y^* < k \\ \infty & \text{if } y^* > k \end{cases} \tag{23}$$

Definition 3. In mathematical or statistical modeling, a threshold model is any model where a threshold value, or set of threshold values, is used to distinguish ranges of values where the behavior predicted by the model varies in some important way.
With the above statements, one can have the following lemma:

Lemma 1.*Suppose that (18) is satisfied.*
(a) If $r \leq a$, then the trivial equilibrium $E_0 = (0, 0)$ is the only equilibrium.
(b) If $a < r \leq r_t$ (i.e $a < r$ and $y^* \geq \bar{y}$), then there are two equilibria E_0 and $E_1 = (\bar{y}, 0)$, where $\bar{y} = \frac{k}{r}(r - a)$
(c) If $r_t < r$ (i.e $a < r$ and $y^* < \bar{y}$), then there are three equilibria, E_0, E_1 and additional equilibrium $E^* = (y^*, z^*)$ with $z^* = \frac{r(k-y^*)-ak}{pk}$.

We call E^* the immune control equilibrium(ICE).
Here we would like to determine the type of the equilibria (E_0, E_1 and E^*) for the system (12).

3.1 Global dynamics of (12)

Let (y^*, z^*) be an equilibrium of (12). The associated characteristic equation of (12) is given by

$$g_0(\lambda) = \lambda^2 + c_1\lambda + c_0 = 0 \tag{24}$$

where

$$c_1 = -(r - \frac{2r}{k}y^* - a - pz^* + \frac{cy^*}{1 + dy^*} - qy^* - b) \tag{25}$$

and

$$c_0 = (r - \frac{2r}{k}y^* - a - pz^*)(\frac{cy^*}{1+dy^*} - qy^* - b) + py^*(\frac{cz^*}{(1+dy^*)^2} - qz^*) \tag{26}$$

At $E_0 = (0, 0)$, two roots of the characteristic equation are $\lambda_1 = -b < 0$ and $\lambda_2 = -(a - r)$. Therefore, E_0 is stable if $r \leq a$. Otherwise, if $r > a$, then E_0 is a saddle point.
At E_1, $y^* = \bar{y}$, $z^* = 0$, a direct calculation implies that E_1 is stable.
At E^* we have

$$c_1 = \frac{ry^*}{k} > 0 \tag{27}$$

and

$$c_0 = py^*z^*(-q + \frac{c}{(1+dy^*)^2}) = py^*z^*g_1(y). \tag{28}$$

If $g_1(y) = 0$, then $\tilde{y} = \frac{\sqrt{c}-\sqrt{q}}{d\sqrt{q}}$. Substituting $\tilde{y}$ in $g(y)$ where

$$g(y) = (y - \frac{c - q - bd}{2qd})^2 \tag{29}$$

we have

$$\begin{aligned} g(\tilde{y}) &= (\frac{\sqrt{c} - \sqrt{q}}{d\sqrt{q}} - \frac{c - q - bd}{2qd})^2 \\ &= (\frac{\sqrt{q} + \sqrt{bd} - \sqrt{q} - \sqrt{bd}}{d\sqrt{q}})^2 = 0 \end{aligned} \tag{30}$$

$\tilde{y} = y^*$, therefore $c_0 = 0$
In this case since $\lambda_1 + \lambda_2 = -c_1 < 0$, $\lambda_1\lambda_2 = 0$ which gives us

$$\lambda_1 = 0 \tag{31}$$

and

$$\lambda_2 = -c_1 = -\frac{r}{k}\sqrt{\frac{b}{qd}} \tag{32}$$

The system (12) under condition (18) for the equilibrium E^* has one negative eigenvalue and one zero eigenvalue. Next we check the conditions for the

Transcritical bifurcation. For this purpose, we use Sotomayor theorem (see Theorem 2.3). In the following, we calculate the Jacobian matrix, second derivative of the Jacobian matrix and also eigenvector ν corresponding to eigenvalue $\lambda_1 = 0$ for A and ω,the eigenvector of $\lambda_1 = 0$, corresponding to A^T.
The Jacobian matrix of (12) is

$$A = \begin{bmatrix} r - \frac{2r}{k}y - a - pz & -py \\ \frac{cz}{(1+dy)^2} - qz & \frac{cy}{1+dy} - qy - b \end{bmatrix} \tag{33}$$

where condition (18) implies that

$$\frac{cz}{(1+dy)^2} - qz = \frac{cy}{1+dy} - qy - b = 0 \tag{34}$$

Therefore,

$$A = \begin{bmatrix} r - \frac{2r}{k}y - a - pz & -py \\ 0 & 0 \end{bmatrix} \tag{35}$$

The Jacobian matrix A at E^* will be

$$A_{E^*} = \begin{bmatrix} -\frac{r}{k}\sqrt{\frac{b}{qd}} & -p\sqrt{\frac{b}{qd}} \\ 0 & 0 \end{bmatrix} \tag{36}$$

By a direct calculation, the eigenvectors ν and ω are

$$\nu = (\nu_1, \nu_2) = (-\frac{kp}{r}, 1) \tag{37}$$

$$\omega = (\omega_1, \omega_2) = (0, 1) \tag{38}$$

$$D^2 f(E^*)(\nu,\nu) = \begin{bmatrix} \frac{\partial^2 f_1(E^*)}{\partial y^2}\nu_1\nu_1 + \frac{\partial^2 f_1(E^*)}{\partial y \partial z}\nu_1\nu_2 + \frac{\partial^2 f_1(E^*)}{\partial y \partial z}\nu_2\nu_1 + \frac{\partial^2 f_1(E^*)}{\partial z^2}\nu_2\nu_2 \\ \frac{\partial^2 f_2(E^*)}{\partial y^2}\nu_1\nu_1 + \frac{\partial^2 f_2(E^*)}{\partial y \partial z}\nu_1\nu_2 + \frac{\partial^2 f_2(E^*)}{\partial y \partial z}\nu_2\nu_1 + \frac{\partial^2 f_2(E^*)}{\partial z^2}\nu_2\nu_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \sigma \end{bmatrix} \tag{39}$$

If $\sigma \neq 0$ implies that $D^2 f(E^*)(\nu,\nu) \neq 0$. Also, one should have

$$f_r(E^*) = \begin{pmatrix} \sqrt{\frac{b}{qd}}(1 - \frac{\sqrt{\frac{b}{qd}}}{k}) \\ 0 \end{pmatrix} \tag{40}$$

From (36) and (38), one can obtain

$$w^T f_r(E^*) = 0 \tag{41}$$

The above calculations and results lead to the conclusion that conditions (8) are valid. *Therefore, by the Sotomayor theorem, the system (10) undergoes transcritical bifurcation.*

4 Example (numerical simulation)

The parameters data are choosen such that the Figures 1-5 are in consistent with [2,4,5,9,12,14,15]. Since we are dealing with AIDS, the following Figures show the regions of weak immune response.

We try to find a region for parameter r with respect to parameter k. From (19), $r > \frac{ak}{k-y^*}$, but $y^* = \sqrt{\frac{b}{qd}}$, therefore

$$r > \frac{ak}{k - \sqrt{\frac{b}{qd}}} > 0 \tag{42}$$

so

$$k > \sqrt{\frac{b}{qd}} \tag{43}$$

Thus, the parameter region is obtained in Figure 1:

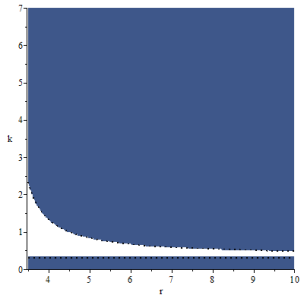

Figure 1: Parameter region r with respect to k by considering $a = 3$, $b = 2$, $q = 9$ and $d = 2$

We use numerical techniques to determine the system (12) with condition (18).

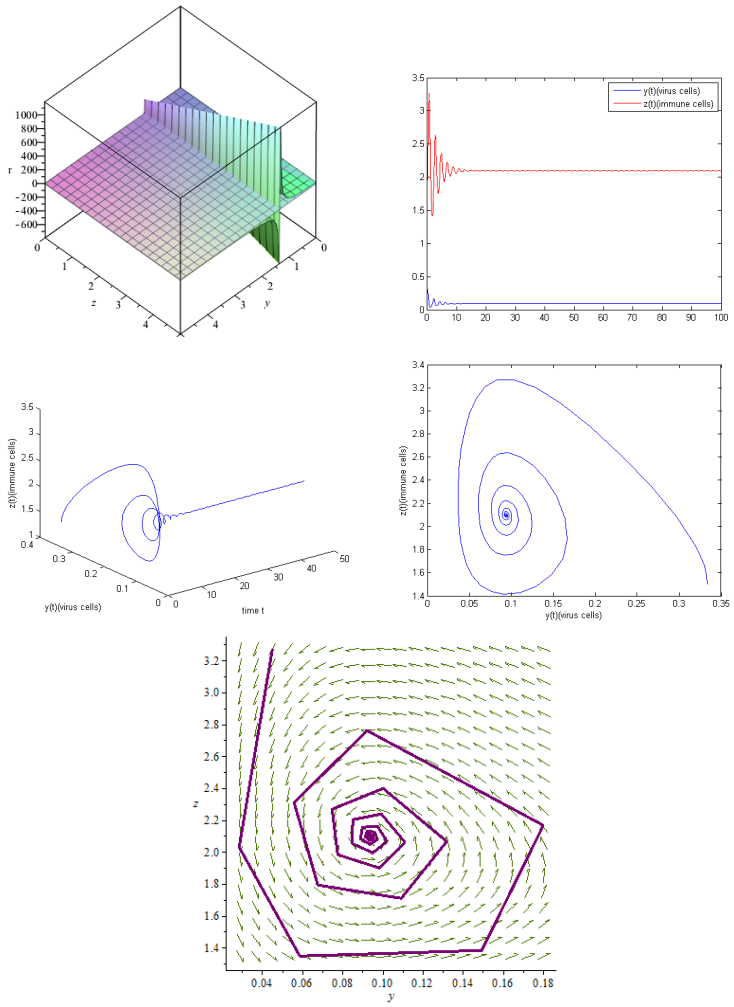

Figure 2: $a = p = 3$, $k = 4/3$, $q = 9$, $b = 2$ and $r = 10$ with initial condition $(\frac{1}{3}, \frac{3}{2})$

In Figure 2, first we obtain the parameter r with respect to y and z. Next by considering the values $a = p = 3$, $k = 4/3$, $q = 9$, $b = 2$ and $r = 10$, the stability regions of the orbits are investigated. Therefore, system (12) is in a steady state; this means that however the immune response of the body is so weak that is still can defend against the disease.
Figure 3 shows that after 100 days, immune cells could not control the growth of virus cells and so the system (12) is unstable.

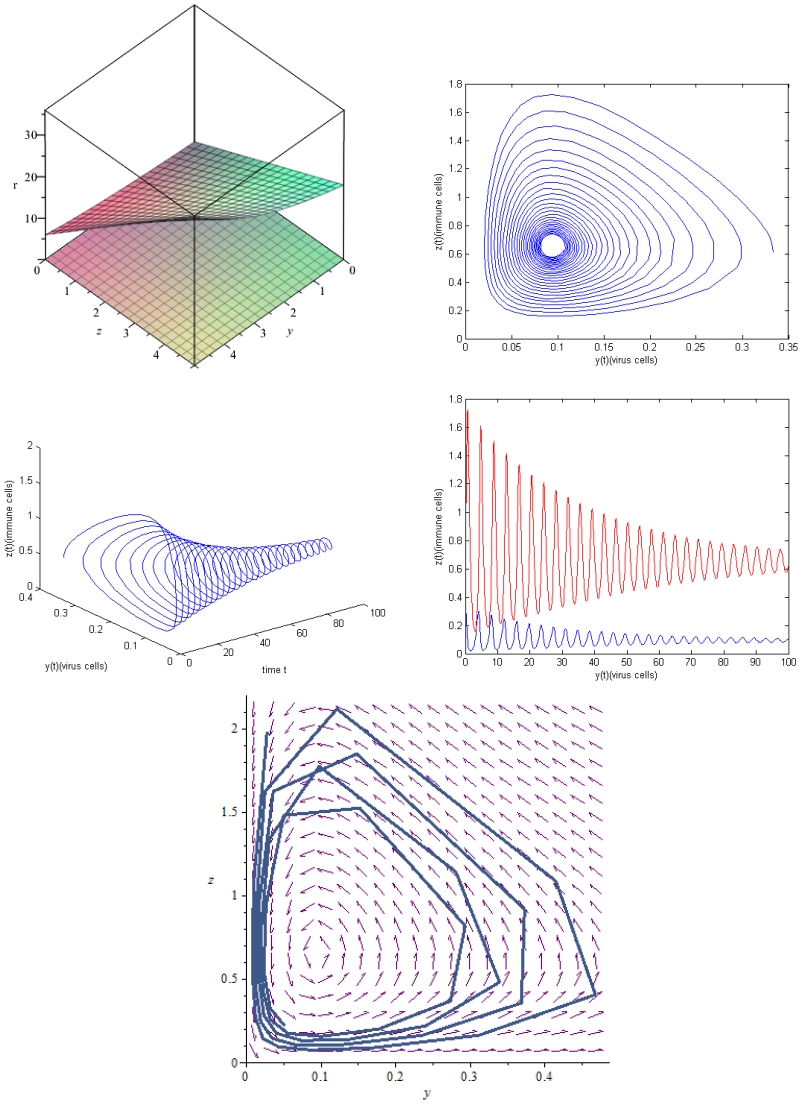

Figure 3: $a = p = 3$, $k = 10$, $q = 9$, $b = 2$ and $r = 5$ with initial condition $(\frac{1}{3}, \frac{2}{3})$

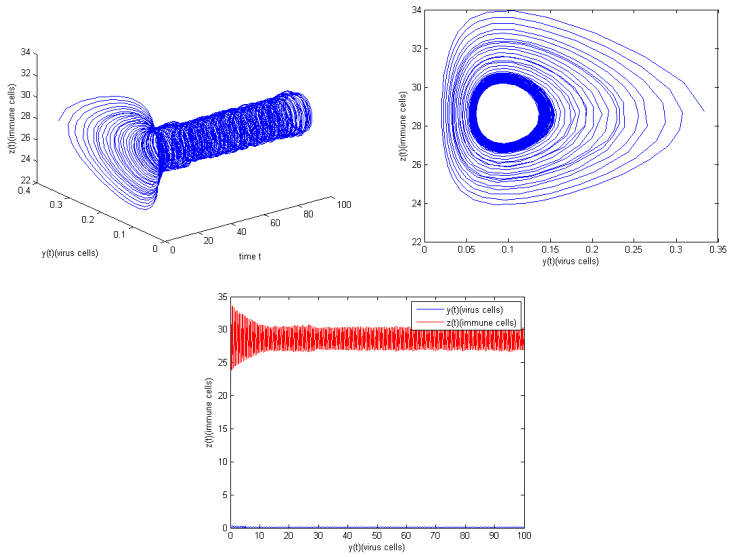

Figure 4: $a = 4$, $p = 3$, $k = 40$, $q = 9$, $b = d = 2$, $c = 36$ and $r = 5$ with initial condition $(\frac{1}{3}, \frac{115}{4})$

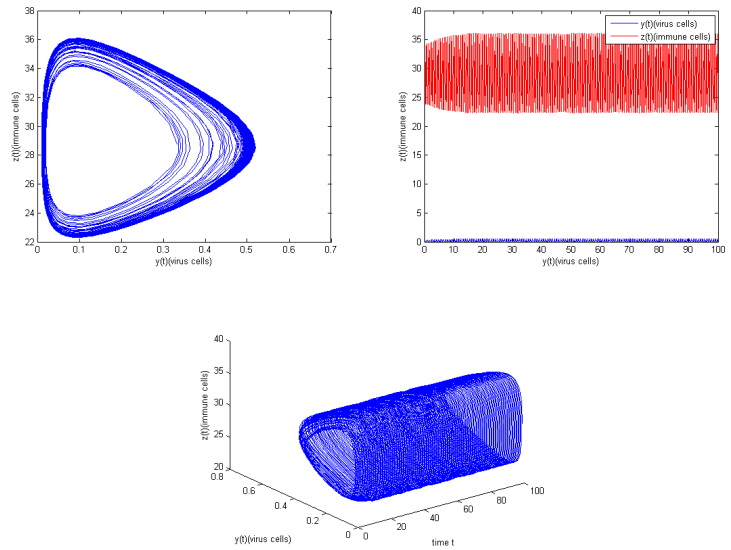

Figure 5: $a = 4$, $p = 3$, $k = 400$, $q = 9$, $b = d = 2$, $c = 36$ and $r = 5$ with initial condition $(\frac{1}{3}, \frac{115}{4})$

Results Figure 1 shows the region which one can choose r with respect to k. In Figure 2, k is small, so one can see the stable regions. Figures 3, 4 and 5 show an unstable region corresponding to an increase in the virus ($k \geq 10$). *Compares*: In this study by assuming condition (4), for system (12), we paid attention to AIDS. According to our knowledge, this condition that causes a complex dynamic is not considered in any related previous works [2, 9, 12]. The new condition was lead to an one zero eigenvalue and as a consequence, by applying Sotomayor theorem, to transcritical bifurcation. Therefore, we determined the stable and unstable regions (by different given values, small and large for k) by using transcritical bifurcation.

5 Conclusion

In this paper, we analyzed system (12), with condition (4), at the equilibrium corresponding to only one zero eigenvalue (co-dimension one bifurcation). In order to determine transcritical bifurcation, we applied condition (10) in Sotomayer theorem (see theorem 2.3). One can notice that as we mentioned in the results of our investigation, the difference between this study and others [2, 9, 12]. From the biological point of view, the stable and unstable regions correspond to the viral population load. Moreover Figures 3, 4 and 5 showed that the virus population of AIDS increases for the value of $k \geq 10$.

References

1. Gleria, I., Neto, A.R. and Canabarro, A. *Nonlinear models for the delayed immune response to a viral infection*, Brazilian Journal of Physics, 45(4):450–456, 2015.

2. Guckenheimer, J. and Holmes, P. *Nonlinear oscillation. Dynamical Systems, and Bifurcations of Vector Fields*, Applied Mathematical Sciences, 42, 1983.

3. Komarova, N.L., Barnes, E., Klenerman, P. and Wodarz, D. *Boosting immunity by antiviral drug therapy: a simple relationship among timing, efficacy, and success*, PNAS, 100:1855–1860, 2003.

4. Kuznetsov, Y.A. *Elements of applied bifurcation theory*, volume 112. Springer Science & Business Media, 2013.

5. Lenhart, S. and Workman, J.T. *An introduction to optimal control applied to immunology*, Modeling and Simulation of Biological Networks, 64:85, 2007.

6. Nowak, M. and May, R. *Virus dynamics: mathematical principles of immunology and virology*, Oxford University Press, Oxford, 2001.

7. Perko, L. *Differential Equations and Dynamical Systems*, (Texts in Applied Mathematics), volume Third edition of Texts in applied mathematics . Springer, 2006.

8. Shu, H., Wang, L. and Watmough, J. *Sustained and transient oscillations and chaos inducedby delayed antiviral immune response in an immunosuppressive infection model*, J. Math. Biol., 68:477–503, 2014.

9. Strogatz, S.H. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*, Westview press, 2014.

10. Tang, B., Xiao, Y., Cheke, R.A. and Wang, N. *Piecewise virus-immune dynamic model with hiv-1 rna-guided therapy*, Journal of theoretical biology, 377:36–46, 2015.

11. Wiggins, S. *Introduction to applied nonlinear dynamical systems and chaos*, volume 2. Springer Science & Business Media, 2003.

12. Wodarz, D. *Killer cell dynamics mathematical and computational apporoches to immunology*, Interdisciplinary Applied Mathematics. Springer, 2007.

13. Wodarz, D., M.A., N., and C.R.M., B. *The dynamics of htlv-i and the ctl response*, Immunol Today, 20:220–227, 1999.

Iranian Journal of Numerical Analysis and Optimization
Vol. 6, No. 2, (2016), pp 17-37

Two numerical methods for nonlinear constrained quadratic optimal control problems using linear B-spline functions

Y. Edrisi-Tabriz, M. Lakestani* and A. Heydari

Abstract

This paper presents two numerical methods for solving the nonlinear constrained optimal control problems including quadratic performance index. The methods are based upon linear B-spline functions. The properties of B-spline functions are presented. Two operational matrices of integration are introduced for related procedures. These matrices are then utilized to reduce the solution of the nonlinear constrained optimal control to a nonlinear programming one to which existing well-developed algorithms may be applied. Illustrative examples are included to demonstrate the validity and applicability of the presented techniques.

Keywords: Optimal control problem; Linear B-spline function; Integration matrix; Collocation method.

1 Introduction

Solving an optimal control problem is not easy. Because of the complexity of most applications, optimal control problems are most often solved numerically. Numerical methods for solving optimal control problems date back nearly five decades to the 1950s with the work of Bellman [2–4]. Numerical methods for solving optimal control problems are divided into two major classes: direct methods and indirect methods.

*Corresponding author
Received 25 October 2014; revised 8 April 2015; accepted 6 Desember 2015
Y. Edrisi-Tabriz
Department of Mathematics, Payame Noor University, Tehran, Iran.
e-mail: yousef_edrisi@pnu.ac.ir

M. Lakestani
Faculty of Mathematical Sciences, University of Tabriz, Tabriz, Iran.
e-mail: lakestani@tabrizu.ac.ir

A. Heydari
Department of Mathematics, Payame Noor University, Tehran, Iran.
e-mail: a_heidari@pnu.ac.ir

In an indirect method, the calculus of variations [14, 24] is used to determine the first-order optimality conditions of the original optimal control problem. The indirect approach leads to a multiple-point boundary-value problem that is solved to determine candidate optimal trajectories called extremals. Each of the computed extremals is then examined to see if it is a local minimum, maximum, or a saddle point. Of the locally optimizing solutions, the particular extremal with the lowest cost is chosen.

One of the widely used methods to solve optimal control problems is the direct method. There is a large number of research papers that employ this method to solve optimal control problems (see for example [5, 6, 10, 15, 25, 26, 28] and the references therein). This method converts the optimal control problem into a mathematical programming problem by using either the discretization technique [5, 6] or the parameterization technique [10, 25, 26, 28].

The discretization technique converts the optimal control problem into a nonlinear programming problem with a large number of unknown parameters and a large number of constraints [6]. On the other hand, parameterizing the control variables [10, 28] requires the integration of the state equations. While the simultaneous parameterization of both the state variables and the control variables [28] results in a nonlinear programming problem with a large number of parameters and a large number of equality constraints.

In the last several years, various methods have been proposed to solve these problems. Yen and Nagurka [32] proposed a method based on the state parameterization, using Fourier series, to solve the linear-quadratic optimal control problem (with equal number of state variables and control variables) subject to state and control inequality constraints. Also Razzaghi and Elnagar [30] proposed a method to solve the unconstrained linear-quadratic optimal control problem with equal number of state and control variables. Their approach is based on using the shifted Legendre polynomials to parameterize the derivative of each of the state variables. In [16] Jaddu and Shimemura proposed a method to solve the linear-quadratic and the nonlinear optimal control problems by using Chebyshev polynomials to parameterize some of the state variables, then the remaining state variables and the control variables are determined from the state equations. The approach proposed in [28] is based on approximating the state variables and control variables with hybrid functions.

In this paper, we present two computational methods for solving nonlinear constrained quadratic optimal control problems by using linear B-spline functions. The methods are based on approximating the state variables and the control variables with a semiorthogonal linear B-spline functions [21]. Our methods consists of reducing the optimal control problem to a NLP one by first expanding the state rate $\dot{x}(t)$ and the control $u(t)$ as a linear B-spline functions with unknown coefficients. These functions are introduced. For the approximation of the integral, the operational matrix of integration $\mathbf{I}_\phi$ is

given. Two operational matrices of integration are calculated using (i) dual basis functions and (ii) interpolation basis functions.

The paper is organized as follows: In Section 2 we describe the basic formulation of the linear B-spline functions required for our subsequent development. Section 3 is devoted to the formulation of optimal control problems. Section 4 summarizes the application of these methods to the optimal control problems, and in Section 5, we report our numerical findings and demonstrate the accuracy of the proposed methods. Sections 6 completes this paper with a brief conclusion.

2 Properties of B-spline functions

2.1 Linear B-spline functions on [0,1]

The mth-order cardinal B-spline $N_m(t)$ has the knot sequence $\{\ldots,-1,0,1,\ldots\}$ and consists of polynomials of order m (degree $m-1$) between the knots. Let $N_1(t)=\chi_{[0,1]}(t)$ be the characteristic function of [0,1]. Then for each integer $m\geqslant 2$, the mth-order cardinal B-spline is defined, inductively by [8,13]

$$N_m(t)=(N_{m-1}*N_1)(t)=\int_{-\infty}^{\infty}N_{m-1}(t-\tau)N_1(\tau)d\tau=\int_0^1 N_{m-1}(t-\tau)d\tau. \tag{1}$$

It can be shown [7] that $N_m(t)$ for $m\geqslant 2$ can be achieved using the following formula

$$N_m(t)=\frac{t}{m-1}N_{m-1}(t)+\frac{m-t}{m-1}N_{m-1}(t-1),$$

recursively, and $\mathrm{supp}[N_m(t)]=[0,m]$.

The explicit expressions of $N_2(t)$ (linear B-spline function) are [7,8,13]

$$N_2(t)=\begin{cases}t & t\in[0,1],\\ 2-t & t\in[1,2],\\ 0 & elsewhere.\end{cases} \tag{2}$$

Suppose $N_{j,k}(t)=N_2(2^jt-k), j,k\in\mathbb{Z}$ and $B_{j,k}=\mathrm{supp}[N_{j,k}(t)]=\mathrm{clos}\{t: N_{j,k}(t)\neq 0\}$. It is easy to see that

$$B_{j,k}=[2^{-j}k,2^{-j}(2+k)],\quad j,k\in\mathbb{Z}.$$

To use these functions on $[0,1]$,

$$S_j=\{k: B_{j,k}\cap[0,1]\neq\emptyset\},\quad j\in\mathbb{Z}.$$

It is easy to see that $\min\{S_j\}=-1$ and $\max\{S_j\}=2^j-1, j\in\mathbb{Z}$.

The support of $N_{j,k}(t)$ may be out of interval [0,1], we need that these functions intrinsically defined on [0,1] so we put

$$\phi_{j,k}(t) = N_{j,k}(t)\chi_{[0,1]}(t), \quad j \in \mathbb{Z}, \quad k \in S_j. \tag{3}$$

2.2 The function approximation

Suppose $\Phi_j(t)$ is a $(2^j + 1)$-vector as

$$\Phi_j(t) = [\phi_{j,-1}(t), \phi_{j,0}(t), \ldots, \phi_{j,2^j-1}(t)]^T, \quad j \in \mathbb{Z}. \tag{4}$$

For a fixed $j = M$, a function $f(t) \in L^2[0,1]$ may be represented by the linear B-spline functions as

$$f(t) \simeq \sum_{k=-1}^{2^M-1} s_k \phi_{M,k}(t) = S^T \Phi_M(t), \tag{5}$$

where

$$S = [s_{-1}, s_0, \ldots, s_{2^M-1}]^T \tag{6}$$

and

$$s_k = f(\frac{k+1}{2^M}), \quad k = -1, \ldots, 2^M - 1. \tag{7}$$

Note that the functions $\phi_{M,k}(t)$ satisfy in the relation

$$\phi_{M,k}(\frac{i+1}{2^M}) = \delta_{k,i} = \begin{cases} 1, k = i, \\ 0, k \neq i, \end{cases} \quad i = -1, \ldots, 2^M - 1.$$

So we have

$$\Phi_M(t_i) = e_i, \quad t_i = \frac{i+1}{2^M}, \quad i = -1, \ldots, 2^M - 1, \tag{8}$$

where e_i is the ith column of unit matrix of order $2^M + 1$ [21].

2.3 Two operational matrices of integration

Suppose

$$\Phi_M^{\int}(t) = \int_0^t \Phi_M(\tau) d\tau, \tag{9}$$

then the integration of vectors Φ_M in (4) can be expressed as

$$\Phi_M^{\int} = \mathbf{I}_\phi \Phi_M, \tag{10}$$

where $\mathbf{I}_\phi$ is $(2^M+1)\times(2^M+1)$ operational matrix of integration for the linear B-spline functions on $[0,1]$. We construct $\mathbf{I}_\phi$ using the following two methods:

Method 1.

$$\mathbf{I}_\phi = \int_0^1 \Phi_M^{\int}(t)\widetilde{\Phi}_M^T(t)dt, \tag{11}$$

where $\widetilde{\Phi}_M$ is the vector of dual basis of Φ_M which can be obtained using the linear combinations of $\phi_{j,k}$ [22, 23] as

$$\widetilde{\Phi}_M = \mathbf{P}^{-1}\Phi_M, \tag{12}$$

where

$$\mathbf{P} = \int_0^1 \Phi_M(t)\Phi_M^T(t)dt = 2^{-M}\begin{bmatrix} \frac{1}{3} & \frac{1}{6} & & & \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ & & & \frac{1}{6} & \frac{1}{3} \end{bmatrix}. \tag{13}$$

Replacing (12) in (11) we get

$$\mathbf{I}_\phi = \left(\int_0^1 \Phi_M^{\int}(t)\Phi_M^T(t)dt\right)\mathbf{P}^{-1} = \mathbf{E}(\mathbf{P}^{-1}), \tag{14}$$

where

$$\mathbf{E} = \int_0^1 \Phi_M^{\int}(t)\Phi_M^T(t)dt. \tag{15}$$

By using Eqs. (9) and (15) we obtain

$$\mathbf{E} = 2^{-(2M+1)}\begin{bmatrix} \frac{1}{4} & \frac{11}{12} & 1 & \cdots & 1 & \frac{1}{2} \\ \frac{1}{12} & 1 & \frac{23}{12} & 2 & \cdots & 2 & 1 \\ & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ & & \ddots & \ddots & \ddots & 2 & \vdots \\ & & & \ddots & 1 & \frac{23}{12} & 1 \\ & & & & \frac{1}{12} & 1 & \frac{11}{12} \\ & & & & & \frac{1}{12} & \frac{1}{4} \end{bmatrix}.$$

Method 2.

In this method, we approximate $\Phi_M^{\int}$ using linear B-spline functions and then construct $\mathbf{I}_\phi$. Suppose

$$\Phi_M^{\int} = \left[\, L_1(t)\ L_2(t)\ \cdots\ L_{2^M+1}(t)\,\right]^T, \tag{16}$$

where using Eq. (4) we have

$$L_i(t) = \int_0^t \phi_{M,i-2}(\tau)d\tau, \qquad i = 1, \dots 2^M + 1, \quad M \in \mathbb{Z}.$$

Finally from Eq. (7) we get

$$(\mathbf{I}_\phi)_{ij} = L_i(\frac{j-1}{2^M}), \qquad i = 1, \dots 2^M + 1, \quad j = 1, \dots 2^M + 1, \tag{17}$$

where $(\mathbf{I}_\phi)_{ij}$ denotes the ij-th element of matrix $\mathbf{I}_\phi$. Final form of this matrix is as follows:

$$\mathbf{I}_\phi = 2^{-(M+1)} \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ & 1 & 2 & \cdots & 2 \\ & & 1 & \ddots & \vdots \\ & & & \ddots & 2 \\ & & & & 1 \end{bmatrix}. \tag{18}$$

3 Problem statement

The problem we are treating is to find the optimal control $\mathbf{u}^*(t)$ and the corresponding optimal state trajectory $\mathbf{x}^*(t)$ that minimizes the performance index

$$J = \frac{1}{2}\mathbf{x}^T(t_f)\mathbf{Z}\mathbf{x}(t_f) + \frac{1}{2}\int_{t_0}^{t_f} \left(\mathbf{x}^T(t)\mathbf{Q}(t)\mathbf{x}(t) + \mathbf{u}^T(t)\mathbf{R}(t)\mathbf{u}(t)\right) dt, \tag{19}$$

subject to

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t), \tag{20}$$

$$\Psi(\mathbf{x}(t_0), t_0, \mathbf{x}(t_f), t_f) = 0, \tag{21}$$

$$\mathbf{g}_i(\mathbf{x}(t), \mathbf{u}(t), t) \leqslant 0, \quad i = 1, 2, \dots, w, \tag{22}$$

where $\mathbf{Z}$ and $\mathbf{Q}(t)$ are positive semidefinite matrices, $\mathbf{R}(t)$ is a positive definite matrix, t_0 and t_f are known initial and terminal time respectively, $\mathbf{x}(t) = (x_i(t))_{i=1}^{l}$ is the state vector, $\mathbf{u}(t) = (u_j(t))_{j=1}^{q}$ is the control vector and $\mathbf{f}, \mathbf{g}_i$ $(i = 1, 2, \dots, w)$ are nonlinear functions. This problem is defined on the time interval $t \in [t_0, t_f]$. Certain numerical techniques (like B-spline functions) require a fixed time interval, such as $[0, 1]$. The independent variable can be mapped to the general interval $\tau \in [0, 1]$ via the affine transformation

$$\tau = \frac{t - t_0}{t_f - t_0}. \tag{23}$$

Note that this mapping is still valid with free initial and final times. Using Eq. (23), this problem can be redefined as follows:

Minimize the performance index

$$J = \frac{1}{2}\mathbf{x}^T(1)\mathbf{Z}\mathbf{x}(1) + \frac{1}{2}(t_f - t_0)\int_0^1 \left(\mathbf{x}^T(\tau)\mathbf{Q}(\tau)\mathbf{x}(\tau) + \mathbf{u}^T(\tau)\mathbf{R}(\tau)\mathbf{u}(\tau)\right) d\tau, \tag{24}$$

subject to

$$\frac{d\mathbf{x}}{d\tau} = (t_f - t_0)\mathbf{f}(\mathbf{x}(\tau), \mathbf{u}(\tau), \tau; t_0, t_f), \tag{25}$$

$$\Psi(\mathbf{x}(0), t_0, \mathbf{x}(1), t_f) = 0, \tag{26}$$

$$\mathbf{g}_i(\mathbf{x}(\tau), \mathbf{u}(\tau), \tau; t_0, t_f) \leqslant 0, \quad i = 1, 2, \ldots, w, \quad \tau \in [0, 1]. \tag{27}$$

4 The proposed method

Let

$$\hat{\Phi}_{M,l}(t) = I_l \otimes \Phi_M(t), \tag{28}$$

$$\hat{\Phi}_{M,q}(t) = I_q \otimes \Phi_M(t), \tag{29}$$

where I_l and I_q are $l \times l$ and $q \times q$ dimensional identity matrices, $\Phi_M(t)$ is $(2^M + 1)$-vector, $\otimes$ denotes Kronecker product [20] and $\hat{\Phi}_{M,l}(t)$ and $\hat{\Phi}_{M,q}(t)$ are matrices of order $l(2^M + 1) \times l$ and $q(2^M + 1) \times q$. Assume that each of $\dot{\mathbf{x}}_i(t)$, $i = 1, 2, \ldots, l$, and each of $\mathbf{u}_j(t)$, $j = 1, 2, \ldots, q$, can be written in terms of linear B-spline functions as

$$\dot{\mathbf{x}}_i(t) \simeq \Phi_M^T(t)\mathbf{X}_i,$$
$$\mathbf{u}_j(t) \simeq \Phi_M^T(t)\mathbf{U}_j.$$

Then using Eqs. (28) and (29) we have

$$\dot{\mathbf{x}}(t) \simeq \hat{\Phi}_{M,l}^T(t)\mathbf{X}, \tag{30}$$

$$\mathbf{u}(t) \simeq \hat{\Phi}_{M,q}^T(t)\mathbf{U}, \tag{31}$$

where X and U are vectors of orders $l(2^M + 1)$ and $q(2^M + 1)$, respectively, given by

$$\mathbf{X} = \left[\mathbf{X}_1^T, \mathbf{X}_2^T, \ldots, \mathbf{X}_l^T\right]^T,$$
$$\mathbf{U} = \left[\mathbf{U}_1^T, \mathbf{U}_2^T, \ldots, \mathbf{U}_q^T\right]^T.$$

Similarly we have

$$\mathbf{x}(0) \simeq \hat{\Phi}_{M,l}^T(t)\mathbf{A}_0, \tag{32}$$

where $\mathbf{A}_0$ is a vector of order $l(2^M+1)$ given by

$$\mathbf{A}_0 = \left[a_1^T, a_2^T, \ldots, a_l^T\right]^T.$$

By integrating Eq. (30) from 0 to t we get

$$\mathbf{x}(t)-\mathbf{x}(0) = \int_0^t \hat{\Phi}_{M,l}^T(\tau)\mathbf{X}d\tau \simeq (I_l \otimes \Phi_M^T(t))(I_l \otimes \mathbf{I}_\phi^T)\mathbf{X} = \hat{\Phi}_{M,l}^T(t)\hat{\mathbf{I}}_\phi^T\mathbf{X}, \tag{33}$$

where $\mathbf{I}_\phi$ is an operational matrix of integration given in Eq. (14). From Eqs. (32) and (33) we obtain

$$\mathbf{x}(t) \simeq \hat{\Phi}_{M,l}^T(t)(\mathbf{A}_0 + \hat{\mathbf{I}}_\phi^T\mathbf{X}). \tag{34}$$

4.1 The performance index approximation

By substituting Eqs. (31)-(34) in Eq. (24) we get

$$\begin{aligned} J =& \frac{1}{2}(\mathbf{A}_0 + \mathbf{I}_\phi^T\mathbf{X})^T(\hat{\Phi}_{M,l}(1)\mathbf{Z}\hat{\Phi}_{M,l}^T(1))(\mathbf{A}_0 + \mathbf{I}_\phi^T\mathbf{X}) \\ &+ \frac{1}{2}(t_f - t_0)(\mathbf{A}_0 + \mathbf{I}_\phi^T\mathbf{X})^T\left(\int_0^1 \hat{\Phi}_{M,l}(t)\mathbf{Q}(t)\hat{\Phi}_{M,l}^T(t)dt\right)(\mathbf{A}_0 + \mathbf{I}_\phi^T\mathbf{X}) \\ &+ \frac{1}{2}(t_f - t_0)\mathbf{U}^T\left(\int_0^1 \hat{\Phi}_{M,q}(t)\mathbf{R}(t)\hat{\Phi}_{M,q}^T(t)dt\right)\mathbf{U}. \end{aligned} \tag{35}$$

Eq. (35) can be computed more efficiently by writing J as

$$\begin{aligned} J =& \frac{1}{2}(\mathbf{A}_0 + \mathbf{I}_\phi^T\mathbf{X})^T\left(\mathbf{Z} \otimes \Phi_M(1)\Phi_M^T(1)\right)(\mathbf{A}_0 + \mathbf{I}_\phi^T\mathbf{X}) \\ &+ \frac{1}{2}(t_f - t_0)(\mathbf{A}_0 + \mathbf{I}_\phi^T\mathbf{X})^T\left(\int_0^1 \mathbf{Q}(t) \otimes \Phi_M(t)\Phi_M^T(t)dt\right)(\mathbf{A}_0 + \mathbf{I}_\phi^T\mathbf{X}) \\ &+ \frac{1}{2}(t_f - t_0)\mathbf{U}^T\left(\int_0^1 \mathbf{R}(t) \otimes \Phi_M(t)\Phi_M^T(t)dt\right)\mathbf{U}. \end{aligned} \tag{36}$$

For problems with time-varying performance index, $\mathbf{Q}(t)$ and $\mathbf{R}(t)$ are functions of time and

$$\int_0^1 \mathbf{Q}(t) \otimes \Phi_M(t)\Phi_M^T(t)dt, \qquad \int_0^1 \mathbf{R}(t) \otimes \Phi_M(t)\Phi_M^T(t)dt$$

can be evaluated numerically. For time-invariant problems, $Q(t)$ and $R(t)$ are constant matrices and can be removed from the integrals. In this case,

Eq. (36) can be rewritten as

$$\begin{aligned} J(\mathbf{X},\mathbf{U}) =& \frac{1}{2}(\mathbf{A}_0+\mathbf{I}_\phi^T\mathbf{X})^T\left(\mathbf{Z}\otimes\Phi_M(1)\Phi_M^T(1)\right)(\mathbf{A}_0+\mathbf{I}_\phi^T\mathbf{X}) \\ &+\frac{1}{2}(t_f-t_0)(\mathbf{A}_0+\mathbf{I}_\phi^T\mathbf{X})^T(\mathbf{Q}\otimes\mathbf{P})(\mathbf{A}_0+\mathbf{I}_\phi^T\mathbf{X}) \\ &+\frac{1}{2}(t_f-t_0)\mathbf{U}^T(\mathbf{R}\otimes\mathbf{P})\mathbf{U}. \end{aligned} \tag{37}$$

4.2 The system constraints approximation

We approximate the system constraints as follows:

Using Eqs. (30), (31) and (34) the system constraints (25), (26) and (27) became

$$\hat{\Phi}_{M,l}^T(t)\mathbf{X} = (t_f-t_0)\mathbf{f}(\hat{\Phi}_{M,l}^T(t)(A_0+\hat{\mathbf{I}}_\phi^T\mathbf{X}),\hat{\Phi}_{M,q}^T(t)\mathbf{U},t;t_0,t_f), \tag{38}$$

$$\Psi(\hat{\Phi}_{M,l}^T(0)(A_0+\hat{\mathbf{I}}_\phi^T\mathbf{X}),t_0,\hat{\Phi}_{M,l}^T(1)(A_0+\hat{\mathbf{I}}_\phi^T\mathbf{X}),t_f)=0, \tag{39}$$

$$\mathbf{g}_i(\hat{\Phi}_{M,l}^T(t)(A_0+\hat{\mathbf{I}}_\phi^T\mathbf{X}),\hat{\Phi}_{M,q}^T(t)\mathbf{U},t;t_0,t_f)\leqslant 0, \qquad i=1,2,\ldots,w. \tag{40}$$

We collocate Eqs. (38) and (40) at Newton-cotes nodes t_k,

$$t_k=\frac{k-1}{2^M}, \quad k=1,2,\ldots,2^M+1. \tag{41}$$

The optimal control problem has now been reduced to a parameter optimization problem which can be stated as follows:

Find $\mathbf{X}$ and $\mathbf{U}$ so that $J(\mathbf{X},\mathbf{U})$ is minimized (or maximized) subject to Eq. (39) and

$$\hat{\Phi}_{M,l}^T(t_k)\mathbf{X} = (t_f-t_0)\mathbf{f}(\hat{\Phi}_{M,l}^T(t_k)(A_0+\hat{\mathbf{I}}_\phi^T\mathbf{X}),\hat{\Phi}_{M,q}^T(t_k)\mathbf{U},t_k), \tag{42}$$

$$\mathbf{g}_i(\hat{\Phi}_{M,l}^T(t_k)(A_0+\hat{\mathbf{I}}_\phi^T\mathbf{X}),\hat{\Phi}_{M,q}^T(t_k)\mathbf{U},t_k;t_0,t_f)\leqslant 0, \tag{43}$$
$$i=1,2,\ldots,w, \quad k=1,2,\ldots,2^M+1.$$

Many well-developed nonlinear programming techniques can be used to solve this extremum problem (see, e.g. [1, 9, 11]).

5 Illustrative examples

This section is devoted to numerical examples. All problems were programmed in MAPLE, running on a Pentium 4, 2.4-GHz PC with 4 GB of RAM. Also we solved the obtained nonlinear programming that is minimize

(or maximize) $J(\mathbf{X}, \mathbf{U})$ subject to Eqs. (39), (42) and (43), using "NLPsolve" command in MAPLE program. To illustrate our techniques, we present five numerical examples and make a comparison with some of the results in the literatures.

Example 1. This example is adapted from [18]. Find the control vector $u(t)$ which minimizes

$$J = \frac{1}{2}\int_0^1 \left(x_1^2(t) + u^2(t)\right) dt, \tag{44}$$

subject to

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t), \tag{45}$$

$$\begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 10 \end{bmatrix}, \tag{46}$$

and subject to the following inequality control constraint

$$|u(t)| \leqslant 1. \tag{47}$$

In Table 1, the minimum of J using the rationalized Haar functions [29], hybrid of block-pulse and Legendre polynomials [25], hybrid of block-pulse and Bernoulli polynomials [28] and present two methods are listed. In Figure 1, the control and state variables with the absolute value of constraint's errors for $M = 8$, are reported.

Table 1: Estimated values of J for Example 1

Method	J	CPUTime
Rationalized Haar functions [29]		
$K = 4$	8.07473	0.389
$K = 8$	8.07065	0.546
Hybrid of block-pulse and Legendre [25]		
$N = 4, M_1 = 3$	8.07059	1.592
$N = 4, M_1 = 4$	8.07056	4.304
Hybrid of block-pulse and Bernoulli [28]		
$N = 4, M = 2$	8.07058	0.858
$N = 4, M = 3$	8.07055	1.155
Present method 1		
$M = 6$	8.07056208507474	1.075
$M = 7$	8.07056206359357	1.453
$M = 8$	8.07056204949560	1.722
Present method 2		
$M = 6$	8.07043910532066	0.665
$M = 7$	8.07053132323846	1.009
$M = 8$	8.07055438812380	1.341

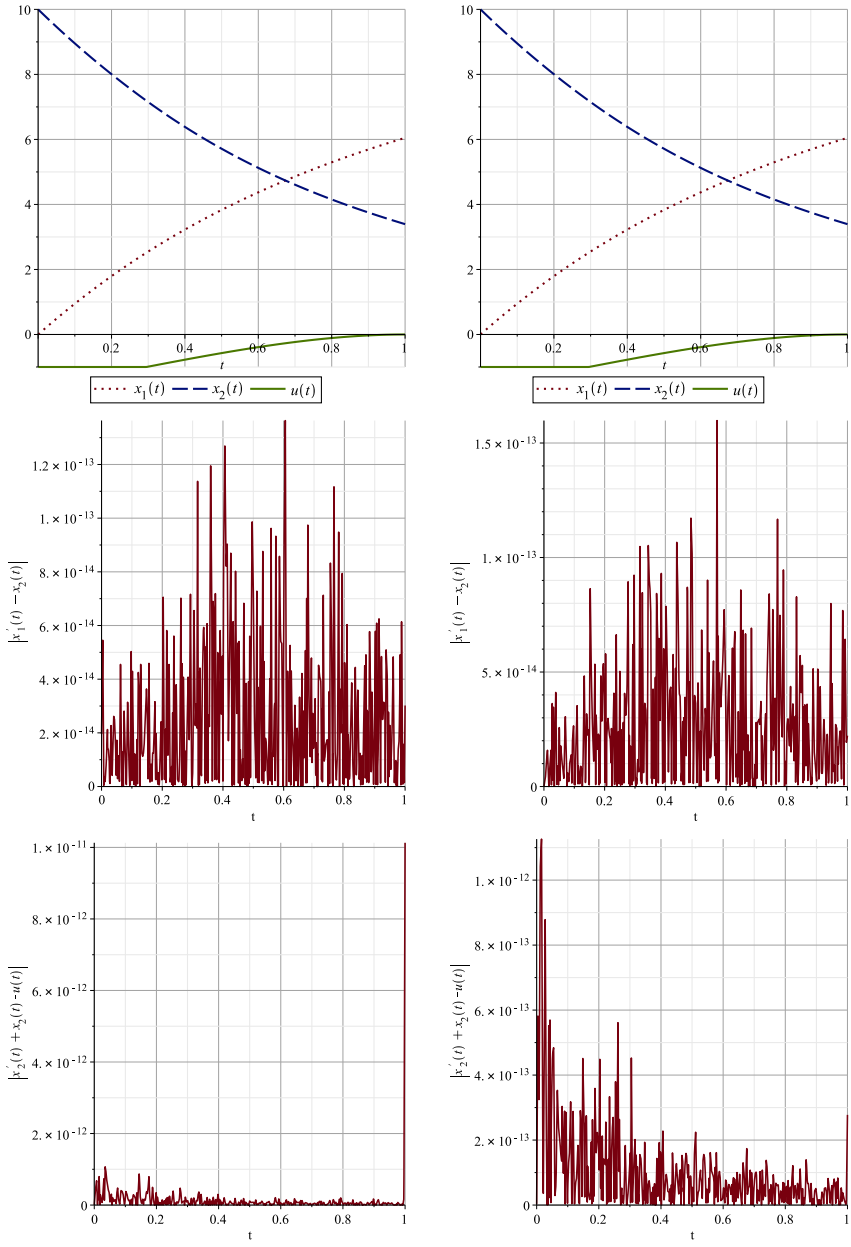

Figure 1: State and control variables and the constraint errors $|\dot{x}_1(t) - x_2(t)|$ and $|\dot{x}_2(t) + x_2(t) - u(t)|$ for Example 1 using Method 1 (left) and using Method 2 (right) with $M = 8$

Example 2. Consider the Breakwell problem [12]. The performance index to be minimized is given by

$$J = \frac{1}{2}\int_0^1 u^2(t)dt, \tag{48}$$

subject to the state equations

$$\dot{x}_1(t) = x_2(t), \qquad \dot{x}_2(t) = u(t), \tag{49}$$

with the endpoint conditions

$$x_1(0) = x_1(1) = 0, \qquad x_2(0) = -x_2(1) = 1, \tag{50}$$

and the state constraint

$$x_1(t) \leqslant 0.1. \tag{51}$$

The exact solution to this problem is given by

$$u^*(t) = \begin{cases} \frac{200}{9}t - \frac{20}{3}, & t \in [0, 0.3], \\ 0, & t \in [0.3, 0.7], \\ -\frac{200}{9}t + \frac{140}{9}, & t \in [0.7, 1]. \end{cases} \tag{52}$$

This example was studied by using pseudospectral method [12] and ChFD scheme [27]. Here we applied the proposed method to solve this problem. Absolute errors between approximation and exact value of the performance index are reported in Table 2. The approximate solutions of $x_1(t)$, $x_2(t)$ and $u(t)$, obtained by linear B-spline functions using method 2 with $M = 9$ and the exact solutions together error bounds $|x_1^*(t) - x_1(t)|$, $|x_2^*(t) - x_2(t)|$ and $|u^*(t) - u(t)|$ are plotted in Figure 2. This results show that accuracy of our method in comparison with ChFD scheme [27] whose result are plotted in Figure 3.

Table 2: Errors of the estimated and exact values of the performance index, $|J - J^*|$, for Example 2

	Method 1		Method 2	
M	$\|J - J^*\|$	CPUTime	$\|J - J^*\|$	CPUTime
6	3.67×10^{-2}	0.053	5.93×10^{-3}	0.163
7	1.86×10^{-2}	0.181	1.48×10^{-3}	0.401
8	9.34×10^{-3}	1.034	3.74×10^{-4}	1.377
9	4.68×10^{-3}	7.662	9.38×10^{-5}	7.400

Example 3. This example is adapted from [19] and also studied by using rationalized Haar approach [26], hybrid of block-pulse and Legendre polynomials [25], hybrid of block-pulse and Bernoulli polynomials [28] and interpolating scaling functions [10]. Find the control vector u(t) which minimizes

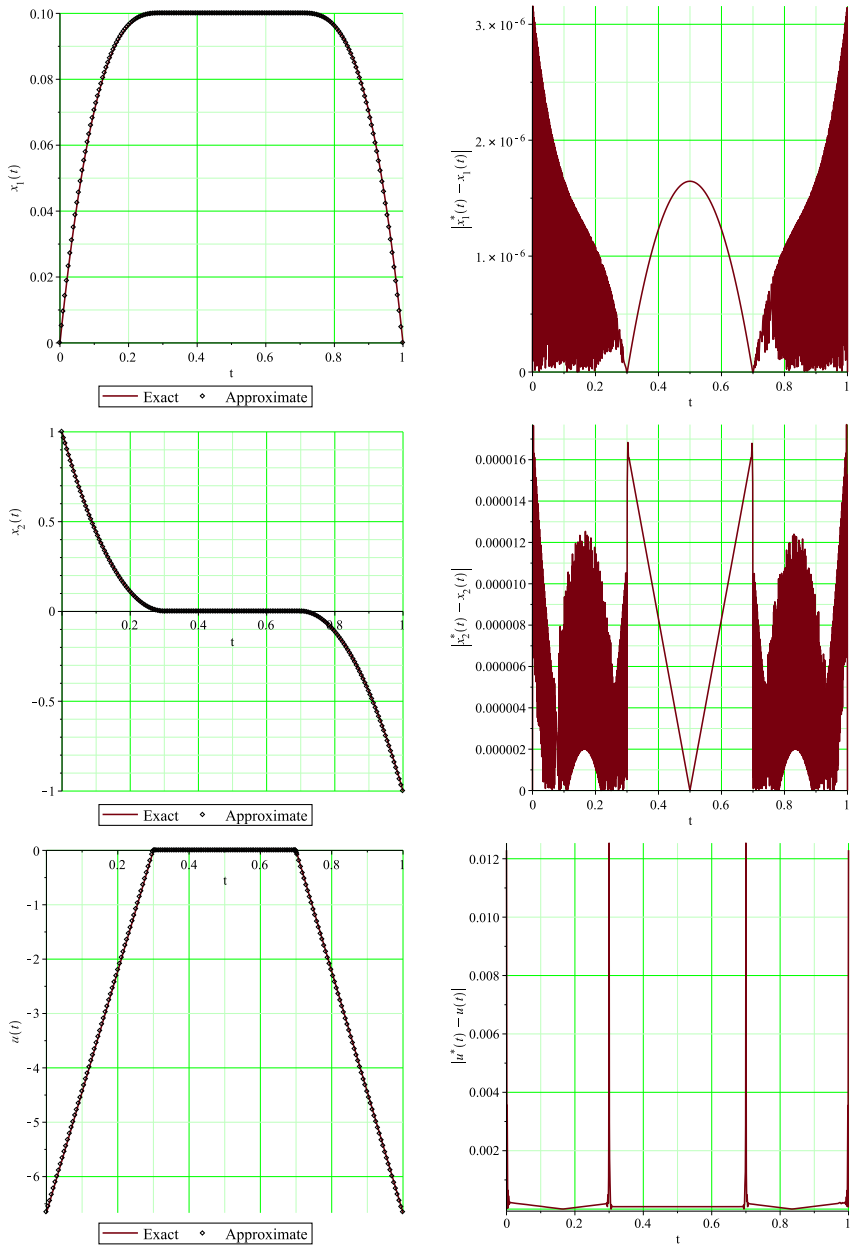

Figure 2: Exact value, approximation of optimal control, state variables and error bounds using method 2 for Example 2 with $M = 9$

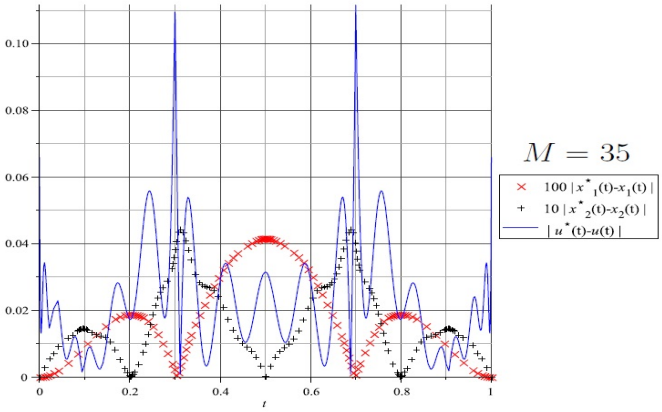

Figure 3: Exact value and approximation errors of $|x_1^*(t) - x_1(t)|$, $|x_2^*(t) - x_2(t)|$ and $|u^*(t) - u(t)|$ using ChFD scheme [27] for Example 2 with $M = 35$

$$J = \int_0^1 \left(x_1^2(t) + x_2^2(t) + 0.005u^2(t)\right) dt, \tag{53}$$

subject to

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t), \tag{54}$$

$$\begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \tag{55}$$

and the following state variable inequality constraint

$$x_2(t) \leqslant r(t), \tag{56}$$

where

$$r(t) = 8(t - 0.5)^2 - 0.5, \quad 0 \leqslant t \leqslant 1.$$

The computational result for $x_2(t)$ using method 2 for $M = 8$ together with $r(t)$ are given in Fig. 4. In Table 3, we compare the minimum of J using the proposed two methods with other solutions in the literature.

Example 4. We consider the optimal maneuvers of a rigid asymmetric spacecraft [17]. This example is studied by using quasilinearization and Chebyshev polynomials [15] and hybrid of block-pulse and Bernoulli polynomials [28]. The system state equations are

Table 3: Results for Example 3

Method	J	CPUTime
Rationalized Haar functions [26]		
$K = 64, w = 100$	0.170115	1.877
$K = 128, w = 100$	0.170103	1.983
Hybrid of block-pulse and Legendre [25]		
$N = 4, M_1 = 3$	0.17013645	0.951
$N = 4, M_1 = 4$	0.17013640	1.545
Hybrid of block-pulse and Bernoulli [28]		
$N = 4, M = 3$	0.1700305	0.756
$N = 4, M = 4$	0.1700301	0.921
Interpolating scaling functions [10]		
$n = 4, r = 5$	0.16982646	2.251
$n = 5, r = 5$	0.16982636	3.175
Present method 1		
$M = 6$	0.169672402102247	1.512
$M = 7$	0.169782602033829	1.607
$M = 8$	0.169811048165412	1.985
Present method 2		
$M = 6$	0.170071967582200	0.599
$M = 7$	0.169885295276034	1.003
$M = 8$	0.169837051920398	1.141

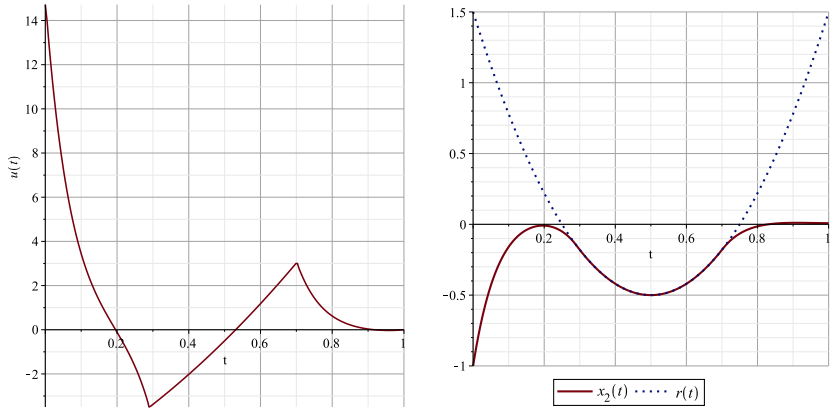

Figure 4: Control and state variables and constraint boundary for Example 3 using method 2 with $M = 8$

$$\dot{x}_1(\tau) = -\frac{I_3 - I_2}{I_1} x_2(\tau) x_3(\tau) + \frac{u_1(\tau)}{I_1},$$
$$\dot{x}_2(\tau) = -\frac{I_1 - I_3}{I_2} x_1(\tau) x_3(\tau) + \frac{u_2(\tau)}{I_2},$$
$$\dot{x}_3(\tau) = -\frac{I_2 - I_1}{I_3} x_1(\tau) x_2(\tau) + \frac{u_3(\tau)}{I_3},$$
$$x_1(\tau) - (5 \times 10^{-6}\tau^2 - 5 \times 10^{-4}\tau + 0.016) \leqslant 0,$$
$$x_1(100) = x_2(100) = x_3(100) = 0,$$

where $I_1 = 86.24$, $I_2 = 85.07$, $I_3 = 113.59$. The performance index to be minimized, starting from the initial states $x_1(0) = 0.01$, $x_2(0) = 0.005$ and $x_3(0) = 0.001$ is

$$J = \frac{1}{2}\int_0^{100} \left(u_1^2(\tau) + u_2^2(\tau) + u_3^2(\tau)\right) d\tau.$$

We use transformation $\tau = 100t, 0 \leqslant t \leqslant 1$, in order to use our proposed method. In Table 4, the results for J using linear B-spline functions, hybrid of block-pulse and Bernoulli polynomials [28] and quasilinearization and Chebyshev polynomials [15] are listed. Optimal control and state variables and constraint boundary using method 2, for $M = 7$, are shown in Figure 5.

Table 4: Results for Example 4

Method	J	CPUTime
Quasilinearization and Chebyshev polynomials [15]		
$N = 6$	0.00536584	0.07
$N = 8$	0.00534427	0.10
$N = 10$	0.00534063	0.12
Quasilinearization and Chebyshev polynomials [15] with using 2 subintervals		
$M2 = 10$	0.00530902	0.36
Hybrid of block-pulse and Bernoulli [28]		
$N = 6, M = 3$	0.00531097	1.89
$N = 6, M = 4$	0.00530263	2.12
$N = 6, M = 5$	0.00530213	2.74
Present method 1		
$M = 5$	0.00527460682730895	0.55
$M = 6$	0.00529464663721832	0.67
$M = 7$	0.00530275422863559	0.71
Present method 2		
$M = 5$	0.00530817821841613	0.06
$M = 6$	0.00530851397972318	0.10
$M = 7$	0.00530847110421464	0.13

Example 5. Consider the problem of transferring containers from a ship to a cargo truck [31]. The container crane is driven by a hoist motor and a trolley drive motor. The aim is to minimize the swing during and at the end of the transfer. After appropriate normalization, we summarize the problem as follows:

$$J = 4.5\int_0^1 \left(x_3^2(t) + x_6^2(t)\right) dt$$

subject to

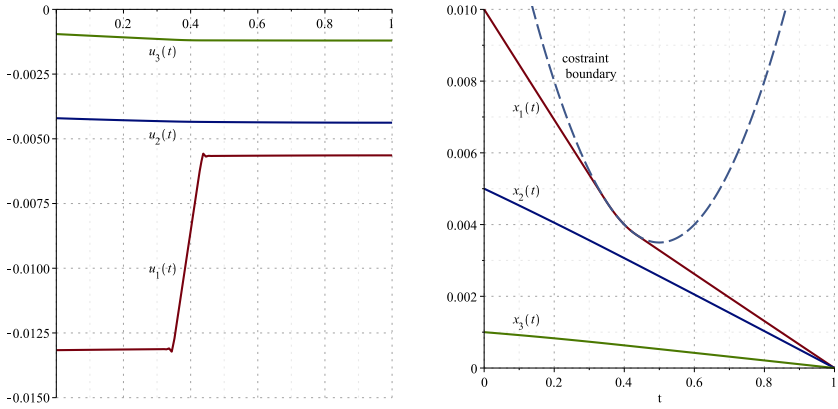

Figure 5: Control and state variables and constraint boundary for Example 4 using method 2 with $M = 7$

$$
\begin{aligned}
\dot{x}_1(t) &= 9x_4(t),\\
\dot{x}_2(t) &= 9x_5(t),\\
\dot{x}_3(t) &= 9x_6(t),\\
\dot{x}_4(t) &= 9(u_1(t) + 17.2656x_3(t)),\\
\dot{x}_5(t) &= 9u_2(t),\\
\dot{x}_6(t) &= \frac{-9(u_1(t) + 27.0756x_3(t) + 2x_5(t)x_6(t))}{x_2(t)},
\end{aligned}
$$

where

$$
\begin{aligned}
x(0) &= [0, 22, 0, 0, -1, 0]^T,\\
x(1) &= [10, 14, 0, 2.5, 0, 0]^T,
\end{aligned}
$$

and

$$
\begin{aligned}
&|u_1(t)| \leqslant 2.83374, \quad t \in [0, 1],\\
&-0.80865 \leqslant u_2(t) \leqslant 0.71265, \quad t \in [0, 1],
\end{aligned}
$$

with continuous state inequality constraints,

$$
\begin{aligned}
|x_4(t)| &\leqslant 2.5, \quad t \in [0, 1],\\
|x_5(t)| &\leqslant 1.0, \quad t \in [0, 1].
\end{aligned}
$$

In Table 5, we compare the solution obtained using the proposed two methods with the method of [9] and [28].

Table 5: Results for Example 5

Method	J	CPUTime
Method of [9]		
$m = 5$	0.5366×10^{-2}	2.589
$m = 7$	0.53614×10^{-2}	2.607
$m = 9$	$0.53610895 \times 10^{-2}$	3.002
$m = 11$	$0.5361102700 \times 10^{-2}$	3.021
Hybrid of block-pulse and Bernoulli [28]		
$N = 2, M = 2$	0.593000×10^{-2}	1.904
$N = 2, M = 3$	0.528915×10^{-2}	2.125
$N = 2, M = 4$	0.521421×10^{-2}	2.305
$N = 2, M = 5$	0.521411×10^{-2}	2.663
Present method 1		
$M = 5$	$0.498574175174882 \times 10^{-2}$	1.815
$M = 6$	$0.503885802644245 \times 10^{-2}$	1.963
$M = 7$	$0.511514185733751 \times 10^{-2}$	2.025
Present method 2		
$M = 5$	$0.516049181578648 \times 10^{-2}$	1.579
$M = 6$	$0.515021009757565 \times 10^{-2}$	1.708
$M = 7$	$0.515021009428266 \times 10^{-2}$	1.869

6 Conclusion

In this paper we presented two numerical methods for solving nonlinear constrained quadratic optimal control problems. Two methods are based upon the linear B-spline functions. Also several test problems were used to see the applicability and efficiency of the method. The obtained results show that when the state variables are unknown at the endpoints, then method 1 is more accurate than method 2 but in all problems method 2 is faster than method 1. In total, our two methods are more accurate than existing methods in the literature.

Acknowledgements

Authors are grateful to the anonymous referees and the editors for their constructive comments.

References

1. Avrile, M. *Nonlinear programming: Analysis and Methods*, Englewood Cliffs, NJ: Prentice-Hall, 1976.

2. Bellman, R. *Dynamic Programming*, Princeton, NJ: Princeton University Press, 1957.

3. Bellman, R., Kalaba, R. and Kotkin, B. *Polynomial Approximation - A New Computational Technique in Dynamic Programming: Allocation Processes*, Mathematics of Computation, 17 (1963), 155-161.

4. Bellman, R. and Dreyfus, S. E. *Applied Dynamic Programming*, Princeton University Press, 1971.

5. Betts, J. *Issues in the direct transcription of optimal control problem to sparse nonlinear programs*, in: Bulirsch, R., Kraft, D. (Eds.), Computational Optimal Control, Germany: Birkhauser, 115 (1994), 3-17.

6. Betts, J. *Survey of numerical methods for trajectory optimization*, J. Guidance Control Dynamics, 21(2) (1998), 193-207.

7. de Boor, C. *A practical guide to spline*, Springer Verlag, New York, 1978.

8. Chui, C.K. *An introduction to wavelets*, San Diego, Calif: Academic Press, 1992.

9. Elnegar, G.N. and Kazemi, M.A. *Pseudospectral Chebyshev optimal control of constrained nonlinear dynamical systems*, Comput. Optim. Applica. 11 (1998), 195-217.

10. Foroozandeh, Z. and Shamsi, M. *Solution of nonlinear optimal control problems by the interpolating scaling functions*, Acta Astronautica, 72 (2012), 21-26.

11. Gill, P.E. and Murray, W. *Linearly constrained problems including linear and quadratic programming, In: Jacobs D, editor, The State of the Art in Numerical Analysis*, London, New York: Academic Press, (1977), 313-363.

12. Gong, Q., Kang, W. and Ross, I.M. *A pseudospectral method for the optimal control of constrained feedback linearizable systems*, IEEE Trans Automat Control, 51(7) (2006), 1115-1129.

13. Goswami, J.C. and Chan, A.K. *Fundamentals of wavelets: theory, algorithms, and applications*, John Wiley & Sons, Inc., 1999.

14. Hull, D.G. *Optimal Control Theory for Applications*, New York: Springer-Verlag, 2003.

15. Jaddu, H. *Direct solution of nonlinear optimal control problems using quasilinearization and Chebyshev polynomials*, Journal of the Franklin Institute, 339 (2002), 479-498.

16. Jaddu, H. and Shimemura, E. *Computation of optimal control trajectories using Chebyshev polynomials: parameterization and quadratic programming*, Optimal Control Appl Methods, 20 (1999), 21-42.

17. Junkins, J.L. and Turner, J.D. *Optimal Spacecraft Rotational Maneuvers*, Amsterdam: Elsevier, 1986.

18. Kirk, D.E. *Optimal control theory*, Englewood Cliffs, NJ: Prentice Hall, 1970.

19. Kleiman, D.L., Fortmann, T. and Athans, M. *On the design of linear systems with piecewise-constant feedback gains*, IEEE Trans Automat Control, 13 (1968), 354-361.

20. Lancaster, P. *Theory of Matrices*, New York: Academic Press, 1969.

21. Lakestani, M., Dehghan, M. and Irandoust-pakchin, S. *The construction of operational matrix of fractional derivatives using B-spline functions*, Commun Nonlinear Sci Numer Simulat, 17 (2012), 1149-1162.

22. Lakestani, M., Razzaghi, M. and Dehghan, M. *Solution of nonlinear fredholm-hammerstein integral equations by using semiorthogonal spline wavelets*, Hindawi Publishing Corporation Mathematical Problems in Engineering, 1 (2005), 113-121.

23. Lakestani, M., Razzaghi, M. and Dehghan, M. *Semiorthogonal spline wavelets approximation for fredholm integro-differential equations*, Hindawi Publishing Corporation Mathematical Problems in Engineering, 1 (2006), 1-12.

24. Leitman, G. *The Calculus of Variations and Optimal Control*, New York: Springer, 1981.

25. Marzban, H.R. and Razzaghi, M. *Hybrid functions approach for linearly constrained quadratic optimal control problems*, Appl Math Modell, 27 (2003), 471-485.

26. Marzban, H.R. and Razzaghi, M. *Rationalized Haar approach for nonlinear constrined optimal control problems*, Appl Math Modell, 34 (2010), 174-183.

27. Marzban, H.R. and Hoseini, S.M. *A composite Chebyshev finite difference method for nonlinear optimal control problems*, Commun Nonlinear Sci Numer Simulat, 18 (2013), 1347-1361.

28. Mashayekhi, S., Ordokhani, Y. and Razzaghi M. *Hybrid functions approach for nonlinear constrained optimal control problems*, Commun Nonlinear Sci Numer Simulat, 17 (2012), 1831-1843.

29. Ordokhani, Y. and Razzaghi, M. *Linear quadratic optimal control problems with inequality constraints via rationalized Haar functions*, Dyn Contin Discrete Impul Syst Ser B, 12 (2005), 761-773.

30. Razzaghi, M. and Elnagar, G.*Linear quadratic optimal control problems via shifted Legendre state parameterization* , Internat. J. Systems Sci., 25 (1994), 393-399.

31. Teo, K.L. and Wong, K.H. *Nonlinearly constrained optimal control problems*, J. Austral Math Soc. Ser B, 33 (1992), 507-530.

32. Yen, V. and Nagurka, M. *Optimal control of linearly constrained linear systems via state parameterization*, Optimal Control Appl Methods, 13 (1992), 155-167.

Iranian Journal of Numerical Analysis and Optimization
Vol. 6, No. 2, (2016), pp 39-50

Convergence of approximate solution of delay Volterra integral equations

M. Zarebnia * and L. Shiri

Abstract

In this paper, sinc-collocation method is discussed to solve Volterra functional integral equations with delay function $\theta(t)$. Also the existence and uniqueness of numerical solutions for these equations are provided. This method improves conventional results and achieves exponential convergence. Numerical results are included to confirm the efficiency and accuracy of the method.

Keywords: Volterra functional integral equations; delay function; sinc-collocation.

1 Introduction

Delay integral equations arise widely in scientific fields such as physics, biology, ecology, control theory, etc. Due to the practical application of these equations, they must be solved successfully with efficient numerical approaches. In recent years, there have been extensive studies in convergence properties and stability analyses of these numerical methods, see, for example, [10]. The numerical solutions of integral equations with delays have also been discussed by several authors such as Brunner [1], Li and Kuang [5], Linz and Wang [6].

Sinc methods for approximating the solutions of Volterra integral equations have received considerable attention mainly due to their high accuracy. These approximations converge rapidly to the exact solutions as the number of sinc points increases. Systematic introduction of these methods can be found in [9]. In [11] sinc-collocation method is emplyed to solve Volterra

*Corresponding author
Received 29 June 2015; revised 15 Desember 2015; accepted 6 January 2016
M. Zarebnia
Department of Applied Mathematics, Faculty of Mathematical Sciences, University of Mohaghegh Ardabili, Ardabil, Iran. e-mail: zarebnia@uma.ac.ir

L. Shiri
Department of Applied Mathematics, Faculty of Mathematical Sciences, University of Mohaghegh Ardabili, Ardabil, Iran. e-mail: l.shiri@uma.ac.ir

integral equations with smooth kernels. The analytical and numerical techniques used in these works can be extended to delay integral equations.

The main objective of the current study is to implement the sinc-collocation method for Volterra functional integral equation of the form

$$y(t) = g(t) + (\mathcal{V}y)(t) + (\mathcal{V}_\theta y)(t), \qquad t \in I := [0, T]. \tag{1}$$

The Volterra delay integral operators $\mathcal{V}$ and $\mathcal{V}_\theta$ (from $C(I) \to C(I)$) describing these equations are defined by

$$(\mathcal{V}y)(t) := \int_0^t K_1(t, s)y(s)ds$$

and

$$(\mathcal{V}_\theta y)(t) := \int_0^{\theta(t)} K_2(t, s)y(s)ds,$$

respectively, and the delay function θ is subject to the following conditions:
(D1) $\theta(0) = 0$, and θ is strictly increasing on the interval I;
(D2) $\theta(t) \leq \overline{q}t$, $t \in I$, for some $\overline{q} \in (0, 1)$;
(D3) $\theta \in C^d(I)$ for some $d \geq 0$.

We will refer to a θ that satisfies (D1) as a vanishing delay function (or, in short, a *vanishing delay*). The linear case, $\theta(t) = qt = t - (1 - q)t =: t - \tau(t)$ $(0 < q < 1)$ (proportional delay) is also known as the pantograph delay function [4]. In this paper we consider vanishing delay but our methods can be use with nonvanishing delay too.

The layout of this paper is as follows. In Section 2, the solvability of equation (1) is stated. Section 3 outlines some of the main properties of sinc function that is necessary for the formulation of the delay integral equation. Sinc-collocation method is considered in Section 4. In section 5, we analyze the existence and uniqueness of numerical solutions. In Section 6, the order of scheme convergence using the new approach is described. Finally, Section 7 contains the numerical experiments.

2 Existence and uniqueness of solutions

In the present section, we state the solvability of integral equations with vanishing delay. The following theorem generalizes Volterras 1897 classical result on the existence and uniqueness of solutions for the equation (1) with $\theta(t) = qt$ $(0 < q < 1)$.

Theorem 1. *[3] Assume that the given functions g, K_1 and K_2 in (1) satisfy 1) $g \in C(I)$, $K_1 \in C(D)$, and $K_2 \in C(D_\theta)$, where*

$$D := \{(t,s) : 0 \le s \le t \le T\}, \quad D_\theta := \{(t,s) : 0 \le s \le \theta(t), t \in I\};$$

2) $\theta(t)$ is subject to the assumptions (D1)-(D3).
Then for each $g \in C(I)$ there exists a unique function $y \in C(I)$ which solves the equation (1) on I.

3 Review of the sinc approximation

In this section, we will review sinc function properties, sinc quadrature rule, and the sinc method. These are discussed thoroughly in [9]. For any $h > 0$, the sinc basis functions are given by

$$S(j,h)(z) = \mathrm{sinc}(\frac{z - jh}{h}), \qquad j = 0, \pm 1, \pm 2, \ldots,$$

where

$$\mathrm{sinc}(z) = \begin{cases} \frac{\sin(\pi z)}{\pi z}, \, z \neq 0; \\ 1, \, z = 0. \end{cases}$$

The sinc function form for the interpolating point $z_k = kh$ is given by

$$S(j,h)(kh) = \begin{cases} 1, \, k = j; \\ 0, \, k \neq j. \end{cases}$$

They are based in the infinite strip D_d in the complex plane

$$D_d = \{w = u + iv : |v| < d\}.$$

To construct approximation on the interval $[0, T]$, we consider the conformal map

$$\phi(z) = \ln(\frac{z}{T - z}).$$

The map ϕ carries the eye-shaped region

$$D = \big\{z \in \mathcal{C} : |\arg(\frac{z}{T - z})| < d\big\}.$$

The function

$$z = \phi^{-1}(w) = \frac{Te^w}{1 + e^w}$$

is an inverse mapping of $w = \phi(z)$. We define the range of ϕ^{-1} on the real line as

$$\Gamma = \{\psi(u) = \phi^{-1}(u) \in D : -\infty < u < \infty\}.$$

The sinc grid points $z_k \in (0,T)$ in D will be denoted by x_k because they are real. For the evenly spaced nodes $\{kh\}_{k=-\infty}^{\infty}$ on the real line, the image which corresponds to these nodes is denoted by

$$x_k = \phi^{-1}(kh) = \frac{Te^{kh}}{1+e^{kh}}, \qquad k = \pm 1, \pm 2, \ldots.$$

Definition 1. Let D be a simply connected domain which satisfies $(a,b) \subset D$ and α and c_1 be a positive constant. Then $\mathcal{L}_\alpha(D)$ denotes the family of all functions $u \in \mathrm{Hol}(D)$ which satisfy

$$|u(z)| \leqslant c_1 |Q(z)|^\alpha \tag{2}$$

for all z in D where $Q(z) = (z-a)(b-z)$.

The next theorem shows the exponential convergence of the sinc approximation.

Theorem 2. *Let $u \in \mathcal{L}_\alpha(D)$, let N be a positive integer, and let h be selected by the formula $h = \sqrt{\frac{\pi d}{\alpha N}}$, then there exists positive constant c_2, independent of N, such that*

$$\sup_{t\in(a,b)} |u(t) - \sum_{j=-N}^{N} u(t_j) S(j,h)(\phi(t))| \leqslant c_2 \sqrt{N} e^{-\sqrt{\pi d \alpha N}}.$$

The error analysis of the sinc indefinite integration has been given in [7].

Theorem 3. *Let $uQ \in \mathcal{L}_\alpha(D)$ for d with $0 < d < \pi$. Let $h = \sqrt{\frac{\pi d}{\alpha N}}$. Then there exists a constant c_2, which is independent of N, such that*

$$\sup_{t\in(a,b)} \left| \int_a^t u(s)ds - h \sum_{j=-N}^{N} \frac{u(t_j)}{\phi'(t_j)} J(j,h)(\phi_{SE}(t)) \right| \leqslant c_3 e^{-\sqrt{(\pi d \alpha N)}}, \tag{3}$$

where

$$J(j,h)(x) = \frac{1}{2} + \int_0^{\frac{x}{h}-j} \frac{\sin(\pi t)}{\pi t} dt.$$

4 Sinc-collocation method

In this section, we apply sinc-collocation method to solve equation (1) which we state again for the convenience of the reader:

$$y(t) = g(t) + \int_0^t K_1(t,s)y(s)ds + \int_0^{\theta(t)} K_2(t,s)y(s)ds$$

if $t = 0$ we have $y(0) = g(0)$. For ease of calculation, we employ the transformation

$$u(t) = y(t) - \frac{T-t}{T}g(0),$$

in this case $u(0) = 0$. Then the above problem becomes

$$u(t) = f(t) + \int_0^t K_1(t,s)u(s)ds + \int_0^{\theta(t)} K_2(t,s)u(s)ds \tag{4}$$

where

$$f(t) := g(t) - \frac{1}{T}(T-t)g(0) + \frac{1}{T}g(0)\left\{\int_0^t K_1(t,s)(T-s)ds + \int_0^{\theta(t)} K_2(t,s)(T-s)ds\right\}.$$

Now, let $u(x)$ be the exact solution of (4) that is approximated by the following expansion

$$u_n(t) = \sum_{j=-N}^{N} u(t_j)S(j,h)o\phi(t) + u(t_{N+1})w(t), \tag{5}$$

we choose $w(t)$ so that above formula interpolate function u at the points t_j, so

$$w(t) = \frac{1}{T}\left(t - \sum_{j=-N}^{N} t_j S(j,h)(\phi(t))\right)$$

where the points t_j are defined by

$$t_j = \begin{cases} \phi^{-1}(jh),\ j = -N, \ldots, N; \\ T,\ j = N+1. \end{cases}$$

By replacing approximate solution (5) in $t = t_k$ in the equation (4), it follows that

$$\begin{aligned} &\sum_{j=-N}^{N} u_j S(j,h)(\phi(t_k)) + u_{N+1}w(t_k) \\ &= \sum_{j=-N}^{N} u_j \int_0^{t_k} K_1(t_k,s)S(j,h)(\phi(s))ds + u_{N+1}\int_0^{t_k} K_1(t_k,s)w(s)ds \\ &+ \sum_{j=-N}^{N} u_j \int_0^{\theta(t_k)} K_2(t_k,s)S(j,h)(\phi(s))ds + u_{N+1}\int_0^{\theta(t_k)} K_2(t_k,s)w(s)ds \\ &+ f(t_k). \end{aligned} \tag{6}$$

We are interested in approximating the integral in above equation by the quadrature formula presented in (3). Then by using Theorem 3, we obtain

$$\int_0^{t_k} K_1(t_k,s)S(j,h)o\phi(s)ds \approx h\frac{K_1(t_k,t_j)}{\phi'(t_j)}J(j,h)(\phi(t_k)), \quad k=-N,\ldots,N+1.$$

From definition of t_k we can write

$$J(j,h)(\phi(t_k)) = \begin{cases} J(j,h)(kh),\ k=-N,\ldots,N; \\ 1,\ k=N+1. \end{cases}$$

The analogue of above equation we have

$$\int_0^{\theta(t_k)} K_2(t,s)S(j,h)o\phi(s)ds \approx h\frac{K_2(t_k,t_j)}{\phi'(t_j)}J(j,h)(\phi_k), \quad k=-N,\ldots,N+1$$

in which $\phi_k := \phi(\theta(t_k))$, in next section these formula will be discussed. Finally, let

$$\begin{aligned} K_{1k} &= \int_0^{t_k} K_1(t_k,s)w(s)ds, \\ K_{2k} &= \int_0^{\theta(t_k)} K_2(t_k,s)w(s)ds, \\ b_k &= K_{1k}+K_{2k}, \qquad k=-N,\ldots,N+1. \end{aligned} \tag{7}$$

By using relation (3), we can approximate b_k in the following form

$$\begin{aligned} K_{1,k} &:= \int_0^{t_k} K_1(t_k,s)w(s)ds \\ &= \frac{1}{T}\int_0^{t_k} K_1(t,s)\left(s-\sum_{j=-N}^{N} t_j S(j,h)(\phi(s))\right)ds \\ &= \frac{1}{T}\left(\int_0^{t_k} sK_1(t,s)ds - \sum_{j=-N}^{N} t_j \int_0^{t_k} K_1(t,s)S(j,h)(\phi(s))ds\right) \\ &= \frac{1}{T}\left(\int_0^{t_k} sK_1(t,s)ds - h\sum_{j=-N}^{N} t_j K_1(t_k,t_j)\frac{1}{\phi'(t_j)}J(j,h)(\phi(t_k))\right) \end{aligned} \tag{8}$$

Thus equation (6) is written as

$$\begin{aligned} &u_k - h\sum_{j=-N}^{N}\frac{1}{\phi'(t_j)}\left\{K_1(t_k,t_j)J(j,h)(\phi(t_k)) + K_2(t_k,t_j)J(j,h)(\phi_k)\right\}u_j \\ &-b_k u_{N+1} = f(t_k). \end{aligned} \tag{9}$$

This linear system of equations is equivalent to (4). By solving this system, the unknown coefficients u_j are determined. We rewirte the linear system (7) in matrix form

$$[\mathcal{I} - \mathcal{A}]\mathbf{U}_N = \mathbf{F} \tag{10}$$

where

$$\mathcal{A}_{k,j} = \frac{h}{\phi'(t_j)} \left\{ K_1(t_k, t_j) J(j,h)(\phi(t_k)) + K_2(t_k, t_j) J(j,h)(\phi_k) \right\},$$
$$k = -N, \dots N+1, \quad j = -N, \dots N,$$

$$\mathcal{A}_{k,N+1} = b_k, \qquad k = -N, \dots N+1,$$
$$\mathbf{U}_N = [u_{-N}, \dots, u_{N+1}]^t, \qquad \mathbf{F} = [f(t_{-N}), \dots, f(t_{N+1})]^t.$$

5 Existence and uniqueness of the sinc-collocation solution

In this section, we study the existence and uniqueness of the solution to (8).

Lemma 1. *For $x \in \mathbb{R}$, the function $J(j,h)(x)$ is bounded by*

$$|J(j,h)(x)| \leqslant 1.1.$$

Theorem 4. *Assume that K_1, K_2 and f in the Volterra integral equation (4) are continuous on their respective domains D, D_θ and I. Then there exists an $\overline{h} > 0$ so that for any $h \in (0, \overline{h})$ the linear algebraic system (8) has a unique solution $\boldsymbol{U}_N$.*

Proof. We know that

$$\mathcal{A}_{k,j} = \frac{h}{\phi'(t_j)} \left\{ K_1(t_k, t_j) J(j,h)(\phi(t_k)) + K_2(t_k, t_j) J(j,h)(\phi_k) \right\},$$
$$k = -N, \dots N+1, \quad j = -N, \dots N,$$
$$\mathcal{A}_{k,N+1} = b_k, \qquad k = -N, \dots N+1,$$

so we can write

$$\|\mathcal{A}\|_\infty = \max_{k=-N,\dots N+1} \Big\{ h \sum_{j=-N}^{N} \frac{1}{\phi'(t_j)} \left| J(j,h)(\phi(t_k)) K_1(t_k, t_j) + K_2(t_k, t_j) J(j,h)(\phi_k) \right| + |b_k| \Big\}.$$

Using Lemma 1, we have

$$\|\mathcal{A}\|_\infty \leqslant \max_{k=-N,\dots N+1} \{1.1h \sum_{j=-N}^{N} \frac{1}{\phi'(t_j)} |K_1(t_k,t_j)+K_2(t_k,t_j)| + |b_k|\}.$$

Theorem 3 and continuty K_1 and K_2 and equations (7) and (8) give

$$|b_k| \leqslant ce^{-\sqrt{\pi d\alpha N}}.$$

Therefore

$$\|\mathcal{A}\|_\infty \leqslant 1.1h \sum_{j=-N}^{N} \frac{1}{\phi'(t_j)} |K_1(t_k,t_j)+K_2(t_k,t_j)| + ce^{-\sqrt{\pi d\alpha N}}.$$

Thus the elements of the matrice $\mathcal{A}$ are all bounded. The Neumann Lemma then shows that the inverse of the matrix $\mathcal{I}-\mathcal{A}$ exists whenever $\|\mathcal{A}\|_\infty < 1$. This clearly holds whenever h is sufficiently small. In other words, there is an $\overline{h} > 0$ so that for any $h < \overline{h}$ matrix $\mathcal{A}$ has a uniformly bounded inverse. The assertion of Theorem 4 now follows. □

6 Convergence analysis

The convergence of the sinc-collocation method which is introduced in the previous sections is discussed in the present section. It is assumed that u is the exact solution of Eq. (4) and $\mathcal{U}_N$ is an approximation of the sinc method. Firstly, we state the following lemma which is used subsequently.

Lemma 2. *([8]) Let $h > 0$. Then it holds that*

$$\sup_{x\in\mathbb{R}} \sum_{j=-N}^{N} |S(j,h)(x)| \leqslant \frac{2}{\pi}(3+\ln N).$$

In the following theorem, we will find an upper bound for the error.

Theorem 5. *Let $\mathcal{U}_N(x)$ is the approximate solution of integral equation (4). Then there exists a constant c_5 independent of N such that*

$$\sup_{x\in(0,T)} |u(x)-\mathcal{U}_N(x)| \leqslant c_5\sqrt{N}\ln N e^{-\sqrt{\pi d\alpha N}}. \tag{11}$$

Proof. For collocation error $e := u - \mathcal{U}_N$

$$\sup_{t\in(0,T)}\left|e(t)-\sum_{j=-N}^{N}S(j,h)(\phi(t))e_j+e_{N+1}w(t)\right|$$

$$\leqslant \sup_{t\in(0,T)}\left|e(t)-\sum_{j=-N}^{N}S(j,h)(\phi(t))e_j\right|+|e_{N+1}|\sup_{t\in(0,T)}|w(t)|$$

$$\leqslant c_1\sqrt{N}e^{-\sqrt{\pi d\alpha N}}+|e_{N+1}|\frac{1}{T}\sup_{t\in(0,T)}\left|t-\sum_{j=-N}^{N}t_jS(j,h)(\phi(t))\right|$$

$$\leqslant c_1\sqrt{N}e^{-\sqrt{\pi d\alpha N}}+|e_{N+1}|\frac{1}{T}c_1'\sqrt{N}e^{-\sqrt{\pi d\alpha N}}$$

$$\leqslant c\sqrt{N}e^{-\sqrt{\pi d\alpha N}},$$

so we can write $e(t)=\sum_{j=-N}^{N}S(j,h)(\phi(t))e_j+e_{N+1}w(t)+c\sqrt{N}e^{-\sqrt{\pi d\alpha N}}$.

For $t=t_j$ it satisfies the error equation

$$e(t)=(\mathcal{V}e)(t)+(\mathcal{V}_\theta e)(t).$$

The contribution of $\mathcal{V}$ in the above error equation is described by

$$(\mathcal{V}e)(t_j)=\int_0^{t_j}K_1(t_j,s)e(s)ds$$

$$=\int_0^{t_j}K_1(t_j,s)\left\{\sum_{k=-N}^{N}S(j,h)(\phi(s))e_k+e_{N+1}w(s)+c\sqrt{N}e^{-\sqrt{\pi d\alpha N}}\right\}$$

$$=h\sum_{k=-N}^{N}\frac{K_1(t_j,t_k)}{\phi'(t_k)}J(j,h)(\phi(t_k))e_k+K_{1j}e_{N+1}+c_2e^{-\sqrt{\pi d\alpha N}}+t_jc\sqrt{N}e^{-\sqrt{\pi d\alpha N}}$$

$$=h\sum_{k=-N}^{N}\frac{K_1(t_j,t_k)}{\phi'(t_k)}J(j,h)(\phi(t_k))e_k+K_{1j}e_{N+1}+c'\sqrt{N}e^{-\sqrt{\pi d\alpha N}},$$

also for delay operator $\mathcal{V}_\theta$ we have

$$(\mathcal{V}_\theta e)(t_j)=\int_0^{\theta(t_j)}K_2(t_j,s)e(s)ds$$

$$=h\sum_{k=-N}^{N}\frac{K_2(t_j,t_k)}{\phi'(t_k)}J(j,h)(\phi_k)e_k+K_{2j}e_{N+1}+c''\sqrt{N}e^{-\sqrt{\pi d\alpha N}}.$$

Thus, the representation of e_j has the form

$$e_j = h \sum_{k=-N}^{N} \frac{1}{\phi'(t_k)} \{K_1(t_j, t_k)J(j,h)(\phi(t_k)) + K_2(t_j, t_k)J(j,h)(\phi_k)\} e_k + b_j e_{N+1} + c\sqrt{N}e^{-\sqrt{\pi d\alpha N}}$$

we may write the collocation equation as

$$[\mathcal{I} - \mathcal{A}]\mathbf{e} = c\sqrt{N}e^{-\sqrt{\pi d\alpha N}}\mathbf{I}.$$

Here, $\mathcal{I}$ and $\mathbf{I}$ denotes the identify matrix and constant vector 1, repectively. It thus follows from Theorem 4 that uniform bound exists for $(\mathcal{I} - \mathcal{A})^{-1}$, so

$$\|\mathbf{e}\| \leqslant c\sqrt{N}e^{-\sqrt{\pi d\alpha N}}. \tag{12}$$

Hence, by Lemma 2 and (12) we can obtain the upper bound (11). □

7 Illustrative examples

In this section, the theoretical results of the previous sections are used for two numerical examples. The numerical experiments are implemented in *Matlab*.

Example 1. The pantograph Volterra integral equation

$$y(t) = g(t) + \int_{\theta(t)}^{t} K(t,s)y(s)ds$$

with $\theta(t) = qt$, $k(t,s) = ts$, and $g(t) = (-t^2 + t + 1)e^t + t(qt - 1)e^{qt}$, has the exact solution $y(t) = e^t$. The results are shown in Table 1.

Table 1: Values of $\|E\|_\infty$ for Example 1

$N \setminus q$	0.01	0.09	0.1	0.5	0.99
10	3.7266×10^{-6}	4.5684×10^{-6}	4.1900×10^{-6}	8.8432×10^{-5}	1.7511×10^{-5}
30	5.5006×10^{-10}	3.6148×10^{-9}	5.2450×10^{-9}	4.0946×10^{-7}	7.5530×10^{-8}
50	8.6331×10^{-13}	1.0118×10^{-10}	1.3788×10^{-10}	1.0852×10^{-8}	1.1792×10^{-9}
70	7.9936×10^{-15}	5.4152×10^{-12}	7.3683×10^{-12}	5.8153×10^{-10}	1.0688×10^{-10}
90	3.2196×10^{-15}	4.3609×10^{-13}	5.9374×10^{-13}	4.6795×10^{-11}	8.6006×10^{-12}

Example 2. Consider

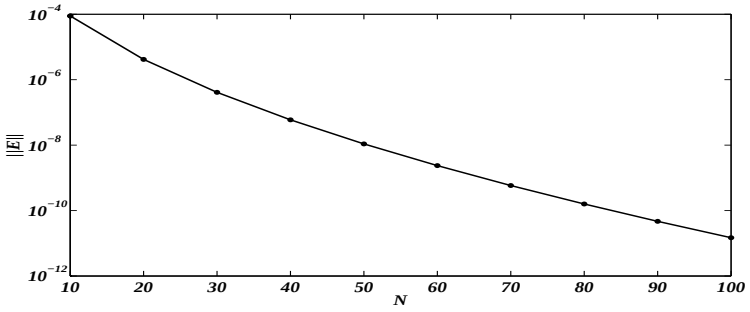

Figure 1: The errors for $q = 0.5$ in Example 1

$$y(t) = g(t) + \int_0^{t^r} K(t,s)y(s)ds$$

with $k(t,s) = s - t$. We choose $g(t)$ so that its exact solution is $y(t) = t - t^2$. Table 2 shows the numerical results.

Table 2: Values of $\|E\|_\infty$ for Example 2

$N \setminus r$	0.01	0.09	0.1	0.5	0.99
10	1.1098×10^{-6}	4.6373×10^{-6}	5.3418×10^{-6}	5.2686×10^{-6}	5.5796×10^{-6}
30	4.0592×10^{-10}	8.1311×10^{-10}	8.1496×10^{-10}	7.9947×10^{-10}	8.1273×10^{-10}
50	8.2652×10^{-13}	1.3631×10^{-12}	1.4027×10^{-12}	1.3289×10^{-12}	1.3898×10^{-12}
70	4.8580×10^{-15}	7.1871×10^{-15}	7.0991×10^{-15}	6.8972×10^{-15}	7.0499×10^{-15}
90	1.4498×10^{-16}	2.2706×10^{-16}	1.9428×10^{-16}	1.6653×10^{-16}	3.3306×10^{-16}

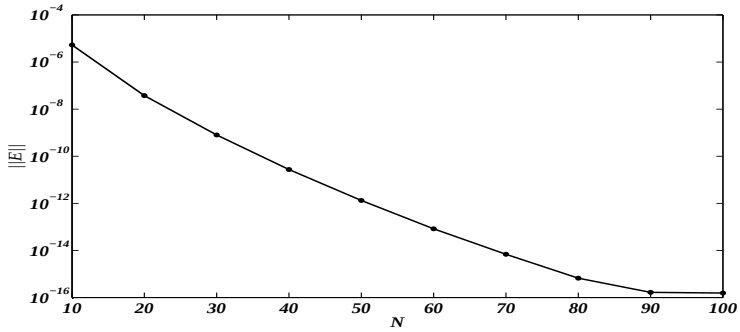

Figure 2: The errors for $r = 0.5$ in Example 2

8 Conclusion

Several methods has been presented for the numerical solution of equation (1) in the special cases for example $\theta(t) = qt$ [2]. We propose a numerical algorithm in order to solve the delay integral equation (Eq. (1)) where θ is general function. Our method has been shown theoretically and numerically to be extremely accurate and achieve exponential convergence with respect to N.

References

1. Brunner, H. *Iterated collocation methods for Volterra integral equations with delay arguments*, Math. Comput. 62 (1994), 581–599.
2. Brunner, H., *Collocation Methods for Volterra Integral and Related Functional Equations*, Cambridge 2004.
3. Denisov, A.M. and Lorenzi, A. *Existence results and regularisation techniques for severely ill-posed integrofunctional equations*, Boll. Un. Mat. Ital. 11 (1997), 713–731.
4. Iserles, A. *On the generalized pantograph functional differential equation*, European J. Appl. Math. 4 (1993), 1–38.
5. Li, Y.K. and Kuang, Y. *Periodic solutions of periodic delay Lotka-Volterra equations and systems*, J. Math. Anal. Appl. 255 (2001), 260–280.
6. Linz, P. and Wang, R.L.C. *Error bounds for the solution of Volterra and delay equations*, Appl. Numer. Math. 9 (1992), 201–207.
7. Okayama, T., Matsuo, T. and Sugihara, M. *Error estimates with explicit constants for Sinc approximation, Sinc quadrature and Sinc indefinite integration*, Mathematical Engineering Technical Reports 2009-01, The University of Tokyo, 2009.
8. Stenger, F. *Handbook of Sinc Numerical Methods*, Springer, New York 2011.
9. Stenger, F. *Numerical Methods Based on Sinc and Analytic Functions*, Springer, New York 1993.
10. Xie, H.H., Zhang, R. and Brunner, H. *Collocation Methods for General Volterra Functional Integral Equations with Vanishing Delays*, SIAM J. Sci. Comp. 33 (2011), 3303–3332.
11. Zarebnia, M. and Rashidinia, J. *Convergence of the Sinc method applied to Volterra integral equations*, Appl. Appl. Math, 5 (2010), 198–216.

Iranian Journal of Numerical Analysis and Optimization
Vol. 6, No. 2, (2016), pp 51-64

Controlling semi-convergence phenomenon in non-stationary simultaneous iterative methods

T. Nikazad* and M. Karimpour

Abstract

When applying the non-stationary simultaneous iterative methods for solving an ill-posed set of linear equations, the error usually initially decreases but after some iterations, depending on the amount of noise in the data, and the degree of ill-posedness, it starts to increase. This phenomenon is called semi-convergence. We study the semi-convergence behavior of the non-stationary simultaneous iterative methods and obtain an upper bound for data error (noise error). Based on this bound, we propose new ways to specify the relaxation parameters to control the semi-convergence. The performance of our strategies is shown by examples taken from tomographic imaging.

Keywords: Simultaneous iterative methods; Semi-convergence; Relaxation parameters; Tomographic imaging.

1 Introduction

A mark-point in the history of medical imaging, was the discovery by Wilhelm Röntgen in 1895 of x-rays [10, 22]. The problem of generating medical images from measurements of the radiation around the body of a patient was considered much later. Hounsfield patented the first CT-scanner in 1972 (and was awarded, together with Cormack, in 1979 the Nobel Prize for this invention). This reconstruction problem belongs to the class of inverse problems, which are characterized by the fact that the information of interest is not directly available for measurements. The imaging device (the camera) provides measurements of a transformation of this information. In practice, these measurements are both imperfect (sampling) and inexact (noise).

*Corresponding author
Received 8 April 2015; revised 6 December 2015; accepted 23 Feburary 2016
T. Nikazad
School of Mathematics, Iran University of Science and Technology, Tehran , Iran. e-mail: tnikazad@iust.ac.ir

M. Karimpour
School of Mathematics, Iran University of Science and Technology, Tehran , Iran. e-mail: mkarimpoursb@yahoo.com

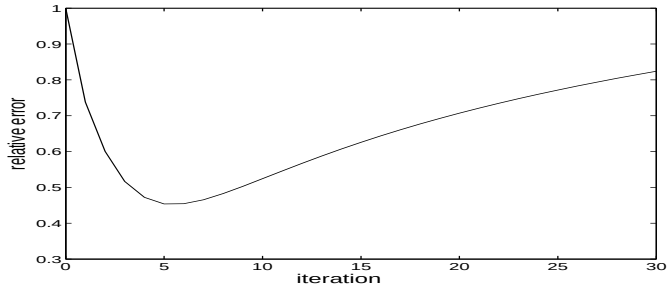

Figure 1: Semi-convergence phenomenon

The mathematical basis for tomographic imaging was laid down by Johann Radon already in 1917 [20]. The word tomography means "reconstruction from slices". It is applied in Computerized (Computed) Tomography (CT) to obtain cross-sectional images of patients. Fundamentally, tomographic imaging deals with reconstructing an image from its projections. The relationship between the unknown distribution (or object) and the physical quantity which can be measured (the projections) is referred to as the forward problem. For several imaging techniques, such as CT, the simplest model for the forward problem involves using the Radon transform R, see [1, 16, 18]. If χ denotes the unknown distribution and β the quantity measured by the imaging device, we have

$$R\chi = \beta.$$

The discrete problem, which is based on expanding χ in a finite series of basis-functions, can be written as

$$Ax \simeq b, \tag{1}$$

where the vector b is a sampled version of β and the vector x, in the case of pixel-(2D) or voxel-(3D) basis, is a finite representation of the unknown object. The matrix $A \in \mathbb{R}^{m\times n}$, typically large and sparse, is a discretization of the Radon transform. An approximative solution to this linear system could be computed by iterative methods, which only require matrix-vector products and hence do not alter the structure of A.

Initially the iteration vectors approach a regularized solution while continuing the iteration often leads to iteration vectors corrupted by noise. This phenomenon is called semi-convergence by Natterer [18]; for analysis of the phenomenon, see, e.g., [1,2,9,11,13,19,21]. The typical overall error behavior is shown in Figure 1.

The Algebraic Reconstruction Technique (ART) is a fully sequential method, and has a long history and rich literature. Originally it was proposed by Kaczmarz [15], and independently, for use in image reconstruction

by [13]. The vector of unknowns is up-dated at each equation of the system, after which the next equation is addressed. In the simultaneous algorithms the current iterate is first projected on all sets to obtain intermediate points, and then the next iterate is made by an averaging process, as convex combination, of intermediate points. The prototype of these algorithms is the well-known Cimmino method [5]. We now explain block-iterative method. The basic idea of a block-iterative algorithm is to partition the data A and b of the system (1) into blocks of equations (rows) and treat each block according to the rule used in the simultaneous algorithm for the whole system, passing, e.g., cyclically over all the blocks, see Figure 2.

An iteration vector of the non-stationary simultaneous iterative method (SIM) is defined as follows

$$x_{k+1} = x_k + \lambda_k A^T M(b - Ax_k), \quad k = 0, 1, \cdots \tag{2}$$

with $x_0 \in \mathbb{R}^n$ where $\{\lambda_k\}_{k=1}^{\infty}$ are relaxation parameters and M is a given symmetric positive definite (SPD) matrix which depends on the particular method. In some papers in image reconstruction from projections, the term "simultaneous iterative reconstruction technique (SIRT)" is used for "SIM"; see, e.g., [7, 8, 21]. Several well-known simultaneous methods can be written as (2) for appropriate choices of the matrix M. With $M = I$ we get the classical Landweber method [17]. Choosing $M = \frac{1}{m}\text{diag}(1/\|a_i\|^2)$ where a_i denotes the ith row of A leads to Cimmino's method [5]. The CAV method [4] uses $M = \text{diag}(1/\sum_{j=1}^{n} N_j a_{ij}^2)$ where N_j is the number of non-zeroes in the jth column of A.

We study semi-convergence behavior of the non-stationary SIM, when applied to noisy data. Our main focus is to propose some techniques for updating relaxation parameters to control the data error. Having a reliable stopping rule leads to stop the iterative method in an iteration which makes a proper approximation of the sought solution. Otherwise, we may stop the iteration process early or far from a proper iteration index. For this reason we introduce relaxation parameters to postpone the semi-convergence phenomenon.

The iteration index of an iterative method may be considered as a regularizing parameter. We explain this a bit more. Let x^* be the sought solution using exact data and let $\bar{x}_k$ and x_k denote the iterate using noisy and exact data respectively. Then we have

$$\|\bar{x}_k - x^*\| \leq \|\bar{x}_k - x_k\| + \|x_k - x^*\|. \tag{3}$$

Therefore, the error decomposes into two components, the data error (or noise error) and the approximation error (or iteration error). The semi-convergence of the iteration interplays between these two error terms.

The semi-convergence behavior of the SIM with constant relaxation parameter is analyzed in [8] where the related M-matrix is a symmetric positive definite (SPD) matrix. Based on this stationary, they suggest two strate-

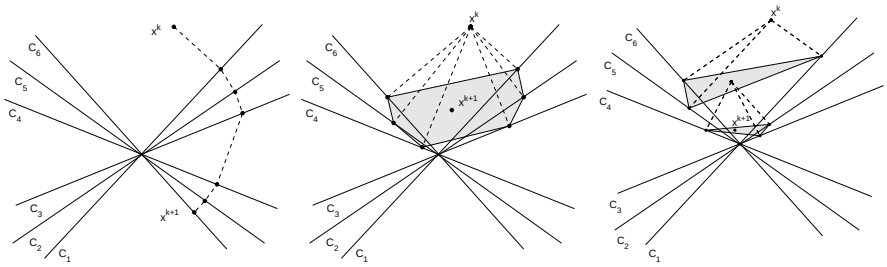

Figure 2: (right to left) sequential method, simultaneous method and sequential block-iterative method

gies for picking relaxation parameters to control the upper bound of data error. The obtained sequence of relaxation parameters is nonnegative and nonascending.

Later in [7], the projected version of the non-stationary SIM is studied where the M-matrix is again assumed SPD. As [8], they consider nonascending sequence of relaxation parameters and emphasize both strategies of [8]. In [7], using nonexpansivity of the projection operator leads to assuming two cases, i.e., the full column-rank problem ($rank(A) = n$) and the rank-deficient problem ($rank(A) < n$) which is handled by a slightly modified problem. Furthermore, they present upper bounds for noise error and iteration error where $rank(A) = n$, see [7, Theorems 3.3 and 3.8] respectively. Also those bounds can be achieved for the modified problem with an unknown regularization parameter (see [7, (3.22),(3.23)]) under some assumptions [7, Lemma 3.9]. In section 2, we give an analysis of non-stationary SIM without having any restriction on $rank(A)$. Additional to strategies given in [8] and [7], we introduce another strategy for choosing relaxation parameters which is able to make more reduction in noise error upper bound comparing with the old strategies.

In Section 3, we consider SIM and give its semi-convergence analysis with three strategies for picking relaxation parameters. We demonstrate the performance of our strategies by examples taken from tomographic imaging in Section 4.

2 Simultaneous iterative algorithm

In this section we give an analysis of the non-stationary SIM without assuming any restriction on $rank(A)$.

Let $\|x\| = \sqrt{x^T x}$ and $\|x\|_M = \sqrt{x^T M x}$ denote the 2-norm and a weighted Euclidean norm respectively. Also, let $M^{1/2}$ and $\rho(Q)$ denote the square root of M and the spectral radius of Q respectively. For $W \in \mathbb{R}^{m\times n}$, we

use $N(W)$ and $R(W)$ to denote the null space and range of W respectively. The orthogonal projection from $\mathbb{R}^n$ onto $N(W)$ is denoted by $P(W)$. Also the orthogonal complement of a subspace K of $\mathbb{R}^n$ is denoted by $K^{\perp}$. Here $x_M(A,b)$ denotes a solution of $\min \|Ax-b\|_M$ with the minimal Euclidean norm.

The convergence analysis of SIM can be obtained in, e.g., [14, Theorem II.3] and [3].

Theorem 1. *Let $\rho = \rho(A^T MA)$ and assume that $0 \leq \epsilon \leq \lambda_k \leq (2-\epsilon)/\rho$. If $\epsilon > 0$, or $\epsilon = 0$ and $\Sigma_{k=0}^{\infty} \min(\rho\lambda_k, 2-\rho\lambda_k) = \infty$, then the iterates of (2) converge to $x_M(A,b) + P(A)x_0$.*

2.1 The error in the *k*-th iteration

As we mentioned before, in this section, we give the same upper bound as [7, Theorems 3.3 and 3.5] but without any restriction on $rank(A)$. Based on our analysis, we give another strategy for choosing relaxation parameters. This strategy is capable to reduce noise error upper bound more than the old strategies given in [8] and [7].

Let $B = A^T MA$, $x^* = x_M(A,b)$ and consider the singular value decomposition (SVD) of $M^{1/2}A$ as

$$M^{1/2}A = U\Sigma V^T$$

where $\Sigma = \mathrm{diag}(\sigma_1, ..., \sigma_p, 0, ..., 0) \in \mathbb{R}^{m\times n}$ with $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_p > 0$ and p is the rank of A. Let $z_k = x_k - x^*$. Using (2) we have

$$z_{k+1} = z_k + \lambda_k A^T M(b - Az_k - Ax^*) = (I - \lambda_k B)z_k$$

which leads to

$$z_k = \prod_{i=0}^{k-1} (I - \lambda_{k-1-i} B)z_0.$$

Since $z_0 = x_0 - x^*$, we obtain

$$x_k = x^* + \prod_{i=0}^{k-1} (I - \lambda_{k-1-i} B)(x_0 - x^*). \tag{4}$$

Using the orthogonal decomposition theorem, we have $\mathbb{R}^n = N(B) \oplus N(B)^{\perp}$ and $N(B)^{\perp} = R(B)$. Therefore we get $x_0 = \hat{x}_0 + P(B)x_0$ where $\hat{x}_0 \in N(B)^{\perp}$ and $P(B)x_0 \in N(B)$. Thus we can rewrite (4) as

$$x_k = x^* + \prod_{i=0}^{k-1}(I - \lambda_{k-1-i}B)\left(\hat{x}_0 + P(B)x_0 - x^*\right). \tag{5}$$

Since $BP(B)x_0 = 0$ and $P(B) = P(A)$, we obtain

$$\prod_{i=0}^{k-1}(I - \lambda_{k-1-i}B)P(B)x_0 = P(B)x_0 = P(A)x_0. \tag{6}$$

Since $\hat{x}_0,\ x^* \in N(B)^{\perp}$, we can rewrite $\hat{x}_0 - x^*$ as

$$\hat{x}_0 - x^* = \sum_{j=1}^{p} c_j v_j \tag{7}$$

where c_j and v_j are scalar and the j-th column of V respectively. Using (5), (6), (7) and the SVD of B, we obtain the following expression for the k-th iteration

$$\begin{aligned} x_k &= x^* + P(A)x_0 + \prod_{i=0}^{k-1}(I - \lambda_{k-1-i}V\Sigma^T\Sigma V^T)(\sum_{j=1}^{p} c_j v_j) = x^* + P(A)x_0 + \\ &+ V\,\mathrm{diag}\left(\prod_{i=0}^{k-1}(1-\lambda_i\sigma_1^2), \ldots, \prod_{i=0}^{k-1}(1-\lambda_i\sigma_p^2), 1, \ldots, 1\right) V^T \sum_{j=1}^{p} c_j v_j \\ &= x^* + P(A)x_0 + \sum_{j=1}^{p}\prod_{i=0}^{k-1}(1-\lambda_i\sigma_j^2)c_j v_j. \end{aligned} \tag{8}$$

Let $\overline{b} = b + \delta b$ and

$$\overline{x}_{k+1} = \overline{x}_k + \lambda_k A^T M(\overline{b} - A\overline{x}_k) \tag{9}$$

where δb is the perturbation consisting of additive noise. Setting $\overline{z}_k = \overline{x}_k - x^*$, we get

$$\begin{aligned} \overline{z}_k &= \overline{z}_{k-1} + \lambda_{k-1}A^T M(b + \delta b - A\overline{z}_{k-1} - Ax^*) \\ &= (I - \lambda_{k-1}B)\overline{z}_{k-1} + \lambda_{k-1}A^T M\delta b \\ &= \prod_{i=0}^{k-1}(I - \lambda_{k-1-i}B)z_0 + \sum_{i=0}^{k-2}\prod_{j=i+1}^{k-1}(I - \lambda_{k-1-j}B)\lambda_i A^T M\delta b + \\ &+ \lambda_{k-1}A^T M\delta b. \end{aligned} \tag{10}$$

Since $\overline{x}_0 = x_0$, similar to (8), we have

$$\overline{x}_k = x^* + P(A)x_0 + \sum_{j=1}^{p}\prod_{i=0}^{k-1}(1-\lambda_i\sigma_j^2)c_j v_j + \\ + \sum_{i=0}^{k-2}\prod_{j=i+1}^{k-1}(I-\lambda_{k-1-j}B)\lambda_i A^T M\delta b + \lambda_{k-1}A^T M\delta b. \tag{11}$$

We now assume that the sequence of relaxation parameters is nonnegative and nonascending, i.e.,

$$0 < \lambda_{i+1} \leq \lambda_i \tag{12}$$

and consider the following function introduced in [8]

$$\Psi^k(\sigma,\lambda) = \frac{1-(1-\lambda\sigma^2)^k}{\sigma}. \tag{13}$$

Theorem 2. *Let* $\omega = \|M^{1/2}\delta b\|$ *and* $0 < \lambda_k \leq \frac{1}{\sigma_1^2}$. *The noise error of SIM is bounded above by*

$$\|\overline{x}_k - x_k\| \leq \frac{\omega\lambda_0\sigma_1}{\lambda_{k-1}\sigma_p}\Psi^k(\sigma_p,\lambda_{k-1}). \tag{14}$$

Proof. By subtracting (8) and (11), we obtain

$$\overline{x}_k - x_k = \sum_{i=0}^{k-2}\prod_{j=i+1}^{k-1}(I-\lambda_{k-1-j}B)\lambda_i A^T M\delta b + \lambda_{k-1}A^T M\delta b. \tag{15}$$

Therefore we have

$$\|\overline{x}_k - x_k\| \leq \sum_{i=0}^{k-2}\lambda_i\left\|\prod_{j=i+1}^{k-1}(I-\lambda_{k-1-j}B)A^T M\delta b\right\| + \lambda_{k-1}\|A^T M\delta b\|. \tag{16}$$

Using the SVD of $M^{1/2}A$, we get that

$$\prod_{j=i+1}^{k-1}(I-\lambda_{k-1-j}B)A^T M^{1/2} = V\prod_{j=i+1}^{k-1}(I-\lambda_{k-1-j}\Sigma^T\Sigma)V^T V\Sigma^T U^T \\ = VW_{i,k}U^T$$

where

$$W_{i,k} = \operatorname{diag}\left(\prod_{j=i+1}^{k-1}(1-\lambda_j\sigma_1^2)\sigma_1,\ldots,\prod_{j=i+1}^{k-1}(1-\lambda_j\sigma_p^2)\sigma_p,0,\ldots,0\right).$$

Using (12) and $0 < \lambda_k \leq \frac{1}{\sigma_1^2}$, we obtain

$$\left\| \prod_{j=i+1}^{k-1} (I - \lambda_{k-1-j} B) A^T M^{1/2} \right\| \leq \|W_{i,k}\| = \max_{1 \leq s \leq p} \left\| \prod_{j=i+1}^{k-1} (1 - \lambda_j \sigma_s^2) \sigma_s \right\| \leq \sigma_1 (1 - \lambda_{k-1} \sigma_p^2)^{k-1-i}. \tag{17}$$

Since $\|A^T M \delta b\| \leq \sigma_1 \omega$, we conclude that, using (12),(17) and the assumptions of theorem,

$$\begin{aligned} \|\overline{x}_k - x_k\| &\leq \sum_{i=0}^{k-2} \lambda_0 \omega \left\| \prod_{j=i+1}^{k-1} (I - \lambda_{k-1-j} B) A^T M^{1/2} \right\| + \lambda_0 \sigma_1 \omega \\ &\leq \sum_{i=0}^{k-2} \lambda_0 \omega \sigma_1 (1 - \lambda_{k-1} \sigma_p^2)^{k-1-i} + \lambda_0 \sigma_1 \omega \\ &= \sum_{s=0}^{k-1} \lambda_0 \omega \sigma_1 (1 - \lambda_{k-1} \sigma_p^2)^s \\ &= \frac{\omega \lambda_0 \sigma_1}{\lambda_{k-1} \sigma_p} \frac{1 - (1 - \lambda_{k-1} \sigma_p^2)^k}{\sigma_p} \\ &= \frac{\omega \lambda_0 \sigma_1}{\lambda_{k-1} \sigma_p} \Psi^k(\sigma_p, \lambda_{k-1}). \end{aligned}$$

This completes the proof. □

Remark 1. To obtain a similar result as (14) where the projected case of (2) is employed, we refer to [7, Theorem 3.3] where it is assumed $rank(A) = n$.

Similar to [8], we consider the equation

$$g_{k-1}(y) = (2k-1)y^{k-1} - (y^{k-2} + \cdots + y + 1) = 0 \tag{18}$$

which has a unique real root $\zeta_k \in (0,1)$. The roots satisfy $0 < \zeta_k < \zeta_{k+1} < 1$ and $\lim_{k \to \infty} \zeta_k = 1$ (see [8, Propositions 2.3, 2.4]), and they can easily be precalculated, see Table 1.

Again, let σ_1 denote the largest singular value of $M^{1/2}A$. Then we have the following alternative upper bound for the noise error.

Theorem 3. *Assume that $\sigma_1 \leq 1/\sqrt{\lambda_{k-1}}$; then*

$$\|x_k - \bar{x}_k\| \leq \frac{\omega \lambda_0 \sigma_1}{\sqrt{\lambda_{k-1}} \sigma_n} \frac{1 - \zeta_k^k}{\sqrt{1 - \zeta_k}}, \tag{19}$$

where ζ_k is the unique root in $(0,1)$ of (18).

Table 1: The unique root $\zeta_k \in (0,1)$ of $g_{k-1}(y) = 0$, cf. (18), as function of the iteration index k

k	ζ_k	k	ζ_k	k	ζ_k	k	ζ_k	k	ζ_k	k	ζ_k
2	0.3333	7	0.8156	12	0.8936	17	0.9252	22	0.9424	27	0.9531
3	0.5583	8	0.8392	13	0.9019	18	0.9294	23	0.9449	28	0.9548
4	0.6719	9	0.8574	14	0.9090	19	0.9332	24	0.9472	29	0.9564
5	0.7394	10	0.8719	15	0.9151	20	0.9366	25	0.9493	30	0.9578
6	0.7840	11	0.8837	16	0.9205	21	0.9396	26	0.9513	31	0.9592

Proof. Using [8, Proposition 2.3] we obtain the following bound for the function $\Psi^k(\sigma, \lambda)$ appearing in (14):

$$\begin{aligned}\max_{1\le i\le n} \Psi^k(\sigma_i, \lambda_{k-1}) &\le \max_{0<\sigma\le\sigma_1} \Psi^k(\sigma, \lambda_{k-1}) \\ &\le \max_{0<\sigma\le 1/\sqrt{\lambda_{k-1}}} \Psi^k(\sigma, \lambda_{k-1}) \le \sqrt{\lambda_{k-1}}\frac{1-\zeta_k^k}{\sqrt{1-\zeta_k}}. \quad (20)\end{aligned}$$

The assumption in the theorem implies

$$\sigma_1 \le 1/\sqrt{\lambda_{k-1}} \quad \Leftrightarrow \quad \lambda_{k-1} \le 1/\sigma_1^2. \quad (21)$$

Then by (14) and (20), and assuming (21), we obtain the bound in (19). □

Remark 2. The case $\lambda_{k-1} \in (1/\sigma_1^2, 2/\sigma_1^2)$ is discussed in [8, Remark 2.2].

3 Choice of relaxation parameters

Using (19), we propose following strategies for choosing relaxation parameters:

$$\Psi_1 - rule: \quad \lambda_k = \begin{cases} \frac{\sqrt{2}}{\sigma_1^2}, & \text{for } k = 0, 1 \\ \frac{2}{\sigma_1^2}(1-\zeta_k), & \text{for } k \ge 2, \end{cases} \quad (22)$$

$$\Psi_2 - rule: \quad \lambda_k = \begin{cases} \frac{\sqrt{2}}{\sigma_1^2}, & \text{for } k = 0, 1 \\ \frac{2}{\sigma_1^2}(1-\zeta_k)(1-\zeta_k^k)^{-2}, & \text{for } k \ge 2 \end{cases} \quad (23)$$

$$\Psi_3 - rule: \quad \lambda_k = \begin{cases} \frac{\sqrt{2}}{\sigma_1^2}, & \text{for } k = 0, 1 \\ \frac{2}{\sigma_1^2}(1-\zeta_k)^{r-1}(1-\zeta_k^k)^2, & \text{for } k \ge 2 \end{cases} \quad (24)$$

where $\{\zeta_k\}_{k\ge 2}$ are the roots of (18) and $1 < r \le 2$.

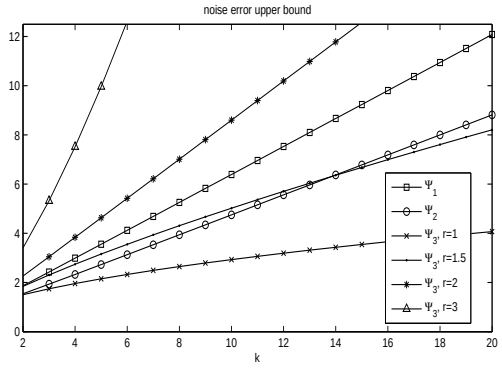

Figure 3: Noise error upper bound (25) for different strategies with the factor ω/σ_p omitted

Remark 3. Using (19) and strategies (22-24), we have the following upper bounds for noise error

$$\|\overline{x}_k - x_k\| \leq \begin{cases} \frac{\omega}{\sigma_p}(1-\zeta_k^k)(1-\zeta_k)^{-1}, & \Psi_1 \\ \frac{\omega}{\sigma_p}(1-\zeta_k^k)^2(1-\zeta_k)^{-1}, & \Psi_2 \\ \frac{\omega}{\sigma_p}(1-\zeta_k)^{-r/2}, & \Psi_3 \end{cases} \tag{25}$$

for $k \geq 2$.

Figure 3 shows the behavior of noise error upper bound (25) for different strategies. As it seen, Ψ_3 with $r = 1$ and Ψ_3 with $r = 3$ give the smallest and largest upper bounds respectively. Furthermore, Ψ_3 with $r = 1.5$ gives smaller upper bound than Ψ_1 and Ψ_2.

Remark 4. It is easy to check that, using [8, Theorems 3.1 and 3.3], the both strategies (22) and (23) satisfy all conditions of Theorem 1. Therefore, the sequence x_k generated by (2) converges to $x_M(A,b) + P(A)x_0$.

Next we will check that the relaxation parameters defined in (24) satisfy all conditions of Theorem 1.

Proposition 1 *The sequence generated by (2) with strategies (24) converges to* $x_M(A,b)$.

Proof. Since $\rho = \sigma_1^2$, we have $0 \leq \rho\lambda_k \leq 2$. Using [8, (2.17), (3.10)], we obtain that

$$\sum_{k\geq 2}(1-\zeta_k)^{r-1}(1-\zeta_k^k)^2 > \sum_{k\geq 2}\left(1-\frac{2k}{2k+1}\right)^{r-1}\left(1-\frac{k-1}{2k-1}\right)^2$$
$$= \sum_{k\geq 2}\left(\frac{1}{2k+1}\right)^{r-1}\left(\frac{k}{2k-1}\right)^2$$
$$> \sum_{k\geq 2}\frac{k^2}{(2k+1)^{r+1}}. \tag{26}$$

It is clear that (26) diverges if $r \leq 2$. Therefore, we have $\Sigma_{k\geq 2}\lambda_k = \infty$. It is easy to check that $\min\{\rho\lambda_k, 2-\rho\lambda_k\} = \rho\lambda_k$ for k sufficiently large. Thus, all conditions of Theorem 1 hold and consequently the sequence x_k generated by (2) converges to $x_M(A,b) + P(A)x_0$. □

4 Numerical results

In this section we give two examples of computerized tomography field. We used 5% and 10% white Gaussian noises to produce noisy data. The constant optimal relaxation parameter λ_{opt} refers to the strategy when a constant value of the relaxation parameter is used, chosen such that it gives rise to the smallest relative error within 20 iterations. For the choices of M matrix in SIM, we always use Cimmino's method. We compare our results with cgne which is a Krylov-type method. The method cgne is sometimes also called cgls. This method is scaled by $M^{1/2}$, i.e., using $M^{1/2}A$, $M^{1/2}b$ instead of A, b.

As we mentioned before, Theorems 2 and 3 and the strategies (22-24) are based on (12). Note that the convergence analysis is not based on the nonascending property. Since $\zeta_k < \zeta_{k+1}$ and using [8, Proposition 3.3], both strategies (22) and (23) satisfy (12). For the the third strategy (24) we have

$$(1-\zeta_k)^{r-1}(1-\zeta_k^k)^2 > (1-\zeta_{k+1})^{r-1}(1-\zeta_k^k)^2$$
$$> (1-\zeta_{k+1})^{r-1}(1-\zeta_{k+1}^{k+1})^2$$

provided that

$$\zeta_{k+1}^{k+1} > \zeta_k^k. \tag{27}$$

We do not have any mathematical proof which shows (27) holds but our numerical tests verifies (27) where $r \geq 1$.

Our first tests are based on the standard head phantom from [13]. We report some numerical tests with an example taken from the field of tomographic image reconstruction from projections, using the SNARK09 software package [6]. The phantom is discretized into 63×63 pixels, and 16 projections (evenly distributed between 0 and 174 degrees) with 99 rays per projection

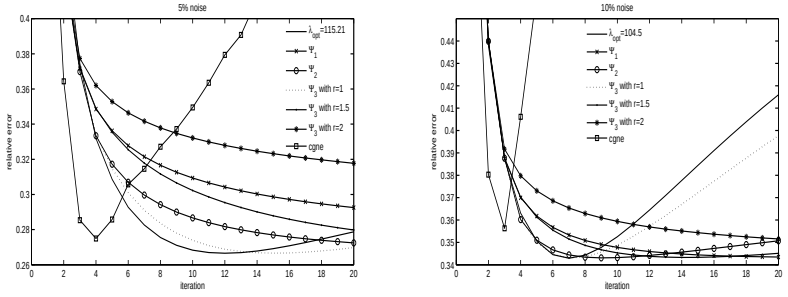

Figure 4: Relative error histories in SIM using small phantom with different relaxation strategies

are used. The resulting matrix A has dimension 1584×3969, so that the system of equations is highly underdetermined. Figure 4 shows the error histories for SIM, using the optimal fixed relaxation parameter as well as Ψ_1, Ψ_2 and Ψ_3 strategies with noisy data.

Based on behavior of noise error upper bound, see Figure 3, using Ψ_3 with $r = 1$ gives the smallest upper bound. This fact is confirmed by Figure 4 (left) where 5% noise is used. But using 10% noise, Figure 4 (right), leads to fast semi-convergence. The reason could be the large value of ω in (19) which is eliminated in all strategies. However, the results of Ψ_3 rule with $r = 1.5$ and Ψ_1 rule are proper where 10% noise is used.

In our second example we used the (matlab-based) package AIRtools [12] to produce the phantom, the matrix and the right-hand side (with and without noise). We again used 5% and 10% white Gaussian noises. The phantom is now discretized into 365×365 pixels. We take 88 projections (evenly distributed between 0 and 179 degrees) with 516 rays per projection. The resulting projection matrix A has dimension 40892×133225, so that again the system of equations is underdetermined. Figure 4 shows the relative error histories of SIM with noisy data. As it is seen, this figure shows that the results of Ψ_3 rule with $r = 1$ are close to the results of optimal rule.

For both noise levels and phantoms, cgne is the fastest method. However it also shows a distinctive semi-convergence behavior making it more dependent on a reliable stopping rule than SIM with our strategies.

Acknowledgments

We thank Tommy Elfving and Per Christian Hansen for their valuable comments. We wish to thank two anonymous referees for constructive criticism and helpful suggestions which improved our paper.

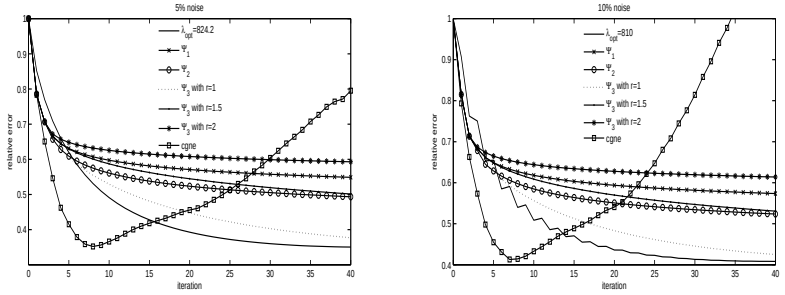

Figure 5: Relative error histories in SIM using the big phantom

References

1. Bertero, M. and Boccacci, P. *Introduction to inverse problems in imaging.* CRC press, 1998.

2. Brianzi, P., Benedetto, F.D. and Estatico, C. *Improvement of space-invariant image deblurring by preconditioned landweber iterations.* SIAM Journal on Scientific Computing, 30(3):1430–1458, 2008.

3. Censor, Y. and Elfving, T. *Block-iterative algorithms with diagonally scaled oblique projections for the linear feasibility problem.* SIAM Journal on Matrix Analysis and Applications, 24(1):40–58, 2002.

4. Censor, Y., Gordon, D. and Gordon, R. *Component averaging: An efficient iterative parallel algorithm for large and sparse unstructured problems.* Parallel computing, 27(6):777–808, 2001.

5. Cimmino, G. and Ricerche, C.N.D. *Calcolo approssimato per le soluzioni dei sistemi di equazioni lineari.* Istituto per le applicazioni del calcolo, 1938.

6. R. Davidi, G. T. Herman, and J. Klukowska. Snark09: A programming system for the reconstruction of 2d images from 1d projections. The CUNY Institute for Software Design and Development, New York, 2009.

7. Tommy Elfving, Per Christian Hansen, and Touraj Nikazad. Semiconvergence and relaxation parameters for projected SIRT algorithms. SIAM Journal on Scientific Computing, 34(4):A2000–A2017, 2012.

8. Elfving, T., Nikazad, T. and Hansen, P.C. *Semiconvergence and relaxation parameters for a class of SIRT algorithms,* Electronic Transactions on Numerical Analysis, 37:321–336, 2010.

9. Engl, H.W., Hanke, M. and Neubauer, A. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

10. Guy C. and Ffytche, D. *An introduction to the principles of medical imaging,* World Scientific, 2005.

11. Hansen, P.C. *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*, volume 4. Siam, 1998.

12. Hansen, P.C. and Hansen, M.S.*AIR tools-a MATLAB package of algebraic iterative reconstruction methods*, Journal of Computational and Applied Mathematics, 236(8):2167–2178, 2012.

13. Herman, G.T. *Fundamentals of computerized tomography: image reconstruction from projections*, Springer Science & Business Media, 2009.

14. Jiang, M. and Wang, G. *Convergence studies on iterative algorithms for image reconstruction*, Medical Imaging, IEEE Transactions on, 22(5):569–579, 2003.

15. Kaczmarz, S. *Angenherte auflsung von systemen linearer gleichungen.* Bulletin International de lAcademie Polonaise des Sciences et des Lettres, 35:355–357, 1937.

16. Kak, A.C. and Slaney, M. *Principles of computerized tomographic imaging*, volume 33. Siam, 1988.

17. Landweber, L. *An iteration formula for Fredholm integral equations of the first kind,* American journal of mathematics, pages 615–624, 1951.

18. Natterer, F. *The mathematics of computerized tomography,* John Wiley, New York, 1986.

19. Piana, M. and Bertero, M. *Projected landweber method and preconditioning,* Inverse Problems, 13(2):441, 1997.

20. Radon, J. *ber die bestimmung von funktionen durch ihre integralwerte lngs gewisser mannigfaltigkeiten,* Classic papers in modern diagnostic radiology, 5, 2005.

21. Sluis, A. and Vorst, H. *Sirt-and cg-type methods for the iterative solution of sparse linear least-squares problems,* Linear Algebra and its Applications, 130:257–303, 1990.

22. Webb, S. *From the watching of shadows: the origins of radiological tomography,* CRC Press, 1990.

Iranian Journal of Numerical Analysis and Optimization
Vol. 6, No. 2, (2016), pp 65-84

Application of modified hat functions for solving nonlinear quadratic integral equations

F. Mirzaee* and E. Hadadiyan

Abstract

A numerical method to solve nonlinear quadratic integral equations (QIE) is presented in this work. The method is based upon modification of hat functions (MHFs) and their operational matrices. By using this approach and the collocation points, solving the nonlinear QIE reduces to solve a nonlinear system of algebraic equations. The proposed method does not need any integration for obtaining the constant coefficients. Hence, it can be applied in a simple and fast technique. Convergence analysis and associated theorems are considered. Some numerical examples illustrate the accuracy and computational efficiency of the proposed method.

Keywords: Modification of hat functions; Nonlinear quadratic integral equation; Vector forms; Operational matrix; Error analysis.

1 Introduction

Over the last years, the integral equations have been used increasingly in different areas of applied science. This tendency could be explained by the deduction of knowledge models which describe real physical phenomena. For details, we refer to [1, 2, 4-9, 12-18, 23, 25, 26]. In particular, quadratic integral equations (QIEs) have many useful applications in the real world. For example, QIEs are often applicable in the theory of radiative transfer, the kinetic theory of gases, the theory of neutron transport, the queuing theory and the traffic theory. The QIEs can be very often encountered in many applications. The quadratic integral equations have been studied in

*Corresponding author
Received 7 May 2015; revised 22 November 2015; accepted 23 Feburary 2016
F. Mirzaee
Faculty of Mathematical Sciences and Statistics, Malayer University, Malayer, Iran. e-mail: f.mirzaee@malayeru.ac.ir

E. Hadadiyan
Faculty of Mathematical Sciences and Statistics, Malayer University, Malayer, Iran. e-mail: e.hadadiyan@gmail.com

several papers and monographs [1, 2, 4-26, 29, 30]. In this paper, we study the numerical solution of a QIE:

$$f(x) = g(x) + \left(\int_0^x k_1(x,y)U_1(y,f(y))dy\right)\left(\int_0^x k_2(x,y)U_2(y,f(y))dy\right), \tag{1}$$

where $x \in D = [0,1]$, $g(x) \in C^3(D)$, $U_1(x,f(x)), U_2(x,f(x)) \in C^3(D \times \mathbb{R})$ and $k_1(x,y), k_2(x,y) \in C^3(D \times D)$ are known functions, $f(x) \in C^3(D)$ is an unknown function and we will obtain the approximate solution in the truncated MHFs series form

$$f_m(x) = \sum_{i=0}^{m} f_i h_i(x),$$

so that f_i; $i = 0, 1, \ldots, m$, are the unknown MHFs coefficients and $h_i(x)$; $i = 0, 1, \ldots, m$, are the MHFs.

To mention some recent works on QIEs, see e.g., [1,8,12,14,15,25,26] and for some applications we refer readers to [18, 23]. Existence, uniqueness and some other properties of the solution to these problems were established in [33]. It should be recalled that nonlinear QIEs have been treated extensively with the measure of noncompactness and a fixed point theorem of Darbo type. This approach seems to be too restrictive. Furthermore, in most of the above investigations, some additional assumptions in terms of the measure of noncompactness were imposed on $g(x)$.

The plan for this paper is as follows: In Section 2, we describe MHFs and their properties. In Section 3, we will apply these sets of MHFs for approximating the solution of QIEs. In Section 4, theorems are proved for convergence analysis. Numerical results are given in Section 5 to illustrate the efficiency and the accuracy of our algorithms. Finally, Section 6 concludes the paper.

2 Modification of hat functions

The purpose of this section is to collect a number of definitions and lemmas concerning MHFs. we first construct the set of MHFs.

An $(m+1)$-set of MHFs consists of $(m+1)$ functions which are defined over district D as follows [3, 29]

$$h_0(x) = \begin{cases} \frac{1}{2h^2}(x-h)(x-2h) & 0 \leqslant x \leqslant 2h, \\ 0 & \text{otherwise}, \end{cases}$$

if i is odd and $1 \leqslant i \leqslant m-1$,

$$h_i(x) = \begin{cases} \frac{-1}{h^2}(x-(i-1)h)(x-(i+1)h) & (i-1)h \leqslant x \leqslant (i+1)h, \\ \\ 0 & \text{otherwise}, \end{cases}$$

if i is even and $2 \leqslant i \leqslant m-2$,

$$h_i(x) = \begin{cases} \frac{1}{2h^2}(x-(i-1)h)(x-(i-2)h) & (i-2)h \leqslant x \leqslant ih, \\ \\ \frac{1}{2h^2}(x-(i+1)h)(x-(i+2)h) & ih \leqslant x \leqslant (i+2)h, \\ \\ 0 & \text{otherwise}, \end{cases}$$

and

$$h_m(x) = \begin{cases} \frac{1}{2h^2}(x-(1-h))(x-(1-2h)) & 1-2h \leqslant x \leqslant 1, \\ \\ 0 & \text{otherwise}, \end{cases}$$

where $m \geqslant 2$ is an even integer and $h = \frac{1}{m}$. It is obvious that

$$h_i(jh) = \begin{cases} 1 & i = j, \\ \\ 0 & i \neq j, \end{cases} \tag{2}$$

$$h_i(x)h_j(x) = \begin{cases} 0 & i \text{ is even and } |i-j| \geqslant 3, \\ \\ 0 & i \text{ is odd and } |i-j| \geqslant 2, \end{cases} \tag{3}$$

and

$$\sum_{i=0}^{m} h_i(x) = 1.$$

Let us write the MHFs vector $H(x)$ as follows

$$H(x) = [h_0(x), h_1(x), \ldots, h_m(x)]^T; \quad x \in D. \tag{4}$$

An arbitrary function $f(x)$ defined over D can be expanded by the MHFs as

$$f(x) \simeq F^T H(x) = H^T(x)F,$$

where

$$F = [f_0, f_1, \ldots, f_m]^T,$$

and

$$f_i = f(ih); \qquad i = 0, 1, \ldots, m.$$

Similarly an arbitrary function of two variables, $k(x, y)$ on district $D \times D$ may be approximated with respect to MHFs such as

$$k(x,y) \simeq H^T(x) K H(y),$$

where $H(x)$ and $H(y)$ are MHFs vector of dimension $(m+1)$ and K is the $(m+1)\times(m+1)$ MHFs coefficients matrix.

According to (2) and expanding $\int_0^x h_i(y)dy,\ i=0,1,\ldots,m$ by MHFs, integration of vector $H(x)$ defined in (4) can be expressed as

$$\int_0^x H(y)dy \simeq PH(x), \tag{5}$$

where P is the $(m+1)\times(m+1)$ matrix as follows

$$P = \frac{h}{12}\begin{pmatrix} 0 & 5 & 4 & 4 & 4 & 4 & 4 & \ldots & 4 & 4 \\ 0 & 8 & 16 & 16 & 16 & 16 & 16 & \ldots & 16 & 16 \\ 0 & -1 & 4 & 9 & 8 & 8 & 8 & \ldots & 8 & 8 \\ 0 & 0 & 8 & 16 & 16 & 16 & 16 & \ldots & 16 & 16 \\ 0 & 0 & 0 & -1 & 4 & 9 & 8 & \ldots & 8 & 8 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 8 & 16 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & -1 & 4 \end{pmatrix}.$$

By considering (2), (3) and expanding entries of $H(x)H^T(x)$ by MHFs, we obtain

$$H(x)H^T(x) \simeq diag(H^T(x)),$$

so we have

$$H(x)H^T(x)F \simeq \widetilde{F}H(x), \tag{6}$$

where F be an $(m+1)$-vector and $\widetilde{F}$ is an $(m+1)\times(m+1)$ diagonal matrix. Also, if A is an $(m+1)\times(m+1)$-matrix, we have

$$H^T(x)AH(x) \simeq H^T(x)\widehat{A}, \tag{7}$$

where $\widehat{A}$ is an $(m+1)$-vector with elements equal to the diagonal entries of matrix A.

3 Basic idea

In this section, we will provide the basic idea. This idea includes of approximating the solution of nonlinear quadratic integral equations (1). To solve this equation, we first consider the approximations

$$\begin{aligned} w_1(x) &= U_1(x, f(x)) \\ &= U_1\left(x, g(x) + \left(\int_0^x k_1(x,y)w_1(y)dy\right)\left(\int_0^x k_2(x,y)w_2(y)dy\right)\right), \\ w_2(x) &= U_2(x, f(x)) \\ &= U_2\left(x, g(x) + \left(\int_0^x k_1(x,y)w_1(y)dy\right)\left(\int_0^x k_2(x,y)w_2(y)dy\right)\right). \end{aligned} \tag{8}$$

We approximate function $w_1(x), w_2(x), k_1(x,y)$ and $k_2(x,y)$ by MHFs,

$$\begin{cases} w_1(x) \simeq W_1^T H(x) = H^T(x)W_1, \\ w_2(x) \simeq W_2^T H(x) = H^T(x)W_2, \\ k_1(x,y) \simeq H^T(x)K1H(y) = H^T(y)K1^T H(x), \\ k_2(x,y) \simeq H^T(x)K2H(y) = H^T(y)K2^T H(x), \end{cases} \tag{9}$$

where W_1, W_2, $K1$ and $K2$ are MHFs coefficients of $w_1(x), w_2(x), k_1(x,y)$ and $k_2(x,y)$, respectively. Substituting (9) in (8), we get

$$\begin{aligned} H^T(x)W_1 \simeq U_1\Big(x, g(x) + &\Big(\int_0^x H^T(x)K1H(y)H^T(y)W_1dy\Big) \\ &\times\Big(\int_0^x H^T(x)K2H(y)H^T(y)W_2dy\Big)\Big), \\ H^T(x)W_2 \simeq U_2\Big(x, g(x) + &\Big(\int_0^x H^T(x)K1H(y)H^T(y)W_1dy\Big) \\ &\times\Big(\int_0^x H^T(x)K2H(y)H^T(y)W_2dy\Big)\Big). \end{aligned}$$

Using (5) and (6), yields

$$\begin{aligned} H^T(x)W_1 &\simeq U_1\left(x, g(x) + \left(H^T(x)K1\widetilde{W_1}PH(x)\right)\left(H^T(x)K2\widetilde{W_2}PH(x)\right)\right), \\ H^T(x)W_2 &\simeq U_2\left(x, g(x) + \left(H^T(x)K1\widetilde{W_1}PH(x)\right)\left(H^T(x)K2\widetilde{W_2}PH(x)\right)\right), \end{aligned}$$

where $\widetilde{W_i} = diag(W_i),\ i = 1, 2$, are an $(m+1)\times(m+1)$ diagonal matrices. From (7), we have

$$\begin{aligned} H^T(x)W_1 &\simeq U_1\left(x, g(x) + \left(H^T(x)\widehat{K1\widetilde{W_1}P}\right)\left(H^T(x)\widehat{K2\widetilde{W_2}P}\right)\right), \\ H^T(x)W_2 &\simeq U_2\left(x, g(x) + \left(H^T(x)\widehat{K1\widetilde{W_1}P}\right)\left(H^T(x)\widehat{K2\widetilde{W_2}P}\right)\right), \end{aligned} \tag{10}$$

where $\widehat{Ki\widetilde{W}_iP}$, $i = 1, 2$, be an $(m+1)$-vector with elements equal to the diagonal entries of matrix $Ki\widetilde{W}_iP$. We can rewrite $\widehat{Ki\widetilde{W}_iP}$, $i = 1, 2$, as follows

$$\widehat{Ki\widetilde{W}_iP} = AiW_i, \quad i = 1, 2, \tag{11}$$

where

$$Ai_{pq} = Ki_{pq}P_{qp}, \quad p, q = 0, 1, \ldots, m.$$

Substituting (11) into (10) and replacing $\simeq$ by $=$, we obtain

$$H^T(x)W_1 = U_1\left(x, g(x) + H^T(x)A1W_1H^T(x)A2W_2\right),$$

$$H^T(x)W_2 = U_2\left(x, g(x) + H^T(x)A1W_1H^T(x)A2W_2\right).$$

Now, using Newton-Cotes nodes as

$$x_i = \frac{2i-1}{2(m+1)}, \qquad i = 1, 2, \ldots, m+1,$$

then

$$H^T(x_i)W_1 = U_1\left(x_i, g(x_i) + H^T(x_i)A1W_1H^T(x_i)A2W_2\right),$$

$$H^T(x_i)W_2 = U_2\left(x_i, g(x_i) + H^T(x_i)A1W_1H^T(x_i)A2W_2\right).$$

We have a system of $(m+1)^2$ nonlinear equations and $(m+1)^2$ unknowns. After solving the above nonlinear system, we can find W_1 and W_2 and then

$$f(x) \simeq f_m(x) = g(x) + H^T(x)A1W_1H^T(x)A2W_2.$$

4 Convergence analysis

In this section, for confirming the accuracy of the proposed scheme in the previous section analytically, we provide an upper bound for difference between the exact solution of (1) and our approximated solution. We show that the MHFs method, is convergent of order $O(h^3)$. We define

$$||g|| = \sup_{x \in D} |g(x)|.$$

Theorem 1. *Suppose $x_i = ih$, $i = 0, 1, \ldots, m$, $g(x) \in C^3(D)$ and $g_m(x)$ be the MHFs expansions of $g(x)$ that defined as $g_m(x) = \sum_{i=0}^{m} g(x_i)h_i(x)$. Also, assume that $e_m = \|g - g_m\|$ where $x \in D$, then*

$$e_m = O(h^3).$$

Proof. According to definition of MHFs, the $g_m(x)$ is the quadratic polynomial interpolation on $[x_{i-2}, x_i)$. Therefore, for the interpolation error on $[x_{i-2}, x_i)$, we have [31]

$$e_{i,m}(x) = g(x) - g_m(x) = \frac{1}{6}\frac{d^3 g(\xi_i)}{dx^3} \prod_{i'=i-2}^{i} (x - x_{i'}); \quad i = 2, 4, \ldots, m,$$

where $x, \xi_i \in [x_{i-2}, x_i)$. Let $v(x) = \prod_{i'=i-2}^{i}(x - x_{i'})$, so

$$\|e_{i,m}\| = \frac{1}{6} \sup_{x\in[x_{i-2},x_i)} \left| \frac{d^3 g(\xi_i)}{dx^3}\ v(x) \right| , \quad i = 2, 4, \ldots, m,$$

or

$$\|e_{i,m}\| \leq \frac{1}{6} \sup_{x\in[x_{i-2},x_i)} \left| \frac{d^3 g(\xi_i)}{dx^3} \right|\ |v(x)| , \quad i = 2, 4, \ldots, m.$$

On the other hand, we have

$$e_m = \sup_{x\in D} |g(x) - g_m(x)| = \max_{i=2,4,\ldots,m} \sup_{x\in[x_{i-2},x_i)} |g(x) - g_m(x)|$$

$$= \max_{i=2,4,\ldots,m} \|e_{i,m}\|.$$

So

$$e_m \leq \frac{1}{6} \max_{i=2,4,\ldots,m} \sup_{x\in[x_{i-2},x_i)} \left| \frac{d^3 g(\xi_i)}{dx^3} \right|\ |v(x)| .$$

Since $|v(x)| \leq \sup_{x\in[x_{i-2},x_i)} \left|\prod_{i'=i-2}^{i}(x - x_{i'})\right|$ and the maximum value of $\left|\prod_{i'=i-2}^{i}(x - x_{i'})\right|$ is obtained at $x = (i - 1 - \frac{\sqrt{3}}{3})h$, we have

$$|v(x)| \leq \frac{2\sqrt{3}h^3}{9}, \qquad \forall x \in [x_{i-2}, x_i).$$

Therefore, it is not difficult to verify that

$$e_m \leq \frac{h^3}{9\sqrt{3}} \| \frac{d^3 g}{dx^3} \| = Ch^3. \tag{12}$$

So

$$e_m = O(h^3).$$

This completes the proof. □

Theorem 2. *Let* $x_i = y_i = ih,\ i = 0, 1, \ldots, m,\ k(x, y) \in C^3(D \times D)$ *and*

$$k_m(x,y)=\sum_{i=0}^{m}\sum_{j=0}^{m}k(x_i,y_j)h_i(x)h_j(y),$$

be the MHFs expansions of $k(x,y)$. *Then*

$$e_m=O(h^3),$$

where $e_m=\|k-k_m\|$ *and* $(x,y)\in D\times D$.

Proof. $k_m(x,y)$ is the quadratic polynomial interpolation of $k(x,y)$ on $\Omega_{ij}=[x_{i-1},x_i)\times[y_{j-1},y_j)$. Therefore for the interpolation error on Ω_{ij}, we have [27]

$$\begin{aligned}e_{ij,m}(x,y)&=k(x,y)-k_m(x,y)\\&=\frac{1}{6}\frac{\partial^3k(\xi_i,y)}{\partial x^3}\prod_{i'=i-2}^{i}(x-x_{i'})+\frac{1}{6}\frac{\partial^3k(x,\eta_j)}{\partial y^3}\prod_{j'=j-2}^{j}(y-y_{j'})\\&\quad-\frac{1}{36}\frac{\partial^6k(\xi_i',\eta_j')}{\partial x^3\partial y^3}\prod_{i'=i-2}^{i}(x-x_{i'})\prod_{j'=j-2}^{j}(y-y_{j'}),\end{aligned}$$

where $i,j=2,4,\ldots,m,\ x,\xi_i,\xi_i'\in[x_{i-1},x_i)$ and $y,\eta_j,\eta_j'\in[y_{j-1},y_j)$. So we have

$$\|e_{ij,m}\|=\frac{1}{6}\sup_{(x,y)\in\Omega_{ij}}\left|\frac{\partial^3k(\xi_i,y)}{\partial x^3}v(x)+\frac{\partial^3k(x,\eta_j)}{\partial y^3}u(y)-\frac{1}{6}\frac{\partial^6k(\xi_i',\eta_j')}{\partial x^3\partial y^3}v(x)u(y)\right|,$$

or

$$\begin{aligned}\|e_{ij,m}\|\le\frac{1}{6}\sup_{(x,y)\in\Omega_{ij}}\Bigg\{&\left|\frac{\partial^3k(\xi_i,y)}{\partial x^3}\right|\ |v(x)|+\left|\frac{\partial^3k(x,\eta_j)}{\partial y^3}\right|\ |u(y)|\\&+\frac{1}{6}\left|\frac{\partial^6k(\xi_i',\eta_j')}{\partial x^3\partial y^3}\right|\ |v(x)|\ |u(y)|\Bigg\},\end{aligned}$$

where $i,j=2,4,\ldots,m,\ v(x)=\prod_{i'=i-2}^{i}(x-x_{i'})$ and $u(y)=\prod_{j'=j-2}^{j}(y-y_{j'})$. On the other hand, we have

$$\begin{aligned}e_m&=\sup_{(x,y)\in D\times D}|k(x,y)-k_m(x,y)|\\&=\max_{i,j=2,4,\ldots,m}\sup_{(x,y)\in\Omega_{ij}}|k(x,y)-k_m(x,y)|=\max_{i,j=2,4,\ldots,m}\|e_{ij,m}\|.\end{aligned}$$

So

$$e_m\le\frac{1}{6}\max_{i,j=2,4,\ldots,m}\sup_{(x,y)\in\Omega_{ij}}\left\{\left|\frac{\partial^3k(\xi_i,y)}{\partial x^3}\right|\ |v(x)|+\left|\frac{\partial^3k(x,\eta_j)}{\partial y^3}\right|\ |u(y)|\right.$$

$$+ \frac{1}{6}\left|\frac{\partial^6 k(\xi_i', \eta_j')}{\partial x^3 \partial y^3}\right| \; |v(x)| \; |u(y)|\Big\} \; .$$

We know,

$$|v(x)| \leq \frac{2\sqrt{3}h^3}{9}, \qquad \forall x \in [x_{i-2}, x_i),$$

and

$$|u(y)| \leq \frac{2\sqrt{3}h^3}{9}, \qquad \forall y \in [y_{j-2}, y_j).$$

Therefore, it is not difficult to verify that

$$e_m \leq \frac{h^3}{9\sqrt{3}}\|\frac{\partial^3 k}{\partial x^3}\| + \frac{h^3}{9\sqrt{3}}\|\frac{\partial^3 k}{\partial y^3}\| + \frac{h^6}{243}\|\frac{\partial^6 k}{\partial x^3 \partial y^3}\| = Ch^3, \tag{13}$$

so

$$e_m = O(h^3).$$

This completes the proof. □

Let $f_m(x), g_m(x), k_{i,m}(x,y)$ and $U_{i,m}(x, f(x))$, for $i = 1, 2$, are the MHFs expansions of $f(x), g(x), k_i(x,y)$ and $U_i(x, f(x))$, respectively. According to Theorems 1, 2 and expression (1), we have

$$\begin{aligned} f_m(x) + O(h^3) &= g_m(x) + O(h^3) \\ &+ \left(\int_0^x \left(k_{1,m}(x,y) + O(h^3)\right)\left(U_{1,m}(y, f_m(y) + O(h^3)) + O(h^3)\right) dy\right) \\ &\times \left(\int_0^x \left(k_{2,m}(x,y) + O(h^3)\right)\left(U_{2,m}(y, f_m(y) + O(h^3)) + O(h^3)\right) dy\right), \end{aligned}$$

where $x \in D$. By ignoring the terms included $O(h^3)$, we have

$$\begin{aligned} f_m(x) = g_m(x) &+ \left(\int_0^x k_{1,m}(x,y) U_{1,m}(y, f_m(y)) dy\right) \\ &\times \left(\int_0^x k_{2,m}(x,y) U_{2,m}(y, f_m(y)) dy\right), \end{aligned} \tag{14}$$

where $x \in D$. Now, assume the following hypotheses:

(M1) Suppose that the error of MHFs is denoted by

$$E_m = \|f - f_m\|.$$

(M2) $\|f\| \leqslant M$.

(M3) The nonlinear term $U_1(x,y)$ and $U_2(x,y)$ satisfies in the Lipschitz and linear growth condition such that

$$|U_1(x, y_1) - U_1(x, y_2)| + |U_2(x, y_1) - U_2(x, y_2)| \leqslant L|y_1 - y_2| \; ,$$

where $(x, y_1), (x, y_2) \in D \times \mathbb{R}$, and

$$|U_1(x, y)| + |U_2(x, y)| \leqslant L(1 + |y|), \quad (x, y) \in D \times \mathbb{R} .$$

(M4) Let

$$\|k_1\| \leqslant M_1,$$

and

$$\|k_2\| \leqslant M_2.$$

(M5) Let

$$e_{1,m} = \|k_1 - k_{1,m}\| \leqslant C_1' h^3,$$

and

$$e_{2,m} = \|k_2 - k_{2,m}\| \leqslant C_2' h^3,$$

where C_1' and C_2' are constants that can be defined as coefficient C in (13).

(M6) Let $2L(C_2' h^3 + M_2)(M_1 + C_1' h^3)(L(1 + M) + C_1 h^3) < 1$, where

$$\|U_1 - U_{1,m}\| \leqslant C_1 h^3,$$

and C_1 is constant that can be defined as coefficient C in (12).

Theorem 3. *Suppose $f(x)$ and $f_m(x)$ be the exact and approximate solutions of (1) respectively. Also, above assumptions of (M1)-(M6) are satisfied, then we have*

$$E_m = O(h^3).$$

Proof. Assume that $w_{i,m}(x)$ and $w_i(x)$ be the approximate and exact solution of (8). we define

$$w_{1,m}(x) = U_{1,m}(x, f_m(x)), \qquad w_{2,m}(x) = U_{2,m}(x, f_m(x)) ,$$

and

$$\hat{w}_{1,m}(x) = U_1(x, f_m(x)), \qquad \hat{w}_{2,m}(x) = U_2(x, f_m(x)) .$$

According to (1) and (14), we have

$$\begin{aligned} f(x) - f_m(x) &= g(x) - g_m(x) \\ &+ \left(\int_0^x k_1(x, y) w_1(y, f(y)) dy \right) \left(\int_0^x k_2(x, y) w_2(y, f(y)) dy \right) \\ &- \left(\int_0^x k_{1,m}(x, y) w_{1,m}(y, f_m(y)) dy \right) \left(\int_0^x k_{2,m}(x, y) w_{2,m}(y, f_m(y)) dy \right) \\ &+ O(h^3), \end{aligned}$$

Therefore,

$$f(x)-f_m(x) = g(x)-g_m(x)+\left(\int_0^x k_1(x,y)w_1(y,f(y))dy\right)$$
$$\times\left[\left(\int_0^x k_2(x,y)w_2(y,f(y))dy\right)-\left(\int_0^x k_{2,m}(x,y)w_{2,m}(y,f_m(y))dy\right)\right]$$
$$+\left(\int_0^x k_{2,m}(x,y)w_{2,m}(y,f_m(y))dy\right)$$
$$\times\left[\left(\int_0^x k_1(x,y)w_1(y,f(y))dy\right)-\left(\int_0^x k_{1,m}(x,y)w_{1,m}(y,f_m(y))dy\right)\right]$$
$$+ O(h^3).$$

So

$$E_m \leqslant e_m + \|x\|^2\ \|k_1\|\ \|w_1\|\ \|k_2w_2 - k_{2,m}w_{2,m}\|$$
$$+\ \|x\|^2\ \|k_{2,m}\|\ \|w_{2,m}\|\ \|k_1w_1 - k_{1,m}w_{1,m}\| + O(h^3).$$

Since $x \in D$, then $\|x\|^2 \leqslant 1$. So

$$E_m \leqslant e_m + \|k_1\|\ \|w_1\|\ \|k_2w_2 - k_{2,m}w_{2,m}\|$$
$$+\ \|k_{2,m}\|\ \|w_{2,m}\|\ \|k_1w_1 - k_{1,m}w_{1,m}\| + O(h^3). \tag{15}$$

We have

$$\|w_i - w_{i,m}\| \leqslant \|w_i - \hat{w}_{i,m}\| + \|\hat{w}_{i,m} - w_{i,m}\| \leqslant LE_m + C_ih^3, \quad i = 1, 2, \tag{16}$$

where C_1 and C_2 are constants that can be defined as coefficient C in (12). Also, we have

$$\|w_{i,m}\| \leqslant \|w_i - w_{i,m}\| + \|w_i\| \leqslant LE_m + C_ih^3 + L(1+M)\ , \tag{17}$$

and

$$\|k_{2,m}\| \leqslant \|k_2 - k_{2,m}\| + \|k_2\| \leqslant C_2'h^3 + M_2\ . \tag{18}$$

Now, according to Theorem 2 and inequalities (14), (15) and assumptions (M4)-(M5), we have

$$\|k_iw_i - k_{i,m}w_{i,m}\| \leqslant \|k_i\|\ \|w_i - w_{i,m}\| + \|w_{i,m}\|\ \|k_i - k_{i,m}\|$$
$$\leqslant M_i(LE_m+C_ih^3)+C_i'h^3(LE_m+C_ih^3+L(1+M))\ . \tag{19}$$

From Theorem 1 and inequalities (15)-(17) and assumptions (M3)-(M4), we can rewrite (15), as follows

$$E_m \leqslant Ch^3 + M_1L(1+M)\Big(L(M_2+C_2'h^3)E_m + C_2h^3(M_2+C_2'h^3)$$
$$+LC_2'h^3(1+M)\Big) + \left(C_2'h^3+M_2\right)\Big(LE_m + C_1h^3 + L(1+M)\Big)$$

$$\times\Big(L(M_1+C_1'h^3)E_m+C_1h^3(M_1+C_1'h^3)+LC_1'h^3(1+M)\Big)+O(h^3),$$

where C is defined in (12). Without loss of generality, ignoring the term included E_m^2 and h^6, we have

$$E_m \leqslant \frac{\Big(C+\Big((M_1LC_2'+M_2LC_1')(1+M)+M_1M_2(C_1+C_2)\Big)L(1+M)\Big)h^3}{1-2L(M_1+C_1'h^3)(M_2+C_2'h^3)(L(1+M)+C_1h^3)}$$

$$+O(h^3).$$

This completes the proof. □

5 Numerical examples

To illustrate the accuracy and efficiency of proposed method, some examples are provided. The algorithms associated with the numerical method were performed using Matlab. We have checked that when more points are used the accuracy improves significantly.

Example 1. Consider the following nonlinear QIE [33]

$$f(x)=\left(x^2+\frac{x^{15}}{1350}\right)+\left(\int_0^x yf^2(y)dy\right)\left(\int_0^x \frac{y^2}{25}f^3(y)dy\right),\quad x\in[0,1], \tag{20}$$

with the exact solution $f(x)=x^2$.

Table 1 and Figure 1 illustrate the error results for this example. Also, we compare the maximum absolute error computed by the present method, repeated trapezoidal (RT) method [33] and Adomian decomposition (AD) method [33] in Table 2.

Example 2. Consider the following nonlinear QIE [33]

$$f(x)=x-(e^x-1)\left(\frac{x^3}{30}+\frac{x^5}{50}\right)+\left(\int_0^x \frac{y^2+1}{10}f^2(y)dy\right)\left(\int_0^x e^{f(y)}dy\right), \tag{21}$$

where $x\in[0,1]$ with the exact solution $f(x)=x$.

Table 3 and Figure 2 illustrate the error results for this example. Also, we compare the maximum absolute error computed by the present method, RT method [33] and AD method [33] in Table 4.

Example 3. Consider the following nonlinear QIE

Table 1: Absolute error for $m = 8, 16, 32$ of $f(x)$ of Equation (18)

Nodes x	Present method		
	m=8	m=16	m=32
x = 0.0	0	0	0
x = 0.1	8.8569950e-13	1.9081958e-17	0
x = 0.2	2.9361930e-13	1.0984269e-14	6.3490879e-15
x = 0.3	2.7433902e-11	4.7872678e-12	3.7166104e-13
x = 0.4	1.4692486e-10	2.1893068e-10	1.1033702e-11
x = 0.5	1.4339294e-09	1.2865681e-10	1.0435374e-11
x = 0.6	1.4664141e-07	1.0643045e-08	6.9546968e-10
x = 0.7	4.2146640e-07	2.6710975e-08	1.5046137e-08
x = 0.8	6.8445266e-06	3.6807176e-07	5.1551159e-08
x = 0.9	3.1263755e-06	2.7156875e-06	1.6706609e-07
x = 1.0	4.2731139e-06	6.9580496e-07	1.1520629e-07

Table 2: Table 2: Comparison of the absolute errors of Example 1

Methods	Maximum error
RT Method	
m = 10	6.38458E-5
m = 100	6.30669E-7
m = 1000	6.30590E-9
AD Method	
q = 5	3.62460E-5
q = 10	9.23545E-7
q = 15	2.35318E-8
Present method	
m = 10	1.15435E-5
m = 100	5.72511E-9
m = 1000	1.05104E-11

Table 3: Absolute error for $m = 8, 16, 32$ of $f(x)$ of Equation (19)

Nodes x	Present method		
	m=8	m=16	m=32
x = 0.0	0	0	0
x = 0.1	1.7137234e-6	3.3597747e-7	2.0934872e-8
x = 0.2	6.0344397e-6	3.7285951e-7	1.1336895e-7
x = 0.3	1.7380694e-5	1.0109144e-6	2.0167842e-7
x = 0.4	8.0824578e-6	3.4573577e-6	2.0948964e-7
x = 0.5	1.1099971e-6	7.1694685e-8	5.0373049e-9
x = 0.6	4.5992070e-5	7.5092211e-6	4.7997155e-7
x = 0.7	8.4403496e-5	5.4418147e-6	1.7002589e-6
x = 0.8	1.9036412e-4	1.1238576e-5	2.3184972e-6
x = 0.9	7.8122715e-5	3.0755442e-5	1.8469227e-6
x = 1.0	1.9011556e-5	2.0920818e-6	2.8234706e-6

Table 4: Comparison of the absolute errors of Example 2

Methods	Maximum error
RT Method	
m = 10	1.07275E-3
m = 100	8.44338E-7
m = 1000	8.44337E-9
AD Method	
q = 5	8.44492E-5
q = 10	8.44338E-7
q = 15	8.44337E-8
Present method	
m = 10	1.25539E-4
m = 100	1.27663E-8
m = 1000	2.35536E-11

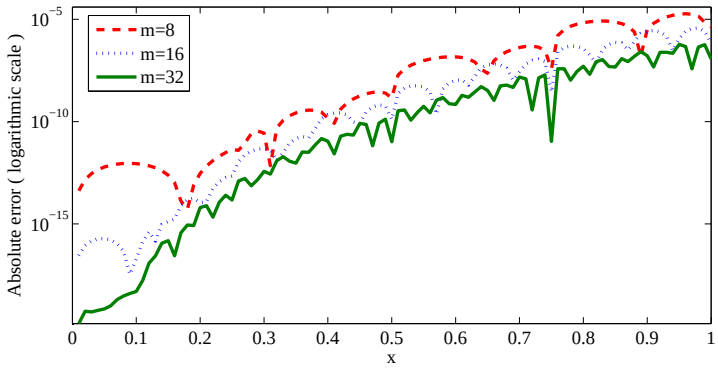

Figure 1: Absolute errors (on logarithmic scale) for Example 1, with $m = 8,\ 16,\ 32$

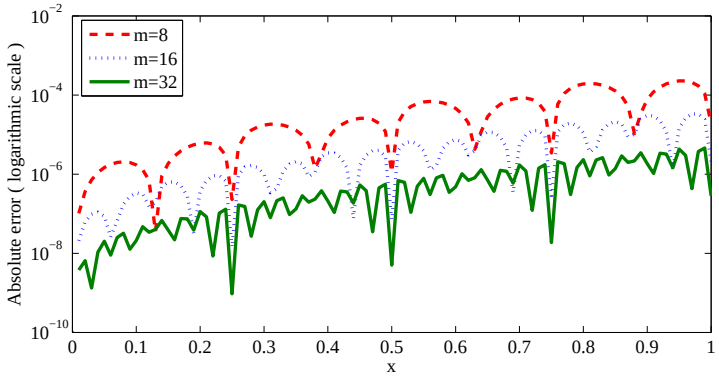

Figure 2: Absolute errors (on logarithmic scale) for Example 2, with $m = 8,\ 16,\ 32$

$$f(x) = g(x) + \left(\int_0^x (y^2+1)f^2(y)dy\right)\left(\int_0^x cos(y)e^{f(y)}dy\right), \quad x \in [0,1], \tag{22}$$

where

$$g(x) = \sin(x) + \left(\frac{x^3}{6} + \frac{x}{2} - \frac{\sin(2x)(1+2x^2)}{8} - \frac{x\cos(2x)}{4}\right)\left(1 - e^{\sin(x)}\right),$$

and the exact solution is $f(x) = \sin(x)$.

Table 5 and Figure 3 illustrate the error results for this example. Also, we compare the maximum absolute error computed by the present method, block-pulse functions (BPFs) method [30] and hat functions (HFs) method [28] in Table 6.

Table 5: Absolute error for $m = 8, 16, 32$ of $f(x)$ of Equation (20)

Nodes x	Present method		
	m=8	m=16	m=32
x = 0.0	0	0	0
x = 0.1	1.6976811e-5	3.1978554e-6	1.9245960e-7
x = 0.2	5.3369972e-5	3.0671554e-6	1.0777490e-6
x = 0.3	1.4115880e-4	9.0002882e-6	1.5689730e-6
x = 0.4	5.0416870e-5	2.5604305e-5	1.6783596e-6
x = 0.5	1.5197939e-5	2.3033416e-6	2.9173156e-7
x = 0.6	2.5006880e-4	3.6295409e-5	2.0906182e-6
x = 0.7	3.2816702e-4	1.8134154e-5	6.9241732e-6
x = 0.8	5.4658908e-4	3.4864263e-5	6.6704623e-6
x = 0.9	7.9549634e-5	5.0605385e-5	2.3233806e-6
x = 1.0	1.9488394e-4	4.0158163e-5	6.8004887e-6

Table 6: Comparison of the absolute errors of Example 3

Methods	Maximum error
BPFs Method	
m = 8	1.15609E-1
m = 16	6.36185E-2
m = 32	3.32452E-2
HFs Method	
m = 8	2.22432E-2
m = 16	5.79464E-3
m = 32	1.51873E-3
Present method	
m = 8	5.57591E-4
m = 16	7.12826E-5
m = 32	1.10246E-5

6 Conclusion

The MHFs method have been proposed for solving nonlinear quadratic integral equations. One of the advantages of this method is that the numerical solution of the nonlinear QIEs can be converted into a system of algebraic equations using the operational matrices. Furthermore, it is proved that MHFs method is convergence and the order of convergence is $O(h^3)$. The

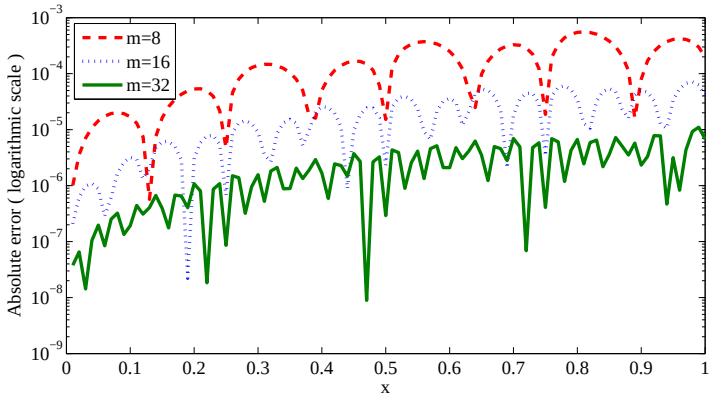

Figure 3: Absolute errors (on logarithmic scale) for Example 3, with $m = 8,\ 16,\ 32$

proposed method does not need any integration for obtaining the constant coefficients hence, it can be applied in a simple and fast technique. The comparison of the obtained results with those based on other methods shows that the present method is a powerful mathematical tool for finding the numerical solutions of such equations.

Acknowledgements

The authors are very thankful to the reviewers and the editor of this paper for their constructive comments and nice suggestions, which helped to improve the paper.

References

1. Argyros, I.K. *On a class of quadratic integral equations with perturbations*, Funct. Approx., 20 (1992) 51-63.

2. Argyros, I.K. *Quadratic equations and applications to Chandrasekhar's and related equations*, Bull. Aust. Math. Soc., 32 (1985) 275-292.

3. Atkinson, K.E. *The numerical solution of integral equations of the second kind*, Cambridge University Press, Cambridge, 1997.

4. Banaś, J., Caballero, J., Rocha, J. and Sadarangani, K *Monotonic solutions of a class of quadratic integral equations of Volterra type*, Comput. Math. Appl., 49 (2005) 943-952.

5. Banaś, J., Lecko, M. and El-Sayed, W.G. *Existence theorems of some quadratic integral equation*, J. Math. Anal. Appl., 227 (1998) 276-279.

6. Banaś, J. and Martinon, A. *Monotonic solutions of a quadratic integral equation of Volterra type*, Comput. Math. Appl., 47 (2004) 271-279.

7. Banaś, J., Rocha Martin, J. and Sadarangani, K. *On the solution of a quadratic integral equation of Hammerstein type*, Math. Comp. Model., 43 (2006) 97-104.

8. Banaś, J. and Rzepka, B.*Monotonic solutions of a quadratic integral equations of fractional order*, J. Math. Anal. Appl., 332 (2007) 1370-1378.

9. Banaś, J. and Rzepka, B.*Nondecreasing solutions of a quadratic singular Volterra integral equation*, Math. Comput. Model., 49 (2009) 488-496.

10. Benchohra, M. and Darwish, M.A. *On quadratic integral equations of urysohn type in fréchet Spaces*, Acta Math. Univ. Comenianae, 79(1) (2010) 105-110.

11. Curtain, R.F. and Pritchard, A.J. *Functional analysis in modern applied mathematics*, Vol. 132. London: Academic press, 1977.

12. Darwish, M.A.*On monotonic solutions of a singular quadratic integral equation with supremum*, Dyn. Syst. Appl., 17 (2008) 539-550.

13. El-Borai, M.M., El-Sayed, W.G. and Abbas, M.I.*Monotonic solutions of a class of quadratic singular integral equations of Volterra type*, Int. J. Contemp. Math. Sci., 2(2) (2007) 89-102.

14. El-Sayed, A.M.A. and Hashem, H.H.G. *Carathéodory type theorem for a nonlinear quadratic integral equation*, Math. Sci. Res. J., 12(4) (2008) 71-95.

15. El-Sayed, A.M.A. and Hashem, H.H.G. *Integrable and continuous solutions of a nonlinear quadratic integral equation*, EJQTDE, 25 (2008) 1-10.

16. El-Sayed, A.M.A. and Hashem, H.H.G. *Monotonic positive solution of nonlinear quadratic Hammerstein and Urysohn functional integral equations*, Commentationes Math., 48(2) (2008) 199-207.

17. El-Sayed, A.M.A. and Hashem, H.H.G. *Monotonic solutions of functional integral and differential equations of fractional order*, EJQTDE, 7 (2009) 1-8.

18. El-Sayed, A.M.A. and Hashem, H.H.G. *Monotonic positive solution of a nonlinear quadratic functional integral equation*, Appl. Math. Comput., 216 (2010) 2576-2580.

19. El-Sayed, A.M.A. and Hashem, H.H.G. *Existence results for nonlinear quadratic functional integral equations of fractional orders*, Miskolc Math. Notes, 14(1) (2013) 79-88.

20. El-Sayed, A.M.A. and Hashem, H.H.G. *Existence results for coupled systems of quadratic integral equations of fractional orders*, Optim. Lett., 7 (2013) 1251-1260.

21. El-Sayed, A.M.A., Hashem, H.H.G. and Omar, Y.M.Y. *Positive continuous solution of a quadratic integral equation of fractional orders*, Math. Sci. Lett., 2(1) (2013) 19-27.

22. El-Sayed, A.M.A., Hashem, H.H.G. and Ziada, E.A.A. *Picard and Adomian decomposition methods for a quadratic integral equation of fractional order*, Comp. Appl. Math., 33 (2014) 95-109.

23. El-Sayed, A.M.A., Hashem, H.H.G. and Ziada, E.A.A. *Picard and Adomian Methods for quadratic integral equation*, Comp. Appl. Math., 29(3) (2010) 447-463.

24. El-Sayed, A.M.A., Mohamed, M.Sh. and Mohamed, F.F.S. *Existence of positive continuous solution of a quadratic integral equation of fractional orders*, J. Fract. Calc. Appl., 1(9) (2011) 1-7.

25. El-Sayed, W.G. and Rzepka, B. *Nondecreasing solutions of a quadratic integral equation of Urysohn Type*, Comput. Math. Appl., 51 (2006) 1065-1074.

26. El-Sayed, A.M.A., Saleh, M.M. and Ziada, E.A.A. *Numerical and analytic solution for a nonlinear quadratic integral equation*, Math. Sci. Res. J., 12(8) (2008) 183-191.

27. Gasca, M. and Sauer, T. *On the history of multivariate polynomial interpolation*, J. Comput. Appl. Math., 122 (2000) 23-35.

28. Mirzaee, F. and Hadadiyan, E. *Application of two-dimensional hat functions for solving space-time integral equations*, J. Appl. Math. Comput., (2015) In press.

29. Mirzaee, F. and Hadadiyan, E.*Numerical solution of linear Fredholm integral equations via two-dimensional modification of hat functions*, Appl. Math. Comput., 250 (2015) 805-816.

30. Mirzaee, F., Hadadiyan, E. and Bimesl, S. *Numerical solution for three-dimensional nonlinear mixed Volterra-Fredholm integral equations via three-dimensional block-pulse functions*, Appl. Math. Comput., 237 (2014) 168-175.

31. Nemati, S., Lima, P.M. and Ordokhani, Y. *Numerical solution of a class of two-dimensional nonlinear Volterra integral equations using Legendre polynomials*, J. Comput. Appl. Math., 242 (2013) 53-69.

32. Salem, H.A.H. *On the quadratic integral equations and their applications*, Comput. Math. Appl., 62(8) (2011) 2931-2943.

33. Ziada, E.A.A. *Adomian solution of a nonlinear quadratic integral equation*, J. Egy. Math. Soci., 21 (2013) 52-56.

Iranian Journal of Numerical Analysis and Optimization
Vol. 6, No. 2, (2016), pp 85-97

A matrix method for system of integro-differential equations by using generalized Laguerre polynomials

M. Matinfar* and A. Riahifar

Abstract

The purpose of this research is to present a matrix method for solving system of linear Fredholm integro-differential equations(FIDEs) of the second kind on unbounded domain with degenerate kernels in terms of generalized Laguerre polynomials(GLPs). The method is based on the approximation of the truncated generalized Laguerre series. Then the system of (FIDEs) along with initial conditions are transformed into the matrix equations, which corresponds to a system of linear algebraic equations with the unknown generalized Laguerre coefficients. Combining these matrix equations and then solving the system yields the generalized Laguerre coefficients of the solution function. In addition, several numerical examples are given to demonstrate the validity, efficiency and applicability of the technique.

Keywords: Systems of linear Fredholm integro-differential equations; Unbounded domain; Generalized Laguerre polynomials; Operational matrix of integration.

1 Introduction

The main object of this paper is to approximate the solution system of Fredholm integro-differential equations of the second kind on a semi-infinite domain of the following form:

$$U'(x) = F(x) + \rho \int_0^\infty w(t)K(x,t)U(t)dt, \quad x \in \mathbb{R}_+, \tag{1}$$

along with initial condition $U(0) = A$, where $\rho \in R$, and

*Corresponding author
Received 14 March 2015; revised 6 January 2015; accepted 23 Feburary 2016
M. Matinfar
Department of Mathematics, Faculty of Mathematical Sciences, University of Mazandaran, Babolsar, Iran. e-mail: m.matinfar@umz.ac.ir

A. Riahifar
Department of Mathematics, Faculty of Mathematical Sciences, University of Mazandaran, Babolsar, Iran. e-mail: Abbas.Riahifar@yahoo.com

$$U(x) = [u_1(x), u_2(x), ..., u_m(x)]^T,$$
$$F(x) = [f_1(x), f_2(x), ..., f_m(x)]^T,$$
$$K(x,t) = [k_{ij}], \quad i,j = 1,2,...,m,$$
$$A = [a_1, a_2, ..., a_m]^T. \tag{2}$$

In system (1), $w(t) = t^\alpha e^{-t} (\alpha > -1)$ and $K(x,t)$ a function of two variables x and t, is called the kernel that might have singularity in the region $D = \{(x,t) : 0 \leq x,t < \infty\}$ and $F(x)$ is continuous function and A is fixed constant vector, and $U(x)$ is the unknown vector function of the solution that will be determined. The considered equation arises in a number of important problem of elasticity theory, neutron transport, particle scattering and the theory of mixed-type equations [11,13,17]. System of linear Fredholm integro-differential equations of the second kind on unbounded domain can not be analytically solved easily. Therefore, it is required to obtain the approximate solutions. It's the reason of great interest for solving these equations. But numerical methods includes Quadrature, Petrov-Galerkin, Nystrom and Galerkin methods with Laguerre polynomial as a bases function for solving infinite boundary integral and integro-differential equations are used ago that their analysis may be found in [1, 7, 9, 10, 12, 16]. On the other hand, there are several numerical techniques for solving fractional differential equations (FDEs) on the half line using generalized Laguerre polynomials [2–6]. However, method of solution for equation (1) is too rear in the literature. In the present work, we are going to use the operational matrix of generalized Laguerre polynomials to find the approximate solutions for the system of FIDEs on the half-line. Next sections of this paper are organized as follows: In Section 2, we describe some necessary definitions and give some relevant properties of the GPLs which is required for our subsequent development. Section 3, is devoted to the approximation of the function $f(x)$ and also the kernel function $k(x,t)$ by using GPLs basis. Also the upper bound of the approximation error is presented. In Section 4, we obtain the operational matrix of integration by GPLs. In Section 5, we implemented the matrix method on the system of linear Fredholm integral-differential equations on unbounded domain and convert them to a linear algebraic system of equations. In Section 6, presented numerical examples that shows the efficiency and accuracy of the proposed method. Also a tall conclusion is given in Section 7.

2 Preliminaries and basic definitions

In this part, for the reader's convenience, we give some basic definitions and properties of the generalized Laguerre polynomials, which are used further in this article.

Let $\mathbb{R}_+ := \Lambda = [0, \infty)$ and $w^{(\alpha)}(x) = x^\alpha e^{-x}$ be a weight function on Λ in the usual sense. We define the following:

$$L^2_{w^{(\alpha)}}(\Lambda) = \{v : v \ \ is\ measurable\ on \ \ \Lambda \ \ and \ \ \|v\|_{w^{(\alpha)}} < \infty\}, \tag{3}$$

equipped with the following inner product and norm:

$$(u, v)_{w^{(\alpha)}} = \int_\Lambda u(x)v(x)w^{(\alpha)}(x)dx, \quad \|v\|_{w^{(\alpha)}} = (v, v)^{\frac{1}{2}}_{w^{(\alpha)}}. \tag{4}$$

Next, suppose $L^{(\alpha)}_n(x)$ be the generalized Laguerre polynomials of degree n, defined by the following:

$$L^{(\alpha)}_n(x) = \frac{1}{n!}x^{-\alpha}e^x\partial^n_x\left(e^{-x}x^{n+\alpha}\right), \qquad n = 0, 1, ... \tag{5}$$

$L^{(\alpha)}_n(x)$(generalized Laguerre polynomials) are the nth eigenfunction of the Sturm-Liouville problem:

$$x^{-\alpha}e^x\left(x^{\alpha+1}e^{-x}\left(L^{(\alpha)}_n(x)\right)'\right)' + \lambda_n L^{(\alpha)}_n(x) = 0, \quad x \in \Lambda, \tag{6}$$

with the eigenvalues $\lambda_n = n$ [8,14].
Generalized Laguerre polynomials are orthogonal in $L^2_{w^{(\alpha)}}(\Lambda)$ Hilbert space with the weight function $w^{(\alpha)}(x) = x^\alpha e^{-x}$ satisfy in the following relation

$$\int_0^\infty x^\alpha e^{-x}L^{(\alpha)}_n(x)L^{(\alpha)}_m(x)dx = \gamma^\alpha_n\delta_{n,m}, \qquad \forall n, m \geq 0, \tag{7}$$

where $\delta_{n,m}$ is the Kronecher delta function and $\gamma^\alpha_n = \frac{\Gamma(n+\alpha+1)}{\Gamma(n+1)}$. The explicit form of these polynomials is in the form

$$L^{(\alpha)}_n(x) = \sum_{i=0}^n E^\alpha_i x^i, \tag{8}$$

where

$$E^\alpha_i = \frac{\binom{n+\alpha}{n-i}(-1)^i}{i!}. \tag{9}$$

These polynomials are satisfied in the following recurrence formula

$$L^{(\alpha)}_0(x) = 1, \quad L^{(\alpha)}_1(x) = 1 + \alpha - x,$$

$$L^{(\alpha)}_{n+1}(x) = \frac{1}{n+1}\left[(2n + \alpha + 1 - x)L^{(\alpha)}_n(x) - (n + \alpha)L^{(\alpha)}_{n-1}(x)\right], \ n = 1, 2, ... \tag{10}$$

The case $\alpha = 0$ leads to the classical Laguerre polynomials, which are used most frequently in practice and will simply be denoted by $L_n(x)$. An important property of the Laguerre polynomials is the following derivative relation [10]:

$$\left(L_n^{(\alpha)}(x)\right)' = \sum_{i=0}^{n-1} L_i^{(\alpha)}(x). \tag{11}$$

Further, $\left(L_i^{(\alpha)}(x)\right)^{(k)}$ are orthogonal with respect to the weight function $w^{(\alpha+k)}(x)$. i.e.

$$\int_0^\infty (L_i^{(\alpha)})^{(k)}(x)(L_j^{(\alpha)})^{(k)}(x)w^{(\alpha+k)}(x)dx = \gamma_{n-k}^{\alpha+k}\delta_{i,j}, \qquad \forall i,j \geq 0, \tag{12}$$

where $\gamma_{n-k}^{\alpha+k}$ is defined in (7).

3 Approximation of functions by using GLPs

An arbitrary function $f(x) \in L^2_{w^{(\alpha)}}(\Lambda)$ may be expanded into generalized Laguerre polynomials as:

$$f(x) = \sum_{i=0}^{\infty} f_i^{(\alpha)} L_i^{(\alpha)}(x), \tag{13}$$

where the generalized Laguerre coefficients $f_i^{(\alpha)}$ are given by

$$f_i^{(\alpha)} = \int_0^\infty \frac{L_i^{(\alpha)}(x)}{\binom{i+\alpha}{i}} \cdot \frac{x^\alpha e^{-x}}{\Gamma(\alpha+1)} \cdot f(x)dx, \qquad i = 0, 1, ... \tag{14}$$

The series converges in the associated Hilbert space $L^2_{w^{(\alpha)}}(\Lambda)$, iff

$$\|f\|^2_{L^2} := \int_0^\infty \frac{x^\alpha e^{-x}}{\Gamma(\alpha+1)}|f(x)|^2dx = \sum_{i=0}^{\infty} \binom{i+\alpha}{i} |f_i^{(\alpha)}|^2 < \infty. \tag{15}$$

In practice, only the first $(n+1)$ terms of generalized Laguerre polynomials are considered. Then we have

$$f(x) \simeq \sum_{i=0}^{n} f_i^{(\alpha)} L_i^{(\alpha)}(x) = F^T L_x, \tag{16}$$

where the generalized Laguerre coefficient vector F and generalized Laguerre vector L_x are given by as follows:

$$F = [f_0^{(\alpha)},\ f_1^{(\alpha)}, \ldots,\ f_n^{(\alpha)}]^T,\ L_x = [L_0^{(\alpha)}(x),\ L_1^{(\alpha)}(x), \ldots,\ L_n^{(\alpha)}(x)]^T. \quad (17)$$

Now in the following lemma we present an upper bound to estimate the error.

Lemma 1. *Suppose that the function $f : \Lambda \to \mathbb{R}$ is $n+1$ times continuously differentiable (i.e. $f \in C^{n+1}(\Lambda)$), and $Y = Span\{L_0^{(\alpha)}(x),\ L_1^{(\alpha)}(x), \ldots,\ L_n^{(\alpha)}(x)\}$. If $F^T L_x$ be the best approximation f out of Y then mean error bound is presented as follows:*

$$\|f - F^T L_x\|_{L^2_{w^{(\alpha)}}(\Lambda)} \leq \frac{N\sqrt{(2n+\alpha+2)!}}{(n+1)!}, \quad (18)$$

where $N = \max_{x\in\Lambda} |f^{(n+1)}(x)|$.

Proof. We know that the power basis $\{1, x, ..., x^n\}$ forms a basis for the space of all polynomials of degree less than or equal to n. Therefore, we define $y_1(x) = f(0) + xf'(0) + \frac{x^2}{2!}f''(0) + \cdots + \frac{x^n}{n!}f^{(n)}(0)$. From Taylor expansion we have

$$|f(x) - y_1(x)| \leq |f^{(n+1)}(\eta_x)\frac{x^{n+1}}{(n+1)!}|, \quad (19)$$

where $\eta_x \in (0, \infty)$. Since $F^T L_x$ is the best approximation f out of Y, $y_1 \in Y$ and using (19) we have

$$\|f - F^T L_x\|^2_{L^2_{w^{(\alpha)}}(\Lambda)} \leq \|f - y_1\|^2_{L^2_{w^{(\alpha)}}(\Lambda)} \leq \frac{N^2(2n+\alpha+2)!}{(n+1)!^2}. \quad (20)$$

Then, by taking square roots we have the above bound. □

This Lemma shows that the error vanishes as $n \to \infty$.
We can also approximate the function of two variables, $k(x,t) \in L^2_{w^{(\alpha)}}(\Lambda^2)$ as follows:

$$k(x,t) \simeq \sum_{i=0}^{n}\sum_{j=0}^{n} L_i^{(\alpha)}(x)k_{ij}^{(\alpha)}L_j^{(\alpha)}(t) = L_x^T K L_t. \quad (21)$$

Here the entries of matrix $K = [k_{ij}^{(\alpha)}]_{(n+1)\times(n+1)}$ will be obtained by

$$k_{ij}^{(\alpha)} = \frac{(L_i^{(\alpha)}(x), (k(x,t), L_j^{(\alpha)}(t)))}{(L_i^{(\alpha)}(x), L_i^{(\alpha)}(x))(L_j^{(\alpha)}(t), L_j^{(\alpha)}(t))}, \quad for\ \ i,j = 0,1,...,n, \quad (22)$$

so that, $(.,.)$ denotes the inner product.

4 Operational matrix of integration, development and applications

The main objective of this part is to obtain the operational matrix of the integration by GPLs.

Theorem 1. *Suppose L_x be the generalized Laguerre vector defined in (17) then,*

$$\int_0^x L_t dt \simeq PL_x, \tag{23}$$

where P is the $(n+1)\times(n+1)$ operational matrix for integration as follows:

$$P = \begin{bmatrix} \Omega(0,0,\alpha) & \Omega(0,1,\alpha) & \Omega(0,2,\alpha) & \cdots & \Omega(0,n,\alpha) \\ \Omega(1,0,\alpha) & \Omega(1,1,\alpha) & \Omega(1,2,\alpha) & \cdots & \Omega(1,n,\alpha) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Omega(i,0,\alpha) & \Omega(i,1,\alpha) & \Omega(i,2,\alpha) & \cdots & \Omega(i,n,\alpha) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Omega(n,0,\alpha) & \Omega(n,1,\alpha) & \Omega(n,2,\alpha) & \cdots & \Omega(n,n,\alpha) \end{bmatrix}, \tag{24}$$

where

$$\Omega(i,j,\alpha) = \sum_{k=0}^{i}\sum_{r=0}^{j} \frac{(-1)^{k+r} j!\Gamma(i+\alpha+1)\Gamma(k+\alpha+r+2)}{(i-k)!(j-r)!(k+1)!r!\Gamma(k+\alpha+1)\Gamma(r+\alpha+1)}. \tag{25}$$

Proof. (See [4]). □

Also, we can see the extent of this theorem for solving fractional differential equations. For example, see [2, 5, 6].

5 Implementation of the matrix method

In this section, we solve the system of linear Fredholm integro-differential equations of the second kind on unbounded domain (1). To this end, we consider the *ith* equation of (1) as follows:

$$u_i'(x) = f_i(x) + \rho \int_0^\infty t^\alpha e^{-t} \sum_{j=1}^{m} k_{ij}(x,t)u_j(t)dt, u_i(0) = a_i,\ i = 1, ..., m, \tag{26}$$

where $f_i \in L^2_{w^{(\alpha)}}(\Lambda)$, $k_{ij} \in L^2_{w^{(\alpha)}}(\Lambda^2)$, and $u_i'(x)$ represents the first order derivative of $u_i(x)$ with respect to x, a_i are constants that give the initial con-

ditions and u_i is an unknown function. In order to approximate the solution of equation (26), we approximate functions $f_i(x)$, $u_i(x)$ and $k_{ij}(x,t)$ with respect to generalized Laguerre polynomials as mentioned in the previous section as follows:

$$f_i(x) \simeq F_i^T L_x, \;\; u_i'(x) \simeq C_i'^T L_x, \;\; u_i(0) \simeq C_{i0}^T L_x, \;\; k_{ij}(x,t) \simeq L_x^T K_{ij} L_t, \quad (27)$$

where F_i, C_i' for $i = 1, ..., m$ are known $(n+1) \times 1$ vectors and K_{ij} for $i, j = 1, 2, ..., m$ are known $(n+1) \times (n+1)$ matrices. Then, for $i = 1, ..., m$, we have:

$$u_i(x) = \int_0^x u_i'(t)dt + u_i(0) \simeq \int_0^x C_i'^T L_t dt + C_{i0}^T L_x \simeq (C_i'^T P + C_{i0}^T) L_x, \quad (28)$$

where P is a $(n+1) \times (n+1)$ operational matrix of integration given in (23). By substituting the approximations (27) and (28) into equation (26), we get the following:

$$\begin{aligned} L_x^T C_i' &= L_x^T F_i + \rho \int_0^\infty t^\alpha e^{-t} \sum_{j=1}^m L_x^T K_{ij} L_t L_t^T (p^T C_j' + C_{j0}) dt \\ &= L_x^T F_i + \rho L_x^T \sum_{j=1}^m K_{ij} \{ \int_0^\infty t^\alpha e^{-t} L_t L_t^T dt \} (p^T C_j' + C_{j0}) \\ &= L_x^T F_i + \rho L_x^T \sum_{j=1}^m K_{ij} Q (p^T C_j' + C_{j0}). \end{aligned} \quad (29)$$

Then we have following system of linear equations:

$$C_i' = F_i + \rho \sum_{j=1}^m K_{ij} Q (P^T C_j' + C_{j0}), \qquad i = 1, ..., m, \quad (30)$$

where

$$Q = \int_0^\infty t^\alpha e^{-t} L_t L_t^T dt = [q_{ij}^{(\alpha)}], \qquad i, j = 0, 1, ..., n, \quad (31)$$

and Q is a $(n+1) \times (n+1)$ matrix with elements

$$q_{ij}^{(\alpha)} = \int_0^\infty t^\alpha e^{-t} L_i^{(\alpha)}(t) L_j^{(\alpha)}(t) dt, \qquad i, j = 0, 1, ..., n. \quad (32)$$

By solving the linear system of algebraic equations (30), we can achieve the vector C_i' for $i = 1, ..., m$, then we will have

$$C_i^T = C_i'^T P + C_{i0}^T \Longrightarrow u_i(x) \simeq C_i^T L_x, \quad i = 1, ..., m. \quad (33)$$

That are the approximate solution for our system of (1).

6 Numerical Examples

In this section, we give several illustrative examples for demonstrate the efficiency of our proposed method to approximate the solutions system of Fredholm integro-differential equations of the second kind along with initial condition on a semi-infinite domain. For each example, we find the approximate solutions using different degree of generalized Laguerre polynomials. The results obtained by the present method reveal that the proposed method is very effective and convenient for system (1) on the half line. In all examples, the package of Matlab (2013) has been used to solve the test problems considered in this paper.

Example 1. For the first example, consider the following of Fredholm integral-differential equation on unbounded domain (constructed):

$$u'(x) = -\frac{247131410303000045}{36028797018963968}x^2 - \frac{38903199231847830919}{144115188075855872} + \int_0^\infty t^{\frac{1}{2}} e^{-t}(x^2+t^2)u(t)dt, \quad u(0)=1. \tag{34}$$

Exact solution of this problem is $u(x) = x^3 - 2x + 1$. If we apply the technique described in the section 5, with $\alpha = \frac{1}{2}$ and $n = 3$, then the approximate solution can be expanded as follows:

$$u(x) \simeq \sum_{i=0}^{3} c_i^{(\alpha)}\ L_i^{(\alpha)}(x) = C^T L_x, \tag{35}$$

where

$$C = [c_0^{(\alpha)},\ c_1^{(\alpha)},\ c_2^{(\alpha)},\ c_3^{(\alpha)}]^T. \tag{36}$$

Hence, from Eqs. (16), (21), (23), and (31), we find the matrices

$$F = \begin{bmatrix} -125363/424 \\ 17011/496 \\ -6434/469 \\ 0 \end{bmatrix},\ K = \begin{bmatrix} 15/2 & -5 & 2 & 0 \\ -5 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},\ P = \begin{bmatrix} 3/2 & -1 & 0 & 0 \\ 3/8 & 1 & -1 & 0 \\ 5/16 & 0 & 1 & -1 \\ 35/128 & 0 & 0 & 1 \end{bmatrix},$$

$$Q = \begin{bmatrix} 148/167 & 0 & 0 & 0 \\ 0 & 222/167 & 0 & 0 \\ 0 & 0 & 555/334 & 0 \\ 0 & 0 & 0 & 2053/1059 \end{bmatrix}.$$

Next, we substitute these matrices into equation (30) and then simplify to obtain

$$\begin{bmatrix} c_0^{(\alpha)'} \\ c_1^{(\alpha)'} \\ c_2^{(\alpha)'} \\ c_3^{(\alpha)'} \end{bmatrix} = \begin{bmatrix} -119/5475 & -19/7262 & -574/2251 & -865/5409 \\ 161/2349 & 347/1578 & 93/632 & -247/3501 \\ -181/6602 & 552/1769 & 1487/1580 & 193/6839 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} -13873/48 \\ 4211/141 \\ -6427/538 \\ 0 \end{bmatrix}. \tag{37}$$

By solving the linear system (37), we have the following:

$$c_0^{(\alpha)'} = \frac{37}{4},\ c_1^{(\alpha)'} = -15,\ c_2^{(\alpha)'} = 6,\ c_3^{(\alpha)'} = 0. \tag{38}$$

By substituting the obtained coefficients in (33) the solution of (34) becomes

$$u(x) \simeq \frac{89}{8} L_0^{(\alpha)}(x) - \frac{97}{4} L_1^{(\alpha)}(x) + 21 L_2^{(\alpha)}(x) - 6 L_3^{(\alpha)}(x), \tag{39}$$

or briefly

$$u(x) \simeq x^3 - 2x + 1, \tag{40}$$

which is the exact solution. Also, if we choose $n \geq 4$, we get the same approximate solution as obtained in equation (40). Numerical results will not be presented since the exact solution is obtained.

Example 2. As the second example, we consider the following system of linear Fredholm integro-differential equations on unbounded domain (constructed):

$$\begin{aligned} u_1'(x) &= f_1(x) + \int_0^\infty t^{\frac{1}{2}} e^{-t} (2x + t^2)(u_1(t) + u_2(t)) dt, \\ u_2'(x) &= f_2(x) + \int_0^\infty t^{\frac{1}{2}} e^{-t} (t - x^2)(u_1(t) - u_2(t)) dt, \end{aligned} \tag{41}$$

where $f_1(x) = 3x^2 - \frac{87307746120759955}{2251799813685248} x - \frac{6631788499575074881}{18014398509481984}$ and

$f_2(x) = \frac{98782478468059837}{9007199254740992} x^2 + 2x - \frac{853121404951425865}{18014398509481984}$.

Subject to the initial conditions $u_1(0) = 1$ and $u_2(0) = 1$. The exact solutions of this problem are $u_1(x) = x^3 + 2x + 1$ and $u_2(x) = x^2 + 1$. If we apply the technique described in this paper and solve equation (41) with $\alpha = \frac{1}{2}$ and $n = 3$. For this system we get:

$$\begin{aligned} u_1(x) &= \frac{137}{8} L_0^{(\alpha)}(x) - \frac{113}{4} L_1^{(\alpha)}(x) + 21 L_2^{(\alpha)}(x) - 6 L_3^{(\alpha)}(x) = x^3 + 2x + 1, \\ u_2(x) &= \frac{19}{4} L_0^{(\alpha)}(x) - 5 L_1^{(\alpha)}(x) + 2 L_2^{(\alpha)}(x) + (0) L_3^{(\alpha)}(x) = x^2 + 1, \end{aligned} \tag{42}$$

which is the exact solution. Also, if we choose $n \geq 4$, we get the same approximate solution as obtained in equation (42). Numerical results will not

be presented since the exact solution is obtained.

Example 3. As the third example, consider the following system of linear Fredholm integro-differential equations on unbounded domain (constructed):

$$u_1'(x) = f_1(x) + \int_0^\infty te^{-t-x}(sin(t-x)u_1(t) + tu_2(t))dt,$$

$$u_2'(x) = f_2(x) + \int_0^\infty te^{-t}(xtu_1(t) - e^{-x}u_2(t))dt, \tag{43}$$

with $f_1(x) = 1 - \frac{1}{4}(1 + 2sinx + 2cosx)e^{-x}$ and $f_2(x) = -6x - \frac{5}{4}e^{-x}$ and with the exact solutions $u_1(x) = x$, $u_2(x) = e^{-x}$ and boundary conditions $u_1(0) = 0$ and $u_2(0) = 1$. We apply the generalized Laguerre series approach and solve equation (43). Table 1 shows the absolute values of error $|e| = |u_2(x) - \overline{u_2}(x)|$, where $u_2(x)$ is the exact solution of equation (43) and $\overline{u_2}(x)$ is the approximate of $u_2(x)$ for $n = 20$, and $n = 30$ with $\alpha = 1$ using the described method in equally divided interval $[0, 1]$. Note that absolute

Table 1: Absolute errors for Example 3

i	x_i	$n = 20$	$n = 30$
0	0.0	$5.4836e-006$	$7.6834e-009$
1	0.1	$1.2376e-006$	$3.6294e-010$
2	0.2	$4.7027e-007$	$1.1165e-009$
3	0.3	$8.2669e-007$	$5.9280e-010$
4	0.4	$5.7470e-007$	$1.8341e-010$
5	0.5	$1.4911e-007$	$5.9768e-010$
6	0.6	$2.2241e-007$	$5.8849e-010$
7	0.7	$4.4523e-007$	$3.0976e-010$
8	0.8	$5.0477e-007$	$4.8167e-011$
9	0.9	$4.2952e-007$	$3.3937e-010$
10	1.0	$2.6633e-007$	$4.8646e-010$

errors for $u_1(x)$ is zero.
Corollary: If the exact solution of the system (1) be a polynomial, then the proposed method will obtain the real solution.

Example 4. Our last example is following of linear Fredholm integro-differential equation on a semi infinite interval (constructed):

$$u'(x) = e^{-x} - \frac{7}{4}\sqrt{x} + \int_0^\infty t^{\frac{1}{2}}e^{-t}\sqrt{xt}u(t)dt, \quad u(0) = 1. \tag{44}$$

With the exact solution $u(x) = 2 - e^{-x}$. In Table 2, the numerical results of the presented method at some selected nods for $n = 10$, and $n = 12$ are displayed.

Table 2: Absolute errors for Example 4

i	x_i	$n = 10$	$n = 12$
0	0.0	$1.3000e-003$	$3.6116e-004$
1	0.1	$4.4620e-004$	$9.3426e-005$
2	0.2	$7.1845e-005$	$4.5500e-005$
3	0.3	$3.2705e-004$	$9.9666e-005$
4	0.4	$4.0407e-004$	$1.0182e-004$
5	0.5	$3.6812e-004$	$7.5666e-005$
6	0.6	$2.6834e-004$	$3.7765e-005$
7	0.7	$1.4076e-004$	$8.8208e-007$
8	0.8	$1.0787e-005$	$3.3527e-005$
9	0.9	$1.0468e-004$	$5.6602e-005$
10	1.0	$1.9540e-004$	$6.8843e-005$

7 Conclusion

Obtaining the analytic solutions for system of linear Fredholm integro-differential equations of the second kind, along with initial conditions on unbounded domain are usually difficult. In many cases, it is required to approximate solutions. For this reason, a new matrix approach which is based on the generalized Laguerre operational matrix of integration is proposed. The solution procedure is very simple by means of generalized Laguerre polynomials expansion and only in a few terms lead to high accurate solutions. The main goal of the presented technique was deriving an approximation to the solution system of linear Fredholm integro-differential equations on unbounded domain. To illustrate the method and its efficiency, four examples were provided. In the first and second examples, we obtained the exact solution. Another considerable advantage of the method is that the nth-order approximation gives the exact solution when the solution is polynomial of degree equal to or less than n. If the solution is not polynomial, generalized Laguerre series approximation converges to the exact solution as n increases.

Acknowledgements

The authors are very grateful to the anonymous referees and editor for their comments.

References

1. Akhavan, S. *Numerical solution of singular Fredholm integro-differential equations of the second kind via Petrov-Galerkin method by using Legendre multiwavelet*, Journal of mathematics and computer science. 9 (2014), 321-331.

2. Bhrawy, A.H., Alghamdi, M.A. and Taha, T.M. *A new modified generalized Laguerre operational matrix of fractional integration for solving fractional differential equations on the half line, Advances in Difference Equations.* (2012), 0:179. doi: 10.1186/1687-1847-2012-179.

3. Bhrawy, A.H., Al-Zahrani, A.A., Alhamed, Y.A. and Baleanu, D. *A new generalized Laguerre-Gauss collocation scheme for numerical solution of generalized fractional pantograph equations*, Romanian Reports Of Physics. 59 (7-8) (2014), 646-657.

4. Bhrawy, A.H., Baleanu, D., Assas, L.M. and Tenreiro Machado, J.A. *On a generalized Laguerre operational matrix of fractional integration*, Mathematical Problems in Engineering. (2013), Article ID 569286, 7 pages.

5. Bhrawy, A.H. and Taha, T.M. *An operational matrix of fractional integration of the Laguerre polynomials and its application on a semi-infinite interval*, Mathematical Sciences. (2012), 6:41. doi: 10.1186/2251-7456-6-41.

6. Bhrawy, A.H., Tharwat, M.M. and Alghamdi, M.A. *A new operational matrix of fractional integration for shifted jacobi polynomials*, Bulletin of the Malaysian Mathematical Sciences Society. 37 (4) (2014), 983-995.

7. De Bonis, M.C. and Mastroianni, M.G. *Nystrom method for systems of integral equations on the real semiaxis*, IMA Journal of Numerical Analysis. 29 (2009), 632-650.

8. Funaro, D. *Polynomial Approximations of Differential Equations*, Springer Verlag, 1992.

9. Maalek Ghaini, F.M., Tavassoli Kajani, F. and Ghasemi, M. *Solving boundary integral equation using Laguerre polynomials*, World Applied Sciences Journal. 7 (1) (2009), 102-104.

10. Mastroianni, G. and Milovanovic, G.V. *Some numerical methods for second kind Fredholm integral equations on the real semiaxis*, IMA Journal of Numerical Analysis. 29 (2009), 1046-1066.

11. Muskhelishvili, N.I. *Singular integral equations*, Noordhoff, Holland, 1953.

12. Nik Long, N. M. A., Eshkuvatov, Z. K., Yaghobifar, M. and Hasan, M. *Numerical solution of infinite boundary integral equation by using Galerkin method with Laguerre polynomials*, World Academy of Science Engineering and Technology. 47 (2008), 334-337.

13. Sanikidze, D.G. *On the numerical solution of a class of singular integral equations on an infinite interval, Differential Equations.* 41 (9) (2005), 1353-1358.

14. Shen, J., Tang, T. and Wang, L.L. *Spectral Methods Algorithms*, Analysis and Applications, Springer, 2011.

15. Shen, J. and Wang, L.L. *Some Recent Advances on Spectral Methods for Unbounded Domains*, J. Commun. comput. Phys. 5 (2009), 195-241.

16. Sloan, I.H. *Quadrature methods for integral equations of the second kind over infinite intervals*, Mathematics of Computation. 36 (154) (1981), 511-523.

17. Volterra, V. *Theory of functionnals of integro-differential equations*, Dover, New York, 1959.

Iranian Journal of Numerical Analysis and Optimization
Vol. 6, No. 2, (2016), pp 99-120

Global error estimation of linear multistep methods through the Runge-Kutta methods

J. Farzi*

Abstract

In this paper, we study the global truncation error of the linear multistep methods (LMM) in terms of local truncation error of the corresponding Runge-Kutta schemes. The key idea is the representation of LMM with a corresponding Runge-Kutta method. For this, we need to consider the multiple step of a linear multistep method as a single step in the corresponding Runge-Kutta method. Therefore, the global error estimation of a LMM through the Runge-Kutta method will be provided. In this estimation, we do not take into account the effects of roundoff errors. The numerical illustrations show the accuracy and efficiency of the given estimation.

Keywords: Linear multistep methods; Runge-Kutta methods; Local truncation error; Global error; Error estimation.

1 Introduction

The error estimation is one of the major issues in designing numerical algorithms. In the study of the linear multistep methods (LMM)

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

for solving an ordinary differential system

$$\begin{cases} y' = f(x, y), \\ y(x_0) = y_0, \end{cases} \tag{2}$$

where, $f : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$, there is a challenging issue of estimation of global truncation error (GTE) or simply global error. In spite the local truncation error (LTE), the estimation of GTE is much more complicated. The

*Corresponding author
Received 25 December 2015; revised 9 March 2016; accepted 11 2016
J. Farzi
Department of Mathematics, Sahand University of Technology, P.O. Box 51335-1996, Tabriz, Iran. e-mail: farzi@sut.ac.ir

LTE estimations have been studied in some special cases, e.g., predictor-corrector (PC) and embedded Runge-Kutta methods. In the case of PC methods that predictor is an Adams-Bashforth scheme of order p and corrector is an Adams-Moulton scheme of the same order, the Milne estimation provides an estimation for LTE of the resulted PC method [9, 10]. Recently, Cao and Petzold have developed an estimation for global error with adjoint method [3]. However, many efforts have previously been reported for providing GTE bounds [6, 7]. These bounds are of no practical value and the LTE form estimation of GTE, given in this paper, is more simpler than the usual theoretical bounds. For more extensive discussion on linear multistep methods and their important subclauses, like backward difference formula (BDF) schemes see [1, 2, 8–10, 12]. The more important application of the LMMs is in the time discretization of time dependent partial differential equations. The LMMs with strong stability preserving property (SSP) have a major role in this context [4, 5].

In this paper, we represent a multiple step of a LMM as a single step of a new Runge-Kutta method. Then, we accomplish the GTE estimation of the LMM by estimation of LTE of the corresponding new Runge-Kutta method.

This paper has been organized as follow: In Section 2 a very short review of the Runge-Kutta schemes is presented. Then, in Section 3 the main idea of the paper is given for one-step methods with a rather detailed study of the LTE and stability region of the new method. In Section 4 the idea of previous section is generalized to formulate a general LMM in the form of a new Runge-Kutta method. The new RK method has a popular structure in view of Butcher array. We take into account a starting procedure, subsequently, the method order and its LTE determined, which is the GTE of the original scheme. Finally, in Section 5 we present some numerical tests including a fourth order total variation bounded (TVB) scheme to illustrate efficiency of the given theory.

2 A review of Runge-Kutta methods

In this section, we shortly review the main concepts of Runge-Kutta methods that are required in the rest of the paper. For more details we refer to [2, 9]. The s-stage Runge-Kutta method for the problem (2) is defined as follow

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i k_i, \tag{3}$$

where

$$k_i = f(x_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j), \quad i = 1, 2, \ldots, s. \tag{4}$$

Traditionally the Runge-Kutta methods is represented by the following Butcher array

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} \tag{5}$$

where,

$$c = [c_1, c_2, \ldots, c_s]^T, \quad b = [b_1, b_2, \ldots, b_s]^T, \quad A = [a_{ij}]_{i,j=1}^{s}. \tag{6}$$

The derivation of a Runge-Kutta method of an arbitrary order is a crucial work without using the advanced concepts of elementary differentials and most related rooted trees. In fact, there is a 1-1 corresponding between elementary differential (that are defined by Fréchet derivative) and the rooted trees. Therefore, using the rooted trees one can easily define the order condition and LTE of a Runge-Kutta method. The local truncation error of the p^{th} order Runge-Kutta method (5) is given by

$$\text{LTE} = \frac{h^{p+1}}{(p+1)!} \sum_{r(t)=p+1} \alpha(t)[1 - \gamma(t)\psi(t)]F(t) + O(h^{p+2}), \tag{7}$$

where,

$$\alpha(t) = \frac{r(t)!}{\sigma(t)\gamma(t)}$$

and $r(t), \sigma(t)$ and $\gamma(t)$ are order, symmetry and density of a tree t. The function $F(t)$ is defined on the set T of all trees which corresponds between the rooted trees and elementary differentials. The elementary differentials are evaluated at the value $y(x_n)$. The $\psi(t)$ function is also defined on the set of all trees T. For example we have,

$$F(\bullet) = f,$$

$$F(\;\vert\;) = \{f\} = f^{(1)}(f),$$

$$F(\;\curlyvee\;) = \{f^2\} = f^{(2)}(f, f),$$

$$F(\;\vert\;) = \{{}_2f\}_2 = f^{(1)}(f^{(1)}(f)),$$

where, $f^{(M)}(K1, K2, \ldots, K_M)$, $K_t \in \mathbb{R}^m, t = 1, 2, \ldots, M$ is the Mth order Fréchet derivative of f. For more detailed definition of these functions see [2, 9].

The following theorem gives the coefficients of the linear combination of $y^{(q)}$ for a general q in terms of elementary differentials [9]:

Theorem 1. *Let y be the solution of the autonomous problem (1). Then*

$$y^{(q)} = \sum_{r(t)=q} \alpha(t)F(t). \tag{8}$$

The stability function of a Runge-Kutta method is given by

$$R(\hat{h}) = \frac{\det\left[I - \hat{h}A + \hat{h}eb^T\right]}{\det\left[I - \hat{h}A\right]}, \tag{9}$$

where, $\hat{h} = \lambda h$, here λ is typically an eigenvalue of the jacobian matrix of f that is equivalently the eigenvalues of the linearized equation.

3 One step methods

In this section, we firstly study the situation for the one-step methods. The idea will be extended to a general linear multistep methods in the next sections. Consider the following linear one step method

$$y_{n+1} = y_n + h(\beta_0 f_n + \beta_1 f_{n+1}) \tag{10}$$

for a consistent method we have $\beta_0 + \beta_1 = 1$. Applying the above rule on N successive steps to advance the solution from x_0 to x_N we obtain

$$\begin{aligned} y_1 &= y_0 + h(\beta_0 f_0 + \beta_1 f_1) \\ y_2 &= y_1 + h(\beta_0 f_1 + \beta_1 f_2) \\ &\vdots \\ y_N &= y_{N-1} + h(\beta_0 f_{N-1} + \beta_1 f_N). \end{aligned}$$

Introducing the following slopes

$$k_1 = f(x_0, y_0), k_2 = f(x_1, y_1), \ldots, k_{N+1} = f(x_N, y_N),$$

we can write (10) as a $(N+1)$-stage Runge-Kutta method with the steplength $H = Nh$,

$$y_{m+1} = y_m + \frac{H}{N}(\beta_0 k_1 + k_2 + \cdots + k_N + \beta_1 k_{N+1})$$

where, $t_m = x_0, t_{m+1} = t_m + H = x_N$ and,

$$
\begin{aligned}
k_1 &= f(t_m, y_m)\\
k_2 &= f(t_m + \frac{H}{N}, y_m + \frac{H}{N}(\beta_0 k_1 + \beta_1 k_2))\\
&\vdots\\
k_{N+1} &= f(t_m + H, y_m + \frac{H}{N}(\beta_0 k_1 + k_2 + \cdots + k_N + \beta_1 k_{N+1})).
\end{aligned}
$$

Thus, the corresponding Butcher array takes the following form

$$
\begin{array}{c|cccccc}
0 & 0 & & & & & \\
\frac{1}{N} & \frac{\beta_0}{N} & \frac{\beta_1}{N} & & & & \\
\frac{2}{N} & \frac{\beta_0}{N} & \frac{1}{N} & \frac{\beta_1}{N} & & & \\
\vdots & \vdots & \vdots & & \ddots & & \\
\frac{N}{N} & \frac{\beta_0}{N} & \frac{1}{N} & \cdots & \frac{1}{N} & \frac{\beta_1}{N} \\
\hline
 & \frac{\beta_0}{N} & \frac{1}{N} & \cdots & \frac{1}{N} & \frac{\beta_1}{N}
\end{array}
\tag{11}
$$

3.1 Local truncation error

In this section we obtain the LTE and the order of the reduced method (11). To determine the order of the deduced RK scheme we verify the order conditions for the Butcher array (11):

$$\psi(\bullet) = \sum_{i=1}^{N+1} b_i = 1, \tag{12}$$

$$\psi(\left|\begin{smallmatrix}\bullet\\ \bullet\end{smallmatrix}\right.) = \sum_{i=1}^{N+1} b_i c_i = \frac{1}{2}. \tag{13}$$

Since we have $\beta_0 + \beta_1 = 1$, substituting the data from (11) we observe that the first condition always holds

$$\sum_{i=1}^{N+1} b_i = 1 \Rightarrow \frac{\beta_0}{N} + \frac{1}{N} + \cdots + \frac{1}{N} + \frac{\beta_1}{N} = 1.$$

However, for the second condition to be valid we have,

$$
\begin{aligned}
\frac{\beta_0}{N}.0 + \frac{1}{N}(\frac{1}{N} + \frac{2}{N} + \cdots + \frac{N-1}{N}) + \frac{\beta_1}{N}.\frac{N}{N}&\\
= \frac{N^2 - N + 2\beta_1 N}{2N^2} = \frac{1}{2},& \text{ only if } \beta_1 = \frac{1}{2}.
\end{aligned}
$$

where, we have used $\beta_0 + \beta_1 = 1$. Therefore, the only possible second order one step method is the trapezoidal rule.

Now, we can find the general form of the local truncation error for both first (forward and backward euler) and second order (Trapezoidal) methods.

For the second order method we have, $\beta_1 = \frac{1}{2}$,

$\psi($ $t_1 =$  $) = \sum_{i=1}^{N+1} b_i c_i^2 = \frac{2N^2+1}{6N^2},$

$\psi($ $t_2 =$ $) = \sum_{i,j=1}^{N+1} b_i a_{ij} c_j = \frac{1}{N}[\frac{1}{2}, 1, \ldots, 1, \frac{1}{2}]Ac = \frac{2N^2+1}{12N^2}.$

It is evident that in the limit the above order conditions, as $N \to \infty$, tend to the following infinite dimensional exact order conditions:

$$\sum_{i=1}^{\infty} b_i c_i^2 = \frac{1}{3},$$
$$\sum_{i,j=1}^{\infty} b_i a_{ij} c_j = \frac{1}{6}.$$

The corresponding elementary differentials with the rooted trees t_1 and t_2 are,

$$F(t_1) = f^{(2)}(f, f), \quad F(t_2) = f^{(1)}(f^{(1)}(f)).$$

To specify the LTE of second order scheme we need a representation of $y^{(3)}$ in terms of elementary differentials. According to Theorem 1, we have

$$y^{(3)} = F(t_1) + F(t_2), \tag{14}$$

therefore, the principle term in local truncation error (PLTE) for $\beta_1 = \frac{1}{2}$, where $p = 2$ reads

$$\begin{aligned} \text{PLTE} &= \frac{H^3}{3!} \sum_{r(t)=3} \alpha(t)[1 - \gamma(t)\psi(t)]F(t) \qquad (15) \\ &= \frac{H^3}{3!}(-\frac{1}{2N^2})(F(t_1) + F(t_2)) \\ &= -\frac{1}{12} N h^3 y^{(3)}(x_0) \\ &= -\frac{1}{12} h^2 (x_N - x_0) y^{(3)}(x_0), \end{aligned}$$

and then,

$$\mathrm{LTE} = -\frac{1}{12}h^2(x_N - x_0)y^{(3)}(x_0) + O(h^3). \tag{16}$$

We can also regard (16) as the global error in N steps of the trapezoidal rule. Similarly, for $\beta_1 \neq \frac{1}{2}$ we obtain

$$\mathrm{LTE} = \frac{1}{2}N(N - 1 + 2\beta_1)h^2y^{(2)}(x_0) + O(h^3). \tag{17}$$

3.2 Stability regions

To construct the stability function of (11) we simply note that

$$\hat{H}A = \frac{\hat{H}}{N}\begin{pmatrix} \beta_0 & \beta_1 & & \\ & & \ddots & \\ \beta_0 & 1 & \dots 1 & \beta_1 \end{pmatrix}, \quad \hat{H}eb^T = \frac{\hat{H}}{N}\begin{pmatrix} \beta_0 & 1 & \dots & 1 & \beta_1 \\ \beta_0 & 1 & & 1 & \beta_1 \\ \vdots & \vdots & & \vdots & \vdots \\ \beta_0 & 1 & \dots & 1 & \beta_1 \end{pmatrix},$$

thus we have

$$I - \hat{H}A + \hat{H}eb^T = \begin{pmatrix} 1 + \hat{H}\frac{\beta_0}{N} & \frac{\hat{H}}{N} & \dots & \frac{\hat{H}}{N} & \frac{\hat{H}}{N}\beta_1 \\ & 1 + \hat{H}\frac{\beta_0}{N} & & \frac{\hat{H}}{N} & \frac{\hat{H}}{N}\beta_1 \\ & & \ddots & & \vdots \\ & & & 1 + \hat{H}\frac{\beta_0}{N} & \frac{\hat{H}}{N}\beta_1 \\ & & & & 1 \end{pmatrix},$$

and thereby,

$$\det[I - \hat{H}A + \hat{H}eb^T] = (1 + \hat{H}\frac{\beta_0}{N})^N.$$

Similarly, we can show that the following relation is also valid

$$\det[I - \hat{H}A] = (1 - \hat{H}\frac{\beta_1}{N})^N.$$

Inserting the above results into (9) we obtain

$$R(\hat{H}) = \frac{(1 + \hat{H}\frac{\beta_0}{N})^N}{(1 - \hat{H}\frac{\beta_1}{N})^N}. \tag{18}$$

which is the stability function of (11).

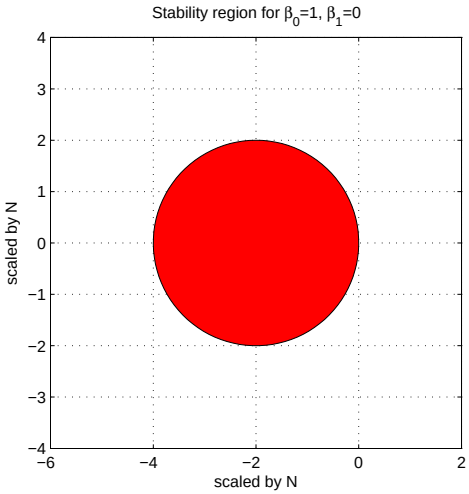

Figure 1: The stability region of the new RK method with $\beta_0 = 1, \beta_1 = 0$

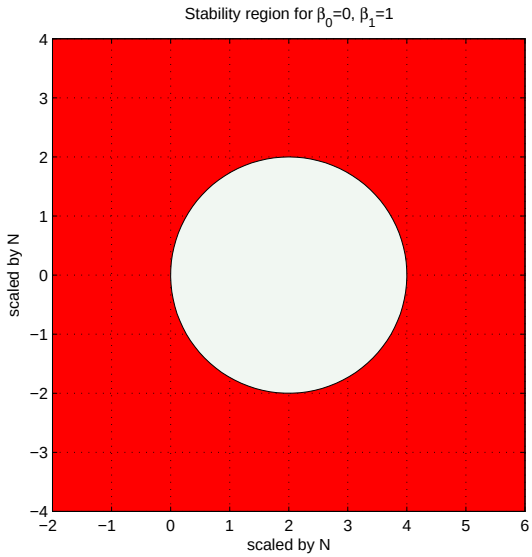

Figure 2: The stability region of the new RK method with $\beta_0 = 0, \beta_1 = 1$

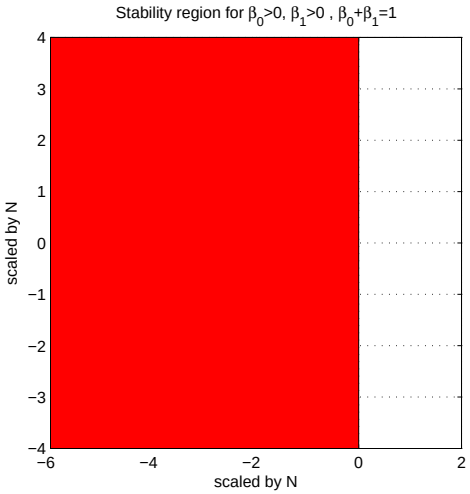

Figure 3: The stability region of the new RK method with $\beta_0 > 0, \beta_1 > 0, \beta_0 + \beta_1 = 1$

In Figures 1, 2, 3, the absolute stability regions of the given RK method (11) have been demonstrated for various values of β_0 and β_1. We observe that in the case $\beta_0 > 0, \beta_1 > 0, \beta_0 + \beta_1 = 1$ the new RK method is A-stable and in the cases $\beta_0 = 0, \beta_1 = 1$ and $\beta_0 = 1, \beta_1 = 0$ the absolute stability regions, in terms of $\hat{h}$, tend to the same regions of the forward and backward Euler methods, respectively.

4 Linear multistep methods

In this section, we present the general theory for an arbitrary linear multistep method. We demonstrated the main idea by using the one-step and multi-step starting procedures. In both cases, we obtain the corresponding Runge-Kutta scheme and the LTE of this method provide an estimation of the global truncation error of the given LMM.

4.1 A single step method as starting procedure

Now, we represent a general linear multistep method (LMM) in the form of a Runge-Kutta scheme. For simplicity, we use the trapezoidal rule as starting procedure of the LMM. We show the starting values $\{y_{n+j}\}_{j=1}^{k-1}$ as a linear combination of the y_n and $\{f_{n+j}\}_{j=0}^{k-1}$. Therefore, starting with

$$y_{n+1} = y_n + \frac{h}{2}(f_n + f_{n+1}),$$

we find

$$y_{n+j} = y_n + \frac{h}{2}(f_n + 2f_{n+1} + \cdots + 2f_{n+j-1} + f_{n+j}), \quad j = 1, 2, \ldots \tag{19}$$

therefore, we have

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = y_{n+k} + (\sum_{j=0}^{k-1} \alpha_j) y_n + \frac{h}{2}\{(\sum_{j=1}^{k-1} \alpha_j) f_n + (\alpha_1 + 2\sum_{j=2}^{k-1} \alpha_j) f_{n+1} + \cdots + (\alpha_{k-2} + 2\alpha_{k-1}) f_{n+k-2} + \alpha_{k-1} f_{n+k-1})\}. \tag{20}$$

Substituting (20) into the LMM (1), we obtain

$$\begin{aligned} y_{n+k} &= -\sum_{j=0}^{k-1} \alpha_j y_{n+j} + h \sum_{j=0}^{k-1} \beta_j f_{n+j} \\ &= -(\sum_{j=0}^{k-1} \alpha_j) y_n + h\{(\beta_0 - \frac{1}{2}\sum_{j=1}^{k-1} \alpha_j) f_n + (\beta_1 - \frac{1}{2}\alpha_1 - \sum_{j=2}^{k-1} \alpha_j) f_{n+1} \\ &\quad + \cdots + (\beta_{k-2} - \frac{1}{2}\alpha_{k-2} - \alpha_{k-1}) f_{n+k-2} \\ &\quad + (\beta_{k-1} - \frac{1}{2}\alpha_{k-1}) f_{n+k-1} + \beta_k f_{n+k})\}, \end{aligned}$$

or,

$$y_{n+k} = y_n + h \sum_{i=1}^{k+1} b_i f_{n+i-1}, \tag{21}$$

where,

$$\begin{aligned} b_1 &= \beta_0 - \frac{1}{2}\sum_{j=1}^{k-1} \alpha_j, \\ b_i &= \beta_{i-1} - \frac{1}{2}\alpha_{i-1} - \sum_{j=i}^{k-1} \alpha_j, \quad 2 \le i \le k-1 \\ b_k &= \beta_{k-1} - \frac{1}{2}\alpha_{k-1} \\ b_{k+1} &= \beta_k. \end{aligned}$$

It is straightforward to show that

$$\sum_{i=1}^{k+1} b_i = k.$$

On defining

$$K_{i+1} = f(x_{n+i}, y_{n+i}) = f(x_n + ih, y_n + \frac{h}{2}(f_n + 2f_{n+1} + \cdots + 2f_{n+i-1} + f_{n+i})),$$

the scheme (21) can be represented in the form of the following RK method

$$y_{m+1} = y_m + H \sum_{i=1}^{k+1} b'_i K_i,$$

where,

$$K_1 = f(t_m, y_m),$$
$$K_i = f(t_m + c_i H, y_m + H \sum_{j=1}^{i} a_{ij} K_j), \quad i = 2, 3, \ldots, k+1.$$

We will change the subscript n to m in order to show that when LMM runs from y_n to y_{n+k} the RK scheme runs just a single step from $y_m = y_n$ to $y_{m+1} = y_{n+k}$. Therefore, we set

$$t_m = x_n, t_{m+1} = x_n + H,$$
$$a_{ij} = \begin{cases} \frac{1}{2k}, j = 1, i, \\ \frac{1}{k}, \quad 2 \leq j \leq i-1, \\ 0, \quad j > i \text{ or } i = 1. \end{cases}$$
$$c_i = \frac{i-1}{k},$$
$$b'_i = \frac{b_i}{k},$$
$$H = kh.$$

In this case, the first order condition holds

$$\sum_{i=1}^{k+1} b'_i = 1,$$

but, the second order condition no longer holds

$$\sum_{i=1}^{k+1} b'_i c_i = \frac{k+1}{2k}.$$

In $k \to \infty$ this condition turns out to be the exact order condition. Again, the principle local truncation error (PLTE), where $p = 1$ reads

$$\text{PLTE} = \frac{H^2}{2!} \sum_{r(t)=2} \alpha(t)[1 - \gamma(t)\psi(t)]F(t) = -\frac{1}{2}h(x_k - x_0)y^{(2)}(x_0),$$

and therefore, we have

$$\text{LTE} = -\frac{1}{2}h(x_k - x_0)y^{(2)}(x_0) + O(h^2),$$

which is the global error in k times application of a k-step LMM with Trapezoidal rule as starting procedure.

4.2 A Runge-Kutta method as starting procedure

Now, we have considered the general explicit Runge-Kutta method (5)-(6) as the starting procedure for the linear multistep method and we find the Runge-Kutta representation of the given LMM (1). Based on this starting procedure, we have obtained approximate values for $y_{n+j}, j = 1, \ldots, k-1$ as follows

$$\begin{aligned} y_{n+j} &= y_n + jh \sum_{i=1}^{s} b_i k_i^{(j)}, \\ k_i^{(j)} &= f(x_n + jc_i h, y_n + jh \sum_{l=1}^{s+1} a_{il} k_l^{(j)}), \quad i = 1, \ldots, s+1, \end{aligned}$$

where $c_{s+1} = 1, a_{i,s+1} = 0, i = 1, 2, \ldots, s+1, a_{s+1,j} = b_j, j = 1, 2, \ldots, s$. For an implicit method corresponding to the y_{n+k}, we define

$$k_{s+1}^{(k)} = f(x_n + kc_{s+1}h, y_n + kh \sum_{l=1}^{s+1} a_{s+1,l} k_l^{(j)}).$$

By inserting these approximations into the LMM (1) we obtain

$$\begin{aligned} y_{n+k} &= -\sum_{j=0}^{k-1} \alpha_j y_{n+j} + h \sum_{j=0}^{k} \beta_j f_{n+j} \\ &= -\sum_{j=0}^{k-1} \alpha_j y_n - h \sum_{j=1}^{k-1} \sum_{i=1}^{s} \alpha_j b_i k_i^{(j)} + h \sum_{j=0}^{k} \beta_j k_{s+1}^{(j)}, \end{aligned}$$

there is $s(k-1)+2$ different $k_i^{(j)}$ in the above representation, however we do not distinguish them and consider $(s+1)(k-1)+1$ moments. The advantage of ignoring the similarity in the moments is that the resulted Butcher array is simpler to work and it is convenient to prove the theorems. Let,

$$t_m = x_n, H = kh, t_{m+1} = t_m + H = x_{n+k}$$

and

$$\bar{c}_{i+(j-1)(s+1)} = \frac{j}{k} c_i, \quad i = 1, \ldots, s+1, j = 1, \ldots, k-1,$$
$$\bar{c}_{(s+1)(k-1)+1} = \frac{1}{k},$$

and

$$b_1^{(1)} = -\frac{1}{k}(\alpha_1 b_1 - \beta_0),$$
$$b_i^{(j)} = -\frac{1}{k} j \alpha_j b_i, \qquad i = 1, \ldots, s, \ j = 1, \ldots, k-1,$$
$$b_{s+1}^{(j)} = \frac{1}{k}\beta_j, j = 1, \ldots, k,$$

the vector $\bar{b}$ consist of these values:

$$\bar{b} = [b^{(1)}, b^{(2)}, \ldots, b^{(k-1)}, b_{s+1}^{(k)}].$$

Therefore, we find the new Runge-Kutta scheme

$$y_{m+1} = y_m + H \sum_{i=1}^{\bar{s}} \bar{b}_i \bar{k}_i,$$
$$k_i^{(j)} = f(t_m + \frac{j}{k} c_i H, y_m + \frac{j}{k} H \sum_{l=1}^{s+1} a_{il} k_l^{(j)}), \quad i = 1, \ldots, s+1,$$

where, $\bar{s} = (s+1)(k-1)+1$ and

$$\bar{k} = [k^{(1)}, k^{(2)}, \ldots, k^{(k-1)}, k_{s+1}^{(k)}].$$

The corresponding Butcher array is given in Table 1. where $\mathbf{0}$ is a $(s+1) \times 1$ zero vector. Introducing D as a $(k-1) \times (k-1)$ diagonal matrix

$$D = \frac{1}{k} \begin{bmatrix} 1 & & & & \\ & 2 & & & \\ & & 3 & & \\ & & & \ddots & \\ & & & & k-1 \end{bmatrix}.$$

The more compact form of the above scheme is resulted.

Table 1: Butcher array of the Runge-Kutta representation of LMM (1) scheme

$\frac{1}{k}$ c	$\frac{1}{k}\begin{bmatrix} A & \mathbf{0} \\ b & \end{bmatrix}$			
$\frac{2}{k}$ c		$\frac{2}{k}\begin{bmatrix} A & \mathbf{0} \\ b & \end{bmatrix}$		
$\vdots$		$\ddots$	$\ddots$	
$\frac{k-1}{k}$ c			$\frac{k-1}{k}\begin{bmatrix} A & \mathbf{0} \\ b & \end{bmatrix}$	
$\frac{k}{k}$		$\bar{b}^T$		
		$\bar{b}^T$		

Table 2: Butcher array of the Runge-Kutta representation of LMM (1) scheme

$\bar{c}$	$D \otimes \begin{bmatrix} A & \mathbf{0} \\ b & \end{bmatrix}$
	$\bar{b}^T$
	$\bar{b}^T$

Now, to verify the order of the new Runge-Kutta scheme suppose that the order of LMM (1) and Runge-Kutta scheme (5)-(4) are p and $\bar{p}$, respectively. Then, we can prove the following theorem.

Theorem 2. *If the Runge-Kutta scheme as starting procedure has order $\bar{p}$ and the order of linear multistep method (1) is p, then the corresponding Runge-Kutta method with Butcher array in Table 2 is of order* $\min\{p, \bar{p}\}$.

Proof. The order condition corresponding to the tree

$t =$ 

with m leaves is

$$\sum_{i=1}^{s} b_i c_i^m = \frac{1}{m+1}, \quad m = 0, 1, \ldots, \bar{p}$$

for any $m \leq \min\{p, \bar{p}\}$ we prove that

$$\sum_{i=1}^{\bar{s}} \bar{b}_i \bar{c}_i^m = \frac{1}{m+1}, \quad m = 0, 1, \ldots, \min\{p, \bar{p}\}.$$

According to the definition of $\bar{c}$ and $\bar{b}$, we have

$$\begin{aligned}\sum_{i=1}^{\bar{s}} \bar{b}_i \bar{c}_i^m &= -\sum_{j=1}^{k-1} \frac{1}{k} j\alpha_j \sum_{i=1}^{s} b_i (\frac{j}{k} c_i)^m + \frac{1}{k} \sum_{j=1}^{k} \beta_j (\frac{j}{k})^m \\ &= -\frac{1}{m+1} \frac{1}{k^{m+1}} \sum_{j=1}^{k-1} j^{m+1}\alpha_j + \frac{1}{k^{m+1}} \sum_{j=1}^{k} j^m \beta_j \\ &= \frac{1}{(m+1)k^{m+1}} \Big(k^{m+1} - \sum_{j=0}^{k} j^{m+1}\alpha_j \Big) + \frac{1}{k^3} \sum_{j=0}^{k} j^m \beta_j \\ &= \frac{1}{m+1}.\end{aligned}$$

note that in the above relations we have used the order conditions for starting Runge-Kutta method as well as the order conditions for linear multistep methods:

$$\frac{1}{m+1} \sum_{j=0}^{k} j^{m+1}\alpha_j = \sum_{j=0}^{k} j^m \beta_j, \quad m = 2, \ldots, p.$$

The rest of the proof is closely related to the block structure of the Butcher array in Table 1. The extracted elements of the Butcher array is demonstrated in the Table 3.

Table 3: Elements of Butcher array of Table 1

$$\begin{array}{c|c} \frac{l}{k}c & \frac{l}{k}\begin{bmatrix} A & \mathbf{0} \\ b & \end{bmatrix} \\ \hline & b^{(l)} \end{array}, \qquad l = 1, 2, \ldots, k-1$$

$$\begin{array}{c|c} 1 & \bar{b}^T \\ \hline & b_{s+1}^{(k)} = \frac{1}{k}\beta_k \end{array}$$

These partial elements will help us to exactly find the effect of them in multiple sums in order conditions.

The last condition to complete the proof of the third order conditions is $\sum_{i,j=1}^{\bar{s}} \bar{b}_i \bar{a}_{ij} \bar{c}_j$. The role of each element in summation, separately, is

$$\sum_{i,j=1}^{s} (-\frac{l}{k}\alpha_l b_i)(\frac{l}{k} a_{ij})(\frac{l}{k} c_j) + (\frac{1}{k}\beta_l) \sum_{i=1}^{s} (\frac{l}{k} b_i)(\frac{l}{k} c_i), \quad l = 1, 2, \ldots, k-1$$

$$\frac{1}{k}\beta_k (\sum_{i=1}^{\bar{s}} \bar{b}_i \bar{c}_i),$$

summing up these terms we obtain

$$\begin{aligned}\sum_{i,j=1}^{\bar{s}} \bar{b}_i \bar{a}_{ij} \bar{c}_j &= \sum_{l=1}^{k-1} \Big(\sum_{i,j=1}^{s} (-\frac{l}{k}\alpha_l b_i)(\frac{l}{k}a_{ij})(\frac{l}{k}c_j) + (\frac{1}{k}\beta_l) \sum_{i=1}^{s} (\frac{l}{k}b_i)(\frac{l}{k}c_i) \Big) \\ &\quad + \frac{1}{k}\beta_k (\sum_{i=1}^{\bar{s}} \bar{b}_i \bar{c}_i) \\ &= \frac{1}{6k^3}(k^3 - \sum_{l=1}^{k} l^3 \alpha_l) + \frac{1}{2k^3} \sum_{l=0}^{k} l^2 \beta_l \\ &= \frac{1}{6}.\end{aligned}$$

similarly we can prove the higher order conditions. □

4.3 Local truncation error of the new Runge-Kutta method

We have shown that the order of the new Runge-Kutta method with Butcher array in Table 2 is $p^* = \min\{p, \bar{p}\}$. Ignoring the effect of the roundoff errors, we can consider LTE of this method as the global error of the given linear multistep method in evaluation of y_{n+k}. The LTE of this scheme now reads

$$\text{LTE} = \frac{h^{p^*+1}}{(p^*+1)!} \sum_{r(t)=p^*+1} \alpha(t)[1 - \gamma(t)\psi(t)]F(t) + O(h^{p^*+2}). \tag{22}$$

5 Numerical illustrations

Example 1. As an example we consider the Heun's third order 3-stage formula

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{3} & \frac{1}{3} & & \\ \frac{2}{3} & 0 & \frac{2}{3} & \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array}$$

as starting procedure for the third order convergent linear multistep method

$$y_{n+3} + \frac{1}{4}y_{n+2} - \frac{1}{2}y_{n+1} - \frac{3}{4}y_n = \frac{h}{8}[19f_{n+2} + 5f_n], \tag{23}$$

the corresponding Runge-Kutta method is

$$\begin{array}{c|ccccccccc}
0 & & & & & & & & & \\
\frac{1}{9} & \frac{1}{9} & & & & & & & & \\
\frac{2}{9} & 0 & \frac{2}{9} & & & & & & & \\
\frac{1}{3} & \frac{1}{12} & 0 & \frac{3}{12} & & & & & & \\
0 & 0 & 0 & 0 & 0 & & & & & \\
\frac{2}{9} & 0 & 0 & 0 & 0 & \frac{2}{9} & & & & \\
\frac{4}{9} & 0 & 0 & 0 & 0 & 0 & \frac{4}{9} & & & \\
\frac{2}{3} & 0 & 0 & 0 & 0 & \frac{2}{12} & 0 & \frac{6}{12} & & \\
1 & \frac{6}{24} & 0 & \frac{3}{24} & 0 & -\frac{1}{24} & 0 & -\frac{3}{24} & \frac{19}{24} & 0 \\
\hline
 & \frac{6}{24} & 0 & \frac{3}{24} & 0 & -\frac{1}{24} & 0 & -\frac{3}{24} & \frac{19}{24} & 0
\end{array} \tag{24}$$

This method is of order $p^* = \min\{3, 3\} = 3$. Substituting the above data in LTE (7) we find that

$$\begin{aligned}
\text{LTE} &= \frac{h^4}{4!} \sum_{r(t)=4} \alpha(t)[1 - \gamma(t)\psi(t)]F(t) + O(h^5) \\
&= \frac{h^4}{4!}\Big(\frac{73}{729}y^{(4)} - \frac{7}{729}\{{}_2f^2\}_2 - \frac{28}{729}\{{}_3f\}_3\Big) + O(h^5),
\end{aligned}$$

where all functions are evaluated at $x = x_n$ and $y = y_n$. Note that this formulation maintains the actual order of both schemes, while the error term is only exact for $\{1, x\}$.

To numerical illustrations, we consider (2) with the following data [9]

$$f(x, y) = [v, v(v-1)/u]^T, \quad x \in [0, 1], \tag{25}$$

where,

$$y = [u, v]^T, \quad y(0) = [1/2, -3]^T.$$

In this test we take $N = 51$ with $h = 1/50$ for 3-step method (23) and $H = Nh = 3/50$ for the corresponding Runge-Kutta scheme (24).

Figure 4 illustrates the global error (accumulation error) of (23) for test problem (25). The estimation of GTE of (23) is shown in Figure 5. As we have proven the LTE of (24) is the GTE of (23). However, to find the true LTE we make localizing assumptions in implementation of (24), i.e., in evaluation of y_{n+1} we assume that $y_n = y(x_n)$. The comparison of the third portions of Figure 4 and Figure 5 justify the efficiency of the given estimation. The negligible difference in the error is due to roundoff errors.

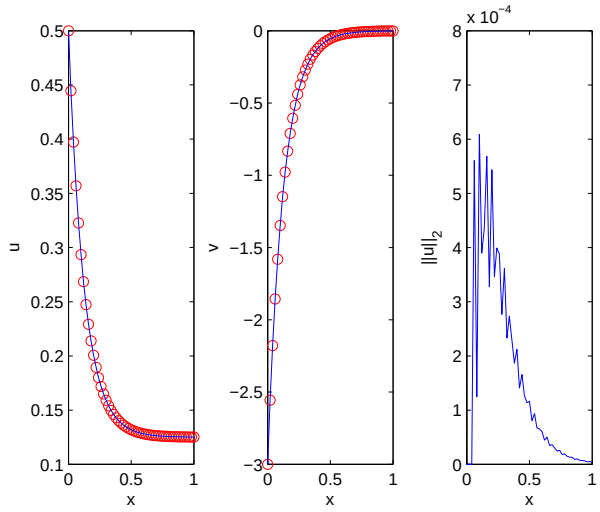

Figure 4: The numerical (red circles) and exact solutions (solid line) of (25) with (23), and the GTE of the method

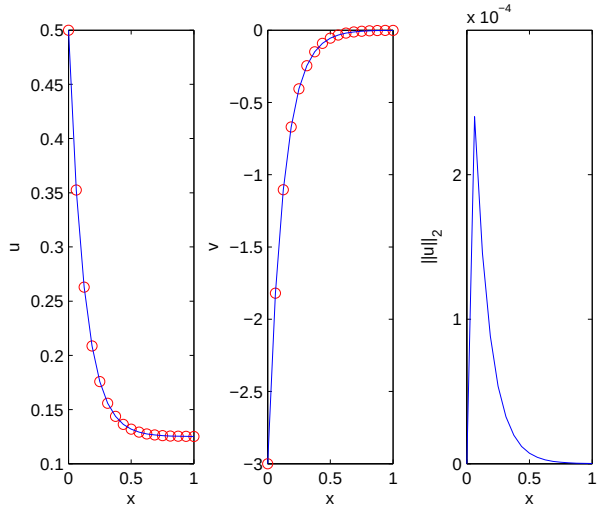

Figure 5: The numerical (red circles) and exact solutions (solid line) of (25) with (24), and the LTE of the RK method

Example 2. In this example, we consider a four-step, fourth-order total variation bounded (TVB(4,4)) linear multistep scheme (1) with the data given in Table 4 [11].

Table 4: The coefficients of the four-step, fourth-order linear multistep scheme (TVB(4,4))

	α_i	β_i
$i=0$	-0.345464734400857	-0.620278703629274
$i=1$	1.494730011212510	2.229909318681302
$i=2$	-2.777506277494861	-3.052866947601049
$i=3$	2.628241000683208	1.618795874276609
$i=4$	1	0

We utilize the fourth order classic Runge-Kutta method as a starting procedure.

$$\begin{array}{c|cccc} 0 & 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array} \tag{26}$$

It turns out that the corresponding Runge-Kutta scheme takes the form given in Butcher array (27), where the elements of b is given in the Table 5.

Table 5: The elements of vector b in Runge-Kutta scheme (27)

	b_i	b_{i+8}
$i=1$	-0.092789258773464	-0.231458856457905
$i=2$	0.124560834267709	-0.763216736900262
$i=3$	0.124560834267709	0.328530125085401
$i=4$	0.062280417133855	0.657060250170802
$i=5$	0.557477329670325	0.657060250170802
$i=6$	-0.231458856457905	0.328530125085401
$i=7$	-0.462917712915810	0.404698968569152
$i=8$	-0.462917712915810	0

$$
\begin{array}{c|ccccccccccc}
0 & 0 & & & & & & & & & & \\
\frac{1}{8} & \frac{1}{8} & & & & & & & & & & \\
\frac{1}{8} & 0 & \frac{1}{8} & & & & & & & & & \\
\frac{1}{4} & 0 & 0 & \frac{1}{4} & & & & & & & & \\
\frac{1}{4} & \frac{1}{24} & \frac{1}{12} & \frac{1}{12} & \frac{1}{24} & 0 & & & & & & \\
0 & & & & & 0 & & & & & & \\
\frac{1}{4} & & & & & \frac{1}{4} & & & & & & \\
\frac{1}{4} & & & & & 0 & \frac{1}{4} & & & & & \\
\frac{1}{2} & & & & & 0 & 0 & \frac{1}{2} & & & & \\
\frac{1}{2} & & & & & \frac{1}{12} & \frac{1}{6} & \frac{1}{6} & \frac{1}{12} & 0 & & \\
0 & & & & & & & & & 0 & & \\
\frac{3}{8} & & & & & & & & & \frac{3}{8} & & \\
\frac{3}{8} & & & & & & & & & 0 & \frac{3}{8} & \\
\frac{3}{4} & & & & & & & & & 0 & 0 & \frac{3}{4} \\
\frac{3}{4} & & & & & & & & & \frac{1}{8} & \frac{1}{4} & \frac{1}{4} & \frac{1}{8} & 0 \\
1 & & & & & & & b & & & & \\
\hline
 & & & & & & & b & & & &
\end{array}
\tag{27}
$$

The test problem is the same as the previous example. For numerical illustrations we take $N = 120$, thus we have $h = 1/120$ and accordingly $H = 4/120 = 1/30$. The numerical results including the GTE of TVB(4,4) scheme and LTE of (27) are presented in Figure 6 and Figure 7, respectively. By comparison, we find the excellent estimation of global truncation error of TVB(4,4) based on new Runge-Kutta method. The maximum error in this estimation is $2.67E-05$.

6 conclusion

In this paper, we have developed an estimation for the global truncation error of a linear multistep method. The global error analysis is more complicated in comparison with the local error analysis. The key idea is the representation of several steps of the LMM as a single step of a corresponding Runge-Kutta method. Therefore, the analysis of global error of a LMM accomplished by estimating the local truncation error the corresponding new Runge-Kutta method. We have demonstrated the theoretical aspects for some important class of linear multistep methods with total variation bounded (TVB) property, which is a crucial property in selecting an appropriate time marching method for solving nonlinear conservation laws [11].

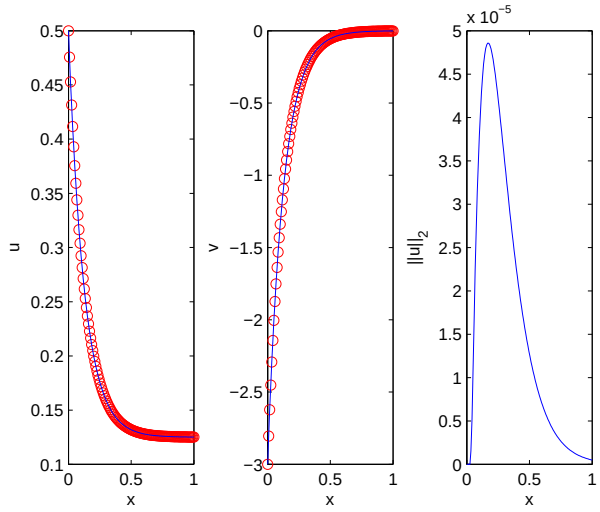

Figure 6: The numerical (circles) and exact solutions (solid line) of (25) with TVB(4,4), and the GTE of the method

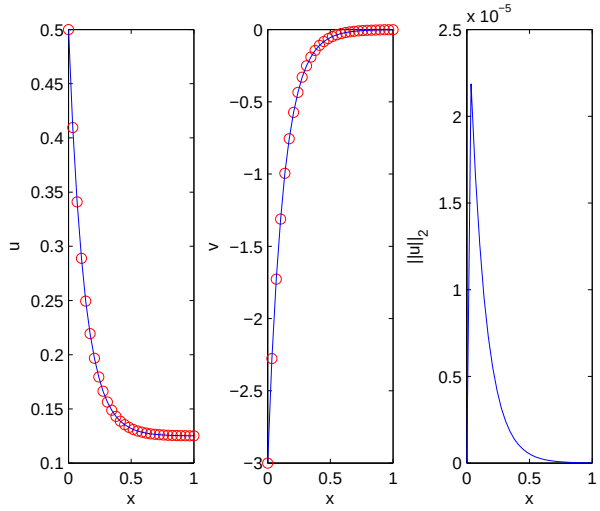

Figure 7: The numerical (circles) and exact solutions (solid line) of (25) with (27), and the LTE of the RK method

References

1. Ascher, U. M. and Petzold, L. R. *Computer methods for ordinary differential equations and differential-algebraic equations*, SIAM, Philadelphia,

2008.

2. Butcher, J. C. *Numerical methods for ordinary differential equations*, 2nd Edition, Wiley, 2008.

3. Cao, Y. and Petzold, L. *A posteriori error estimation and global error control for ordinary differential equations by the adjoint method*, SIAM J. Sci. Comput. 26 (2004), 359-374.

4. Gottlieb, S., Ketcheson, D. I. and Shu, C. W. *High order strong stability preserving time discretizations*, J. Sci. Comput. 38 (2009) 251-289.

5. Hadjimichael, Y., Ketcheson, D., Lóczi, L. and Németh, A. *Strong stability preserving explicit linear multistep methods with variable step size*, Submitted.

6. Henrici, P. *Discrete variable methods in ordinary differential equations*, Wiley, New York, 1962.

7. Henrici, P. *Error propagation for difference methods*, Wiley, New York, 1963.

8. Iserles, A. *A First course in the numerical analysis of differential equations*, Cambridge University Press, 1996.

9. Lambert, J. D., *Numerical methods for ordinarry differential systems: The initial value problem*, Wiley, 1993.

10. Press, W.H., Teukolsky, S.A. and Vetterling, W.T., Flannery, BP. *Numerical recipes: The art of scientific computing*, 3rd ed., New York: Cambridge University Press, 2007.

11. Ruuth, S. J. and Hundsdorfer, W. *High-order linear multistep methods with general monotonicity and boundedness properties*, Journal of Computational Physics, 209 (2005) 226–248.

12. Süli, E., and Mayers, D. *An Introduction to numerical analysis*, Cambridge University Press, 2003.

Persian Translation of Abstracts

انشعاب بحرانی دوسویه در مدل عفونی سرکوب کننده ی سیستم ایمنی بدن

الهام شمس آرا، ریحانه مستولی زاده و زهرا افشارنژاد

دانشگاه فردوسی مشهد، دانشکده علوم ریاضی، گروه ریاضی کاربردی

دریافت مقاله ۲ بهمن ۱۳۹۳، دریافت مقاله اصلاح شده ۱۷ آبان ۱۳۹۴، پذیرش مقاله ۱ آذر ۱۳۹۴

چکیده : در این مقاله، رفتار دینامیکی مدل عفونی سرکوب کننده ی سیستم ایمنی بدن، به ویژه ایدز، مورد تجزیه و تحلیل قرار گرفته است. ما نشان داده ایم که یک مدلسازی ساده ی ریاضی، با فرض پاسخ CTL (ایمنی بدن) به صورت تابع سیگموئید، می تواند منجر به انشعاب بحرانی دو سویه شود. معمولا این شرط، برای بیماری هایی که سیستم ایمنی بدن دچار نقص شده است (مانند ایدز) به وجود می آید و ویروس ها به سلول های ایمنی CD۴+T حمله می کنند. نتایج ما نشان می دهد که تقابل بین پاسخ ایمنی بدن و ویروس ها یHIV بسیار پیچیده است و ناحیه های دینامیکی پایدار و ناپایدار وجود دارند. در آخر با توجه به داده های کلینیکی، مثال هایی توسط شبیه سازی عددی که نشان گر نتایج اساسی است را ارائه می دهیم.

کلمات کلیدی : پاسخ CTL ؛ انشعاب بحرانی دو سویه؛ HAM/TSP .

دو روش عددی برای حل مسائل کنترل بهینه درجه دوم با قیود غیرخطی با استفاده از توابع بی-اسپلاین خطی

یوسف ادریسی تبریز[1] ، *مهرداد لکستانی*[2] *و عقیله حیدری* [1]

[1] دانشگاه پیام نور تهران، گروه ریاضی
[2] دانشگاه تبریز، دانشکده علوم ریاضی

دریافت مقاله ۴ آبان ۱۳۹۳، دریافت مقاله اصلاح شده ۲۰ فروردین ۱۳۹۴، پذیرش مقاله ۶ آذر ۱۳۹۴

چکیده : این مقاله به ارائه ی دو روش عددی برای حل مسائل کنترل بهینه ای می پردازد که دارای تابع هدف درجه دوم می باشند. همچنین ویژگی های توابع بی-اسپلاین بیان می گردد. دو ماتریس عملیاتی انتگرال مرتبط با روش ها معرفی می شوند. سپس از این ماتریسها استفاده می شود تا حل مسئله کنترل بهینه درجه دوم با قیود غیرخطی به حل مسئله برنامه ریزی غیرخطی تبدیل گردد. در انتها چندین مثال کاربردی برای نمایش کارایی و راستی آزمایی روشهای مذکور بیان می شود.

کلمات کلیدی : مسائل کنترل بهینه؛ تابع بی-اسپلاین خطی؛ ماتریس عملیاتی انتگرال؛ روش هم مکانی .

همگرایی جواب تقریبی معادلات انتگرال ولترای تأخیری

محمد ضاربنیا و لیلا شیری

دانشگاه محقق اردبیلی، دانشکده علوم ریاضی، گروه ریاضی کاربردی

دریافت مقاله ۲ تیر ۱۳۹۴، دریافت مقاله اصلاح شده ۲۵ آذر ۱۳۹۴، پذیرش مقاله ۱۷ دی ۱۳۹۴

چکیده : در این مقاله، روش هم محلی سینک برای حل معادلات انتگرال تابعی ولترا با تابع تأخیر بحث شده است. هم چنین وجود و یکتایی جواب های تقریبی برای این معادلات اثبات شده است. این روش نتایج متعارف را بهبود می بخشد و همگرایی نمایی را نتیجه می دهد. نتایج عددی برای تأیید دقت و کارایی روش ارائه شده اند.

کلمات کلیدی : معادلات انتگرال تابعی ولترا؛ تابع تأخیر؛ هم‌محلی سینک.

کنترل پدیده شبه همگرایی در روش های تکراری همزمان غیر ایستا

تورج نیک آزاد و مهدی کریم پور

دانشگاه علم و صنعت ایران، دانشکده ریاضی

دریافت مقاله ۲۸ فروردین ۱۳۹۴، دریافت مقاله اصلاح شده ۱۶ آذر ۱۳۹۴، پذیرش مقاله ۵ اسفند ۱۳۹۴

چکیده : هنگام به کارگیری روش های تکراری همزمان غیرایستا برای حل یک دستگاه بدوضع از معادلات خطی، در ابتدا معمولا خطا کاهش می یابد اما پس از چند تکرار بسته به مقدار اختلال موجود در داده ها و میزان بدوضعی دستگاه، خطا شروع به افزایش می کند. این پدیده شبه همگرایی نامیده می شود. ما رفتار شبه همگرایی رابرای روش های تکراری همزمان غیر ایستا بررسی کرده و یک کران بالا برای خطای داده (خطای اختلال) بدست می آوریم. براساس این کران ما راه های جدیدی برای تعیین پارامترهای آزاد به منظور کنترل شبه همگرایی پیشنهاد می کنیم. کارآمدی راهکارهای ما به وسیله مثال هایی که از تصویر پرتونگاری پزشکی آمده اند مشخص می شوند.

کلمات کلیدی : روش های تکراری همزمان؛ شبه همگرایی؛ پارامترهای آزاد؛ تصویر پرتونگاری پزشکی.

استفاده از توابع کلاهی اصلاح شده برای حل معادلات انتگرال کوادراتور غیر خطی

فرشید میرزایی و الهام حدادیان

، دانشگاه ملایر دانشکده علوم ریاضی و آمار

دریافت مقاله ۲۵ اردیبهشت ۱۳۹۴، دریافت مقاله اصلاح شده ۲ آذر ۱۳۹۴، پذیرش مقاله ۵ اسفند ۱۳۹۴

چکیده : این مقاله یک روش عددی برای حل معادلات انتگرال کوادراتور غیرخطی ارائه می دهد. این روش بر مبنای توابع کلاهی اصلاح شده و ماتریس عملیاتی آنها می باشد. با استفاده از این روش و نقاط هم محلی حل معادلات انتگرال کوادراتور غیرخطی به حل یک دستگاه معادلات جبری غیر خطی کاهش می یابد. روش ارائه شده برای بدست آوردن ضرایب ثابت به انتگرال گیری نیاز ندارد. از اینرو، می توان به عنوان یک تکنیک ساده و سریع مورد استفاده قرار بگیرد. تجزیه و تحلیل همگرایی و قضایای مربوط به آن مورد بررسی قرارگرفته است. با چند مثال عددی کارایی و دقت روش ارائه شده، نشان داده شده است.

کلمات کلیدی : توابع کلاهی اصلاح شده؛ معادلات انتگرال کوادراتور غیرخطی؛ فرم برداری؛ ماتریس عملیاتی؛ تجزیه و تحلیل خطا.

يك روش ماتريسی براي سيستمی از معادلات انتگرال-ديفرانسيل با استفاده از چندجمله ايهای لاگر تعميم يافته

ماشاءاله متين فر و عباس رياحی فر

دانشگاه مازندران، دانشکده علوم رياضی، گروه رياضی

دريافت مقاله ۲۴ اسفند ۱۳۹۳، دريافت مقاله اصلاح شده ۱۶ دی ۱۳۹۴، پذيرش مقاله ۵ اسفند ۱۳۹۴

چکيده : هدف از اين مقاله، ارائه يك روش ماتريسی براي حل سيستمی از معادلات انتگرال-ديفرانسيل فردهلم خطي از نوع دوم روي دامنه بي کران با هسته های جدايی پذير با جملاتی از چندجمله ای های لاگر تعميم يافته می باشد. اين روش مبتنی بر تقريب بوسيله ی سری لاگر تعميم يافته است. سيستم معادلات انتگرال-ديفرانسيل همراه با شرايط اوليه تبديل به معادلاتی ماتريسی می شود که متناظر با سيستمی از معادلات جبری خطی با مجهولاتی از ضرايب لاگر تعميم يافته است. با ترکيب معادلات ماتريسی و سپس حل سيستم مذکور می توان به ضرايب لاگر تعميم يافته دست يافت و لذا تابع جواب بدست می آيد. به علاوه در اين مقاله، چندين مثال عددی براي نشان دادن درستی و کارايی روش ارائه گرديده است.

کلمات کليدی : سيستم معادلات انتگرال-ديفرانسيل فردهلم خطی؛ دامنه ی بی کران؛ چندجمله ای های لاگر تعميم يافته؛ ماتريس عملياتی انتگرال گيری.

تخمین خطای سراسری روشهای چندگامه خطی توسط روشهای رونگه-کوتا

جواد فرضی

تبریز، دانشگاه صنعتی سهند، دانشکده علوم پایه، گروه ریاضی

دریافت مقاله ۵ دی ۱۳۹۴، دریافت مقاله اصلاح شده ۱۹ اسفند ۱۳۹۴، پذیرش مقاله ۲۳ فروردین ۱۳۹۵

چکیده : در این مقاله خطای برشی سراسری روشهای چندگامه خطی را با کمک خطای برشی محلی روشهای رونگه-کوتا مطالعه می کنیم. ایده اصلی، نمایش یک روش چندگامه خطی با یک روش رونگه -کوتای متناظر است. برای این کار باید چندگام روش چندگامه خطی را به عنوان یک گام ساده روش رونگه-کوتای متناظر در نظر بگیریم. بنابراین، خطای برشی سراسری روش چندگامه خطی از طریق روش رونگه-کوتا فراهم می شود. در این تخمین تاثیرات خطاهای گردکردن را در نظر نمی گیریم. نتایج عددی دقت و کارآمدی تخمین ارائه شده را نمایش می دهد

کلمات کلیدی : روشهای چندگامه خطی؛ روشهای رونگه-کوتا؛ خطای برشی محلی؛ خطای سراسری؛ تخمین خطا.

Aims and scope

Iranian Journal of Numerical Analysis and Optimization (IJNAO) is published twice a year by the Department of Applied Mathematics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad. Papers dealing with different aspects of numerical analysis and optimization, theories and their applications in engineering and industry are considered for publication.

Journal Policy

After receiving an article, the editorial committee will assign referees. Refereeing process can be followed via the web site of the Journal.
The manuscripts are accepted for review with the understanding that the work has not been published and also it is not under consideration for publication by any other journal. All submissions should be accompanied by a written declaration signed by the author(s) that the paper has not been published before and has not been submitted for consideration elsewhere.

Instruction for Authors

The Journal publishes all papers in the fields of numerical analysis and optimization. Articles must be written in English.
All submitted papers will be refereed and the authors may be asked to revise their manuscripts according to the referee's reports. The Editorial Board of the Journal keeps the right to accept or reject the papers for publication.
The papers with more than one authors, should determine the corresponding author. The e-mail address of the corresponding author must appear at the end of the manuscript or as a footnote of the first page.
It is strongly recommended to set up the manuscript by Latex or Tex, using the template provided in the web site of the Journal. Manuscripts should be typed double-spaced with wide margins to provide enough room for editorial remarks.
References should be arranged in alphabetical order by the surname of the first author as examples below:
[1] Stoer, J. and Bulirsch, R. *Introduction to Numerical Analysis*, Springer-verlag, New York, 2002.
[2] Brunner, H. *A survey of recent advances in the numerical treatment of Volterra integral and integro-differential equations*, J. Comput. Appl. Math. 8 (1982), 213-229.

Submission of Manuscripts

Authors may submit their manuscripts by either of the following ways:

a) Online submission (pdf or dvi files) via the website of the Journal at:

http://ijnao.um.ac.ir

b) Via journal's email mjms@um.ac.ir

Copyright Agreement

Upon the acceptance of an article by the Journal, the corresponding author will be asked to sign a "Copyright Transfer Agreement" (see the web site) and send it to the Journal address. This will permit the publisher to publish and distribute the work.

Iranian Journal of Numerical Analysis and Optimization

Former : MRJMS

CONTENTS Vol. 6, No.2, pp 1-136, 2016

web site : http://ijnao.um.ac.ir

Email : mjms@um.ac.ir

ISSN : 2423-6977 **Serial Number: 10**