Numerical analysis of an inverse problem governed by a one-dimensional hyperbolic equation

Doctoral Course of Civil and Environmental Engineering,
Intelligent Structures and Mechanics Systems Engineering,
The Graduate School of Advanced Technology and Science,
The University of Tokushima.

March, 2014

ENKHBAYAR AZJARGAL

Contents

Abstract

In the paper a simple inverse problem(Problem 2) governed by a one-dimensional hyperbolic equation is focused on. Its related problems are considered and solved numerically.

Accurate data of a direct problem is very important for solving an inverse problem. So we consider a direct problem(Problem 1). IPNS(Infinite-Precision Numerical Simulation) is applied to obtain accurate numerical results. Examples which have solutions with various smoothness are considered and solved numerically. Numerical results are satisfactory in accuracy and they are successful for showing the smoothness. They also show the possibility of the numerical distinction between the smooth solution and the analytic solution.

Inverse problems are usually ill-posed and difficult to solve numerically. Moreover, the nonuniqueness of the solution makes them even more difficult. This is because when the nonuniqueness holds, the coefficient matrix in the discretized problem may be rank deficit. The use of the multiple-precision arithmetic is necessary for the numerical computation of inverse problems. Thus, we consider the numerical computation of rank deficit of matrices in multiple precision. A simple criterion for determining rank deficit is proposed. Our criterion is applied to some simple matrices. Numerical results for these matrices show that our criterion works well. Also our criterion is applied to an example derived from Problem 2. Numerical results show that the nonuniqueness of the solutions to inverse problems may be determined numerically by using our criterion. They also show that a famous solver for a linear system does not work here.

This study gives basic and important information for solving inverse problems.

1 Introduction

1.1 Background

In mathematical physics, a direct problem is usually considered and obtained from modeling some physical system or phenomena(gravitational field, magnetic field, electromagnetic, seismic, acoustic, heat, ect)[21] with a set of parameters (model parameters) and a set of observable parameters(data). Model parameters charaterize important media properties such as density, magnetic susceptibility, electrical conductivity, wave-speed(density), heat conductivity, ect. The inverse problem can be conceptually formulated as a problem that model parameters are determined from the data. It is considered the "inverse" to the direct problem that the observable data is determined from model parameters. The solution to an inverse problem generally tells us something about a physical parameter that we cannot directly observe. Thus, inverse problems are some of the most important problems in science and mathematics.

There are various statements of an inverse problem for determining a coefficient in a wave equation in applied mathematics. They differ by the character of the data(the given information about the solution of the direct initial-boundary value problem). The theory of inverse problems has intensely been extended. Inverse problems for a hyperbolic equation of viberation of an inhomogeneous string refer to the most developed directions in the theory of inverse problems.

The general approach to solutions of inverse problems for string is based on the classical results of M.G.Krein[24]. Later A.S.Blagoveshchenskii[6] gave a new proof of the results of M.G.Krein without the use of spectral theory, namely some set of integral equations of the Gelfand-Levitan type are constructed by the data, and then the density of string is determined by the solutions of these equations. One of other effective methods is the boundary control method(BC-method)[7]. Several authors[1, 8, 9, 10] have applied the BC-method to one-dimensional inverse problems.

It is very difficult to solve differential or integral equations exactly. Thus, numerical computations which give their approximate solutions play an important role, and the development of computers and numerical methods has enabled their widespread applications. However, it is not always easy to apply these methods to inverse problems. This is because inverse problems are usually ill-posed[31]. We know that J.Hadamard first formulated the concept of the well-posed for problems in mathematical physics. A problem is said to be well-posed in the sense of Hadamard if for all admissible data, it possesses the following three properties:

(1) A solution exists.

(2) The solution is unique.

(3) The solution depends continuously on the data.

If a problem is not well-posed, it is said to be ill-posed. Ill-posed problems are very sensitive to numerical errors. This sensitivity causes the phenomenon of oscillation and destroys the numerical solutions. To avoid oscillation, regularization techniques are usually used. But choosing a suitable regularization is not easy. Moreover, regularization may lead to incorrect information about the solutions[14].

In this paper we consider two problems, one is a direct problem(Problem 1), other is an invese problem(Problem 2) corresponding to Problem 1. We study them numerically.

Problem 1. For a positive number T, find $u(x,t)$ such that

$$\begin{cases} \rho(x)u_{tt}(x,t) - u_{xx}(x,t) = 0, & x \in (0,1), \quad t \in (0,T) \\ u(x,0) = \alpha(x), \quad u_t(x,0) = \beta(x), & x \in [0,1] \\ u(0,t) = f(t), \quad u(1,t) = g(t), & t \in [0,T] \end{cases}$$

where $\rho(x)$ is given as a strictly positive function on $[0,1]$, $\alpha(x), \beta(x), f(t), g(t)$ are given functions.

Problem 1 is a problem of vibrations for inhomogeneous string, $u(x,t)$ describes a vertical displacement of the string and $\rho(x)$ is the density of string. Its inverse problem is given as follows.

Problem 2. For a positive numebr T, find $u(x,t)$ and $\rho(x)$ such that

$$\begin{cases} \rho(x)u_{tt}(x,t) - u_{xx}(x,t) = 0, & x \in (0,1), \quad t \in (0,T) \\ u(x,0) = \alpha(x), \quad u_t(x,0) = \beta(x), & x \in (0,1) \\ u(0,t) = f(t), \quad u(1,t) = g(t), & t \in [0,T] \\ u_x(0,t) = h(t), & t \in (0,T) \end{cases}$$

where $\rho(x)$ is a strictly positive unknown function on $(0,1)$, $\alpha(x), \beta(x), f(t), g(t)$ and $h(t)$ are given functions.

1.2 Motivation and purpose of the study

Motivation is as follows:

- Problem 2 is one of the basic model inverse problems used in many fields of science. It is interesting but very difficult. This is because it is sensitive to errors and the uniqueness of the solution does not always hold. Solving Problem 2 numerically and accurately is our future work. But we want to do something basic and necessary for solving Problem 2.

- Accurate data of a direct problem is very important for solving an inverse problem. So we consider Problem 1.

- Inverse problems are usually ill-posed and they are very sensitive to numerical errors. This sensitivity causes the phenomenon of oscillation and destroys the numerical solutions. To overpass the difficulty, we apply IPNS (Infinite-Precision Numerical Simulation)[17] to inverse problems.

- We have the hypothesis that if the nonuniqueness of the solution holds for an inverse problem, the coefficient matrix in the discretized problem may be rank deficit.

Purpose of the study is as follows:

(1) In order to obtain good numerical results for Problem 2 it is necessary to collect accurate data for Problem 1. So by using IPNS we solve Problem 1 accurately. We also develop a numerical method for determining the smoothness of the solution. This is because the existence of the solution is related to analyticity of data for inverse problems.

(2) The uniqueness of the solution is not expected for an inverse problem. There is a case where the nonuniqueness of the solution holds for Problem 2. We can expect the rank deficit of the coefficient matrix in the discretized problem when the nonuniqueness holds. By using IPNS, we develop a numerical method of determining the nonuniqueness of the solutions to inverse problems.

As for (1), we should remark two things: First, when we develop a numerical method for solving Problem 2, a check for the numerical method is important. If Problem 2 has the data with no sufficient small errors every method does not work. This is because inverse problems are sensitive to errors. So accurate data of Problem 1 is necessary. Secondly, in some cases the smoothness (regularity) of a function plays a very important role in the differential equation investigation. For example, the analyticity of a data is related to the existance of the solution. It is well known that if a date is smooth but not analytic then a partial differential equation has no solution [25]. So the smoothness of the solution of a direct problem is important in the study of an inverse problem. In [12] the smoothness of the solution is analyzed numerically by using the spectral method. It is investigated by the degree of convergence of numerical solutions. However, the spectral method used in [12] is of the Galerkin type. It is not easily applicable to complicated or nonlinear problems. We apply IPNS for investigating the smoothness of the solution of Problem 1. We use the spectral collocation method(SCM)[11]. It is easily applicable to complicated or nonlinear problems. If the solution to PDE is smooth, numerical errors can be reduced arbitrarily by IPNS. This means that the numerical solution converges to the smooth exact solution in arbitrary order.

As for (2), we should remark the following: Problem 2 is a nonlinear inverse problem so solving Problem 2 numerically is difficult. Moreover, the nonuniqueness of the solution makes it even more difficult. This is because when the nonuniqueness holds, the coefficient matrix in the discretized problem may be rank deficit. We want to discuss the relationship between rank deficiency and the nonuniqueness of the solution. But it is not easy to determine rank deficit of a matrix rigorously and numerically. In [19] a simple numerical method for checking the invertibility of a matrix is proposed. GEPP(Gaussian elimination with partial pivoting) and multiple-precision arithmetic are used. Authors of [19] also discussed the relationship between rank deficiency and the nonuniqueness of the solution of a problem governed by a partial differential equation. We propose a simple criterion for determining rank deficit by using multiple-precision arithmetic. GECP(Gaussian elimination with complete pivoting), GECPE(GECP with Equilibration) and SVD(Singular Value Decomposition)[15] are used. With the help of our criterion we investigate numerically the nonuniqueness of the solution to Problem 2.

1.3 Thesis outline

The thesis is organized in 5 chapters. Chapter 1 is this introduction.

In Chapter 2, numerical methods used here are presented. We present about IPNS which consists of the arbitrary-order approximation method and the multiple-precision arithmetic. We present briefly about the spectral collocation methods(SCM) and make the assertion of

analytic expression for derivative matrices of SCM with Chebyshev-Gauss-Lobatto collocation points. Problem 2 is nonlinear inverse problem. Thus, the iterative methods are necessary. We consider the Newton method and give some theorems and assertions togther. We present briefly about Gaussian Elimination with Complete Pivoting (GECP), GECP with Equilibration (GECPE) and Singular Value Decomposition (SVD)[15] for determining the rank of a matrix.

In Chapter 3, we discuss the existence of a unique solution of a one-dimensional hyperbolic equation on the semi-axis. Then we consider a direct problem(Problem 1) and solve it by using IPNS. We also investigate the smoothness of the solution. Examples which have solutions with various smoothness are considered. Numerical results are presented.

In Chapter 4, a simple criterion for determining rank deficit is proposed. The numerical computation of rank deficit of matrices is performed in multiple precision. GECP, GECPE, and SVD are compared. These methods are applied to some examples, including an example derived from an inverse problem. Numerical results are presented.

In Chapter 5, we summarize this doctoral dissertation briefly.

2 Numerical methods

2.1 Infinite-Precision Numerical Simulation

Errors in numerical simulation to PDE(Partial Differential Equation) systems originate from truncation errors in the discretization and rounding errors. IPNS (Infinite-Precision Numerical Simulation)[17] is a method for numerical simulation with ultimate accuracy. The method consists of the arbitrary order approximation (the spectral method) and the multiple-precision arithmetic.

The spectral method is used as a discretization method for the reduction of truncation errors. The use of discretization methods are inevitable in numerical simulation of differential equations. Usual methods, e.g. FDM(Finite Difference Method) or FEM(Finite Element Method), are low order approximations, so they are not suitable for the arbitrary reduction of truncation errors. On the other hand, the spectral method realizes the arbitrary order approximation when PDE has a smooth solution. So it is suitable for the arbitrary reduction of truncation errors. The spectral collocation method is commonly used in IPNS. This is because it deals with nonlinear problems more easily than others(Galerkin or tau spectral methods).

The multiple precision arithmetic[23] is used for the arbitrary reduction of rounding errors. A lot of FORTRAN subroutines about the multiple precision arithmetic are already known. We used Exflib [13] which is fast and compact.

In IPNS, truncation errors and rounding errors are controlled easily. This is very important in numerical simulations in applied mathematics. If the solution is smooth both errors can be reduced arbitrarily.

2.2 Spectral collocation methods

Spectral method may be viewed as a method of weighted residuals(MWR). The trial functions are choosed by high order polynomials; usually orthogonal polynomials (Jacobi, Chebyshev, Legendre and trigonometric polynomials). The choice of the test functions distinguishes between the three most commonly used spectral schemes, namely, the Galerkin, collocation and tau versions. The simplest version is the collocation version, in which the test function are considered by Direc delta functions centered at special, so-called collocation points[30]. The collocation method with Chebyshev polynomials is used here.

2.2.1 Chebyshev polynomials

Chebyshev polynomials of order $k(k = 0, 1, ...)$ is defined by

$$T_k(x) = \cos(k \arccos x), \qquad -1 \leq x \leq 1.$$

Proposition 2.1 For Chebyshev polynomials the following recurrence formula holds.

$$T_{k+1} = 2xT_k(x) - T_{k-1}(x), \qquad k = 1, 2, ...$$

Proof.

$$\begin{aligned} T_{k+1}(x) + T_{k-1}(x) &= \cos((k+1)\arccos x) + \cos((k-1)\arccos x) \\ &= 2\cos(k \arccos x)\cos(\arccos x) = 2xT_k(x). \end{aligned}$$

■

Proposition 2.2 Chebyshev polynomial $T_k(x)$ is represented in the form

$$T_k(x) = \sum_{l=0}^{\left[\frac{k}{2}\right]} (-1)^l C_k^{2l} x^{k-2l} (1-x^2)^l$$

Proof. Using De Moivres formula we have

$$T_k(x) = \cos k\theta = Re(e^{\imath k\theta}) = Re(\cos\theta + \imath \sin\theta)^k$$

where $\theta = \arccos x$. According to binomial theorem,

$$\begin{aligned} T_k(x) &= Re\{C_k^0 \cos^k\theta(\imath\sin\theta)^0 + C_k^1\cos^{k-1}\theta(\imath\sin\theta)^1 + ... + C_k^k\cos^0\theta(\imath\sin\theta)^k\} \\ &= C_k^0\cos^k\theta - C_k^2\cos^{k-2}\theta\sin^2\theta + ... + (-1)^{\left[\frac{k}{2}\right]} C_k^{2\left[\frac{k}{2}\right]} \cos^{k-2\left[\frac{k}{2}\right]}\theta \sin^{2\left[\frac{k}{2}\right]}\theta \\ &= \sum_{l=0}^{\left[\frac{k}{2}\right]} (-1)^l C_k^{2l}\cos^{k-2l}\theta\sin^{2l}\theta. \end{aligned}$$

Putting $\theta = \arccos x$ in the equality above, we obtain

$$T_k(x) = \sum_{l=0}^{\left[\frac{k}{2}\right]} (-1)^l C_k^{2l} x^{k-2l}(1-x^2)^l. \qquad \blacksquare$$

Proposition 2.3 The following properties are satisfied.

(1) $|T_k(x)| \le 1, \qquad -1 \le x \le 1$

(2) $T_k(\pm 1) = (\pm 1)^k$

(3) $T_k'(\pm 1) = (\pm 1)^{k+1} k^2$

(4) $T_k''(\pm 1) = (\pm 1)^k \cdot \dfrac{k^2(k^2-1)}{3}$

(5) $$\int_{-1}^{1} \frac{T_k(x)T_l(x)}{\sqrt{1-x^2}}dx = \begin{cases} 0, & k \ne l \\ \frac{\pi}{2}, & k = l \ne 0 \\ \pi, & k = l = 0 \end{cases}$$

Proof.

(1) $|T_k(x)| = |\cos(k\arccos x)| \le 1, \qquad -1 \le x \le 1.$

(2) $T_k(1) = \cos(k\arccos 1) = \cos(k\cdot 0) = 1,$

$T_k(-1) = \cos(k\arccos(-1)) = \cos(k\cdot\pi) = (-1)^k.$

(3) For $T_k(x) = \cos(k\arccos x), \qquad -1 < x < 1$

$T_k'(x) = -\sin(k\arccos x)\cdot k \cdot \frac{-1}{\sqrt{1-x^2}} = \frac{k}{\sqrt{1-x^2}}\sin(k\arccos x).$

Since $T_k'(x)$ is a continuous function on $(-1,1)$,

$$T_k'(1) = \lim_{x\to 1} T_k'(x) = \lim_{x\to 1} \frac{k\sin(k\arccos x)}{\sqrt{1-x^2}}$$

and

$$T_k'(-1) = \lim_{x\to -1} T_k'(x) = \lim_{x\to -1} \frac{k\sin(k\arccos x)}{\sqrt{1-x^2}}.$$

For the substitute $t = \arccos x$ we can conclude that $t \to 0$ as $x \to 1$ and $t \to \pi$ as $x \to -1$. Therefore

$$T_k'(1) = \lim_{t\to 0} \frac{k\sin kt}{\sin t} = \lim_{t\to 0} \frac{k^2 \cdot \dfrac{\sin kt}{kt}}{\dfrac{\sin t}{t}} = k^2$$

$$T_k'(-1) = \lim_{t\to \pi} \frac{k\sin kt}{\sin t} = \lim_{t\to \pi} \frac{(k\sin kt)'}{(\sin t)'} = \lim_{t\to \pi} \frac{k^2\cos kt}{\cos t} = \frac{(-1)^k k^2}{-1} = (-1)^{k+1}k^2.$$

(4) For $T_k'(x) = \dfrac{k}{\sqrt{1-x^2}}\sin(k\arccos x), \qquad -1 < x < 1$

$$T_k''(x) = \frac{-k}{2(\sqrt{1-x^2})^3}\cdot(-2x)\cdot\sin(k\arccos x) + \frac{k}{\sqrt{1-x^2}}\cdot\cos(k\arccos x)\cdot k\cdot\frac{-1}{\sqrt{1-x^2}}$$

$$= \frac{kx}{(\sqrt{1-x^2})^3}\cdot\sin(k\arccos x) - \frac{k^2}{1-x^2}\cdot\cos(k\arccos x).$$

Since $T_k''(x)$ is a continuous function on $(-1,1)$,

$$T_k''(1) = \lim_{x\to 1} T_k''(x) = \lim_{x\to 1}\left[\frac{kx}{(\sqrt{1-x^2})^3}\cdot\sin(k\arccos x) - \frac{k^2}{1-x^2}\cdot\cos(k\arccos x)\right]$$

and

$$T_k''(-1) = \lim_{x\to -1} T_k''(x) = \lim_{x\to -1}\left[\frac{kx}{(\sqrt{1-x^2})^3}\cdot\sin(k\arccos x) - \frac{k^2}{1-x^2}\cdot\cos(k\arccos x)\right].$$

By the same way we have

$$T_k''(1) = \lim_{t\to 0}\left(\frac{k\cos t\sin kt}{\sin^3 t} - \frac{k^2\cos kt}{\sin^2 t}\right) = \lim_{t\to 0}\left(\frac{k\cos t\sin kt - k^2\sin t\cos kt}{\sin^3 t}\right)$$

$$= \lim_{t\to 0}\frac{(k\cos t\sin kt - k^2\sin t\cos kt)'}{(\sin^3 t)'} = \lim_{t\to 0}\frac{(k^3-k)\sin t\sin kt}{3\sin^2 t\cos t}$$

$$= \lim_{t\to 0}\frac{(k^3-k)\sin kt}{3\sin t\cos t} = \lim_{t\to 0}\frac{k^3-k}{3}\cdot\frac{\sin kt}{\sin t}\cdot\frac{1}{\cos t} = \frac{k^2(k^2-1)}{3}.$$

$$T_k''(-1) = \lim_{t\to \pi}\left(\frac{k\cos t\sin kt}{\sin^3 t} - \frac{k^2\cos kt}{\sin^2 t}\right).$$

By substituting $t = s + \pi$ into the last limit, we have

$$T_k''(-1) = \lim_{s\to 0}\left(\frac{k(-\cos s)(-1)^{k+1}\sin ks}{\sin^3 s} - \frac{k^2(-1)^k\cos ks}{\sin^2 s}\right) = \frac{(-1)^k k^2(k^2-1)}{3}.$$

(5) We use the substitute $x = \cos\theta (0 \le \theta \le \pi)$ then $\dfrac{d\theta}{dx} = -\dfrac{1}{\sqrt{1-x^2}}$.

$$\int_{-1}^{1} \frac{T_k(x)T_l(x)}{\sqrt{1-x^2}}dx = \int_{\pi}^{0} -\cos k\theta \cos l\theta d\theta = \frac{1}{2}\int_{0}^{\pi} (\cos(k+l)\theta + \cos(k-l)\theta)d\theta$$

If $k \neq l$, then we obtain

$$= \frac{1}{2}\left[\frac{1}{k+l}\sin(k+l)\theta + \frac{1}{k-l}\sin(k-l)\theta\right]_0^{\pi} = 0,$$

If $k = l > 0$, then we obtain

$$= \frac{1}{2}\left[\frac{1}{2k}\sin 2k\theta + \theta\right]_0^{\pi} = \frac{\pi}{2},$$

If $k = l = 0$, then we obtain

$$= \frac{1}{2}[2\theta]_0^{\pi} = \pi. \qquad \blacksquare$$

For arbitrary function $f(x)$ on $(-1,1)$, the norm $\|f\|_{L^2_w(-1,1)}$ is defined by

$$\|f\|_{L^2_w(-1,1)} = \left(\int_{-1}^{1} f^2(x)w(x)dx\right)^{-\frac{1}{2}}$$

where $w(x) = (1-x^2)^{-\frac{1}{2}}$. Denote by $L^2_w(-1,1)$ a set of all functions $f(x)$ such that

$$\|f\|_{L^2_w(-1,1)} < \infty.$$

Chebyshev polynomials form an orthogonal set in the Hilbert space $L^2_w(-1,1)$. An arbitrary function f in $L^2_w(-1,1)$ can be expanded as a series of Chebyshev polynomials[11]:

$$f(x) = \sum_{k=0}^{\infty} \bar{f}_k T_k(x)$$

where the coefficients u_k are given by

$$\bar{f}_0 = \frac{1}{\pi}\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}dx \qquad \text{and} \qquad \bar{f}_k = \frac{2}{\pi}\int_{-1}^{1} \frac{f(x)T_k(x)}{\sqrt{1-x^2}}dx, \quad k = 1, 2, ...$$

2.2.2 Spectral method with Chebyshev-Gauss-Lobatto collocation points.

Let $u(x)$ be function on $[-1,1]$. The function $u(x)$ is approximated by the polynomial $u^M(x)$ of form

$$u^M(x) = \sum_{k=0}^{M} \tilde{u}_k T_k(x) \tag{2.1}$$

In spectral collocation method, $u^M(x)$ should be satisfied that

$$u^M(x_j) = u(x_j)$$

on $(x_i)_{0\leq i\leq N}$(Spectral collocation points).

Chebyshev-Gauss-Lobatto(C-G-L) collocation points are given as follows:

$$x_j = \cos\frac{j\pi}{N}, \qquad j = 0, 1, ..., N$$

For C-G-L collocation points $x_j (j = 0, 1, ..., N)$,

$$u_j = u(x_j) = \sum_{k=0}^{M} \tilde{u}_k T_k(x_j), \tag{2.2}$$

and $M = N$ then an inverse relationship is

$$\tilde{u}_k = \frac{2}{N\bar{c}_k} \sum_{j=0}^{N} \frac{1}{\bar{c}_j} u_j T_k(x_j), \qquad k = 0, 1, ..., N \tag{2.3}$$

where

$$\bar{c}_j = \begin{cases} 2, & j = 0, N \\ 1, & 1 \le j \le N-1. \end{cases}$$

Substituting (2.3) into (2.1), we obtain

$$u^N(x) = \sum_{j=0}^{N} \left(\frac{1}{\bar{c}_j} \sum_{k=0}^{N} \frac{2}{N\bar{c}_k} T_k(x_j) T_k(x) \right) u_j. \tag{2.4}$$

The first and the second derivatives of $u^N(x)$ with respect to x are

$$(u^N)'(x) = \sum_{j=0}^{N} \left(\frac{1}{\bar{c}_j} \sum_{k=0}^{N} \frac{2}{N\bar{c}_k} T_k(x_j) T_k'(x) \right) u_j,$$

$$(u^N)''(x) = \sum_{j=0}^{N} \left(\frac{1}{\bar{c}_j} \sum_{k=0}^{N} \frac{2}{N\bar{c}_k} T_k(x_j) T_k''(x) \right) u_j.$$

Substituting C-G-L collocation point $x_l (l = 0, 1, ..., N)$ into the first and the second derivatives, we can write as

$$(u^N)'(x_l) = \sum_{j=0}^{N} (D_x)_{l,j} u_j,$$

$$(u^N)''(x_l) = \sum_{j=0}^{N} (D_{xx})_{l,j} u_j.$$

where $(D_x)_{l,j}$ and $(D_{xx})_{l,j}$ are defined by

$$(D_x)_{l,j} = \frac{1}{N\bar{c}_j} \sum_{k=0}^{N} \frac{2}{\bar{c}_k} T_k(x_j) T_k'(x_l),$$

and

$$(D_{xx})_{l,j} = \frac{1}{N\bar{c}_j} \sum_{k=0}^{N} \frac{2}{\bar{c}_k} T_k(x_j) T_k''(x_l).$$

$(D_x)_{l,j}$ and $(D_{xx})_{l,j}$ are called first and second derivative matrices, respectively. For C-G-L points we have the following propositions.

Proposition 2.4 The first derivative matrix can be given as

$$(D_x)_{l,j} = \begin{cases} \dfrac{\bar{c}_l}{\bar{c}_j}\dfrac{(-1)^{l+j+1}}{x_j - x_l}, & l \neq j, \quad l = 0, 1, ..., N, \quad j = 0, 1, ..., N, \\ -\dfrac{x_j}{2(1-x_j^2)}, & 2 \leq l = j \leq N-1, \\ \dfrac{2N^2+1}{6}, & l = j = 0, \\ -\dfrac{2N^2+1}{6}, & l = j = N. \end{cases}$$

Proof. With aid of (3) in the proof of Proposition 2.3 we can easily obtain that

$$T_k'(x_l) = \frac{k \sin \frac{kl\pi}{N}}{\sin \frac{l\pi}{N}}.$$

If $l \neq j$ then

$$(D_x)_{l,j} = \frac{1}{\bar{c}_j} \sum_{k=0}^{N} \frac{2}{N\bar{c}_k} T_k(x_j) T_k'(x_l) = \frac{1}{\bar{c}_j N \sin \frac{l\pi}{N}} \sum_{k=0}^{N} \frac{2k}{\bar{c}_k} \cos \frac{kj\pi}{N} \sin \frac{kl\pi}{N}$$

$$= \frac{1}{\bar{c}_j N \sin \frac{l\pi}{N}} \sum_{k=1}^{N-1} 2k \cos \frac{kj\pi}{N} \sin \frac{kl\pi}{N} = \frac{1}{\bar{c}_j N \sin \frac{l\pi}{N}} \sum_{k=1}^{N-1} k \left(\sin \frac{k(l+j)\pi}{N} + \sin \frac{k(l-j)\pi}{N} \right).$$

Assume that $i = \sqrt{-1}$, $\varphi = \dfrac{l\pi}{N}$ and $\psi = \dfrac{j\pi}{N}$ then we have

$$(D_x)_{l,j} = \frac{1}{\bar{c}_j N \sin \varphi} \sum_{k=1}^{N-1} k \left(\sin k(\varphi + \psi) + \sin k(\varphi - \psi) \right)$$

$$= \frac{1}{\bar{c}_j N \sin \varphi} \sum_{k=1}^{N-1} k \left(\frac{e^{ik(\varphi+\psi)} - e^{-ik(\varphi+\psi)}}{2i} + \frac{e^{ik(\varphi-\psi)} - e^{-ik(\varphi-\psi)}}{2i} \right).$$

By using the formula:

$$r + 2r^2 + ... + (n-1)r^{n-1} = \frac{(n-1)r^{n+1} + r - nr^n}{(r-1)^2}, \qquad (r \neq 0, r \neq 1)$$

we obtain

$$
\begin{aligned}
(D_x)_{l,j} &= \frac{1}{2i\bar{c}_j N \sin\varphi}\left(\frac{(N-1)e^{i(N+1)(\varphi+\psi)}+e^{i(\varphi+\psi)}-Ne^{iN(\varphi+\psi)}}{(e^{i(\varphi+\psi)}-1)^2}\right)\\
&-\frac{1}{2i\bar{c}_j N \sin\varphi}\left(\frac{(N-1)e^{-i(N+1)(\varphi+\psi)}+e^{-i(\varphi+\psi)}-Ne^{-iN(\varphi+\psi)}}{(e^{-i(\varphi+\psi)}-1)^2}\right)\\
&+\frac{1}{2i\bar{c}_j N \sin\varphi}\left(\frac{(N-1)e^{i(N+1)(\varphi-\psi)}+e^{i(\varphi-\psi)}-Ne^{iN(\varphi-\psi)}}{(e^{i(\varphi-\psi)}-1)^2}\right)\\
&-\frac{1}{2i\bar{c}_j N \sin\varphi}\left(\frac{(N-1)e^{-i(N+1)(\varphi-\psi)}+e^{-i(\varphi-\psi)}-Ne^{-iN(\varphi-\psi)}}{(e^{-i(\varphi-\psi)}-1)^2}\right).
\end{aligned}
$$

By taking into account

$$
\begin{aligned}
e^{iN(\varphi+\psi)} &= \cos(l+j)\pi + i\sin(l+j)\pi = (-1)^{l+j}\\
e^{-iN(\varphi+\psi)} &= \cos(l+j)\pi - i\sin(l+j)\pi = (-1)^{l+j}\\
e^{iN(\varphi-\psi)} &= \cos(l-j)\pi + i\sin(l-j)\pi = (-1)^{l-j}\\
e^{-iN(\varphi-\psi)} &= \cos(l-j)\pi - i\sin(l-j)\pi = (-1)^{l-j}
\end{aligned}
$$

we obtain

$$
\begin{aligned}
(D_x)_{l,j} &= \frac{1}{2i\bar{c}_j N \sin\varphi}\left(\frac{(-1)^{l+j}(N-1)e^{i(\varphi+\psi)}+e^{i(\varphi+\psi)}-(-1)^{l+j}N}{(e^{i(\varphi+\psi)}-1)^2}\right)\\
&-\frac{1}{2i\bar{c}_j N \sin\varphi}\left(\frac{(-1)^{l+j}(N-1)e^{-i(\varphi+\psi)}+e^{-i(\varphi+\psi)}-(-1)^{l-j}N}{(e^{-i(\varphi+\psi)}-1)^2}\right)\\
&+\frac{1}{2i\bar{c}_j N \sin\varphi}\left(\frac{(-1)^{l-j}(N-1)e^{i(\varphi-\psi)}+e^{i(\varphi-\psi)}-(-1)^{l+j}N}{(e^{i(\varphi-\psi)}-1)^2}\right)\\
&-\frac{1}{2i\bar{c}_j N \sin\varphi}\left(\frac{(-1)^{l-j}(N-1)e^{-i(\varphi-\psi)}+e^{-i(\varphi-\psi)}-(-1)^{l-j}N}{(e^{-i(\varphi-\psi)}-1)^2}\right)\\
&= \frac{1}{2i\bar{c}_j N \sin\varphi}\left(\frac{(-1)^{l+j}(N-1)e^{i(\varphi+\psi)}+e^{i(\varphi+\psi)}-(-1)^{l+j}N}{(e^{i(\varphi+\psi)}-1)^2}\right)\\
&-\frac{1}{2i\bar{c}_j N \sin\varphi}\left(\frac{(-1)^{l+j}(N-1)e^{i(\varphi+\psi)}+e^{i(\varphi+\psi)}-(-1)^{l-j}Ne^{2i(\varphi+\psi)}}{(e^{i(\varphi+\psi)}-1)^2}\right)\\
&+\frac{1}{2i\bar{c}_j N \sin\varphi}\left(\frac{(-1)^{l-j}(N-1)e^{i(\varphi-\psi)}+e^{i(\varphi-\psi)}-(-1)^{l+j}N}{(e^{i(\varphi-\psi)}-1)^2}\right)\\
&-\frac{1}{2i\bar{c}_j N \sin\varphi}\left(\frac{(-1)^{l-j}(N-1)e^{i(\varphi-\psi)}+e^{i(\varphi-\psi)}-(-1)^{l-j}Ne^{2i(\varphi-\psi)}}{(e^{i(\varphi-\psi)}-1)^2}\right).
\end{aligned}
$$

By using $(-1)^{l-j} = (-1)^{l+j}$ we obtain

$$\begin{aligned}(D_x)_{l,j} &= \frac{1}{2i\bar{c}_j N \sin\varphi}\left(\frac{(-1)^{l+j}N(e^{2i(\varphi+\psi)}-1)}{(e^{i(\varphi+\psi)}-1)^2}\right) + \frac{1}{2i\bar{c}_j N \sin\varphi}\left(\frac{(-1)^{l+j}N(e^{2i(\varphi-\psi)}-1)}{(e^{i(\varphi-\psi)}-1)^2}\right)\\ &= \frac{(-1)^{l+j}}{2i\bar{c}_j \sin\varphi}\left(\frac{e^{i(\varphi+\psi)}+1}{(e^{i(\varphi+\psi)}-1)} + \frac{e^{i(\varphi-\psi)}+1}{(e^{i(\varphi-\psi)}-1)}\right)\\ &= \frac{(-1)^{l+j}}{2i\bar{c}_j \sin\varphi}\left(\frac{\cos\frac{\varphi+\psi}{2}}{i\sin\frac{\varphi+\psi}{2}} + \frac{\cos\frac{\varphi-\psi}{2}}{i\sin\frac{\varphi-\psi}{2}}\right) = \frac{(-1)^{l+j}}{-2\bar{c}_j \sin\varphi}\left(\frac{\sin\left(\frac{\varphi+\psi}{2}+\frac{\varphi-\psi}{2}\right)}{\sin\frac{\varphi+\psi}{2}\sin\frac{\varphi-\psi}{2}}\right)\\ &= \frac{1}{2\bar{c}_j}\left(\frac{(-1)^{l+j+1}}{\sin\frac{\varphi+\psi}{2}\sin\frac{\varphi-\psi}{2}}\right) = \frac{1}{2\bar{c}_j}\left(\frac{(-1)^{l+j+1}}{\sin\frac{(l+j)\pi}{2N}\sin\frac{(l-j)\pi}{2N}}\right)\\ &= \frac{1}{\bar{c}_j}\left(\frac{(-1)^{l+j+1}}{\cos\frac{j\pi}{N}-\cos\frac{l\pi}{N}}\right) = \frac{1}{\bar{c}_j}\frac{(-1)^{l+j+1}}{x_j - x_l}.\end{aligned}$$

If $1 \le l = j \le N-1$ then

$$\begin{aligned}(D_x)_{j,j} &= \frac{1}{\bar{c}_j}\sum_{k=0}^{N}\frac{2}{N\bar{c}_k}T_k(x_j)T_k'(x_j) = \frac{1}{\bar{c}_j N\sin\frac{j\pi}{N}}\sum_{k=0}^{N}\frac{2k}{\bar{c}_k}\cos\frac{kj\pi}{N}\sin\frac{kj\pi}{N}\\ &= \frac{1}{\bar{c}_j N\sin\frac{j\pi}{N}}\sum_{k=0}^{N}k\sin\frac{2kj\pi}{N} = \frac{1}{\bar{c}_j N\sin\frac{j\pi}{N}}\sum_{k=1}^{N-1}k\sin\frac{2kj\pi}{N}.\end{aligned}$$

By the manner of the previous case we can obtain

$$\begin{aligned}(D_x)_{j,j} &= \frac{1}{2i\bar{c}_j\sin\psi}\left(\frac{e^{2i\psi}+1}{e^{2i\psi}-1}\right) = \frac{1}{2i\bar{c}_j\sin\psi}\frac{\cos\psi}{i\sin\psi} = -\frac{\cos\frac{j\pi}{N}}{2\sin^2\frac{j\pi}{N}}\\ &= -\frac{\cos\frac{j\pi}{N}}{2(1-\cos^2\frac{j\pi}{N})} = -\frac{x_j}{2\left(1-x_j^2\right)}.\end{aligned}$$

If $l = j = 0$ then

$$\begin{aligned}(D_x)_{0,0} &= \frac{1}{\bar{c}_0}\sum_{k=0}^{N}\frac{2}{N\bar{c}_k}T_k(x_0)T_k'(x_0) = \frac{1}{\bar{c}_0}\frac{2}{N\bar{c}_0}T_0(x_0)T_0'(x_0) + \frac{1}{\bar{c}_0}\sum_{k=1}^{N-1}\frac{2}{N\bar{c}_k}T_k(x_0)T_k'(x_0)\\ &\quad + \frac{1}{\bar{c}_0}\frac{2}{N\bar{c}_N}T_N(x_0)T_N'(x_0) = \frac{1}{N}\sum_{k=1}^{N-1}k^2 + \frac{N}{2}\\ &= \frac{1}{N}\cdot\frac{(N-1)N(2N-1)}{6} + \frac{N}{2} = \frac{(N-1)(2N-1)}{6} + \frac{N}{2} = \frac{2N^2+1}{6}.\end{aligned}$$

If $l = j = N$ then

$$(D_x)_{N,N} = \frac{1}{\bar{c}_N}\sum_{k=0}^{N}\frac{2}{N\bar{c}_k}T_k(x_N)T_k'(x_N) = \frac{1}{\bar{c}_N}\frac{2}{N\bar{c}_0}T_0(x_N)T_0'(x_N) + \frac{1}{\bar{c}_N}\sum_{k=1}^{N-1}\frac{2}{N\bar{c}_k}T_k(x_N)T_k'(x_N)$$

$$+\frac{1}{\bar{c}_N}\frac{2}{N\bar{c}_N}T_N(x_N)T_N'(x_N) = \frac{1}{N}\sum_{k=1}^{N-1}(-1)^{2k+1}k^2 + (-1)^{2N+1}\frac{N}{2}$$

$$= -\frac{1}{N}\cdot\frac{(N-1)N(2N-1)}{6} - \frac{N}{2} = -\frac{(N-1)(2N-1)}{6} - \frac{N}{2} = -\frac{2N^2+1}{6}. \qquad \blacksquare$$

Proposition 2.5 The second derivative matrix can be given as

$$(D_{xx})_{l,j} = \begin{cases} \dfrac{(-1)^{l+j}}{\bar{c}_j}\dfrac{x_l^2 + x_l x_j - 2}{(1-x_l^2)(x_l - x_j)^2}, & l \neq j, 1 \leq l \leq N-1 \\ -\dfrac{(N^2-1)(1-x_l^2)+3}{3(1-x_l^2)^2}, & 1 \leq l = j \leq N-1 \\ \dfrac{2}{3}\dfrac{(-1)^j}{\bar{c}_j}\dfrac{(2N^2+1)(1-x_j)-6}{(1-x_j)^2}, & l = 0, 1 \leq j \leq N \\ \dfrac{2}{3}\dfrac{(-1)^{j+N}}{\bar{c}_j}\dfrac{(2N^2+1)(1+x_j)-6}{(1+x_j)^2}, & l = N, 0 \leq j \leq N-1 \\ \dfrac{N^4-1}{15}, & l = j = 0 \quad \text{or} \quad l = j = N \end{cases}$$

Proof. With aid of (3) in the proof of Proposition 2.3 we can verify that

$$T_k''(x_l) = \frac{k\cos\frac{l\pi}{N}}{\sin^3\frac{l\pi}{N}}\cdot\sin\frac{kl\pi}{N} - \frac{k^2}{\sin^2\frac{l\pi}{N}}\cdot\cos\frac{kl\pi}{N}.$$

If $l \neq j, 1 \leq l \leq N-1$ then

$$\begin{aligned}(D_{xx})_{l,j} &= \frac{1}{\bar{c}_j}\sum_{k=0}^{N}\frac{2}{N\bar{c}_k}T_k(x_j)T_k''(x_l) = \frac{\cos\frac{l\pi}{N}}{\bar{c}_j N\sin^3\frac{l\pi}{N}}\sum_{k=0}^{N}\frac{2k}{\bar{c}_k}\cos\frac{kj\pi}{N}\sin\frac{kl\pi}{N} \\ &\quad -\frac{1}{\bar{c}_j N\sin^2\frac{l\pi}{N}}\sum_{k=0}^{N-1}\frac{2k^2}{\bar{c}_k}\cos\frac{kj\pi}{N}\cos\frac{kl\pi}{N} + \frac{(-1)^{l+j+1}N}{\bar{c}_j\sin^2\frac{l\pi}{N}}.\end{aligned}$$

For simplicity $\dfrac{l\pi}{N}$ and $\dfrac{j\pi}{N}$ are denoted by φ and ψ respectively, then we have

$$\begin{aligned}(D_{xx})_{l,j} &= \frac{\cos\varphi}{\bar{c}_j N\sin^3\varphi}\sum_{k=1}^{N-1}k\left(\sin k(\varphi+\psi) + \sin k(\varphi-\psi)\right) + \frac{(-1)^{l+j+1}N}{\bar{c}_j\sin^2\varphi} \\ &\quad -\frac{1}{\bar{c}_j N\sin^2\varphi}\sum_{k=1}^{N-1}k^2\left(\cos k(\varphi+\psi) + \cos k(\varphi-\psi)\right).\end{aligned}$$

By using the calculation of the proof of Proposition 2.4, we obtain

$$
\begin{aligned}
(D_{xx})_{l,j} &= \frac{\cos\varphi}{2\bar{c}_j \sin^2\varphi}\left(\frac{(-1)^{l+j+1}}{\sin\dfrac{\varphi+\psi}{2}\sin\dfrac{\varphi-\psi}{2}}\right)+\frac{(-1)^{l+j+1}N}{\bar{c}_j\sin^2\frac{l\pi}{N}} \\
&\quad -\frac{1}{\bar{c}_j N\sin^2\varphi}\sum_{k=1}^{N-1}k^2\left(\frac{e^{ik(\varphi+\psi)}+e^{-ik(\varphi+\psi)}}{2}+\frac{e^{ik(\varphi-\psi)}+e^{-ik(\varphi-\psi)}}{2}\right)
\end{aligned}
$$

By using the following equalizations:

$$
\begin{aligned}
&r+4r^2+9r^3+...+(n-1)^2r^{n-1} \\
&\qquad =\frac{(n^2-2n+1)r^{n+2}-(2n^2-2n-1)r^{n+1}-r-r^2+n^2r^n}{(r-1)^3} \qquad (r\neq 0, r\neq 1),
\end{aligned}
$$

$$
e^{iN(\varphi+\psi)}=(-1)^{l+j}=e^{-iN(\varphi+\psi)}, \qquad e^{iN(\varphi-\psi)}=(-1)^{l-j}=e^{-iN(\varphi-\psi)}
$$

we obtain

$$
\begin{aligned}
(D_{xx})_{l,j} &= \frac{\cos\varphi}{2\bar{c}_j \sin^2\varphi}\left(\frac{(-1)^{l+j+1}}{\sin\dfrac{\varphi+\psi}{2}\sin\dfrac{\varphi-\psi}{2}}\right)+\frac{(-1)^{l+j+1}N}{\bar{c}_j\sin^2\varphi} \\
&\quad +\frac{1}{2\bar{c}_j N\sin^2\varphi}\left(\frac{(-1)^{l+j+1}(N^2-2N+1)e^{i2(\varphi+\psi)}-(-1)^{l+j+1}(2N^2-2N-1)e^{i(\varphi+\psi)}}{(e^{i(\varphi+\psi)}-1)^3}\right) \\
&\quad +\frac{1}{2\bar{c}_j N\sin^2\varphi}\left(\frac{(-1)^{l+j+1}N^2+e^{i(\varphi+\psi)}+e^{i2(\varphi+\psi)}}{(e^{i(\varphi+\psi)}-1)^3}\right) \\
&\quad +\frac{1}{2\bar{c}_j N\sin^2\varphi}\left(\frac{(-1)^{l+j+1}(N^2-2N+1)e^{-i2(\varphi+\psi)}-(-1)^{l+j+1}(2N^2-2N-1)e^{-i(\varphi+\psi)}}{(e^{-i(\varphi+\psi)}-1)^3}\right) \\
&\quad +\frac{1}{2\bar{c}_j N\sin^2\varphi}\left(\frac{(-1)^{l+j+1}N^2+e^{-i(\varphi+\psi)}+e^{-i2(\varphi+\psi)}}{(e^{-i(\varphi+\psi)}-1)^3}\right) \\
&\quad +\frac{1}{2\bar{c}_j N\sin^2\varphi}\left(\frac{(-1)^{l+j+1}(N^2-2N+1)e^{i2(\varphi-\psi)}-(-1)^{l+j+1}(2N^2-2N-1)e^{i(\varphi-\psi)}}{(e^{i(\varphi-\psi)}-1)^3}\right) \\
&\quad +\frac{1}{2\bar{c}_j N\sin^2\varphi}\left(\frac{(-1)^{l+j+1}N^2+e^{i(\varphi-\psi)}+e^{i2(\varphi-\psi)}}{(e^{i(\varphi-\psi)}-1)^3}\right) \\
&\quad +\frac{1}{2\bar{c}_j N\sin^2\varphi}\left(\frac{(-1)^{l+j+1}(N^2-2N+1)e^{-i2(\varphi-\psi)}-(-1)^{l+j+1}(2N^2-2N-1)e^{-i(\varphi-\psi)}}{(e^{-i(\varphi-\psi)}-1)^3}\right) \\
&\quad +\frac{1}{2\bar{c}_j N\sin^2\varphi}\left(\frac{(-1)^{l+j+1}N^2+e^{-i(\varphi-\psi)}+e^{-i2(\varphi-\psi)}}{(e^{-i(\varphi-\psi)}-1)^3}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\cos\varphi}{2\bar{c}_j \sin^2\varphi}\left(\frac{(-1)^{l+j+1}}{\sin\dfrac{\varphi+\psi}{2}\sin\dfrac{\varphi-\psi}{2}}\right) + \frac{(-1)^{l+j+1}N}{\bar{c}_j \sin^2\varphi} \\
&\quad + \frac{1}{2\bar{c}_j N \sin^2\varphi}\left(\frac{(-1)^{l+j+1}(N^2-2N+1)e^{i2(\varphi+\psi)} - (-1)^{l+j+1}(2N^2-2N-1)e^{i(\varphi+\psi)}}{(e^{i(\varphi+\psi)}-1)^3}\right) \\
&\qquad + \frac{1}{2\bar{c}_j N \sin^2\varphi}\left(\frac{(-1)^{l+j+1}N^2 + e^{i(\varphi+\psi)} + e^{i2(\varphi+\psi)}}{(e^{i(\varphi+\psi)}-1)^3}\right) \\
&\quad - \frac{1}{2\bar{c}_j N \sin^2\varphi}\left(\frac{(-1)^{l+j+1}(N^2-2N+1)e^{i(\varphi+\psi)} - (-1)^{l+j+1}(2N^2-2N-1)e^{i2(\varphi+\psi)}}{(e^{i(\varphi+\psi)}-1)^3}\right) \\
&\qquad - \frac{1}{2\bar{c}_j N \sin^2\varphi}\left(\frac{(-1)^{l+j+1}N^2 e^{i3(\varphi+\psi)} + e^{i2(\varphi+\psi)} + e^{i(\varphi+\psi)}}{(e^{i(\varphi+\psi)}-1)^3}\right) \\
&\quad + \frac{1}{2\bar{c}_j N \sin^2\varphi}\left(\frac{(-1)^{l-j+1}(N^2-2N+1)e^{i2(\varphi-\psi)} - (-1)^{l-j+1}(2N^2-2N-1)e^{i(\varphi-\psi)}}{(e^{i(\varphi-\psi)}-1)^3}\right) \\
&\qquad + \frac{1}{2\bar{c}_j N \sin^2\varphi}\left(\frac{(-1)^{l-j+1}N^2 + e^{i(\varphi-\psi)} + e^{i2(\varphi-\psi)}}{(e^{i(\varphi-\psi)}-1)^3}\right) \\
&\quad - \frac{1}{2\bar{c}_j N \sin^2\varphi}\left(\frac{(-1)^{l-j+1}(N^2-2N+1)e^{i(\varphi-\psi)} - (-1)^{l-j+1}(2N^2-2N-1)e^{i2(\varphi-\psi)}}{(e^{i(\varphi-\psi)}-1)^3}\right) \\
&\qquad - \frac{1}{2\bar{c}_j N \sin^2\varphi}\left(\frac{(-1)^{l-j+1}N^2 e^{i3(\varphi-\psi)} + e^{i2(\varphi-\psi)} + e^{i(\varphi-\psi)}}{(e^{i(\varphi-\psi)}-1)^3}\right) \\
&= \frac{\cos\varphi}{2\bar{c}_j \sin^2\varphi}\left(\frac{(-1)^{l+j+1}}{\sin\frac{\varphi+\psi}{2}\sin\frac{\varphi-\psi}{2}}\right) + \frac{(-1)^{l+j+1}N}{\bar{c}_j \sin^2\varphi} \\
&\quad + \frac{(-1)^{l+j+1}}{2\bar{c}_j N \sin^2\varphi}\left(\frac{(3N^2-4N)e^{i2(\varphi+\psi)} - (3N^2-4N)e^{i(\varphi+\psi)} - N^2 e^{i3(\varphi+\psi)} + N^2}{(e^{i(\varphi+\psi)}-1)^3}\right) \\
&\quad + \frac{(-1)^{l-j+1}}{2\bar{c}_j N \sin^2\varphi}\left(\frac{(3N^2-4N)e^{i2(\varphi-\psi)} - (3N^2-4N)e^{i(\varphi-\psi)} - N^2 e^{i3(\varphi-\psi)} + N^2}{(e^{i(\varphi-\psi)}-1)^3}\right) \\
&= \frac{\cos\varphi}{2\bar{c}_j \sin^2\varphi}\left(\frac{(-1)^{l+j+1}}{\sin\frac{\varphi+\psi}{2}\sin\frac{\varphi-\psi}{2}}\right) + \frac{(-1)^{l+j+1}N}{\bar{c}_j \sin^2\varphi} \\
&\quad + \frac{(-1)^{l+j+1}}{2\bar{c}_j N \sin^2\varphi}\left(\frac{(3N^2-4N)e^{i(\varphi+\psi)} - N^2(e^{i2(\varphi+\psi)} + e^{i(\varphi+\psi)} + 1)}{(e^{i(\varphi+\psi)}-1)^2}\right) \\
&\quad + \frac{(-1)^{l+j+1}}{2\bar{c}_j N \sin^2\varphi}\left(\frac{(3N^2-4N)e^{i(\varphi-\psi)} - N^2(e^{i2(\varphi-\psi)} + e^{i(\varphi-\psi)} + 1)}{(e^{i(\varphi-\psi)}-1)^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\cos\varphi}{2\bar{c}_j \sin^2\varphi}\left(\frac{(-1)^{l+j+1}}{\sin\frac{\varphi+\psi}{2}\sin\frac{\varphi-\psi}{2}}\right) + \frac{(-1)^{l+j+1}N}{\bar{c}_j\sin^2\varphi} \\
&\quad + \frac{(-1)^{l+j+1}}{2\bar{c}_j N\sin^2\varphi}\left(\frac{-4Ne^{i(\varphi+\psi)}}{(e^{i(\varphi+\psi)}-1)^2} - N^2\right) + \frac{(-1)^{l+j+1}}{2\bar{c}_j N\sin^2\varphi}\left(\frac{-4Ne^{i(\varphi-\psi)}}{(e^{i(\varphi-\psi)}-1)^2} - N^2\right) \\
&= \frac{\cos\varphi}{2\bar{c}_j \sin^2\varphi}\left(\frac{(-1)^{l+j+1}}{\sin\frac{\varphi+\psi}{2}\sin\frac{\varphi-\psi}{2}}\right) \\
&\quad + \frac{(-1)^{l+j+1}}{2\bar{c}_j N\sin^2\varphi}\left(\frac{-4Ne^{i(\varphi+\psi)}}{4i^2\sin^2\frac{(\varphi+\psi)}{2}e^{i(\varphi+\psi)}} + \frac{-4Ne^{i(\varphi-\psi)}}{4i^2\sin^2\frac{(\varphi-\psi)}{2}e^{i(\varphi-\psi)}}\right) \\
&= \frac{(-1)^{l+j+1}}{2\bar{c}_j\sin^2\varphi}\left(\frac{\cos\varphi}{\sin\frac{\varphi+\psi}{2}\sin\frac{\varphi-\psi}{2}} + \frac{1}{\sin^2\frac{(\varphi+\psi)}{2}} + \frac{1}{\sin^2\frac{(\varphi-\psi)}{2}}\right) \\
&= \frac{(-1)^{l+j+1}}{2\bar{c}_j\sin^2\varphi}\frac{\cos\varphi\sin\frac{\varphi+\psi}{2}\sin\frac{\varphi-\psi}{2} + \sin^2\frac{\varphi+\psi}{2} + \sin^2\frac{\varphi-\psi}{2}}{\sin^2\frac{\varphi+\psi}{2}\sin^2\frac{\varphi-\psi}{2}} \\
&= \frac{(-1)^{l+j+1}}{2\bar{c}_j(1-x_j^2)}\frac{0.5x_l(x_j-x_l)+1-x_jx_l}{0.25(x_j-x_l)^2} = \frac{(-1)^{l+j}(x_lx_j+x_l^2-2)}{\bar{c}_j(1-x_j^2)(x_j-x_l)^2}.
\end{aligned}
$$

If $1 \le l = j \le N-1$ then

$$
\begin{aligned}
(D_{xx})_{l,j} &= \frac{1}{\bar{c}_j}\sum_{k=0}^{N}\frac{2}{N\bar{c}_k}T_k(x_j)T_k''(x_j) = \frac{\cos\frac{j\pi}{N}}{\bar{c}_j N\sin^3\frac{j\pi}{N}}\sum_{k=0}^{N}\frac{2k}{\bar{c}_k}\cos\frac{kj\pi}{N}\sin\frac{kj\pi}{N} \\
&\quad - \frac{1}{\bar{c}_j N\sin^2\frac{j\pi}{N}}\sum_{k=0}^{N-1}\frac{2k^2}{\bar{c}_k}\cos\frac{kj\pi}{N}\cos\frac{kj\pi}{N} - \frac{N}{\bar{c}_j\sin^2\frac{j\pi}{N}}.
\end{aligned}
$$

By using the result of case $l = j$ in Proposition 2.4, we obtain

$$
(D_{xx})_{j,j} = \frac{\cos\psi}{\sin^2\psi}\left(-\frac{\cos\psi}{2\sin^2\psi}\right) - \frac{1}{\bar{c}_j N\sin^2\psi}\sum_{k=0}^{N-1}\frac{2k^2}{\bar{c}_k}\cos^2\psi - \frac{N}{\bar{c}_j\sin^2\psi}
$$

where $\psi = \dfrac{j\pi}{N}$. By the manner of the previous case we can obtain

$$
\begin{aligned}
(D_{xx})_{j,j} &= \frac{\cos\psi}{\sin^2\psi}\left(-\frac{\cos\psi}{2\sin^2\psi}\right) - \frac{1}{\bar{c}_j N\sin^2\psi}\sum_{k=1}^{N-1}k^2\left(1+\frac{e^{i2\psi}+e^{-i2\psi}}{2}\right) - \frac{N}{\bar{c}_j\sin^2\psi} \\
&= \frac{\cos\psi}{\sin^2\psi}\left(-\frac{\cos\psi}{2\sin^2\psi}\right) - \frac{1}{\bar{c}_j N\sin^2\psi}\cdot\frac{(N-1)N(2N-1)}{6} - \frac{N}{\bar{c}_j\sin^2\psi} \\
&\quad + \frac{1}{2\bar{c}_j N\sin^2\psi}\left(\frac{-(N^2-2N+1)e^{i(2N+4)\psi}+(2N^2-2N-1)e^{i(2N+2)\psi}}{(e^{i2\psi}-1)^3}\right) \\
&\quad + \frac{1}{2\bar{c}_j N\sin^2\psi}\left(\frac{-N^2e^{i2N\psi}+e^{i2\psi}+e^{i4\psi}}{(e^{i2\psi}-1)^3}\right) \\
&\quad + \frac{1}{2\bar{c}_j N\sin^2\psi}\left(\frac{-(N^2-2N+1)e^{-i(2N+4)\psi}+(2N^2-2N-1)e^{-i(2N+2)\psi}}{(e^{-i2\psi}-1)^3}\right) \\
&\quad + \frac{1}{2\bar{c}_j N\sin^2\psi}\left(\frac{-N^2e^{-i2N\psi}+e^{-i2\psi}+e^{-i4\psi}}{(e^{-i2\psi}-1)^3}\right).
\end{aligned}
$$

Taking into account

$$e^{i2N\psi} = \cos 2N\psi + i \sin 2N\psi = \cos 2\pi j + i \sin 2\pi j = 1$$

$$e^{-i2N\psi} = \cos(-2N\psi) + i \sin(-2N\psi) = \cos(-2\pi j) - i \sin 2\pi j = 1$$

we obtain

$$\begin{aligned}
(D_{xx})_{j,j} &= \frac{\cos\psi}{\sin^2\psi}\left(-\frac{\cos\psi}{2\sin^2\psi}\right) - \frac{1}{\bar{c}_j \sin^2\psi}\cdot\frac{2N^2+3N+1}{6} \\
&\quad + \frac{1}{2\bar{c}_j N \sin^2\psi}\left(\frac{-(N^2-2N+1)e^{i4\psi} + (2N^2-2N-1)e^{i2\psi}}{(e^{i2\psi}-1)^3}\right) \\
&\quad + \frac{1}{2\bar{c}_j N \sin^2\psi}\left(\frac{-N^2 + e^{i2\psi} + e^{i4\psi}}{(e^{i2\psi}-1)^3}\right) \\
&\quad + \frac{1}{2\bar{c}_j N \sin^2\psi}\left(\frac{-(N^2-2N+1)e^{-i4\psi} + (2N^2-2N-1)e^{-i2\psi}}{(e^{-i2\psi}-1)^3}\right) \\
&\quad + \frac{1}{2\bar{c}_j N \sin^2\psi}\left(\frac{-N^2 + e^{-i2\psi} + e^{-i4\psi}}{(e^{-i2\psi}-1)^3}\right) \\
&= \frac{\cos\psi}{\sin^2\psi}\left(-\frac{\cos\psi}{2\sin^2\psi}\right) - \frac{1}{\bar{c}_j \sin^2\psi}\cdot\frac{2N^2+3N+1}{6} \\
&\quad + \frac{1}{2\bar{c}_j N \sin^2\psi}\left(\frac{-(N^2-2N)e^{i4\psi} + (2N^2-2N)e^{i2\psi} - N^2}{(e^{i2\psi}-1)^3}\right) \\
&\quad + \frac{1}{2\bar{c}_j N \sin^2\psi}\left(\frac{(N^2-2N)e^{i2\psi} - (2N^2-2N)e^{i4\psi} + N^2 e^{i6\psi}}{(e^{i2\psi}-1)^3}\right) \\
&= \frac{\cos\psi}{\sin^2\psi}\left(-\frac{\cos\psi}{2\sin^2\psi}\right) - \frac{1}{\bar{c}_j \sin^2\psi}\cdot\frac{2N^2+3N+1}{6} \\
&\quad + \frac{1}{2\bar{c}_j N \sin^2\psi}\left(\frac{-(3N^2-4N)e^{i4\psi} + (3N^2-4N)e^{i2\psi} - N^2 + N^2 e^{i6\psi}}{(e^{i2\psi}-1)^3}\right).
\end{aligned}$$

By simplifying we obtain

$$\begin{aligned}
(D_{xx})_{j,j} &= \frac{\cos\psi}{\sin^2\psi}\left(-\frac{\cos\psi}{2\sin^2\psi}\right) - \frac{1}{\bar{c}_j \sin^2\psi}\cdot\frac{2N^2+3N+1}{6} \\
&\quad + \frac{1}{2\bar{c}_j N \sin^2\psi}\left(\frac{-(3N^2-4N)e^{i2\psi} + N^2(e^{i4\psi} + e^{i2\psi} + 1)}{(e^{i2\psi}-1)^2}\right).
\end{aligned}$$

Using the identification:

$$\frac{e^{i2\psi}}{(e^{i2\psi}-1)^2} = \frac{-1}{4\sin^2\psi}$$

and taking into account $\bar{c}_j = 1$ we obtain

$$\begin{aligned}
(D_{xx})_{j,j} &= \frac{\cos\psi}{\sin^2\psi}\left(-\frac{\cos\psi}{2\sin^2\psi}\right) - \frac{1}{\sin^2\psi}\cdot\frac{2N^2+3N+1}{6} \\
&\quad -\frac{1}{2\sin^4\psi} + \frac{N}{2\sin^2\psi} \\
&= \frac{-1}{\sin^2\psi}\left(\frac{1+\cos^2\psi}{2\sin^2\psi}\right) - \frac{1}{\sin^2\psi}\cdot\frac{2N^2+1}{6} \\
&= \frac{-1-x_j^2}{2(1-x_j^2)^2} - \frac{2N^2+1}{6(1-x_j^2)} \\
&= \frac{-3-3x_j^2-(2N^2+1)(1-x_j^2)}{6(1-x_j^2)^2} \\
&= \frac{-6-(2N^2-2)(1-x_j^2)}{6(1-x_j^2)^2} \\
&= -\frac{(N^2-1)(1-x_l^2)+3}{3(1-x_l^2)^2}.
\end{aligned}$$

If $l = 0, 1 \le j \le N$ then

$$(D_{xx})_{0,j} = \frac{1}{\bar{c}_j}\sum_{k=0}^{N}\frac{2}{N\bar{c}_k}T_k(x_j)T_k''(x_0).$$

Putting the (4) of Proposition 2.3 into the above equation, we have

$$(D_{xx})_{0,j} = \frac{2}{3\bar{c}_j N}\sum_{k=0}^{N}\frac{(k^2-1)k^2}{\bar{c}_k}T_k(x_j) = \frac{2}{3\bar{c}_j N}\sum_{k=0}^{N}\frac{(k^2-1)k^2}{\bar{c}_k}\cos\frac{kj\pi}{N}.$$

Then

$$(D_{xx})_{0,j} = \frac{2}{3\bar{c}_j N}\sum_{k=1}^{N-1}(k^4-k^2)\cos\frac{kj\pi}{N} + \frac{N^4-N^2}{3\bar{c}_j N}\cos\frac{Nj\pi}{N}.$$

For simplicity $\psi = \dfrac{j\pi}{N}$,

$$\begin{aligned}
(D_{xx})_{0,j} &= \frac{2}{3\bar{c}_j N}\sum_{k=1}^{N-1}(k^4-k^2)\cos k\psi + \frac{N^4-N^2}{3\bar{c}_j N}\cos\frac{Nj\pi}{N} \\
&= \frac{2}{3\bar{c}_j N}\sum_{k=1}^{N-1}(k^4-k^2)\frac{e^{ik\psi}+e^{-ik\psi}}{2} + \frac{N^3-N}{3\bar{c}_j}\cos j\pi \\
&= \frac{1}{3\bar{c}_j N}\sum_{k=1}^{N-1}(k^4-k^2)(e^{i\psi})^k + \frac{1}{3\bar{c}_j N}\sum_{k=1}^{N-1}(k^4-k^2)(e^{-i\psi})^k \\
&\quad + \frac{(-1)^j(N^3-N)}{3\bar{c}_j}.
\end{aligned}$$

Assume that

$$A=\sum_{k=1}^{N-1}(k^4-k^2)(e^{i\psi})^k, \qquad B=\sum_{k=1}^{N-1}(k^4-k^2)(e^{-i\psi})^k.$$

By using the following equalizations:

$$\begin{aligned}\sum_{k=1}^{N-1}(k^4-k^2)r^k &= 12r^2+72r^3+240r^4...+[(N-1)^4-(N-1)^2]r^{n-1}\\ &=\frac{(N^4-4N^3+5N^2-2N)r^{N+4}-(4N^4-12N^3+2N^2+18N-12)r^{N+3}}{(r-1)^5}\\ &+\frac{(6N^4-12N^3-12N^2+18N+12)r^{N+2}-(4N^4-4N^3-10N^2-2N)r^{N+1}}{(r-1)^5}\\ &+\frac{(N^4-N^2)r^N-12r^3-12r^2}{(r-1)^5} \qquad (r\neq 0, r\neq 1),\end{aligned}$$

$$e^{iN\psi}=(-1)^j=e^{-iN\psi},$$

we first estimate A

$$\begin{aligned}A &= \frac{(N^4-4N^3+5N^2-2N)(-1)^j e^{i4\psi}-(4N^4-12N^3+2N^2+18N-12)(-1)^j e^{i3\psi}}{(e^{i\psi}-1)^5}\\ &+\frac{(6N^4-12N^3-12N^2+18N+12)(-1)^j e^{i2\psi}-(4N^4-4N^3-10N^2-2N)(-1)^j e^{i\psi}}{(e^{i\psi}-1)^5}\\ &+\frac{(N^4-N^2)(-1)^j-12e^{i3\psi}-12e^{i2\psi}}{(e^{i\psi}-1)^5}.\end{aligned}$$

By the same way we estimate B,

$$\begin{aligned}B &= \frac{(N^4-4N^3+5N^2-2N)(-1)^j e^{-i4\psi}-(4N^4-12N^3+2N^2+18N-12)(-1)^j e^{-i3\psi}}{(e^{-i\psi}-1)^5}\\ &+\frac{(6N^4-12N^3-12N^2+18N+12)(-1)^j e^{-i2\psi}-(4N^4-4N^3-10N^2-2N)(-1)^j e^{-i\psi}}{(e^{-i\psi}-1)^5}\\ &+\frac{(N^4-N^2)(-1)^j-12e^{-i3\psi}-12e^{-i2\psi}}{(e^{-i\psi}-1)^5}\\ &= \frac{(N^4-4N^3+5N^2-2N)(-1)^j e^{i\psi}-(4N^4-12N^3+2N^2+18N-12)(-1)^j e^{i2\psi}}{e^{5\psi}(e^{-i\psi}-1)^5}\\ &+\frac{(6N^4-12N^3-12N^2+18N+12)(-1)^j e^{i3\psi}-(4N^4-4N^3-10N^2-2N)(-1)^j e^{i4\psi}}{e^{i5\psi}(e^{-i\psi}-1)^5}\\ &+\frac{(N^4-N^2)(-1)^j e^{i5\psi}-12e^{i2\psi}-12e^{i3\psi}}{e^{i5\psi}(e^{-i\psi}-1)^5}.\end{aligned}$$

Then

$$
\begin{aligned}
A+B &= \frac{(-1)^j(5N^4-8N^3-5N^2-4N)e^{i4\psi}-(-1)^j(10N^4-24N^3-10N^2+36N)e^{i3\psi}}{(e^{i\psi}-1)^5} \\
&\quad+\frac{(-1)^j(10N^4-24N^3-10N^2+36N)e^{i2\psi}-(-1)^j(5N^4-8N^3-5N^2-4N)e^{i\psi}}{(e^{-i\psi}-1)^5} \\
&\quad+\frac{(-1)^j(N^4-N^2)(1-e^{i5\psi})}{(e^{i\psi}-1)^5} \\
&= \frac{(-1)^j(5N^4-8N^3-5N^2-4N)e^{i\psi}(e^{i3\psi}-1)}{(e^{i\psi}-1)^5} \\
&\quad+\frac{(-1)^j(10N^4-24N^3-10N^2+36N)e^{i2\psi}(1-e^{i\psi})}{(e^{-i\psi}-1)^5} \\
&\quad+\frac{(-1)^j(N^4-N^2)(1-e^{i5\psi})}{(e^{i\psi}-1)^5}. \\
&= \frac{(-1)^j(5N^4-8N^3-5N^2-4N)(e^{i3\psi}+e^{i2\psi}+e^{i\psi})}{(e^{i\psi}-1)^4} \\
&\quad+\frac{(-1)^{j+1}(10N^4-24N^3-10N^2+36n)e^{i2\psi}}{(e^{-i\psi}-1)^4} \\
&\quad+\frac{(-1)^{j+1}(N^4-N^2)(e^{i4\psi}+e^{i3\psi}+e^{i2\psi}+e^{i\psi}+1)}{(e^{i\psi}-1)^4} \\
&= \frac{(-1)^j(5N^4-8N^3-5N^2-4N)(e^{i3\psi}+e^{i2\psi}+e^{i\psi})}{(e^{i\psi}-1)^4} \\
&\quad+\frac{(-1)^{j+1}(10N^4-24N^3-10N^2+36N)e^{i2\psi}}{(e^{-i\psi}-1)^4} \\
&\quad+\frac{(-1)^{j+1}(N^4-N^2)[(e^{i\psi}-1)^4+5e^{i3\psi}-5e^{i2\psi}+5e^{i\psi}]}{(e^{i\psi}-1)^4}. \\
&= \frac{(-1)^j(5N^4-8N^3-5N^2-4N)(e^{i3\psi}+e^{i\psi})+(-1)^j(-5N^4+16N^3+5N^2-40N)e^{i2\psi}}{(e^{i\psi}-1)^4} \\
&\quad+(-1)^j(-N^4+N^2)+\frac{(-1)^j(5N^4-5N^2)(-e^{i3\psi}+e^{i2\psi}-e^{i\psi})}{(e^{i\psi}-1)^4} \\
&= \frac{(-1)^j(-8N^3-4N)(e^{i3\psi}+e^{i\psi})+(-1)^j(16N^3-40N)e^{i2\psi}}{(e^{i\psi}-1)^4} \\
&\quad+(-1)^j(-N^4+N^2)+\frac{(-1)^j(5N^4-5N^2)(-e^{i3\psi}+e^{i2\psi}-e^{i\psi})}{(e^{i\psi}-1)^4} \\
&= \frac{(-1)^j(-8N^3-4N)(e^{i3\psi}-2e^{i2\psi}+e^{i\psi})-48(-1)^jNe^{i2\psi}}{(e^{i\psi}-1)^4}+(-1)^j(-N^4+N^2) \\
&= -4(-1)^j(2N^3+N)\frac{e^{i\psi}}{(e^{i\psi}-1)^2}-48(-1)^jN\left(\frac{e^{i\psi}}{(e^{i\psi}-1)^2}\right)^2+(-1)^j(-N^4+N^2).
\end{aligned}
$$

By using the identification:

$$\frac{e^{i\psi}}{(e^{i\psi}-1)^2} = \frac{-1}{4\sin^2\psi/2},$$

we obtain

$$\begin{aligned} A+B &= (-1)^j(2N^3+N)\frac{1}{\sin^2\psi/2} - 3(-1)^j n\frac{1}{\sin^4\psi/2} + (-1)^j(-N^4+N^2) \\ &= \frac{2(-1)^j(2N^3+N)}{1-\cos\psi} - \frac{12(-1)^j N}{(1-\cos\psi)^2} + (-1)^j(-N^4+N^2). \end{aligned}$$

By substituting $\cos\psi = \cos j\pi/N = x_j$ we obtain

$$\begin{aligned} A+B &= \frac{(-1)^j(2N^3+N)}{1-\cos j\pi/N} - \frac{12(-1)^j N}{(1-\cos j\pi/N)^2} + (-1)^j(-N^4+N^2) \\ &= \frac{(-1)^j(2N^3+N)}{1-x_j} - \frac{12(-1)^j N}{(1-x_j)^2} + (-1)^j(-N^4+N^2). \end{aligned}$$

Therefore,

$$\begin{aligned} (D_{xx})_{0,j} &= \frac{1}{3\bar{c}_j N}\left(\frac{(-1)^j(2N^3+N)}{1-x_j} - \frac{12(-1)^j N}{(1-x_j)^2} + (-1)^j(-N^4+N^2)\right) + \frac{(-1)^j(N^3-N)}{3\bar{c}_j} \\ &= \frac{2(-1)^j}{3\bar{c}_j}\frac{(2N^2+1)(1-x_j)-6}{(1-x_j)^2} + \frac{(-1)^j}{3\bar{c}_j}(-N^3+5N) + \frac{(-1)^j(N^3-N)}{3\bar{c}_j} \\ &= \frac{2(-1)^j}{3\bar{c}_j}\frac{(2N^2+1)(1-x_j)-6}{(1-x_j)^2}. \end{aligned}$$

If $l = N, 0 \le j \le N-1$ then

$$(D_{xx})_{N,j} = \frac{1}{\bar{c}_j}\sum_{k=0}^{N}\frac{2}{N\bar{c}_k}T_k(x_j)T_k''(x_N).$$

Putting (4) of Proposition 2.3 into the above equation, we have

$$(D_{xx})_{N,j} = \frac{2}{3\bar{c}_j N}\sum_{k=0}^{N}\frac{(-1)^k(k^2-1)k^2}{\bar{c}_k}T_k(x_j) = \frac{2}{3\bar{c}_j N}\sum_{k=0}^{N}\frac{(-1)^k(k^2-1)k^2}{\bar{c}_k}\cos\frac{kj\pi}{N}.$$

The right hand side of the above equality is similar form to the previous case. So, we omit the proof for $l = 0, 1 \le j \le N$.

If $l = j = 0$ or $l = j = N$ then taking into accoun (2) and (4) of proposition 2.3, we have

$$(D_{xx})_{0,0} = \frac{1}{\bar{c}_0}\sum_{k=0}^{N}\frac{2}{N\bar{c}_k}T_k(x_0)T_k''(x_0) = \frac{1}{N}\sum_{k=0}^{N}\frac{k^4-k^2}{3\bar{c}_k}$$

and

$$(D_{xx})_{N,N} = \frac{1}{\bar{c}_N}\sum_{k=0}^{N}\frac{2}{N\bar{c}_k}T_k(x_N)T_k''(x_N) = \frac{1}{N}\sum_{k=0}^{N}\frac{(-1)^{2k}(k^4-k^2)}{3\bar{c}_k}.$$

Therefore,

$$\begin{aligned}(D_{xx})_{0,0} = (D_{xx})_{N,N} &= \frac{1}{N}\sum_{k=0}^{N}\frac{k^4-k^2}{3\bar{c}_k} = \frac{1}{3N}\sum_{k=1}^{N-1}(k^4-k^2) + \frac{N^4-N^2}{6N} \\ &= \frac{1}{3N}\sum_{k=1}^{N}(k^4-k^2) - \frac{N^4-N^2}{6N} = \frac{1}{3N}\left(\sum_{k=1}^{N}k^4 - \sum_{k=1}^{N}k^2\right) - \frac{N^4-N^2}{6N} \\ &= \frac{1}{3N}\left(\frac{(N+1)N(6N^3+9N^2+N-1)}{30} - \frac{N(N+1)(2N+1)}{6}\right) \\ &\quad -\frac{N^4-N^2}{6N} = \frac{N+1}{3}\cdot\frac{6N^3+9N^2-9N-6}{30} - \frac{N^3-N}{6}. \\ &= \frac{N+1}{3}\cdot\frac{6N^3-6N^2+6N-6}{30} = \frac{N^4+1}{15}. \qquad \blacksquare\end{aligned}$$

2.2.3 Spectral method with Chebyshev-Gauss collocation points.

Chebyshev-Gauss(C-G) collocation points are given as follows:

$$x_j = \cos\frac{(2j+1)\pi}{2N+2}, \qquad j = 0, 1, ..., N$$

For C-G collocation points $x_j (j = 0, 1, ..., N)$,

$$u_j = u(x_j) = \sum_{k=0}^{M}\tilde{u}_k T_k(x_j), \tag{2.5}$$

and $M = N$ then an inverse relationship is

$$\tilde{u}_k = \frac{2}{(N+1)\bar{c}_k}\sum_{j=0}^{N}u_j T_k(x_j), \qquad k = 0, 1, ..., N \tag{2.6}$$

where

$$\bar{c}_j = \begin{cases} 2, & j = 0 \\ 1, & 1 \le j \le N. \end{cases}$$

Substituting (2.6) into (2.1), we obtain

$$u^N(x) = \sum_{j=0}^{N}\left(\frac{1}{N+1}\sum_{k=0}^{N}\frac{2}{\bar{c}_k}T_k(x_j)T_k(x)\right)u_j. \tag{2.7}$$

The first and the second derivatives of $u^N(x)$ on C-G collocation points $x_l (l = 0, 1, ..., N)$ can be represented in the matrix forms as

$$(u^N)'(x_l) = \sum_{j=0}^{N}(D_x)_{l,j}u_j,$$

$$(u^N)''(x_l) = \sum_{j=0}^{N}(D_{xx})_{l,j}u_j.$$

where $(D_x)_{l,j}$ and $(D_{xx})_{l,j}$ are defined by

$$(D_x)_{l,j} = \frac{1}{N+1}\sum_{k=0}^{N}\frac{2}{\bar{c}_k}T_k(x_j)T_k'(x_l),$$

and

$$(D_{xx})_{l,j} = \frac{1}{N+1}\sum_{k=0}^{N}\frac{2}{\bar{c}_k}T_k(x_j)T_k''(x_l).$$

2.2.4 Spectral method with Chebyshev-Gauss-Radau collocation points.

Chebyshev-Gauss-Radau(C-G-R) collocation points are given as follows:

$$x_j = \cos\frac{2j\pi}{2N+1}, \qquad j = 0, 1, ..., N$$

For C-G-R collocation points $x_j (j = 0, 1, ..., N)$,

$$u_j = u(x_j) = \sum_{k=0}^{M}\tilde{u}_k T_k(x_j), \tag{2.8}$$

and $M = N$ then an inverse relationship is

$$\tilde{u}_k = \frac{4}{(2N+1)\bar{c}_k}\sum_{j=0}^{N}\frac{1}{\bar{c}_j}u_j T_k(x_j), \qquad k = 0, 1, ..., N \tag{2.9}$$

where

$$\bar{c}_j = \begin{cases} 2, & j = 0 \\ 1, & 1 \le j \le N. \end{cases}$$

Substituting (2.9) into (2.1), we obtain

$$u^N(x) = \sum_{j=0}^{N}\left(\frac{2}{(2N+1)\bar{c}_j}\sum_{k=0}^{N}\frac{2}{\bar{c}_k}T_k(x_j)T_k(x)\right)u_j. \tag{2.10}$$

The first and the second derivatives of $u^N(x)$ on C-G-R collocation points $x_l (l = 0, 1, ..., N)$ can be represented in the matrix forms as

$$(u^N)'(x_l) = \sum_{j=0}^{N}(D_x)_{l,j}u_j,$$

$$(u^N)''(x_l) = \sum_{j=0}^{N}(D_{xx})_{l,j}u_j.$$

where $(D_x)_{l,j}$ and $(D_{xx})_{l,j}$ are defined by

$$(D_x)_{l,j} = \frac{2}{(2N+1)\bar{c}_j}\sum_{k=0}^{N}\frac{2}{\bar{c}_k}T_k(x_j)T_k'(x_l),$$

and

$$(D_{xx})_{l,j} = \frac{2}{(2N+1)\bar{c}_j} \sum_{k=0}^{N} \frac{2}{\bar{c}_k} T_k(x_j) T_k''(x_l).$$

2.3 Newton method

The Newton method will be used for our inverse problem. Therefore we will make its brief introduction in this subsection. The Newton method is a popular iterative method for finding solutions of a system of nonlinear equations

$$f(x) = 0, \tag{2.11}$$

where 0 is the zero vector, $x = (x_1, ..., x_n)^T$ and $f = (f_1(x_1, ..., x_n), ..., f_n(x_1, ..., x_n))^T$, $f_j : \mathbb{R}^n \to \mathbb{R}, j = 1, ..., n$, are continuously differentiable functions. By $f'(x)$ we denote the Jacobian matrix of f which is given by

$$f'(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \cdots & \frac{\partial f_2}{\partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1}(x) & \frac{\partial f_n}{\partial x_2}(x) & \cdots & \frac{\partial f_n}{\partial x_n}(x) \end{pmatrix}. \tag{2.12}$$

The Newton method is defined by

$$x^{(k+1)} = x^{(k)} - [f'(x^{(k)})]^{-1} f(x^{(k)}), \quad k = 0, 1, ... \tag{2.13}$$

with some $x^{(0)} \in \mathbb{R}^n$. We start from an initial guess $x^{(0)} \in \mathbb{R}^n$ and through the recurrence relation (2.13) we compute step by step approximations $x^{(1)}, x^{(2)}, ...$ to an unknown solution of

$$x = g(x), \tag{2.14}$$

where $g(x) = x - [f'(x)]^{-1} f(x)$. Note that a solution of (2.14) is called a fixed point of g, which is a solution of (2.11).

Assume that the sequence $\{x^{(k)}\}$ converges to $\hat{x}$. Usually, the iteration is terminated when the following criteria holds;

$$\|x^{(k+1)} - x^{(k)}\| \leq \epsilon$$

where $\|\cdot\|$ is a norm defined in $\mathbb{R}^n$ and ϵ is a given small positive number.

Example 2.1 Find numerical solutions to the one-dimensional nonlinear equation

$$f(x) = x^3 - x = 0, \quad x \in \mathbb{R}.$$

Exact solutions to the equation:

$$\hat{x}_1 = -1, \quad \hat{x}_2 = 0, \quad \hat{x}_3 = 1.$$

By applying the Newton method to $f(x) = 0$, we have

$$x^{(k+1)} = x^{(k)} - \frac{(x^{(k)})^3 - x^{(k)}}{3(x^{(k)})^2 - 1} = \frac{2(x^{(k)})^3}{3(x^{(k)})^2 - 1} \qquad (k = 0, 1, ...).$$

Iterations proceed until

$$|x^{(k+1)} - x^{(k)}| \leq 10^{-15}.$$

Table 2.1 shows numerical results for various initial guess $x^{(0)}$.

Table 2.1. Numerical results.

$x^{(0)}$	-0.4	0.52	2
$x^{(1)}$	0.24615386392943328	-1.4894908920373511	1.4545454545454546
$x^{(2)}$	-3.6456696724859017E-002	-1.1685664058975043	1.1510467893775467
$x^{(3)}$	9.7296463254681331E-005	-1.0306213616298114	1.0253259289766978
$x^{(4)}$	-1.8421337995592688E-012	-1.0013127713890331	1.0009084519430513
$x^{(5)}$	0.0000000000000000	-1.0000025771591445	1.0000012353089454
$x^{(6)}$		-1.0000000000099625	1.0000000000022891
$x^{(7)}$		-1.0000000000000000	1.0000000000000000

The following theorem shows that the Newton method converges locally.

Theorem 2.1([28], Theorem 6.14) Let $D \in \mathbb{R}^n$ be open and convex and let $f : D \to \mathbb{R}^n$ be a continuously differentiable. Assume that for some norm $\|\cdot\|$ on $\mathbb{R}^n$ and some $x^{(0)} \in D$ the following conditions hold:

(a) f satisfies

$$\|f'(x) - f'(y)\| \leq \gamma \|x - y\|$$

for all $x, y \in D$ and some constant $\gamma > 0$.

(b) The Jacobian matrix $f'(x)$ is nonsingular for all $x \in D$, and there exists a constant $\beta > 0$ such that

$$\|[f'(x)]^{-1}\| \leq \beta, \qquad x \in D.$$

(c) For the constants $\alpha = \|[f'(x^{(0)})]^{-1} f(x^{(0)})\|$ and $q = \alpha\beta\gamma$ the inequality

$$q < \frac{1}{2}$$

is satisfied.

(d) The closed ball $S_{2\alpha}(x^{(0)}) = \{x : \|x - x^{(0)}\| \leq 2\alpha\}$ is contained in D.

Then

(i) Starting with $x^{(0)}$ the Newton iteration

$$x^{(k+1)} = x^{(k)} - [f'(x^{(k)})]^{-1} f(x^{(k)}) \qquad (k = 0, 1, ...)$$

is well defined and $x^{(k)} \in S_{2\alpha}(x^{(0)})$.

(ii) The sequence $\{x^{(k)}\}$ converges to the solution $\hat{x}$ of the $f(x) = 0$ and for all $k \geq 0$

$$\|x^{(k)} - \hat{x}\| \leq 2\alpha q^{2^k - 1} \qquad (k = 0, 1, ...)$$

(iii) The equation $f(x) = 0$ has a unique solution $\hat{x}$ in $S_{2\alpha}(x^{(0)})$.

Proof. (i) First we will obtain an inquality that is useful for proving. Let $x, y, z \in D$. We can verify that

$$f(y) - f(x) = \int_0^1 f'[\lambda x + (1-\lambda)y](y-x)d\lambda$$

(for more detail see [28], Theorem 6.7). Hence

$$f(y) - f(x) - f'(z)(y-x) = \int_0^1 \{f'[\lambda x + (1-\lambda)y] - f'(z)\}(y-x)d\lambda,$$

and estimating with the aid of condition (a) we find that

$$\begin{aligned} \|f(y) - f(x) - f'(z)(y-x)\| &= \textstyle\int_0^1 \{f'[\lambda x + (1-\lambda)y] - f'(z)\}(y-x)d\lambda \\ &\leq \gamma\|y-x\| \textstyle\int_0^1 \|\lambda x + (1-\lambda)y - z\|d\lambda \\ &= \gamma\|y-x\| \textstyle\int_0^1 \|\lambda(x-z) + (1-\lambda)(y-z)\|d\lambda \\ &\leq \gamma\|y-x\|\{\|x-z\| \textstyle\int_0^1 \lambda d\lambda + \|(y-z)\| \int_0^1 (1-\lambda)d\lambda\} \\ &= \frac{\gamma}{2}\|y-x\|\{\|x-z\| + \|y-z\|\}. \end{aligned} \tag{2.15}$$

By choosing $z = x$ we have

$$\|f(y) - f(x) - f'(x)(y-x)\| \leq \frac{\gamma}{2}\|y-x\|^2 \tag{2.16}$$

for all $x, y \in D$. Choosing $z = x^{(0)}$ in (2.15), it yields

$$\begin{aligned} \|f(y) - f(x) - f'(x^{(0)})(y-x)\| &\leq \frac{\gamma}{2}\|y-x\|\{\|x-x^{(0)}\| + \|y-x^{(0)}\|\} \\ &\leq \gamma\|y-x\|(2\alpha + 2\alpha) = 2\alpha\gamma\|y-x\| \end{aligned}$$

for all $x, y \in S_{2\alpha}(x^{(0)})$.

Now we prove through induction the following two inqualities

$$\|x^{(k)} - x^{(0)}\| \leq 2\alpha, \qquad \|x^{(k)} - x^{(k-1)}\| \leq \alpha q^{2^{k-1}-1}, \qquad k = 1, 2, ... \tag{2.17}$$

This is valid for $k = 1$, since

$$\|x^{(1)} - x^{(0)}\| = \|[f'(x^{(0)})]^{-1} f(x^{(0)})\| = \alpha < 2\alpha.$$

Assume that the inequalities (2.17) are proven up to $k \geq 1$. Then since $x^{(k)} \in S_{2\alpha}(x^{(0)}) \subset D$, by the condition (b) the element $x^{(k+1)}$ is well-defined. The condition (b) is used then the Newton iteration is applied to $x^{(k)}$ and by using the estimate (2.16), we obtain

$$\begin{aligned} \|x^{(k+1)} - x^{(k)}\| &= \|[f'(x^{(k)})]^{-1} f(x^{(k)})\| \leq \beta\|f(x^{(k)})\| \\ &= \beta\|f(x^{(k)}) - f(x^{(k-1)}) - f'(x^{(k-1)})(x^{(k)} - x^{(k-1)})\| \\ &\leq \frac{\beta\gamma}{2}\|x^{(k)} - x^{(k-1)}\|^2, \end{aligned}$$

With the aid of the induction assumption, and the definition of q we can estimate

$$\begin{aligned}\|x^{(k+1)} - x^{(k)}\| = \frac{\beta\gamma}{2}\|x^{(k)} - x^{(k-1)}\|^2 &\leq \frac{\beta\gamma}{2}[aq^{2^{k-1}-1}]^2 \\ &= \frac{\alpha}{2}q^{2^k-1} < \alpha q^{2^k-1}.\end{aligned}$$

From this, with the help of the triangle inequality, the induction assumption and the condition (c), we obtain that

$$\begin{aligned}\|x^{(k+1)} - x^{(0)}\| &\leq \|x^{(k+1)} - x^{(k)}\| + ... + \|x^{(1)} - x^{(0)}\| \\ &= \alpha(1 + q + q^3 + q^7 + ... + q^{2^k-1}) \leq \frac{\alpha}{1-q} \leq 2\alpha.\end{aligned}$$

i.e., the inequality (2.17) also hold for $k+1$. Therefore $x^{(k)} \in S_{2\alpha}(x^{(0)})$ for all $k > 0$.

(ii) We observe that $\{x^{(k)}\}$ is a Cauchy sequence, since $q < 1/2$ and for $m > 0$

$$\begin{aligned}\|x^{(k)} - x^{(k+m)}\| &\leq \|x^{(k)} - x^{(k+1)}\| + ... + \|x^{(k+m-1)} - x^{(k+m)}\| \\ &\leq \alpha(q^{2^k-1} + q^{2^{k+1}-1} + ... + q^{2^{k+m-1}-1}) \\ &= \alpha q^{2^k-1}\left(1 + q^{2^k} + ... + [q^{2^k}]^{2^{m-1}-1}\right) \leq 2\alpha q^{2^k-1}.\end{aligned} \tag{2.18}$$

From the fact that each finite-dimensional normed space is a Banach space, the limit

$$\hat{x} = \lim_{k\to\infty} x^{(k)}$$

exists. By passing to the limit $k \to \infty$ in (2.17) we obtain $\|\hat{x} - x^{(0)}\| \leq 2\alpha$, i.e., $\hat{x} \in S_{2\alpha}(x^{(0)})$ and by passing to the limit $m \to \infty$ in (2.18) we obtain $\|x^{(k)} - \hat{x}\| \leq 2\alpha q^{2^k-1}$ for all $k \geq 0$.

We now show that the limit $\hat{x}$ is a solution to $f(x) = 0$. With the aid of the Newton iteration and the condition (a) we can estimate

$$\begin{aligned}\|f(x^{(k)})\| &= \|f'(x^{(k)})(x^{(k+1)} - x^{(k)})\| \\ &\leq \|f'(x^{(k)}) - f'(x^{(0)}) + f'(x^{(0)})\|\|(x^{(k+1)} - x^{(k)})\| \\ &\leq [\gamma\|x^{(k)} - x^{(0)}\| + \|f'(x^{(0)})\|]\|(x^{(k+1)} - x^{(k)})\| \\ &\leq [2\gamma\alpha + \|f'(x^{(0)})\|]\|(x^{(k+1)} - x^{(k)})\| \to 0, \qquad k \to \infty.\end{aligned}$$

Hence $f(x^{(k)}) \to 0, k \to \infty$, and the continuity of f implies that indeed $f(\hat{x}) = 0$.

(iii) We consider the function $g(x)$ defined by

$$g(x) = x - [f'(x^{(0)})]^{-1}f(x).$$

With the aid of condtions (b), (c) and (2.16), we have

$$\begin{aligned}\|g(x) - g(y)\| &= \|[f'(x^{(0)})]^{-1}\{f(y) - f(x) - f'(x^{(0)})(y - x)\}\| \\ &\leq 2\alpha\beta\gamma\|y - x\| \leq 2q\|y - x\|.\end{aligned}$$

for all $x, y \in S_{2\alpha}(x^{(0)})$, i.e, g is a contraction. Therefore by [28], Theorem 3.44 the function g has at most one fixed point in $S_{2\alpha}(x^{(0)})$. Now uniqueness of the solution of $f(x) = 0$ in $S_{2\alpha}(x^{(0)})$

follows from the equivalence of the equations $g(x) = x$ and $f(x) = 0$. ■

Let $\{x_k\} \in \mathbb{R}^n$ be a convergent sequence with limit $\hat{x}$. The sequence $\{x_k\}$ is said to be convergent of order $p \geq 1$ if there exists a constant $C > 0$ such that

$$\|x_{k+1} - \hat{x}\| \leq C\|x_k - \hat{x}\|^p, \quad k = 1, 2, ...$$

Theorem 2.2([28], Theorem 6.20) Under the assumptions of Theorem 2.1 the Newton method is at least quadratically convergent (the convergent of the second order).

Proof. By using the condition (b) of Theorem 2.1, the inequality 2.16 and $f(\hat{x}) = 0$ we can estimate

$$\begin{aligned}\|\hat{x} - x^{(k+1)}\| &= \|\hat{x} - x^{(k)} + [f'(x^{(k)})]^{-1} f'(x^{(k)})\| \\ &\leq \|[f'(x^{(k)})]^{-1}\| \|f(\hat{x}) - f(x^{(k)}) - f'(x^{(k)})(\hat{x} - x^{(k)})\| \\ &\leq \frac{\beta\gamma}{2}\|\hat{x} - x^{(k)}\|^2. \end{aligned}$$ ■

2.4 Gaussian elimination

Gaussian elimination is an algorithm for solving a system of linear equations. This algorithm converts a system to an equivalent triangular system that can easily be solved. A system of linear equations can be represented in the matrix equation form:

$$Ax = b \tag{2.19}$$

where $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$. Then this algorithm is usually understood as a sequence of operations performed on the coefficient matrix A. This algorithm can be described as the following recursive process: $1 \leq k \leq n-1$,

$$A^{(k)} = G_k A^{(k-1)}, \qquad A^{(0)} = A$$

where

$$G_k = \begin{pmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & & 1 & 0 & & 0 \\ 0 & & -l_{k+1}^{(k)} & 1 & & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -l_n^{(k)} & 0 & \cdots & 1 \end{pmatrix} \quad \text{and} \quad l_i^{(k)} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}, \ k+1 \leq i \leq n.$$

Consequently, $A^{(n-1)}$ is obtained as the result, and it is an upper triangular matrix. In the recursive process, $a_{kk}^{(k-1)} (k = 1, 2, ..., n-1)$ is called the pivot element or the pivot. For this process to be well defined, the pivot elements must be nonzero.

Moreover, if the pivot elements encountered in the recursive process are nonzero then the matrix A can be factorized into a product

$$A = LU$$

of a lower triangular matrix L which all elements on diagonal are equal to 1 and an upper triangular matrix U. L and U are determined as follows:

$$L = G_k^{-1} \cdots G_1^{-1} \quad \text{and} \quad U = A^{(n-1)}.$$

Therefore, we have the following matrix forms:

$$L = \begin{pmatrix} 1 & & & & \\ l_2^{(1)} & 1 & & 0 & \\ l_3^{(1)} & l_3^{(2)} & 1 & & \\ \vdots & \vdots & & \ddots & \\ l_n^{(1)} & l_n^{(2)} & \cdots & l_n^{(n-1)} & 1 \end{pmatrix}, \qquad U = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ & & \cdots & \vdots \\ 0 & & & a_{nn}^{(n-1)} \end{pmatrix}$$

where $l_i^{(k)}$ and $a_{ij}^{(k)}$ are determined by the recursive formulas: $1 \le k \le n-1$,

$$l_i^{(k)} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}, \qquad a_{ij}^{(k)} = a_{ij}^{(k-1)} - l_i^{(k)} a_{kj}^{(k-1)}, \qquad i \ge k+1, j \ge k+1.$$

In practical computation, it is not just zero pivots, and also small pivots are the source of unacceptable terms. Indeed if the pivot $a_{kk}^{(k-1)}$ is small then it entails large multipliers $l_i^{(k)}$. There is a possible loss of significance in the subtraction $a_{ij}^{(k-1)} - l_i^{(k)} a_{kj}^{(k-1)}$, with low-order digits of $a_{ij}^{(k-1)}$ being lost. Losing these digits could correspond to making a relatively large change to the original matrix A. This phenomenon can be seen from the following example[15]: 3-digit floating point arithmetic is used to solve

$$\begin{pmatrix} .001 & 1.00 \\ 1.00 & 2.00 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1.00 \\ 3.00 \end{pmatrix}.$$

Applying Gaussuan elimination we get

$$\hat{L} = \begin{pmatrix} 1 & 0 \\ 1000 & 1 \end{pmatrix}, \qquad \hat{U} = \begin{pmatrix} .001 & 1 \\ 0 & -1000 \end{pmatrix},$$

where $\hat{L}$ and $\hat{U}$ approximate L and U, respectively in 3-digit ploating point arithmetic. and a calculation shows that

$$\hat{L}\hat{U} = \begin{pmatrix} .001 & 1 \\ 1 & 0 \end{pmatrix} = A + \Delta A.$$

If we go on to solve the problem using the $\hat{L}$ and $\hat{U}$, then using the same precision arithmetic we obtain a computed solution $\hat{x} = (0, 1)^T$. This is in contrast to the exact solution $x = (1.002..., .998....)^T$. A way out of this difficulty is to use pivoting strategies.

2.4.1 Gaussian elimination with partial pivoting(GEPP)

In the partial pivoting algorithm the rows are permuted to obtain the pivot elements such that no multiplier is greater than 1 in absolute value. Suppose

$$A = \begin{pmatrix} 0 & 2 & -3 \\ 2 & 2 & 6 \\ 2 & 4 & 5 \end{pmatrix}.$$

The pivot a_{11} should be the element with largest absolute value in the first column. To swep the first row with the second row we use the interchange permutation as follows:

$$P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \text{then} \qquad P_1 A = \begin{pmatrix} 2 & 2 & 6 \\ 0 & 2 & -3 \\ 2 & 4 & 5 \end{pmatrix}.$$

It follows that

$$G_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \qquad \Rightarrow \qquad G_1 P_1 A = \begin{pmatrix} 2 & 2 & 6 \\ 0 & 2 & -3 \\ 0 & 2 & -1 \end{pmatrix}.$$

The pivot a_{22} should be the element with largest absolute value in the second column without a_{12}. Thus, P_2 became the identity matrix. If

$$P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \text{and} \qquad G_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix},$$

then

$$G_2 P_2 G_1 P_1 A = U = \begin{pmatrix} 2 & 2 & 6 \\ 0 & 2 & -3 \\ 0 & 0 & 2 \end{pmatrix}.$$

For general n we have

$$G_{n-1} P_{n-1} \cdots G_1 P_1 A = U$$

where U is an upper trianglar. As a consequence of the partial pivoting, no multiplier is larger than one in absolute value. Moreover, it turns out that the partial pivoting algorithm computes the factorization

$$PA = LU$$

where $P = P_{n-1} \cdots P_1, U$ is an upper triangular, and L is a lower triangular with $|l_{ij}| \leq 1$, $L = G_1^{-1} \cdots G_{n-1}^{-1}$ (see [15] for more detail).

2.4.2 Gaussian elimination with complete pivoting(GECP)

In complete pivoting algorithm, the rows and also the columns are permuted to obtain the pivot elements such that no multiplier is greater than 1 in absolute value. Suppose

$$A = \begin{pmatrix} 0 & 2 & -3 \\ 2 & 2 & 6 \\ 2 & 4 & 5 \end{pmatrix}.$$

The pivot a_{11} should be the element with largest absolute value in the whole matrix. Thus, to swap rows 1 and 2, colums 1 and 3, we use interchange permutations P_1 and Q_1 for interchanging rows and columns, respectively. If

$$P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \text{and} \qquad Q_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

then

$$P_1AQ_1 = \begin{pmatrix} 6 & 2 & 2 \\ -3 & 2 & 0 \\ 5 & 4 & 2 \end{pmatrix}.$$

If

$$G_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ -5/6 & 0 & 1 \end{pmatrix} \quad \Rightarrow \quad G_1P_1AQ_1 = \begin{pmatrix} 6 & 2 & 4 \\ 0 & 3 & 1 \\ 0 & 7/3 & 1/3 \end{pmatrix}.$$

The pivot a_{22} should be the element with largest absolute value in the remained matrix. Thus, we choose $P_2 = Q_2 = I$ then

$$P_2G_1P_1AQ_1Q_2 = \begin{pmatrix} 6 & 2 & 2 \\ 0 & 3 & 1 \\ 0 & 7/3 & 1/3 \end{pmatrix}.$$

If

$$G_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -7/9 & 1 \end{pmatrix} \quad \Rightarrow \quad G_2P_2G_1P_1AQ_1Q_2 = U = \begin{pmatrix} 6 & 2 & 2 \\ 0 & 3 & 1 \\ 0 & 0 & -4/9 \end{pmatrix}.$$

For general n we have

$$G_{n-1}P_{n-1}\cdots G_1P_1AQ_1\cdots Q_{n-1} = U.$$

where U is an upper trianglar. Moreover, It turns out that complete pivoting algorithm computes the factorization

$$PAQ = LU$$

where $P = P_{n-1}\cdots P_1, Q = Q_1\cdots P_{n-1}, U$ is an upper triangular, and L is a lower triangular with $|l_{ij}| \leq 1$, $L = G_1^{-1}\cdots G_{n-1}^{-1}$ (see [15]). The Gaussian elimination with complete pivoting can be used to determine the rank of a matrix.

2.4.3 GECP with equilibration(GECPE)

Equilibration is a procedure before the GECP as follows[22]: Let $A = [a_{i,j}]$ be the coefficient matrix of (2.19). A diagonal matrix V is defined as follows:

$$V = \text{diag}(\max_{1\leq j\leq n}|a_{1,j}|, ..., \max_{1\leq j\leq n}|a_{n,j}|,)^{-1}.$$

The both side of (2.19) is multiplied by V from the left hand side then we have

$$VAx = Vb.$$

Suppose $VA = [c_{i,j}]$, a diagonal matrix W is defined as follows:

$$W = \text{diag}(\max_{1\leq i\leq n}|c_{i,1}|, ..., \max_{1\leq i\leq n}|c_{i,n}|,)^{-1}.$$

With aid of the simple transformation $x = Wy$ we obtain

$$VAWy = Vb, \qquad x = Wy.$$

For the coefficient matrix VAW there is at least one element with absolute value 1 in any row and in any column.

The GECP is applied to VAW after the equilibration. This algorithm is called GECP with equilibration.

2.4.4 Some remarks on the Gaussian elimination

By using GEPP or GECP or GECPE the square matrix A can be factorized into the product of the lower triangular matrix L and the upper triangular matrix U. We should remark the following things:

- The rank of A is detemined by the number of nonzero diagonal elements in the upper triangular matrix U. When A is rank deficit, GECP is better than GEPP to reveal the rank of A

- The determinant of the matrix A is detemined by

$$|\det(A)| = |\prod_{i=1}^{n} a_{ii}^{(i-1)}|$$

  in CEPP or CECP.

$$|\det(A)| = |\det(V^{-1})\left(\prod_{i=1}^{n} u_{ii}\right)\det(W^{-1})|$$

  in CECPE.

- The inverse of an invertible square matrix is computed as follows:
  If GEPP is used then $PA = LU$ and its inverse is given by

$$A^{-1} = U^{-1}L^{-1}P,$$

  If GECP is used then $PAQ = LU$ and its inverse is given by

$$A^{-1} = QU^{-1}L^{-1}P,$$

  If GECPE is used then $PVAWQ = LU$ and its inverse is given by

$$A^{-1} = WQU^{-1}L^{-1}PV.$$

- For GECP the solution to $Ax = b$ in (2.19) is found as follows:

  1. Solve $Lz = Pb$ for z.
  2. Solve $Uy = z$ for y.
  3. Solve $x = Qy$ for x.

  For GEPP assume $Q = I$, it means that first two steps are done.

2.5 Singular value decomposition(SVD)

The singular values of $m \times n$ matrix A are the positive square roots of the nonzero eigenvalues of the associated Gram matrix $A^T A$. The corresponding eigenvectors of $A^T A$ are known as the singular vectors of A.

Theorem 2.3([15], Theorem 2.4.1) If A is a real $m \times n$ matrix, then there exists orthogonal matrices

$$U = [u_1, u_2, ..., u_m] \in \mathbb{R}^{m \times m} \quad \text{and} \quad V = [v_1, v_2, ..., v_n] \in \mathbb{R}^{n \times n}$$

such that

$$U^T A V = D = \text{diag}(\sigma_1, \sigma_2, ..., \sigma_p) \in \mathbb{R}^{m \times n} \tag{2.20}$$

where $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_p \geq 0$ and $p = \min(m, n)$.

It follows from (2.20) that $AV = UD$ and $A^T U = V D^T$. Therefore, the following equalities hold;

$$\begin{aligned} A v_i &= \sigma_i u_i \qquad i = 1, ..., p, \\ A^T u_i &= \sigma_i v_i \qquad i = 1, ..., p. \end{aligned}$$

The σ_i are the singular values of A and the vectors u_i and v_i are the ith left singular vector and the ith right singular vector, respectively.

The SVD reveals a great deal about the structure of a matrix. Assume A is a real $m \times n$ matrix, $p = \min(m, n)$, and $r \leq p$ denotes the number of positive singular values of A, $\sigma_1 \geq ... \geq \sigma_r > \sigma_{r+1} = ... = \sigma_p = 0$. Then

$$\begin{aligned} \text{rank}(A) &= r \\ \text{null}(A) &= \text{span}\{v_{r+1}, ..., v_n\} \\ \text{range}(A) &= \text{span}\{v_1, ..., v_r\} \end{aligned}$$

and we have the SVD expansion

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T.$$

2.5.1 Applications of the SVD

By using the SVD we can compute the following things:

- The rank of a matrix is detemined by the number of non-zero singular values. A is nonsingular if $\sigma_i \neq 0$ for all i.
- If A is an $n \times n$ matrix then the determinant of A is detemined by

$$|\det(A)| = \prod_{i=1}^{n} \sigma_i.$$

- The inverse of a square matrix is computed as follows:

  If an $n \times n$ matrix A is nonsingular and $A = UDV^T$ then its inverse is given by

  $$A^{-1} = VD^{-1}U^T$$

  where

  $$D = \text{diag}(\sigma_1, \sigma_2, ..., \sigma_n) \in \mathbb{R}^{n\times n} \quad \text{and} \quad D^{-1} = \text{diag}(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, ..., \frac{1}{\sigma_n}) \in \mathbb{R}^{n\times n}.$$

  If a $n \times n$ matrix A is singular or ill-conditioned, then its inverse can be approximated by

  $$A^{-1} \approx VD_*^{-1}U^T$$

  where

  $$D_* = \text{diag}(\sigma_1, ..., \sigma_N, 0, ..., 0) \in \mathbb{R}^{n\times n} \quad \text{and} \quad D_*^{-1} = \text{diag}(\frac{1}{\sigma_1}, ..., \frac{1}{\sigma_N}, 0, ..., 0) \in \mathbb{R}^{n\times n},$$

  N is the positive integer such that $\sigma_N > \epsilon$ and $\sigma_{N+1} \leq \epsilon$ for a given small threshold ϵ.

- The ratio $\frac{\sigma_1}{\sigma_n}$ is related to the condition number of the matrix A. σ_1 and σ_n are the largest and the smallest singular values, respectively. The larger the value of the ratio is, the closer A is to being singular.

2.5.2 Computing the SVD

In theory, we can compute the SVD of A by finding the eigenvalues and the eigenvectors of A^TA. In practice, more efficient specialized algorithms are used. We consider the Householder bidiagonalization and a variant of the QR algorithm for the SVD.

(i) Let us reduced a given $n \times n$ matrix A to a bidiagonal form by the Householder method[29] . We first determine the Householder matrix P_1 which annihilates elements in the positions $(2, 1), \cdots, (n, 1)$ in the first column of A. i.e.,

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \quad \rightarrow \quad P_1A = \begin{pmatrix} \bar{a}_{11} & \bar{a}_{12} & \cdots & \bar{a}_{1n} \\ 0 & \bar{a}_{22} & \cdots & \bar{a}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \bar{a}_{n2} & \cdots & \bar{a}_{nn} \end{pmatrix}.$$

The P_1 is given as follows[29]:

$$P_1 = I - \beta ww^T \tag{2.21}$$

with

$$s = \sqrt{\sum_{k=1}^{n} |a_{k1}|^2}, \qquad w = \begin{pmatrix} \pm(|a_{11}| + s) \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}, \qquad \beta = (s(s + |a_{11}|))^{-1}, \tag{2.22}$$

which the sign of $(|a_{11}| + s)$ should be even with the sign of a_{11}. Now we verify that

$$P_1 \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} = \begin{pmatrix} \mp s \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Suppose $a_{11} > 0$, by using (2.21) and (2.22), we obtain the following estimation

$$\begin{aligned} P_1 \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} &= \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} - \beta \begin{pmatrix} a_{11}+s \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} (a_{11}+s, a_{21}, \cdots, a_{n1}) \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} \\ &= \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} - \beta \begin{pmatrix} a_{11}+s \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} (s(a_{11}+s)) = \begin{pmatrix} -s \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \end{aligned}$$

Suppose $a_{11} < 0$, we can prove by same way as above.

We secondly determine the Householder matrix Q_1 of the form

$$Q_1 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{Q}_1 \end{pmatrix}$$

which annihilates the elements in the positions $(1,3), \cdots, (1,n)$ in the first row of $P_1 A Q_1$,

$$P_1 A = \begin{pmatrix} \bar{a}_{11} & \bar{a}_{12} & \cdots & \bar{a}_{1n} \\ 0 & \bar{a}_{22} & \cdots & \bar{a}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \bar{a}_{n2} & \cdots & \bar{a}_{nn} \end{pmatrix} \rightarrow P_1 A Q_1 = A^{(1)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & 0 & \cdots & 0 \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(1)} & a_{n3}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix}.$$

As (2.21) and (2.22) we can determine $\tilde{Q}_1$ to be satisfied

$$(\bar{a}_{12}, \cdots, \bar{a}_{1n}) \tilde{Q}_1 = (a_{12}^{(1)}, 0, \cdots, 0)$$

where $a_{12}^{(1)} = \pm \sqrt{\sum_{k=2}^{n} |a_{1k}|^2}$, and it has opposite sing of $\bar{a}_{12}$.

In general, by using the recursive formula:

$$A^{(j)} = P_j A^{(j-1)} Q_j, \quad j = 1, ..., n-2,$$

the matrix $A^{(0)} := A$ can be reduced into the bidiagonal triangular matrix:

$$B = P_{n-1}A^{(n-2)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & 0 & & & 0 \\ & a_{22}^{(2)} & a_{23}^{(2)} & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & \ddots & a_{n-1n}^{(n-2)} \\ 0 & & & & & a_{nn}^{(n-1)} \end{pmatrix}.$$

A process of reducing to the bidiagonal triangular matrix as follows. If the matrix $A^{(j-1)}$ after $j-1$ steps has the form

$$A^{(j-1)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & 0 & & & \cdots & 0 \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & & & & \\ & & \ddots & \ddots & & & \vdots \\ & & & \ddots & & & \\ \vdots & & & \ddots & \ddots & \cdots & 0 \\ & & & 0 & a_{jj}^{(j-1)} & \cdots & a_{jn}^{(j-1)} \\ & & & \vdots & \vdots & & \vdots \\ 0 & & & 0 & a_{nj}^{(j-1)} & \cdots & a_{nn}^{(j-1)} \end{pmatrix} = \left(\begin{array}{c|c} B_{j-1} & 0 \\ \hline \tilde{a}^{(j-1)} & \tilde{A}^{(j-1)} \end{array} \right),$$

where B_{j-1} is a bidiagonal matrix with $(j-1) \times j$ and $\tilde{a}^{(j-1)} = (a_{jj}^{(j-1)}, \cdots, a_{nj}^{(j-1)})^T$ then we can determine $\tilde{P}_j$ and $\tilde{Q}_j$ for $(\tilde{a}^{(j-1)}, \tilde{A}^{(j-1)})$ which they play the role as P_1 and Q_1. Then we can construct the desired $n \times n$ matrices as

$$P_j = \begin{pmatrix} I_{j-1} & 0 \\ 0 & \tilde{P}_j \end{pmatrix},$$

$$Q_j = \begin{pmatrix} I_{j-1} & 0 \\ 0 & \tilde{Q}_j \end{pmatrix}.$$

Using these formulas, we have

$$A^{(j)} = P_j A^{(j-1)} Q_j = \begin{pmatrix} I_{j-1} & 0 \\ 0 & \tilde{P}_j \end{pmatrix} \left(\begin{array}{c|c} B_{j-1} & 0 \\ \hline \tilde{a}^{(j-1)} & \tilde{A}^{(j-1)} \end{array} \right) \begin{pmatrix} I_{j-1} & 0 \\ 0 & \tilde{Q}_j \end{pmatrix}$$

$$= \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & 0 & & & & \cdots & 0 \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & & & & & \\ & & \ddots & \ddots & & & & \vdots \\ & & & \ddots & & & & \\ \vdots & & & & \ddots & \ddots & & \\ & & & & & a_{jj}^{(j)} & a_{jj+1}^{(j)} & \cdots & 0 \\ & & & & & 0 & a_{j+1j+1}^{(j)} & \cdots & a_{j+1n}^{(j)} \\ & & & & & \vdots & \vdots & & \vdots \\ 0 & & & & & 0 & a_{nj+1}^{(j)} & \cdots & a_{nn}^{(j)} \end{pmatrix}.$$

In such a way, $A^{(n-2)}$ is obtained after $n-2$ steps. Then by multiplying $A^{(n-2)}$ by $P^{(n-1)}$ from the left hand side, we obtain the bidiagonal matrix B; $B = P_{n-1}\cdots P_1 A Q_1 \cdots Q_{n-2}$ and P_i, Q_i are certain Householder matrices.

Since $Q = Q_1 \cdots Q_{n-2}$ is unitary and $B^T B = Q^T A^T A Q$, the matrices B and A have the same singular values. Thus, the remaining problem is to compute the SVD of B. We apply the QR iteration method to the tridiagonal matrix $T = B^T B$.

(ii) We present the QR iteration method[28, 29]. It allows the computation of all eigenvalues and eigenvectors of a real, symmetric, full rank matrix at once.

Let A be a real matrix. Staring $A_1 = A$, one forms matrices Q_i, R_i, A_i according to the following prescription:

$$\begin{array}{ll} A_i = Q_i R_i, & Q_i^T Q_i = I, \quad R_i \text{ is upper triangular matrix,} \\ A_{i+1} = R_i Q_i. & \end{array}$$

From that

$$A_{k+1} = R_k Q_k = Q_k^T Q_k R_k Q_k = Q_k^T A_k Q_k = Q_k^{-1} A_k Q_k,$$

it follows that A_{k+1} is similar to A_k. So they have the same eigenvalues.

Especially for Hessenberg matrices and Hermitian tridiagonal mattrices, Givens rotations are used to compute the QR factorization. Givens rotation is represented by a matrix of the form

$$G(i,j,\theta) = \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & -s & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix}$$

where $c = \cos(\theta)$ and $s = \sin(\theta)$ appear at the intersections ith and jth rows and columns. That is, the non-zero elements of the Givens matrix are given by

$$\begin{array}{ll} g_{k\,k} = 1 & \text{for } k \neq i,\ j \\ g_{i\,i} = c & \\ g_{j\,j} = c & \\ g_{j\,i} = -s & \\ g_{i\,j} = s & \text{for } i > j. \end{array}$$

When a Givens rotation $G(i,j,\theta)$ multiplies a matrix from the left hand side, only rows i and j of the matrix are affected. Thus we restrict attention to the following problem. Let a and b be given, find $c = \cos(\theta)$ and $s = \sin(\theta)$ such that

$$\begin{pmatrix} c & -s \\ s & c \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} r \\ 0 \end{pmatrix}.$$

We obtain the solution in the following form:

$$r = \sqrt{a^2 + b^2}, \qquad c = \frac{a}{r}, \qquad s = -\frac{b}{r}.$$

We should remark some facts related to the QR iteration and tridiagonal matrices[15]:

- Preservation of a form. If $T \in \mathbb{R}^{n\times n}$ is a symmetric tridiagonal, and $T = QR$ is the QR factorization then $T_+ = RQ = Q^T(QR)Q = Q^TTQ$ is also symmetric and tridiagonal.
- Shifts. If $s \in \mathbb{R}$ and $T - sI = QR$ is the QR factorization, then $T_+ = RQ + sI = Q^TTQ$ is also tridiagonal. This is called a shifted QR step.
- Perfect Shifts. If T is irreducible, then the first $n-1$ columns of $T - sI$ are independent regardless of s. Thus, if s is a eigenvalue of T and $QR = T - sI$ is a QR factorization, then $t_{nn} = 0$ and the last column of $T_+ = RQ + sI$ equals the column vector $(0, ..., 0, s)^T$.
- Cost. If $T \in \mathbb{R}^{n\times n}$ is a tridiagonal, then its QR factorization can be computed by applying a sequence of $n-1$ Givens rotations:

$$G_n \cdots G_3G_2T = R \quad \Rightarrow \quad T = QR, \quad Q = G_2^TG_3^T \cdots G_n^T, \quad G_i = G(i, i+1, \theta_i).$$

If s is a good approximate eigenvalue, then the element in the $(n, n-1)$ position of a tridiagonal matrix T will be small after a QR step with shift s. Therefore the choice of a suitable shift is important. One of effective choices is to shift by the eigenvalue λ of the 2×2 matrix

$$\begin{pmatrix} t_{n-1n-1} & t_{n-1n} \\ t_{nn-1} & t_{nn} \end{pmatrix} \tag{2.23}$$

for which $|t_{nn} - \lambda|$ is smallest.

Theorem 2.4([29], pp 374) If the QR method with the shift strategy as (23) is applied to a real, irreducible, symmetric $n \times n$ tridiagonal matrix A, then the elements $a^{(i)}_{n-1n}$ of the ith iteration matrix A_i converge to zero at least quadratically as $i \to \infty$, while $a^{(i)}_{nn}$ converges at least quadratically toward an eigenvalue of A.

For Theorem 2.4, we should remark that in the ith iteration, a shift s_i is choosed by the eigenvalue λ of the 2×2 matrix

$$\begin{pmatrix} a^{(i)}_{n-1n-1} & a^{(i)}_{n-1n} \\ a^{(i)}_{nn-1} & a^{(i)}_{nn} \end{pmatrix}$$

for which $|a^{(i)}_{nn} - \lambda|$ is smallest.

We can determine $\lambda_n \approx a^{(i)}_{nn}$ after a finite number of iterations with sufficient accuracy. Then A_i has the form

$$A_i \approx \begin{pmatrix} & & & * \\ & \tilde{A} & & \vdots \\ & & & * \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix}$$

The remaining eigenvalues of A_i are then obtained by treating the $(n-1) \times (n-1)$ matrix $\tilde{A}$ with further QR steps.

2.5.3 SVD algorithm

The first step is to reduce A to the upper bidiagonal form by the Householder method:

$$P^T A Q = B = \begin{pmatrix} d_1 & f_1 & & \cdots & 0 \\ 0 & d_2 & \ddots & & \vdots \\ & \vdots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & f_{n-1} \\ 0 & \cdots & & 0 & d_{n-1} \end{pmatrix}$$

where P^T and Q are unitary matrices([29] section 6.7,[15] section 5.4).

Since Q is unitary and $B^T B = Q^T A^T A Q$, the matrices B and A have the same singular values. The remaining problem is thus to compute the SVD of B. To this end, an implicit-shift QR step (Algorithm 8.3.2 of [15]) is applied to the tridiagonal matrix $T = B^T B$:

Step 1. Compute the eigenvalue λ of

$$\begin{pmatrix} t_{n-1n-1} & t_{n-1n} \\ t_{nn-1} & t_{nn} \end{pmatrix}$$

that is closer to t_{nn}.

Step 2. Compute $c_1 = \cos(\theta_1)$ and $s_1 = \sin(\theta_1)$ such that

$$\begin{pmatrix} c_1 & s_1 \\ -s_1 & c_1 \end{pmatrix}^T \begin{pmatrix} t_{11} - \lambda \\ t_{21} \end{pmatrix} = \begin{pmatrix} \times \\ 0 \end{pmatrix}$$

and set Givens rotation $G_1 = G(1, 2, \theta_1)$, where $\times$ is a nonzero element.

Step 3. Compute Givens rotations $G_2, \cdots, G_{n-1}$ so that if $Q = G_1 ... G_{n-1}$ then $Q^T T Q$ is tridiagonal and $Q e_1 = G_1 e_1$.

3 Our direct problem and numerical results

3.1 Discussion about a one-dimensional hyperbolic equation on the semi-axis

For a positive number T, find $\tilde{u}(x,t)$ such that

$$\begin{cases} \rho(x)\tilde{u}_{tt}(x,t) - \tilde{u}_{xx}(x,t) = 0, & t \in (0,T), x \in (0,+\infty) \\ \tilde{u}(x,0) = \tilde{u}_t(x,0) = 0, & x \in (0,+\infty) \\ \tilde{u}(0,t) = f(t), & t \in [0,T] \end{cases} \tag{3.1}$$

where $\rho(\cdot)$ is a given strictly positive C^2-smooth function on $[0,+\infty)$. Its solution describes a vertical displacement of the string initiated by a boundary sourse $f(t)$. We denote its solution by $\tilde{u}^f(x,t)$. About this problem we can discuss the follows.

First we introduce some concepts[9] related to this problem. A function $\tau(x)$ is defined by

$$\tau(x) = \int_0^x \rho^{\frac{1}{2}}(s)ds, \qquad x \geq 0 \tag{3.2}$$

The value of $\tau(x_0)$ coincides with the time needed for the wave initiated at the endpoint $x = 0$, and moving along string with the variable velocity $(1/\rho(x))^{-\frac{1}{2}}$, to arrive at the point $x = x_0$. It is a monotonic (strictly growing) positive function, $\tau(0) = 0$. $x(\tau)$ is the inverse function of $\tau(x)$, $\tau\left(x(\xi)\right) = \xi$ and $x\left(\tau(s)\right) = s$ hold for any $\xi, s \geq 0$. It has the following property:

$$x(\tau) = \int_0^\tau \rho^{-\frac{1}{2}}(x(\xi))d\xi, \qquad \tau \geq 0 \tag{3.3}$$

We shall prove the relation (3.3). By derivativing with respect to ξ the both side of the following relation

$$\xi = \tau\left(x(\xi)\right) = \int_0^{x(\xi)} \rho^{\frac{1}{2}}(s)ds,$$

we have

$$1 = x'(\xi)\rho^{\frac{1}{2}}(x(\xi)) \qquad \Rightarrow \qquad x'(\xi) = \rho^{-\frac{1}{2}}(\hat{x}(\xi)).$$

For $\tau > 0$,

$$\int_0^\tau x'(\xi)d\xi = \int_0^\tau \rho^{-\frac{1}{2}}(x(\xi))d\xi \qquad \Rightarrow \qquad x(\tau) = \int_0^\tau \rho^{-\frac{1}{2}}(x(\xi))d\xi.$$

Second we consider the following Goursat problem associated with the problem (3.1). For a positive number T, find $w(\tau,s)$ such that

$$\begin{cases} w_{ss}(\tau,s) - w_{\tau\tau}(\tau,s) + q(\tau)w(\tau,s) = 0, & 0 < \tau < s < T \\ w(0,s) = 0, \qquad w(\tau,s)\mid_{s=\tau} = -\dfrac{1}{2}\displaystyle\int_0^s q(\eta)d\eta \end{cases} \tag{3.4}$$

where

$$q(\tau) = \frac{1}{4}\frac{\rho''(x(\tau))}{\rho^2(x(\tau))} - \frac{5}{16}\frac{\left(\rho'(x(\tau))\right)^2}{\rho^3(x(\tau))} \tag{3.5}$$

In[1, 2] there are results about the smoothness of the soluton of the Goursat problem. However we will show the existence and the uniqueness of the solution of the Goursat problem. So the Goursat problem can be reduced the problem (3.6) by means of the following method[3]: suppose that $u = s + \tau, v = s - \tau$ and

$$V(u,v) = w\left(\frac{u-v}{2}, \frac{u+v}{2}\right) = w(\tau, s).$$

$w_s(\tau,s) = V_u(u,v) + V_v(u,v)$ and $w_{ss}(\tau,s) = V_{uu}(u,v) + 2V_{uv}(u,v) + V_{vv}(u,v)$.
$w_\tau(\tau,s) = V_u(u,v) - V_v(u,v)$ and $w_{\tau\tau}(\tau,s) = V_{uu}(u,v) - 2V_{uv}(u,v) + V_{vv}(u,v)$.
Now, substituting them into (3.4):

$$w_{ss}(\tau,s) - w_{\tau\tau}(\tau,s) + q(x)w(\tau,s) = 4V_{uv}(u,v) + q\left(\frac{u-v}{2}\right)V(u,v) = 0$$

$$w(0,s) = V(u,u) = 0, \qquad w(\tau,s)|_{s=\tau} = w(\tau,\tau) = V(u,0) = -\frac{1}{2}\int_0^{\frac{u}{2}} q(s)ds.$$

It follows from the domain of (3.4) that $0 < u < 2T$, $0 < v < T$ and $u + v = 2s < 2T$.

$$\begin{cases} V_{uv}(u,v) + \dfrac{1}{4}q\left(\dfrac{u-v}{2}\right)V(u,v) = 0, & (u,v) \in Q^T \\ V(u,u) = 0, \quad V(u,0) = -\dfrac{1}{2}\displaystyle\int_0^{\frac{u}{2}} q(s)ds, \end{cases} \tag{3.6}$$

where $Q^T = \{(u,v) \in (0,2T) \times (0,T) | \quad v < u < 2T - v\}$ is a triangle domain, bounded by segments $\beta_{1,T} : v = u, 0 \le u \le T$, $\beta_{2,T} : v = 2T - u, T \le u \le 2T$ and $\beta_{3,T} : v = 0, 0 \le u \le 2T$.

Let (u_0, v_0) be an arbitrary point in the domain Q^T. Denote by $D_{u_0v_0}$ a quadrangle domain with vertices $(u_0,v_0), (u_0,0), (v_0,0)$ and (v_0,v_0). Let us reduce the problem (3.6) to a linear Volterra type integral equation. By taking an integral from the equation of (3.6) over the region $D_{u,v}$ we have

$$\int_0^v \int_v^u V_{u_1v_1}(u_1,v_1)dv_1du_1 = \frac{1}{4}\int_0^v \int_v^u q\left(\frac{u_1-v_1}{2}\right)V(u_1,v_1)dv_1du_1$$

and

$$V(u,v) - V(u,0) - V(v,v) + V(v,0) = \frac{1}{4}\int_0^v \int_v^u q\left(\frac{u_1-v_1}{2}\right)V(u_1,v_1)dv_1du_1$$

By using conditions of (3.6), we have

$$V(u,v) + \frac{1}{2}\int_0^{\frac{u}{2}} q(s)ds - \frac{1}{2}\int_0^{\frac{v}{2}} q(s)ds = \frac{1}{4}\int_0^v \int_v^u q\left(\frac{u_1-v_1}{2}\right)V(u_1,v_1)dv_1du_1.$$

We can write as follows:

$$V(u,v) = -\frac{1}{2}\int_{\frac{v}{2}}^{\frac{u}{2}} q(s)ds - \frac{1}{4}\int_0^v d\eta \int_v^u q\left(\frac{\xi-\eta}{2}\right)V(\xi,\eta)d\xi, \qquad (u,v) \in \bar{Q}^T \tag{3.7}$$

where $\bar{Q}^T$ is the closure of Q^T.

We shall prove the existence of a continuous solution to the linear Volterra type integral equation (3.7) by the standard iteration method([32], pp 201-211). Since $q \in C(\mathbb{R}^+)$

$$P(u,v) = -\frac{1}{2}\int_{v/2}^{u/2} q(s)ds, \qquad \|P\| = \max_{u,v\in\bar{Q}^T} |P(u,v)|.$$

$$\|q\| = \max_{\eta,\xi\in\bar{Q}^T} \left| q\left(\frac{\xi-\eta}{2}\right)\right|.$$

We construct the following sequence:

$$V_{n+1}(u,v) = -\frac{1}{2}\int_{\frac{v}{2}}^{\frac{u}{2}} q(s)ds - \frac{1}{4}\int_0^v \int_v^u q\left(\frac{\xi-\eta}{2}\right) V_n(\xi,\eta)d\xi d\eta \qquad n = 0,1,2,..., \tag{3.8}$$

$$V_0(u,v) = 0$$

The recurrence relation of sequence (3.8) can be written in the form

$$V_n = P - \frac{1}{4}K(V_{n-1}), \qquad n = 1,2,3..., \tag{3.9}$$

$$V_0 = 0$$

were

$$P = P(u,v), \qquad K(V_{n-1})(u,v) = \int_0^v \int_v^u q\left(\frac{\xi-\eta}{2}\right) V_{n-1}(\xi,\eta)d\eta d\xi.$$

and K maps $\bar{Q}^T$ into $\bar{Q}^T$. We shall prove that

$$V_n = \sum_{m=0}^{n-1}\left(-\frac{1}{4}\right)^m K^m(P), \qquad P = K^0P, \quad n = 1,2,3..., \tag{3.10}$$

In fact, for $n = 1$ the formula (3.10) is true. Supposing that this formula is true for all integers less than n we obtain:

$$V_n = P - \frac{1}{4}K(V_{n-1}) = P - \frac{1}{4}K\sum_{m=0}^{n-2}\left(-\frac{1}{4}\right)^m K^m(P) = P - \sum_{m=0}^{n-2}\left(-\frac{1}{4}\right)^{m+1} K^{m+1}(P)$$

$$= \sum_{n=0}^{n-1}\left(-\frac{1}{4}\right)^m K^m(P).$$

Therefore the formula (3.10) is true for all n.

$$\|K(P)\| = \max_{u,v\in\bar{Q}^T}\left|\int_0^v\int_v^u q\left(\frac{\xi-\eta}{2}\right)P(\xi,\eta)d\xi d\eta\right| \le \|P\|\|q\|\int_0^v\int_v^u d\xi d\eta \le \|P\|\|q\|uv.$$

The iterates $K^p(P) \in C(\bar{Q}^T)$ satisfy the inequality

$$K^n(P) \le \|P\|\|q\|^n\frac{u^n}{n!}\frac{v^n}{n!}, \qquad (u,v)\in\bar{Q}^T, \quad n = 1,2,..., \tag{3.11}$$

We shall prove the enquality (3.11) by the mathematical induction with respect to n. The inequality is true for $n = 1$, Assuming that the inquality is true for $n-1$, we shall prove it for n:

$$\|K^n(P)(u,v)\| = \|K(K^{n-1}(P))(u,v)\| = \max_{u,v\in\bar{Q}^T}\left|\int_0^v\int_v^u q\left(\frac{\xi-\eta}{2}\right)(K^{n-1}(P))(\xi,\eta)d\xi d\eta\right| \leq$$

$$\|q\|\|P\|\|q\|^{n-1}\int_0^v\int_v^u \frac{\xi^{n-1}}{(n-1)!}\frac{\eta^{n-1}}{(n-1)!}d\xi d\eta \leq \|P\|\|q\|^n\frac{u^n-v^n}{n!}\frac{v^n}{n!} \leq \|P\|\|q\|^n\frac{u^n}{n!}\frac{v^n}{n!}.$$

It follows from the enquality (3.11) that the series

$$\sum_{m=0}^{n-1}\left(-\frac{1}{4}\right)^m(K^m(P))(u,v), \qquad (u,v)\in\bar{Q}^T$$

is majorized over $\bar{Q}^T$ by the following converging numerical series.

$$\|P\|\sum_{m=0}^{\infty}\left(\frac{1}{4}\right)^m\|q\|^m\frac{u^m}{m!}\frac{v^m}{m!} \leq \|P\|\sum_{m=0}^{\infty}\left(-\frac{1}{4}\right)^m\|q\|^m\frac{u^m}{m!}\frac{u^m}{m!} = \|P\|\sum_{m=0}^{\infty}\left[\frac{(2^{-1}\sqrt{\|q\|}u)^m}{m!}\right]^2$$

$$\leq \|P\|\left[\sum_{m=0}^{\infty}\frac{(2^{-1}\sqrt{\|q\|}u)^m}{m!}\right]^2 = \|P\|e^{\sqrt{\|q\|}u} \leq \|P\|e^{2\sqrt{\|q\|}T}.$$

The series (3.10) converges uniformily with respect to (u,v) in $\bar{Q}^T$ and this converging defines a function $V(u,v)$ which is continuous over $\bar{Q}^T$. By virtue of (3.9), the succissive approximations V_n as $n\to\infty$ uniformly tend to the function V:

$$V_n(u,v) \Longrightarrow V(u,v) = \sum_{m=0}^{\infty}\left(-\frac{1}{4}\right)^m(K^m(P))(u,v), \tag{3.12}$$

So, the inquality

$$\|V(u,v)\| \leq \|P\|e^{2\sqrt{\|q\|}T}$$

is true. Letting $n\to\infty$ in (3.8) and using the uniform convergence of the sequence V_n to V over $\bar{Q}^T$, we obtain the result that the existence of a continuous solution $V(u,v)$ to (3.7).

We shall prove the uniqueness of the solution to (3.7) in the class $C(\bar{Q}^T)$ and will apply the analogous method to that used in [26]. Let us assume that there is a another solution $V^*\in C(\bar{Q}^T)$, i.e.

$$V^*(u,v) = -\frac{1}{2}\int_{\frac{v}{2}}^{\frac{u}{2}} q(s)ds - \frac{1}{4}\int_0^v\int_v^u q\left(\frac{\xi-\eta}{2}\right)V^*(\xi,\eta)d\xi d\eta. \tag{3.13}$$

By subtracting (3.13) from (3.8) we have

$$V_{n+1}(u,v) - V^*(u,v) = -\frac{1}{4}\int_0^v\int_v^u q\left(\frac{\xi-\eta}{2}\right)[V_n(\xi,\eta) - V^*(\xi,\eta)]d\xi d\eta. \tag{1.19}$$

We have a new sequence:

$$V_n(u,v) - V^*(u,v), \qquad n = 0,1,2,...,$$

and we choose a positive constant λ such that $\lambda > \|q\|$. We shall employ the mathematical induction to show that

$$|V_n(u,v) - V^*(u,v)| \le \left(\frac{\|q\|}{4\lambda}\right)^n \max_{\bar{Q}^T} |V^*(u,v)| e^{\sqrt{\lambda}(u+v)}, \qquad (u,v) \in \bar{Q}^T \tag{3.14}$$

for $n = 0, 1, 2, ...$ It is easy to verify that the above inequality is true for $n = 0$. Assume that (3.14) is true for $n = N$. It follows from this assumption that

$$|V_{N+1}(u,v) - V^*(u,v)| = \left|\frac{1}{4}\int_0^v \int_v^u q\left(\frac{\xi-\eta}{2}\right)(V_N(\xi,\eta) - V^*(\xi,\eta))d\xi d\eta\right| \le \frac{1}{4}\int_0^v \int_v^u$$

$$\left|q\left(\frac{\xi-\eta}{2}\right)\right| |V_N(\xi,\eta) - V^*(\xi,\eta)|d\xi d\eta \le \frac{1}{4}\|q\|\left(\frac{\|q\|}{4\lambda}\right)^N \max_{\bar{Q}^T} |V^*(u,v)| \int_0^v \int_v^u e^{\sqrt{\lambda}(\xi+\eta)} d\xi d\eta$$

$$= \max_{\bar{Q}^T} |V^*(u,v)| \left(\frac{\|q\|}{4\lambda}\right)^{N+1} (e^{\sqrt{\lambda}v} - 1)(e^{\sqrt{\lambda}u} - e^{\sqrt{\lambda}v}) \le \max_{\bar{Q}^T} |V^*(u,v)| \left(\frac{\|q\|}{4\lambda}\right)^{N+1} e^{\sqrt{\lambda}(u+v)}.$$

Therefore, (3.14) is true for all $n \ge 0$. By passing to the limit as $n \to \infty$ in (3.14), we obtain $V(u,v) = V^*(u,v)$.

We shall show that for any $f \in C^2[0,T]$ provided $f(0) = f'(0) = f''(0) = 0$, $\tilde{u}^f(x,t)$ defined in (3.15) becomes the classical solution of the problem (3.1).

$$\tilde{u}^f(x,t) = \left(\frac{\varrho(x)}{\varrho(0)}\right)^{-\frac{1}{4}} f(t - \tau(x)) + \int_x^{x(t)} \bar{w}(x,s) f(t - \tau(s)) ds \quad x \ge 0, 0 \le t \le T \tag{3.15}$$

where

$$\bar{w}(x,s) = \left[\left(\frac{\varrho(x)}{\varrho(0)}\right)^{-\frac{1}{4}} \varrho^{\frac{1}{2}}(s)\right] w(\tau(x), \tau(s))$$

and $w(\tau, s)$ is a unique solution to the Grousat problem (3.4). Author of [9] gives the lemma that the unique solution of the problem (3.1) can be represented in the form (3.15). He also gives the outline of its proof. However we consider it in detail.

Remark 3.1. All functions depending on time t are assumed to be extended by zero to $t < 0$.

Let us introduce two Hilbert spaces.
$\tilde{H} = L_{2,\varrho}(\mathbb{R}^+)$,

$$(y,w)_{\tilde{H}} = \int_0^{+\infty} y(x)w(x)\varrho(x)dx.$$

$H = L_2(\mathbb{R}^+)$,

$$(f,g)_H = \int_0^{+\infty} f(\tau)g(\tau)d\tau.$$

The map $A : \tilde{H} \to H$,

$$y(\tau) = (A\tilde{y})(\tau) = \varrho^{\frac{1}{4}}(x(\tau))\tilde{y}(x(\tau)), \qquad \tau \ge 0. \tag{3.16}$$

The map A is a unitary operator from $\tilde{H}$ onto H[9]. Therefore for $f \in H$,

$$\left((A)^{-1}f\right)(x) = \left((A)^{*}f\right)(x) = \varrho^{-\frac{1}{4}}(x)f(\tau(x)), \qquad \tau \geq 0.$$

We consider a fuction $u^f(\tau,t)$: for a fixed $t > 0$

$$u^f(\tau,t) := \varrho^{-\frac{1}{4}}(0)(A\tilde{u}^f(\cdot,t))(\tau) = \left(\frac{\varrho(x(\tau))}{\varrho(0)}\right)^{\frac{1}{4}} \tilde{u}^f(x(\tau),t) \tag{3.17}$$

Substituting (3.15) into the definition of $u^f(\tau,t)$:

$$u^f(\tau,t) = \left(\frac{\varrho(x(\tau))}{\varrho(0)}\right)^{\frac{1}{4}} \left[\left(\frac{\varrho(x(\tau))}{\varrho(0)}\right)^{-\frac{1}{4}} f(t-\tau) + \int_{x(\tau)}^{x(t)} \left(\frac{\varrho(x(\tau))}{\varrho(0)}\right)^{-\frac{1}{4}} \varrho^{\frac{1}{2}}(s)w(\tau,\tau(s))f(t-\tau(s))ds\right],$$

we obtain

$$u^f(\tau,t) = f(t-\tau) + \int_{x(\tau)}^{x(t)} \varrho^{\frac{1}{2}}(s)w(\tau,\tau(s))f(t-\tau(s))ds.$$

Changing the variable $\tau(s) = \eta, \quad s = x(\eta)$ and taking into account

$$ds = (x(\eta))'d\eta = \left(\int_0^{\eta} \varrho^{-\frac{1}{2}}(x(\xi))d\xi\right)' d\eta = \varrho^{-\frac{1}{2}}(s)d\eta,$$

we arrive at the following form

$$u^f(\tau,t) = f(t-\tau) + \int_{\tau}^{t} w(\tau,\eta)f(t-\eta)d\eta, \qquad \tau \geq 0, \quad 0 \leq t \leq T \tag{3.18}$$

By putting $t = 0$ in (3.18),

$$u^f(\tau,0) = f(-\tau) + \int_{\tau}^{0} w(\tau,\eta)f(-\eta)d\eta = 0.$$

From (3.17),

$$0 = u^f(\tau,0) = \left(\frac{\varrho(x(\tau))}{\varrho(0)}\right)^{\frac{1}{4}} \tilde{u}^f(x(\tau),0) \quad \Rightarrow \quad u^f(\tau,0) = \tilde{u}^f(x(\tau),0) = 0.$$

By putting $t = 0$ in the derivative of (3.18),

$$u^f_t(\tau,t) = f'(t-\tau) + w(\tau,t)f(0) + \int_{\tau}^{t} w(\tau,\eta)f'(t-\eta)d\eta \quad \Rightarrow$$

$$u^f_t(\tau,0) = f'(-\tau) + w(\tau,t)f(0) + \int_{\tau}^{0} w(\tau,\eta)f'(-\eta)d\eta = 0.$$

From (3.17),

$$0 = u^f_t(\tau,0) = \left(\frac{\varrho(x(\tau))}{\varrho(0)}\right)^{\frac{1}{4}} \tilde{u}^f_t(x(\tau),0) \quad \Rightarrow \quad u^f_t(\tau,0) = \tilde{u}^f_t(x(\tau),0) = 0.$$

By putting $\tau = 0$ in (3.17),

$$u^f(0,t) = \left(\frac{\varrho(0)}{\varrho(0)}\right)^{\frac{1}{4}} \tilde{u}^f(0,t) = \tilde{u}^f(0,t).$$

By putting $\tau = 0$ in (3.18) and taking into account $w(0,s) = 0$, we obtain

$$\tilde{u}^f(0,t) = u^f(0,t) = f(t) + \int_0^t w(0,\eta) f'(t-\eta) d\eta = f(t).$$

We just proved that $\tilde{u}^f(x,t)$ satisfies the conditons of the problem (3.1).

We shall prove that $\tilde{u}^f(x,t)$ becomes a solution to equation of the problem (3.1). The following equality (3.19) is very important to prove it. The equality

$$u^f_{tt}(\tau,t) - u^f_{\tau\tau}(\tau,t) + \left[\frac{1}{4}\frac{\varrho''(x(\tau))}{\varrho^2(x(\tau))} - \frac{5}{16}\frac{[\varrho'(x(\tau))]^2}{\varrho^3(x(\tau))}\right] u^f(\tau,t) =$$

$$\frac{\varrho^{-\frac{3}{4}}(x(\tau))}{\varrho(0)}\left[\varrho(x(\tau)\tilde{u}^f_{tt}(x(\tau),t) - \tilde{u}^f_{xx}(x(\tau),t)\right]. \tag{3.19}$$

is valid. Let us prove the equality (3.19) and take a derivative from the definition of $u^f(\tau,t)$.

$$u^f_{tt}(\tau,t) = \left(\frac{\varrho(x(\tau))}{\varrho(0)}\right)^{\frac{1}{4}} \tilde{u}^f_{tt}(x(\tau),t). \tag{3.19*}$$

$$u^f_\tau(\tau,t) = \frac{1}{4}\left(\frac{\varrho(x(\tau))}{\varrho(0)}\right)^{-\frac{3}{4}} \frac{\varrho'(x(\tau))}{\varrho(0)} \varrho^{-\frac{1}{2}}(x(\tau))\tilde{u}^f(x(\tau),t) + \left(\frac{\varrho(x(\tau))}{\varrho(0)}\right)^{\frac{1}{4}} \tilde{u}^f_x(x(\tau))\varrho^{-\frac{1}{2}}(x(\tau))$$

$$= u^f_\tau(\tau,t) = \frac{1}{4}\frac{\varrho^{-\frac{5}{4}}(x(\tau))}{\varrho^{\frac{1}{4}}(0)}\varrho'(x(\tau))\tilde{u}^f(x(\tau),t) + \frac{\varrho^{-\frac{1}{4}}(x(\tau))}{\varrho^{\frac{1}{4}}(0)}\tilde{u}^f_x(x(\tau)).$$

$$u^f_{\tau\tau}(\tau,t) = -\frac{5}{16}\frac{\varrho^{-\frac{9}{4}}(x(\tau))}{\varrho^{\frac{1}{4}}(0)}\varrho^{-\frac{1}{2}}(x(\tau))\varrho'(x(\tau))\tilde{u}^f(x(\tau),t) + \frac{1}{4}\frac{\varrho^{-\frac{5}{4}}(x(\tau))}{\varrho^{\frac{1}{4}}(0)}\varrho''(x(\tau))\varrho^{-\frac{1}{2}}(x(\tau))\tilde{u}^f(x(\tau),t)$$

$$+\frac{1}{4}\frac{\varrho^{-\frac{5}{4}}(x(\tau))}{\varrho^{\frac{1}{4}}(0)}\varrho'(x(\tau))\tilde{u}^f_x(x(\tau),t)\varrho^{-\frac{1}{2}}(x(\tau) - \frac{1}{4}\frac{\varrho^{-\frac{5}{4}}(x(\tau))}{\varrho^{\frac{1}{4}}(0)}\varrho'(x(\tau))\varrho^{-\frac{1}{2}}(x(\tau)\tilde{u}^f_x(x(\tau),t)$$

$$+\frac{\varrho^{-\frac{1}{4}}(x(\tau))}{\varrho^{\frac{1}{4}}(0)}\tilde{u}^f_{xx}(x(\tau))\varrho^{-\frac{1}{2}}(x(\tau) = -\frac{5}{16}\frac{\varrho^{-\frac{11}{4}}(x(\tau))}{\varrho^{\frac{1}{4}}(0)}\varrho'(x(\tau))\tilde{u}^f(x(\tau),t) + \frac{1}{4}\frac{\varrho^{-\frac{7}{4}}(x(\tau))}{\varrho^{\frac{1}{4}}(0)}\varrho''(x(\tau))\tilde{u}^f(x(\tau),t)$$

$$+\frac{\varrho^{-\frac{3}{4}}(x(\tau))}{\varrho^{\frac{1}{4}}(0)}\tilde{u}^f_{xx}(x(\tau)) \quad \Rightarrow$$

$$u^f_{\tau\tau}(\tau,t) = \left[-\frac{5}{16}\frac{[\varrho'(x(\tau))]^2}{\varrho^3(x(\tau))} + \frac{1}{4}\frac{\varrho''(x(\tau))}{\varrho^2(x(\tau))}\right]\left(\frac{\varrho(x(\tau))}{\varrho(0)}\right)^{\frac{1}{4}} \tilde{u}^f(x(\tau),t) + \frac{\varrho^{-\frac{3}{4}}(x(\tau))}{\varrho^{\frac{1}{4}}(0)}\tilde{u}^f_{xx}(x(\tau)). \tag{3.19**}$$

Subtracting (3.19**) from (3.19*) gives

$$u^f_{tt}(\tau,t)-u^f_{\tau\tau}(\tau,t)=\left(\frac{\varrho(x(\tau))}{\varrho(0)}\right)^{\frac{1}{4}}\tilde{u}^f_{tt}(x(\tau),t)-\left[\frac{1}{4}\frac{\varrho''(x(\tau))}{\varrho^2(x(\tau))}-\frac{5}{16}\frac{[\varrho'(x(\tau))]^2}{\varrho^3(x(\tau))}\right]u^f(\tau,t)$$

$$-\frac{\varrho^{-\frac{3}{4}}(x(\tau))}{\varrho^{\frac{1}{4}}(0)}\tilde{u}^f_{xx}(x(\tau))=\frac{\varrho^{-\frac{3}{4}}(x(\tau))}{\varrho^{\frac{1}{4}}(0)}\left[\varrho(x(\tau))\tilde{u}^f_{tt}(x(\tau),t)-\tilde{u}^f_{xx}(x(\tau))\right]-q(\tau)u^f(\tau,t).$$

It is sufficient to show that $u^f(\tau,t)$ becomes a solution of the problem:

$$\begin{cases}u_{tt}(\tau,t)-u_{\tau\tau}(\tau,t)+q(\tau)u(\tau,t)=0, & \tau>0,\quad 0<t<T\\ u(\tau,0)=u_t(\tau,0)=0, & \tau\geq 0\\ u(0,t)=f(t), & t\in[0,T]\end{cases}\tag{3.20}$$

Namely, $\tilde{u}^f(x,t)$ becames a solution to the problem (3.1) if and only if $u^f(x,t)$ became a solution to the problem (3.20).

To substitute (3.18) into the equation of the problem (3.20) we should find all derivatives of $u^f(\tau,t)$ up to the second order.

$$u^f_t(\tau,t)=f'(t-\tau)+w(\tau,t)f(0)+\int_\tau^t w(\tau,\eta)f'(t-\eta)d\eta=f'(t-\tau)+\int_\tau^t w(\tau,\eta)f'(t-\eta)d\eta.$$

$$u^f_{tt}(\tau,t)=f''(t-\tau)+w(\tau,t)f'(0)+\int_\tau^t w(\tau,\eta)f''(t-\eta)d\eta=f''(t-\tau)+\int_\tau^t w(\tau,\eta)f''(t-\eta)d\eta.\tag{3.18*}$$

$$u^f_\tau(\tau,t)=-f'(t-\tau)-w(\tau,\tau)f(t-\tau)+\int_\tau^t w_\tau(\tau,\eta)f(t-\eta)d\eta$$

$$u^f_{\tau\tau}(\tau,t)=f''(t-\tau)-\frac{dw(\tau,\tau)}{d\tau}f(t-\tau)+w(\tau,\tau)f'(t-\tau)-w_\tau(\tau,\eta)|_{\eta=\tau}f(t-\tau)+\int_\tau^t w_{\tau\tau}(\tau,\eta)f(t-\eta)d\eta.\tag{3.18**}$$

Substituting (3.18*) and (3.18**) into the equation of the problem (3.20), we have

$$u^f_{tt}(\tau,t)-u^f_{\tau\tau}(\tau,t)+q(\tau)u^f(\tau,t)=\int_\tau^t w(\tau,\eta)f''(t-\eta)d\eta+\frac{dw}{d\tau}(\tau,\tau)f(t-\tau)-w(\tau,\tau)f'(t-\tau)+$$

$$\frac{\partial w}{\partial\tau}(\tau,\eta)|_{\tau=\eta}f(t-\eta)-\int_\tau^t w_{\tau\tau}(\tau,\eta)f(t-\eta)d\eta+q(\tau)f(t-\eta)+q(\tau)\int_\tau^t w(\tau,\eta)f(t-\eta)d\eta.\tag{3.20*}$$

Integrating by parts twice the first term in the right hand side of the expression:

$$\int_\tau^t w(\tau,\eta)f''(t-\eta)d\eta=-w(\tau,\eta)f'(t-\eta)\,|_\tau^t+\int_\tau^t w_\eta(\tau,\eta)f'(t-\eta)d\eta=-w(\tau,\tau)f'(t-\tau)$$

$$+\int_\tau^t w_\eta(\tau,\eta)f'(t-\eta)d\eta=w(\tau,\tau)f'(t-\tau)-w_\eta(\tau,\eta)f(t-\eta)\,|_\tau^t+\int_\tau^t w_{\eta\eta}(\tau,\eta)f(t-\eta)d\eta\quad\Rightarrow$$

$$\int_\tau^t w(\tau,\eta)f''(t-\eta)d\eta=w(\tau,\tau)f'(t-\tau)+w_\eta(\tau,\eta)\,|_{\eta=\tau}\,f(t-\eta)+\int_\tau^t w_{\eta\eta}(\tau,\eta)f(t-\eta)d\eta.\tag{3.20**}$$

By substituting (3.20**) into (3.20*), we obtain

$$u^f_{tt}(\tau,t)-u^f_{\tau\tau}(\tau,t)+q(\tau)u^f(\tau,t)=w(\tau,\tau)f'(t-\tau)+w_\eta(\tau,\eta)\mid_{\eta=\tau} f(t-\eta)+\int_\tau^t w_{\eta\eta}(\tau,\eta)f(t-\eta)d\eta$$

$$+\frac{dw}{d\tau}(\tau,\tau)f(t-\tau)-w(\tau,\tau)f'(t-\tau)+\frac{\partial w}{\partial\tau}(\tau,\eta)|_{\tau=\eta}f(t-\eta)-\int_\tau^t w_{\tau\tau}(\tau,\eta)f(t-\eta)d\eta+q(\tau)f(t-\eta)$$

$$+q(\tau)\int_\tau^t w(\tau,\eta)f(t-\eta)d\eta=\left[\frac{dw(\tau,\tau)}{d\tau}+\left(\frac{\partial w(\tau,\eta)}{\partial\tau}+\frac{\partial w(\tau,\eta)}{\partial\eta}\right)_{\eta=\tau}\right]f(t-\tau)$$

$$+q(\tau)f(t-\tau)+\int_\tau^t\left[w_{\eta\eta}(\tau,\eta)-w_{\tau\tau}(\tau,\eta)+q(\tau)w(\tau,\eta)\right]f(t-\eta)d\eta \quad\Rightarrow$$

$$u^f_{tt}(\tau,t)-u^f_{\tau\tau}(\tau,t)+q(\tau)u^f(\tau,t)=\left[2\frac{dw(\tau,\tau)}{d\tau}+q(\tau)\right]f(t-\tau)$$

$$+\int_\tau^t\left[w_{\eta\eta}(\tau,\eta)-w_{\tau\tau}(\tau,\eta)+q(\tau)w(\tau,\eta)\right]f(t-\eta)d\eta=0.$$

3.2 Our direct problem (Problem 1) and the smoothness of the solution

We consider the following direct problem and solve numerically[4].

Problem 1. For a positive number T, find $u(x,t)$ such that

$$\begin{cases}\rho(x)u_{tt}(x,t)-u_{xx}(x,t)=0, & t\in(0,T), \quad x\in(0,1)\\ u(x,0)=\alpha(x), \qquad u_t(x,0)=\beta(x), & x\in(0,1)\\ u(0,t)=f(t), \qquad u(1,t)=g(t), & t\in[0,T]\end{cases}$$

where $\rho(x)$ is a strictly positive function on $[0,1]$, $\alpha(x),\beta(x),f(t),g(t)$ are given functions.

Remark 3.2. To Problem 1 for a non-negative integer k the k-th order compatibility condition is as follows[12]:

$$L^k\alpha(0)=\frac{\partial^{2k}}{\partial t^{2k}}f(0), \quad L^k\alpha(1)=\frac{\partial^{2k}}{\partial t^{2k}}g(0), \quad L^k\beta(0)=\frac{\partial^{2k+1}}{\partial t^{2k+1}}f(0), \quad L^k\beta(1)=\frac{\partial^{2k+1}}{\partial t^{2k+1}}g(0)$$

where $L\equiv\dfrac{1}{\rho}\dfrac{\partial^2}{\partial x^2}$. The regularity of $u(x,t)$ depends on the regularity of functions $\rho(x),\alpha(x),\beta(x)$, $f(t),g(t)$ and the compatibility condition[16, 27]. Functions become smoother and the higher order compatibility condition holds, then $u(x,t)$ becomes smoother.

For invetigating the smoothness of the solution we solve the following examples numerically where $\rho(x)$ is analytic (constant) or smooth ($C^\infty[0,1]$).

Example 3.1.

$$\begin{cases} \pi^2 u_{tt}(x,t) - u_{xx}(x,t) = 0, & t \in (0,T), \quad x \in (0,1) \\ u(x,0) = \dfrac{\pi^3}{8} x(1-x), \qquad u_t(x,0) = 0, & x \in [0,1] \\ u(0,t) = 0, \qquad u(1,t) = 0, & t \in [0,T] \end{cases}$$

The first order compatibility condition ($k = 1$) is not satisfied. The exact solution is given as follows [12]:

$$u(x,t) = \sum_{n=1}^{\infty} \frac{\cos(2n-1)t}{(2n-1)^3} \sin(2n-1)\pi x$$

Example 3.2.

$$\begin{cases} u_{tt}(x,t) - u_{xx}(x,t) = 0, & t \in (0,T), \quad x \in (0,1) \\ u(x,0) = 0, \qquad u_t(x,0) = 2 - 2x, & x \in [0,1] \\ u(0,t) = t + \sin t, \qquad u(1,t) = 0, & t \in [0,T] \end{cases}$$

The first order compatibility condition ($k = 1$) is not satisfied. The exact solution is given as follows:

$$u(x,t) = (1-x)(t + \sin t) + 2 \sum_{n=1}^{\infty} u_n(t) \sin n\pi x$$

where

$$u_n(t) = \frac{1}{(n\pi)^2} \int_0^t \sin n\pi(t-\tau) \sin \tau d\tau = \frac{1}{(n\pi)^2(1-(n\pi)^2)} \left(-n\pi \sin t + \sin n\pi t\right)$$

Example 3.3.

$$\begin{cases} \rho u_{tt}(x,t) - u_{xx}(x,t) = 0, & t \in (0,T), \quad x \in (0,1) \\ u(x,0) = 0, \qquad u_t(x,0) = 0, & x \in [0,1] \\ u(0,t) = f_\varepsilon(t), \quad u(1,t) = 0, & t \in [0,T] \end{cases}$$

where ρ is a positive constant, $f_\varepsilon(t)$ is given as follows:

$$r(x) = \begin{cases} 0, & x \le 0 \\ \exp\left(-\frac{1}{x}\right), & x > 0 \end{cases} \qquad h(x) = \begin{cases} 0, & x \le 0 \\ \dfrac{r(x)}{r(x) + r(1-x)}, & 0 < x \le 1 \\ 1, & x \ge 1 \end{cases}$$

$$f_\varepsilon(x) = x h\left(\frac{x}{\varepsilon}\right)$$

For any non-negative integer k the k-th order compatibility condition is satisfied. Let n be a non-negative integer. The exact solution is smooth and it is given as follows:

- $2n\sqrt{\rho} \le t < (2n+1)\sqrt{\rho}$

$$u(x,t) = \begin{cases} \left(\sum_{k=0}^{n} f_\varepsilon(t - \sqrt{c}x - 2k\sqrt{\rho})\right) - \left(\sum_{k=1}^{n} f_\varepsilon(t - \sqrt{\rho}(1-x) - (2k-1)\sqrt{\rho})\right), \\ t - \sqrt{\rho}x > 2n\sqrt{\rho} \\ \left(\sum_{k=0}^{n-1} f_\varepsilon(t - \sqrt{\rho}x - 2k\sqrt{\rho})\right) - \left(\sum_{k=1}^{n} f_\varepsilon(t - \sqrt{\rho}(1-x) - (2k-1)\sqrt{\rho})\right), \\ t - \sqrt{\rho}x \le 2n\sqrt{\rho} \end{cases}$$

- $(2n+1)\sqrt{\rho} \le t < 2(n+1)\sqrt{\rho}$

$$u(x,t) = \begin{cases} \left(\sum_{k=0}^{n} f_\varepsilon(t - \sqrt{\rho}x - 2k\sqrt{\rho})\right) - \left(\sum_{k=1}^{n} f_\varepsilon(t - \sqrt{\rho}(1-x) - (2k-1)\sqrt{\rho})\right), \\ t - \sqrt{\rho}(1-x) \le (2n+1)\sqrt{\rho} \\ \left(\sum_{k=0}^{n} f_\varepsilon(t - \sqrt{\rho}x - 2k\sqrt{\rho})\right) - \left(\sum_{k=1}^{n+1} f_\varepsilon(t - \sqrt{\rho}(1-x) - (2k-1)\sqrt{\rho})\right), \\ t - \sqrt{\rho}(1-x) > (2n+1)\sqrt{\rho} \end{cases}$$

Example 3.4.

$$\begin{cases} \rho(x)u_{tt}(x,t) - u_{xx}(x,t) = 0, & t \in (0,T), \quad x \in (0,1) \\ u(x,0) = \sin \pi x, \qquad u_t(x,0) = 0, & x \in [0,1] \\ u(0,t) = 0, \qquad u(1,t) = 0, & t \in [0,T] \end{cases}$$

where $\rho(x) = 1/\left(1 + 0.5\exp\left(-1/x(1-x)\right)\sin \pi x\right)$. For any non-negative integer k the k-th order compatibility condition is satisfied(see Remark 3.3). We do not know about the concrete representation of the exact solution.

Example 3.5.

$$\begin{cases} \rho u_{tt}(x,t) - u_{xx}(x,t) = 0, & t \in (0,T), \quad x \in (0,1) \\ u(x,0) = \sin \pi x, \qquad u_t(x,0) = 0, & x \in [0,1] \\ u(0,t) = 0, \qquad u(1,t) = 0, & t \in [0,T] \end{cases}$$

where ρ is a positive constant. For any non-negative integer k the k-th order compatibility condition is satisfied. The exact solution is analytic and it is given as follows:

$$u(x,t) = \cos\left(\frac{\pi t}{\sqrt{\rho}}\right)\sin \pi x$$

Remark 3.3. Example 3.4 is derived from Example 3.5 as follows. If Problem 1 with a constant ρ has an analytic solution, it can be modified into a problem with a following non-constant $\rho(x)$ which satisfies the k-th order compatibility condition for any non-negative integer k:

$$\rho(x) = \rho_1(x) = \begin{cases} \rho + \exp\left(-\frac{1}{x(1-x)}\right)\varphi(x), & 0 < x < 1 \\ \rho, & x = 0, 1 \end{cases}$$

or

$$\rho(x) = \rho_2(x) = \begin{cases} \dfrac{\rho}{1 + \exp\left(-\frac{1}{x(1-x)}\right)\varphi(x)}, & 0 < x < 1 \\ \rho, & x = 0, 1 \end{cases}$$

where $\varphi(x)$ is an arbitrary function such that $\rho_1(x), \rho_2(x) > 0 \quad (0 < x < 1)$.

Proof. If Problem 1 with a constant ρ has an analytic solution then for any non-negative integer k, k-th order compatibility condition about $\tilde{L} \equiv \dfrac{1}{\rho}\dfrac{\partial^2}{\partial x^2}$ is satisfied.

$$L \equiv \frac{1}{\rho(x)}\frac{\partial^2}{\partial x^2}$$

where $\rho(x) = \rho_i(x), i = 1$ or 2. For simplicity we denote $\dfrac{1}{\rho(x)}$ by $\mu(x)$, we have

$$L \equiv \frac{1}{\rho(x)}\frac{\partial^2}{\partial x^2} = \mu(x)\frac{\partial^2}{\partial x^2}.$$

With0 the help of the equality

$$\lim_{t\to+0} L^k = \tilde{L}^k \quad \text{and} \quad \lim_{t\to1-0} L^k = \tilde{L}^k \qquad \text{for} \quad k = 1, 2, ... \tag{3.21}$$

we shall show that for any non-negative integer k, k-th order compatibility condition about L is satisfied. So first we prove the equality (3.21) by using the mathematical induction method.
(a) When $\rho(x) = \rho_1(x)$, we have

$$\mu(x) = \begin{cases} \dfrac{1}{\rho + \exp\left(-\frac{1}{x(1-x)}\right)\varphi(x)}, & 0 < x < 1 \\ \dfrac{1}{\rho}, & x = 0, 1 \end{cases}$$

When $k = 1$,

$$\lim_{x\to+0}\mu(x) = \lim_{x\to+0}\frac{1}{\rho + \exp\left(-\frac{1}{x(1-x)}\right)\varphi(x)} = \frac{1}{\rho}$$

$$\text{and} \quad \lim_{x\to1-0}\mu(x) = \lim_{x\to1-0}\frac{1}{\rho + \exp\left(-\frac{1}{x(1-x)}\right)\varphi(x)} = \frac{1}{\rho}.$$

So,

$$L = \lim_{x\to+0}\mu(x)\frac{\partial^2}{\partial x^2} = \frac{1}{\rho}\frac{\partial^2}{\partial x^2} = \tilde{L} \quad \text{and} \quad L = \lim_{x\to1-0}\mu(x)\frac{\partial^2}{\partial x^2} = \frac{1}{\rho}\frac{\partial^2}{\partial x^2} = \tilde{L}.$$

When $k = 2$, the first order derivative of $\mu(x)$ is

$$\frac{\partial\mu(x)}{\partial x} = \begin{cases} \dfrac{-1}{\left[\rho + \exp\left(-\frac{1}{x(1-x)}\right)\varphi(x)\right]^2}\left[\dfrac{1-2x}{[x(1-x)]^2}\varphi(x) + \dfrac{\partial\varphi(x)}{\partial x}\right]\exp\left(-\frac{1}{x(1-x)}\right), & 0 < x < 1 \\ 0, & x = 0, 1, \end{cases}$$

and the second order derivative of $\mu(x)$ is

$$\frac{\partial^2\mu(x)}{\partial^2 x} = \frac{2}{\left[\rho + \exp\left(-\frac{1}{x(1-x)}\right)\varphi(x)\right]^3}\left[\frac{1-2x}{[x(1-x)]^2}\varphi(x) + \frac{\partial\varphi(x)}{\partial x}\right]^2 \exp\left(-\tfrac{2}{x(1-x)}\right)$$
$$-\frac{1}{\left[\rho + \exp\left(-\frac{1}{x(1-x)}\right)\varphi(x)\right]^2}\left[\frac{1-6x+12x^2-12x^3+6x^4}{[x(1-x)]^4}\varphi(x)\right]\exp\left(-\tfrac{1}{x(1-x)}\right)$$
$$-\frac{1}{\left[\rho + \exp\left(-\frac{1}{x(1-x)}\right)\varphi(x)\right]^2}\left[\frac{2-4x}{[x(1-x)]^2}\frac{\partial\varphi(x)}{\partial x} + \frac{\partial^2\varphi(x)}{\partial^2 x}\right]^2 \exp\left(-\tfrac{1}{x(1-x)}\right).$$

By using that

$$\lim_{x\to+\infty} P(t)\exp(-t) = 0$$

where $P(t)$ is any polynomial of n order, we obtain

$$\lim_{x\to+0}\frac{\partial\mu(x)}{\partial x} = 0 \quad \text{and} \quad \lim_{x\to1-0}\frac{\partial\mu(x)}{\partial x} = 0,$$
$$\lim_{x\to+0}\frac{\partial^2\mu(x)}{\partial x^2} = 0 \quad \text{and} \quad \lim_{x\to1-0}\frac{\partial^2\mu(x)}{\partial x^2} = 0.$$

L^2 can be represented in the form

$$L^2 = \mu(x)\frac{\partial^2}{\partial x^2}\left(\mu(x)\frac{\partial^2}{\partial x^2}\right) = \mu(x)\left(\frac{\partial^2\mu(x)}{\partial x^2}\frac{\partial^2}{\partial x^2} + 2\frac{\partial\mu(x)}{\partial x}\frac{\partial^3}{\partial x^3} + \mu(x)\frac{\partial^4}{\partial x^4}\right).$$

So we easily obtain

$$\lim_{x\to+0} L^2 = \frac{1}{\rho^2}\frac{\partial^4}{\partial x^4} = \tilde{L}^2 \quad \text{and} \quad \lim_{x\to1-0} L^2 = \frac{1}{\rho^2}\frac{\partial^4}{\partial x^4} = \tilde{L}^2.$$

Assume that the equality (3.21) is true for all integers less than or equal to $k = K$. By using this assumption, when $k = K+1$,

$$\lim_{x\to+0} L^{K+1} = \lim_{x\to+0}\mu(x)\frac{\partial^2}{\partial x^2}\left(\mu(x)\frac{\partial^2 L^{K-1}}{\partial x^2}\right)$$
$$= \lim_{x\to+0}\mu(x)\left(\frac{\partial^2\mu(x)}{\partial x^2}\frac{\partial^2 L^{K-1}}{\partial x^2} + 2\frac{\partial\mu(x)}{\partial x}\frac{\partial^3 L^{K-1}}{\partial x^3} + \mu(x)\frac{\partial^4 L^{K-1}}{\partial x^4}\right) = \frac{1}{\rho^2}\frac{\partial^4\tilde{L}^{K-1}}{\partial x^4} = \tilde{L}^{K+1}.$$

By same way we can obtain $\lim_{x\to1-0} L^{K+1} = \tilde{L}^{K+1}$. Therefore the equality (3.21) is true for all integers.

(b) When $\rho(x) = \rho_2(x)$, we have

$$\mu(x) = \begin{cases} \dfrac{1}{\rho} + \dfrac{1}{\rho}\exp\left(-\frac{1}{x(1-x)}\right)\varphi(x), & 0 < x < 1 \\ \dfrac{1}{\rho}, & x = 0, 1 \end{cases}$$

When $k = 1$,

$$\lim_{x\to+0}\mu(x)=\lim_{x\to+0}\left(\frac{1}{\rho}+\frac{1}{\rho}\exp\left(-\tfrac{1}{x(1-x)}\right)\varphi(x)\right)=\frac{1}{\rho}$$

$$\text{and}\quad\lim_{x\to1-0}\mu(x)=\lim_{x\to1-0}\left(\frac{1}{\rho}+\frac{1}{\rho}\exp\left(-\tfrac{1}{x(1-x)}\right)\varphi(x)\right)=\frac{1}{\rho}.$$

So,

$$L=\lim_{x\to+0}\mu(x)\frac{\partial^2}{\partial x^2}=\frac{1}{\rho}\frac{\partial^2}{\partial x^2}=\tilde{L}\quad\text{and}\quad L=\lim_{x\to1-0}\mu(x)\frac{\partial^2}{\partial x^2}=\frac{1}{\rho}\frac{\partial^2}{\partial x^2}=\tilde{L}.$$

When $k=2$, the first order derivative of $\mu(x)$ is

$$\frac{\partial\mu(x)}{\partial x}=\begin{cases}\dfrac{1}{\rho}\left(\dfrac{1-2x}{[x(1-x)]^2}\varphi(x)+\dfrac{\partial\varphi(x)}{\partial x}\right)\exp\left(-\tfrac{1}{x(1-x)}\right), & 0<x<1\\ 0, & x=0,1,\end{cases}$$

and the second order derivative of $\mu(x)$ is

$$\frac{\partial^2\mu(x)}{\partial^2 x}=\frac{1}{\rho}\left(\frac{1-6x+12x^2-12x^3+6x^4}{[x(1-x)]^4}\varphi(x)+\frac{2-4x}{[x(1-x)]^2}\frac{\partial\varphi(x)}{\partial x}\right)\exp\left(-\tfrac{1}{x(1-x)}\right)$$
$$+\frac{1}{\rho}\frac{\partial^2\varphi(x)}{\partial^2 x}\exp\left(-\tfrac{1}{x(1-x)}\right),\qquad 0<x<1.$$

By using that

$$\lim_{t\to+\infty}P(t)\exp(-t)=0$$

where $P(t)$ is any polynomial of n order, we obtain

$$\lim_{x\to+0}\frac{\partial\mu(x)}{\partial x}=0\quad\text{and}\quad\lim_{x\to1-0}\frac{\partial\mu(x)}{\partial x}=0,$$

$$\lim_{x\to+0}\frac{\partial^2\mu(x)}{\partial x^2}=0\quad\text{and}\quad\lim_{x\to1-0}\frac{\partial^2\mu(x)}{\partial x^2}=0.$$

L^2 can be represented in the form

$$L^2=\mu(x)\frac{\partial^2}{\partial x^2}\left(\mu(x)\frac{\partial^2}{\partial x^2}\right)=\mu(x)\left(\frac{\partial^2\mu(x)}{\partial x^2}\frac{\partial^2}{\partial x^2}+2\frac{\partial\mu(x)}{\partial x}\frac{\partial^3}{\partial x^3}+\mu(x)\frac{\partial^4}{\partial x^4}\right).$$

So we easily obtain

$$\lim_{x\to+0}L^2=\frac{1}{\rho^2}\frac{\partial^4}{\partial x^4}=\tilde{L}^2\quad\text{and}\quad\lim_{x\to1-0}L^2=\frac{1}{\rho^2}\frac{\partial^4}{\partial x^4}=\tilde{L}^2.$$

Assume that the equality (3.21) is true for all integers less than or equal to $k=K$. By using this assumption, when $k=K+1$,

$$\lim_{x\to+0}L^{K+1}=\lim_{x\to+0}\mu(x)\frac{\partial^2}{\partial x^2}\left(\mu(x)\frac{\partial^2L^{K-1}}{\partial x^2}\right)$$
$$=\lim_{x\to+0}\mu(x)\left(\frac{\partial^2\mu(x)}{\partial x^2}\frac{\partial^2L^{K-1}}{\partial x^2}+2\frac{\partial\mu(x)}{\partial x}\frac{\partial^3L^{K-1}}{\partial x^3}+\mu(x)\frac{\partial^4L^{K-1}}{\partial x^4}\right)=\frac{1}{\rho^2}\frac{\partial^4\tilde{L}^{K-1}}{\partial x^4}=\tilde{L}^{K+1}.$$

By same way we can obtain $\lim_{x\to1-0}L^{K+1}=\tilde{L}^{K+1}$. Therefore the equality (3.21) is true for all integers.

From the equality (3.21) and the k-th order compatibility condition about $\tilde{L}$:

$$\tilde{L}^k\alpha(0) = \frac{\partial^{2k}}{\partial t^{2k}}f(0), \quad \tilde{L}^k\alpha(1) = \frac{\partial^{2k}}{\partial t^{2k}}g(0), \quad \tilde{L}^k\beta(0) = \frac{\partial^{2k+1}}{\partial t^{2k+1}}f(0), \quad \tilde{L}^k\beta(1) = \frac{\partial^{2k+1}}{\partial t^{2k+1}}g(0),$$

It follows that the k-th order compatibility condition is satisfied as follows:

$$L^k\alpha(0) = \frac{\partial^{2k}}{\partial t^{2k}}f(0), \quad L^k\alpha(1) = \frac{\partial^{2k}}{\partial t^{2k}}g(0), \quad L^k\beta(0) = \frac{\partial^{2k+1}}{\partial t^{2k+1}}f(0), \quad L^k\beta(1) = \frac{\partial^{2k+1}}{\partial t^{2k+1}}g(0). \quad \blacksquare$$

3.3 Discretization of Problem 1

Problem 1 is discretized by using the Chebyshev-Gauss-Lobatto collocation method then we obtain the following discretized problem

$$\begin{cases} \dfrac{4}{T^2}\rho_i \displaystyle\sum_{l=0}^{N_t}(D_{tt})_{jl}u_{il} - 4\sum_{k=0}^{N_x}(D_{xx})_{ik}u_{kj} = 0, & i = 1, ..., N_x - 1, \quad j = 1, ..., N_t - 1 \\ u_{iN_t} - \alpha_i = 0, & i = 1, ..., N_x - 1 \\ \dfrac{2}{T}\displaystyle\sum_{l=0}^{N_t}(D_t)_{N_t l}u_{il} - \beta_i = 0, & i = 1, ..., N_x - 1 \\ u_{N_x j} - f_j = 0, & j = 0, ..., N_t \\ u_{0j} - g_j = 0, & j = 0, ..., N_t \end{cases}$$

where

$$\begin{array}{ll} u_{ij} = u(x_i, t_j),\ x_i = \dfrac{1}{2}\left(1 + \cos\frac{i\pi}{N}\right),\ t_j = \dfrac{T}{2}\left(1 + \cos\frac{j\pi}{N}\right), & 0 \leqq i \leqq N_x, 0 \leqq j \leqq N_t \\ \rho_i = \rho(x_i),\ \alpha_i = \alpha(x_i),\ \beta_i = \beta(x_i), & 1 \leqq i \leqq N_x - 1, \\ f_j = f(t_j),\ g_j = g(t_j), & 0 \leqq j \leqq N_t, \end{array}$$

$(D_t)_{j,k}, (D_{tt})_{j,k}$ and $(D_{xx})_{j,k}$ are first and second derivative matrices,

N_x and N_t are approximation orders in x and t, respectively.

3.4 Numerical results

In this section numerical results for Examples 3.1–3.5 are shown. For the simplicity we set $N = N_x = N_t$ where N_x and N_t denote the approximation order in SCM about x and t.

The following two types of errors are used to evaluate numerical results. One is Err which is defined as follows:

$$Err = \max_{0 \leq i,j \leq N} |u_{i,j} - u_{i,j}^N|$$

where $u_{i,j}$ and $u_{i,j}^N$ are values of the exact solution and the numerical solution for the approximation order N, respectively, at the point (x_i, t_j)

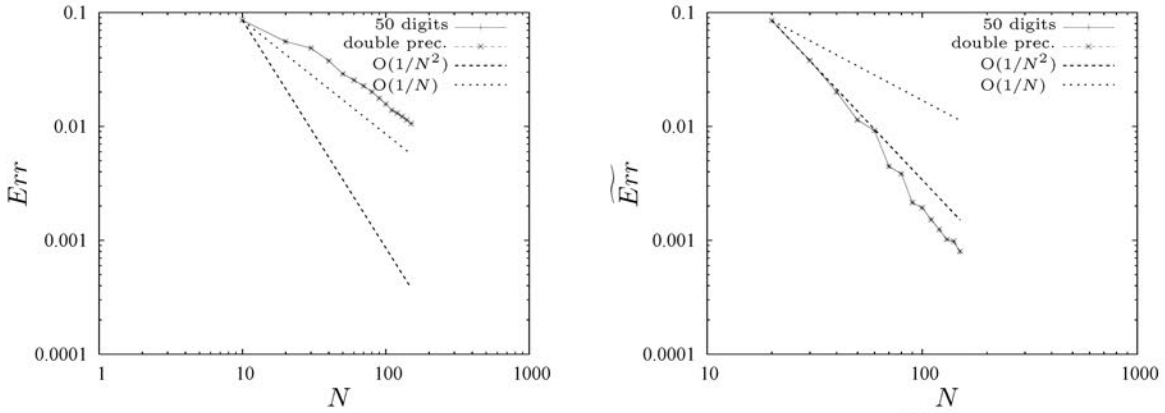
If the exact solution is given in infinite series, then $u_{i,j}$ is calculated by using a partial sum with the truncation error which is small enough.

Another one is $\widetilde{Err}$ which is defined by using numerical solutions as follows:

$$\widetilde{Err} = \max_{0 \leq i,j \leq 100} |\tilde{u}_{i,j}^{N_1} - \tilde{u}_{i,j}^{N_2}|, \qquad N_1 = N, \quad N_2 = N_1 + 10$$

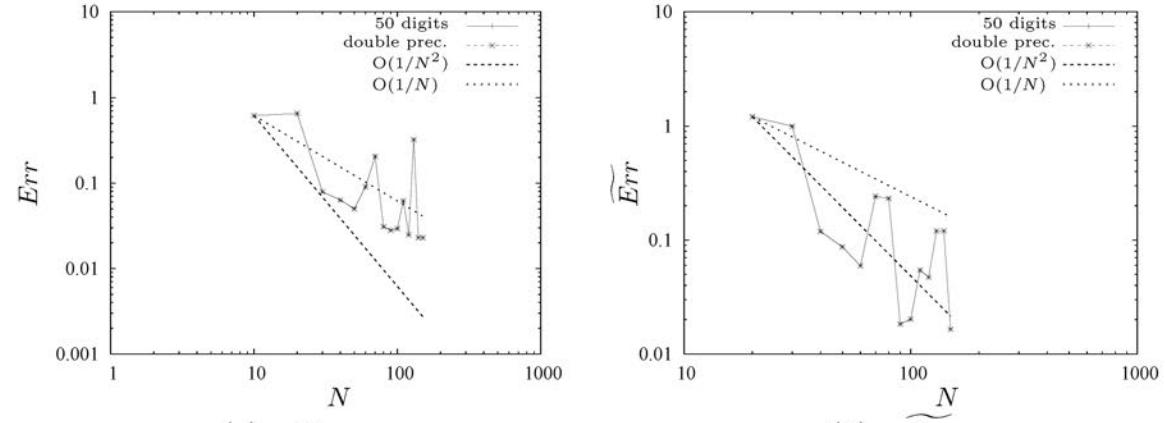
where $\tilde{u}^M_{i,j} = \tilde{u}^M(\tilde{x}_i, \tilde{t}_j) \quad (M = N_1, N_2), \quad \tilde{x}_i = i/100, \quad \tilde{t}_j = jT/100$ and $\tilde{u}^M_{i,j}$ is the interpolated function for $\{u^M_{i,j}\}_{i,j}$ which are values of the numerical solution at (x_i, t_j).

Figs. 3.1.1 and 3.1.2 show numerical results for Example 3.1. The first order ($k = 1$) compatibility condition is not satisfied. Err in Figs. 3.1.1(a) and 3.1.2(a) decays as $\mathrm{O}(N^{-1})$. $\widetilde{Err}$ in Figs. 3.1.1(b) and 3.1.2(b) decays as $\mathrm{O}(N^{-2})$.



(a) Err (b) $\widetilde{Err}$

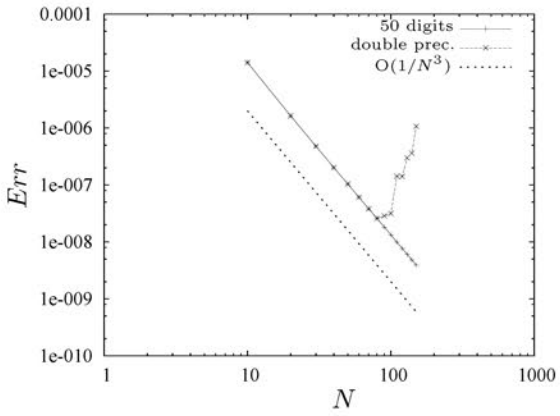
Fig. 3.1.1 Err and $\widetilde{Err}$ in Example 4.1($T = \pi$, log-log scale)



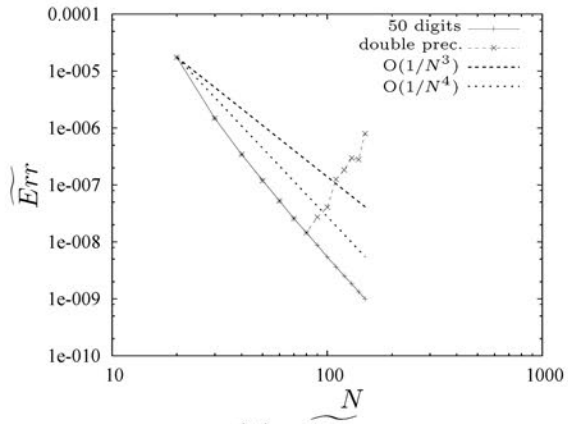
(a) Err (b) $\widetilde{Err}$

Fig. 3.1.2. Err and $\widetilde{Err}$ in Example 4.1($T = 2\pi$, log-log scale).

Figs. 3.2.1 and 3.2.2 show numerical results for Example 3.2. The first order ($k = 1$) compatibility condition is not satisfied. Double precision is not enough. Err in Figs. 3.2.1(a) and 3.2.2(a) decays as $\mathrm{O}(N^{-3})$. $\widetilde{Err}$ in Figs. 3.2.1(b) and 3.2.2(b) decays as $\mathrm{O}(N^{-4})$.

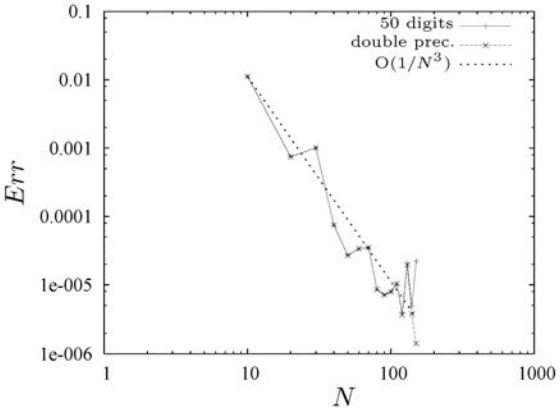


(a) Err

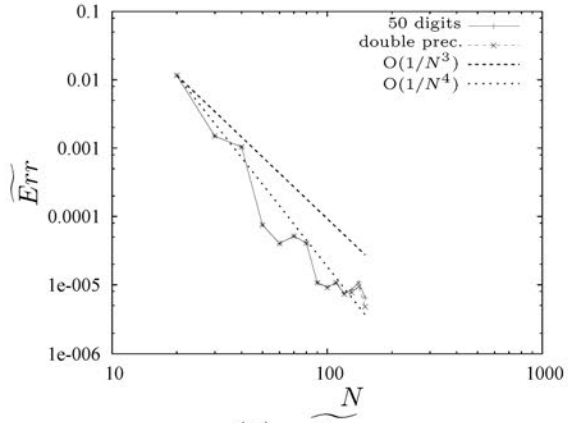


(b) $\widetilde{Err}$

Fig. 3.2.1. Err and $\widetilde{Err}$ in Example 3.2($T = 1$, log-log scale).



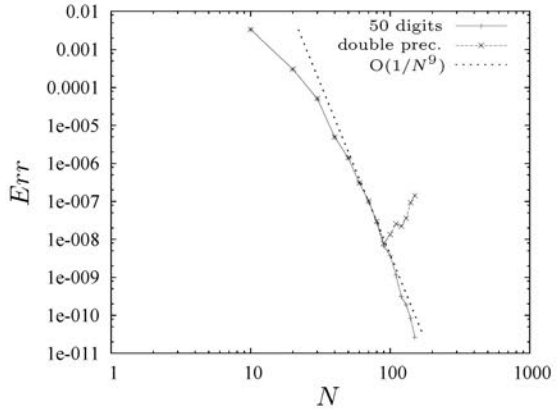
(a) Err



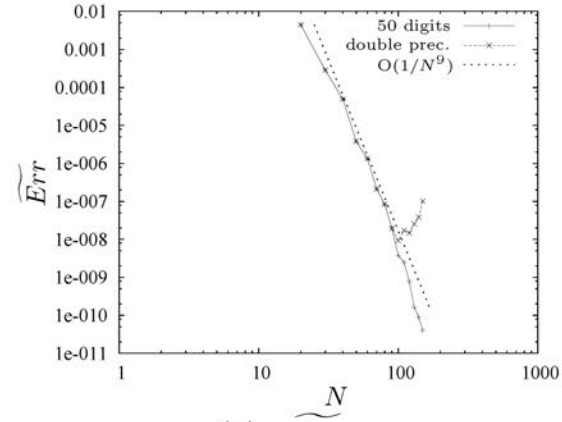
(b) $\widetilde{Err}$

Fig. 3.2.2. Err and $\widetilde{Err}$ in Example 3.2($T = 2$, log-log scale).

Figs. 3.3.1.1∼3.3.2.2 show numerical results for Example 3.3. For any non-negative integer k the k-th order compatibility condition is satisfied. Double precision is not enough. Err in Fig. 3.3.1.1(a) converges superlinearly. Taking into consideration with Fig. 3.3.1.2(a) Err decays almost exponentially. $\widetilde{Err}$ in Figs. 3.3.1.1(b) and 3.3.1.2(b) shows the same behavior. On the other hand, behaviors of Err and $\widetilde{Err}$ in Figs. 3.3.2.1 and 3.3.2.2 are not clear.

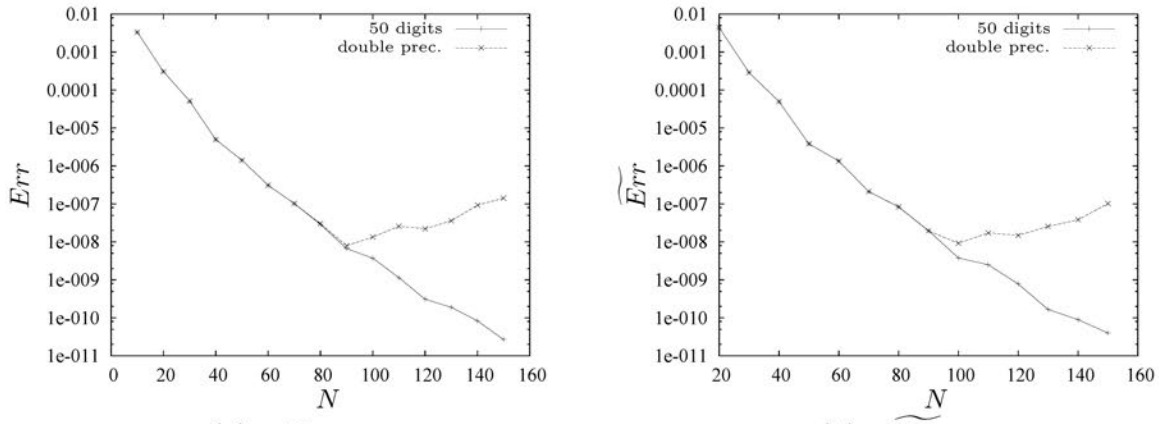


(a) Err



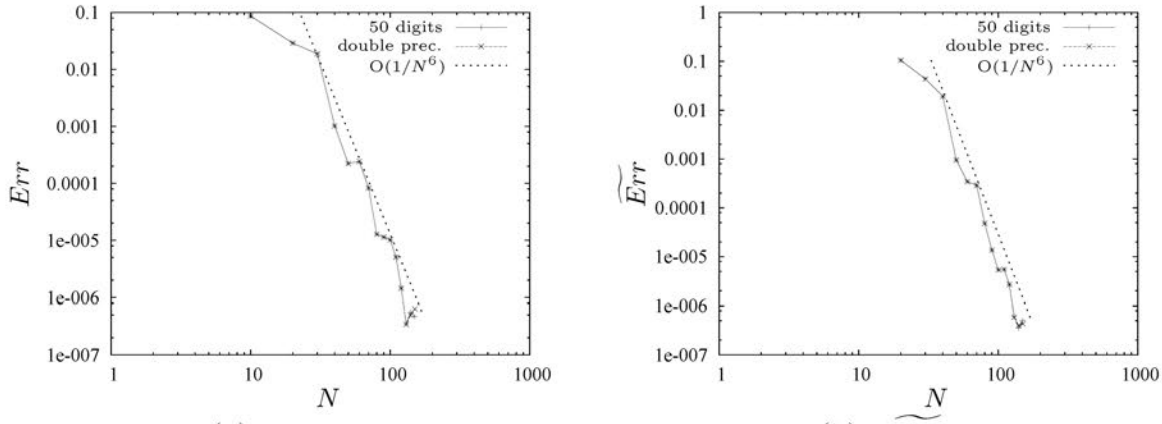
(b) $\widetilde{Err}$

Fig. 3.3.1.1. Err and $\widetilde{Err}$ in Example 3.3($T = 1$, $\rho = 1, \varepsilon = 1$, log-log scale).



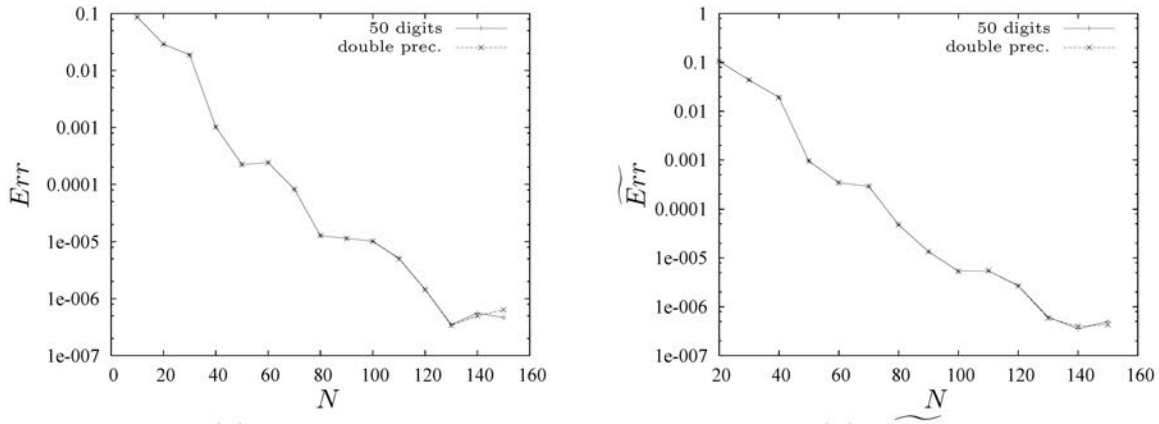
(a) Err (b) $\widetilde{Err}$

Fig. 3.3.1.2. Err and $\widetilde{Err}$ in Example 3.3($T = 1$, $\rho = 1, \varepsilon = 1$, semi-log scale).



(a) Err (b) $\widetilde{Err}$

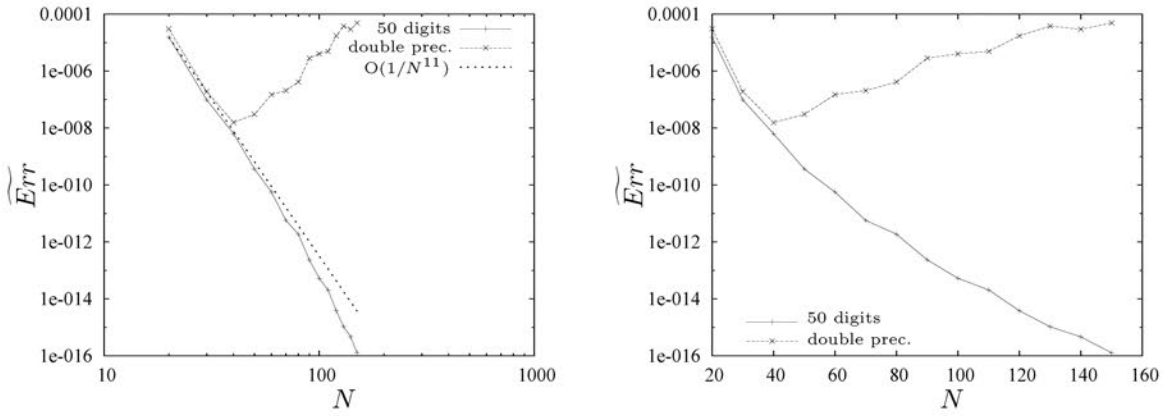
Fig. 3.3.2.1. Err and $\widetilde{Err}$ in Example 3.3($T = 2$, $\rho = 1, \varepsilon = 1$, log-log scale).



(a) Err (b) $\widetilde{Err}$

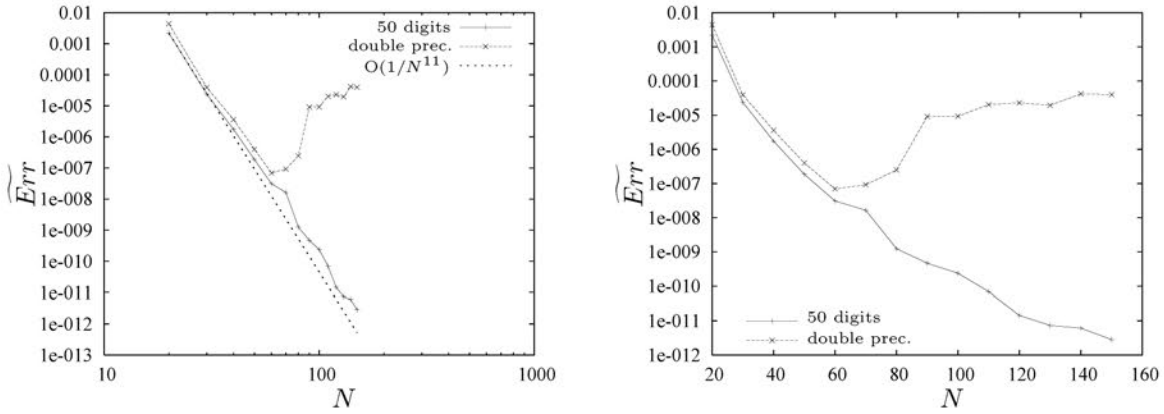
Fig. 3.3.2.2. Err and $\widetilde{Err}$ in Example 3.3($T = 2$, $\rho = 1, \varepsilon = 1$, semi-log scale).

Figs. 3.4.1 and 3.4.2 show numerical results for Example 3.4. For any non-negative integer k the k-th order compatibility condition is satisfied. Double precision is not enough. By a similar argument to Fig. 3.4.1 $\widetilde{Err}$ decays almost exponentially. On the other hand, behaviors of $\widetilde{Err}$ in Fig. 3.4.2 are not clear.



(a) log-log scale (b) semi-log scale

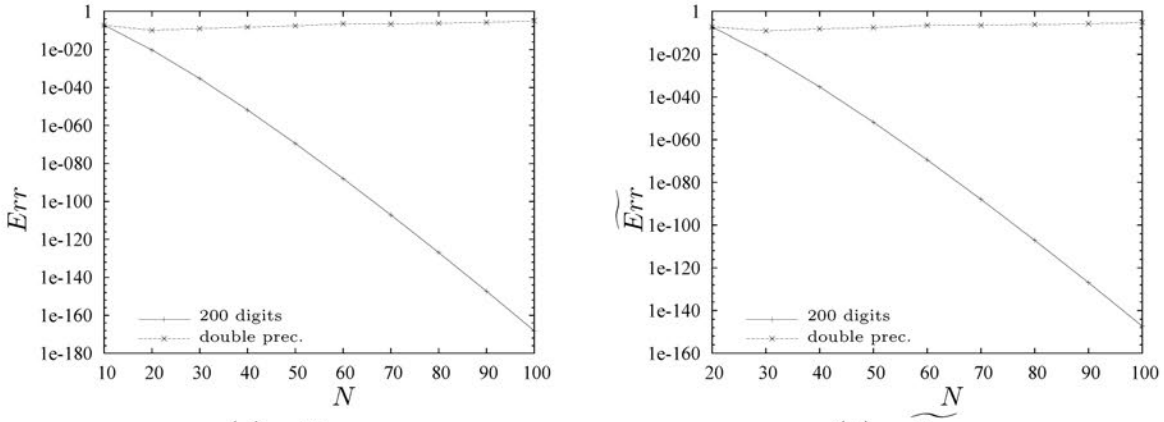
Fig. 3.4.1. $\widetilde{Err}$ in Example 3.4($T = 1$).



(a) log-log scale (b) semi-log scale

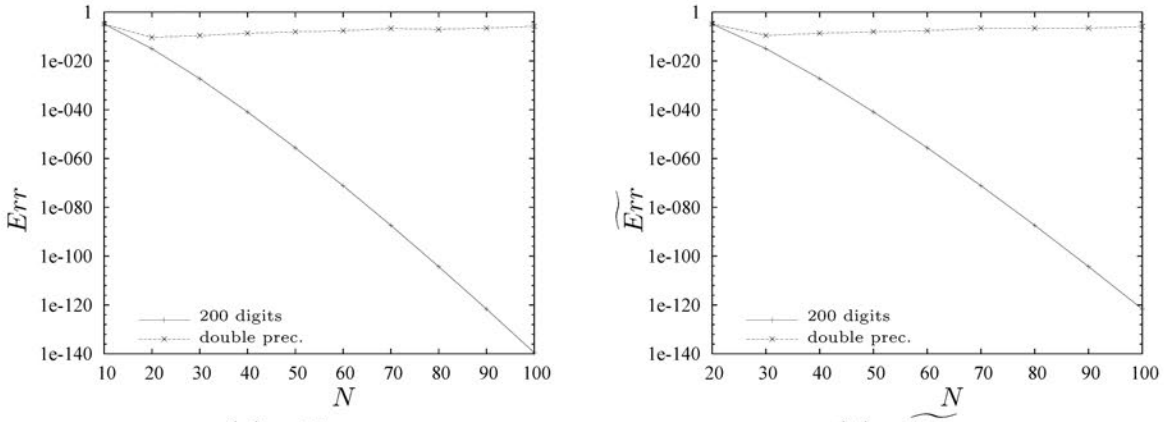
Fig. 3.4.2. $\widetilde{Err}$ in Example 3.4($T = 2$).

Figs. 3.5.1 and 3.5.2 shows numerical results for Example 3.5. For any non-negative integer k the k-th order compatibility condition is satisfied. The exact solution is analytic. Double precision is not enough. Err and $\widetilde{Err}$ decay exponentially. This is often seen when the solution is analytic.



(a) Err (b) $\widetilde{Err}$

Fig. 3.5.1. Err and $\widetilde{Err}$ in Example 3.5($T = 1$, $\rho = 1$, semi-log scale).



(a) Err (b) $\widetilde{Err}$

Fig. 3.5.2. Err and $\widetilde{Err}$ in Example 3.5($T = 2$, $\rho = 1$, semi-log scale).

In all Examples numerical solutions become worse in accuracy and sometimes convergence rate as T increases. This shows that numerical simulation is not so easy for global solutions of hyperbolic equations.

Numerical results show that the smoothness of the solution can be determined numerically by the convergence order of errors. Behaviors of $\widetilde{Err}$ or Err in Example 3.3(the exact solution is smooth but not analytic) and in Example 3.4(the exact solution may be smooth) are almost same(zigzag). On the other hand, those in Example 3.5(the exact solution is analytic) are quite different from those in Example 3.3(or 3.4). This suggests the possibility of the numerical distinction between the smooth solution and the analytic solution.

To check the compatibility condition is difficult if $\rho(x)$ is not constant. This means that theoretical investigation about the regularity (the smoothness) of the solution is difficult. However, by using our numerical results it is easy to know the regularity(the smoothness) of the solution.

4 Our inverse problem and numerical results

4.1 Our inverse problem(Problem 2) and the nonuniqueness of the solution

We consider the following an inverse problem:

Problem 2. For a positive numebr T, find $u(x,t)$ and $\rho(x)$ such that

$$\begin{cases} \rho(x)u_{tt}(x,t) - u_{xx}(x,t) = 0, & x \in (0,1), \quad t \in (0,T) \\ u(x,0) = \alpha(x), \quad u_t(x,0) = \beta(x), & x \in (0,1) \\ u(0,t) = f(t), \quad u(1,t) = g(t), & t \in [0,T] \\ u_x(0,t) = h(t), & t \in (0,T) \end{cases}$$

where ρ is a strictly positive function on $(0,1)$, $\alpha(x), \beta(x), f(t), g(t)$ and $h(t)$ are all known functions.

Remark 4.1. If $\alpha(x) = x, \beta(x) = 1, f(t) = t, g(t) = t+1$, and $h(t) = 1$, then $\forall\rho(x)$, $u(x,t) = x + t$ become solutions. Thus, uniqueness does not hold in this case.

Problem 2 is a nonlinear inverse problem so solving it numerically is difficult. If the nonuniqueness of the solution holds, it will be even more difficult. This is because when the nonuniqueness holds, the coefficient matrix in the discretized problem may be rank deficit.

We shall investigate numerically the nonuniqueness of a solution as for Problem 2. The use of multiple-precision arithmetic is necessary for the numerical computation of inverse problems. The numerical computation of rank deficit of matrices is performed in multiple precision. A simple criterion for determining rank deficit is proposed. Our criterion is applied to some examples, including an example derived from Problem 2[4].

4.2 Our criterion for rank deficit

We used SVD, GECP and GECPE for determining the rank of a matrix. The rank of a matrix is determined by the number of nonzero singular values in SVD or the number of nonzero diagonal elements in the upper triangular matrix in GECP or GECPE.

A natural(theoretical) criterion for determining rank deficit of an $M \times M$ matrix A is

$$RD(A) = \lim_{L\to\infty} \sum_{\substack{i=1 \\ |c_i| \leqq 10^{-sL}}}^{M} 1,$$

where L denotes the number of digits in multiple precision, s is a proper constant, and c_i is the i-th diagonal element in GECP or GECPE or the i-th singular value in SVD.

Our criterion is as follows: we set $s = 9/10$, and we consider the numerical rank deficit:

$$NRD(A;L) = \sum_{\substack{i=1 \\ |c_i| \leqq 10^{-\frac{9}{10}L}}}^{M} 1.$$

For sufficiently large N, $NRD(A;L)$ may be regarded as $RD(A)$.

4.3 Discretization of Problem 2 and application of the Newton method

For Problem 2, we consider the following discretization, obtained by using the spectral collocation method (SCM)(Fig. 4.1).

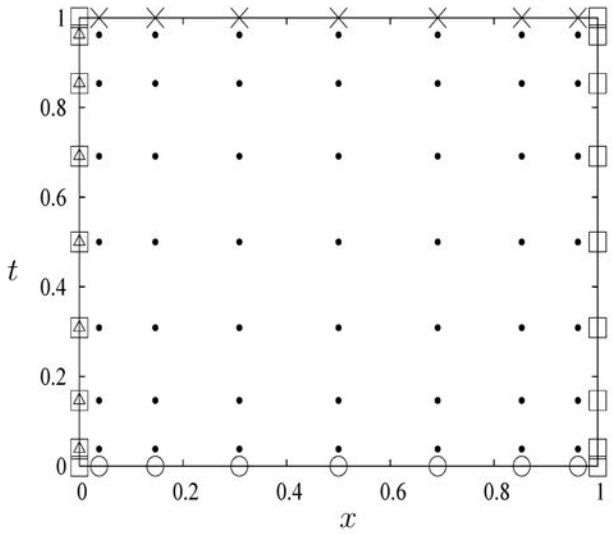


Fig. 4.1. Discretization by SCM for $N = 8$.
(•: hyperbolic equation; □: Dirichlet data; ◯: Cauchy data; △: Neumann data; ×: no condition).

Discretized equations are obtained as follows:

$$\begin{cases} \dfrac{4}{T^2}\rho_i \displaystyle\sum_{l=0}^{N}(D_{tt})_{jl}u_{il} - 4\sum_{k=0}^{N}(D_{xx})_{ik}u_{kj} = 0, & i,j = 1,2,\cdots,N-1, \\ u_{Nj} - f_j = 0, & j = 0,1,\cdots,N, \\ u_{0j} - g_j = 0, & j = 0,1,\cdots,N, \\ u_{iN} - \alpha_i = 0, & i = 1,2,\cdots,N-1, \\ \dfrac{2}{T}\displaystyle\sum_{l=0}^{N}(D_t)_{Nl}u_{il} - \beta_i = 0, & i = 1,2,\cdots,N-1, \\ 2\displaystyle\sum_{k=0}^{N}(D_x)_{Nk}u_{kj} - h_j = 0, & j = 1,2,\cdots,N-1, \end{cases}$$

where

$$\begin{array}{ll} u_{ij} = u(x_i,t_j),\ x_i = \dfrac{1}{2}\left(1+\cos\frac{i\pi}{N}\right),\ t_j = \dfrac{T}{2}\left(1+\cos\frac{j\pi}{N}\right), & 0 \leqq i,j \leqq N, \\ \rho_i = \rho(x_i),\ \alpha_i = \alpha(x_i),\ \beta_i = \beta(x_i), & 1 \leqq i \leqq N-1, \\ f_j = f(t_j),\ g_j = g(t_j), & 0 \leqq j \leqq N, \\ h_j = h(t_j), & 1 \leqq j \leqq N-1. \end{array}$$

They are nonlinear, so we cannot solve them directly. We apply Newton's method, as follows:

$$v^{n+1} = v^n - [F'(v^n)]^{-1} F(v^n), \quad n = 0, 1, \cdots,$$

$$v^n = \begin{pmatrix} u^n_{0,0} \\ u^n_{1,0} \\ \vdots \\ u^n_{N,N} \\ \rho^n_1 \\ \vdots \\ \rho^n_{N-1} \end{pmatrix}, \quad F(v) = \begin{pmatrix} F_1(v) \\ \vdots \\ F_m(v) \\ \vdots \\ F_M \end{pmatrix}, \quad F'(v) = \begin{pmatrix} \frac{\partial F_1}{\partial v_1} & \cdots & \frac{\partial F_1}{\partial v_M} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_M}{\partial v_1} & \cdots & \frac{\partial F_M}{\partial v_M} \end{pmatrix},$$

$$F_m(v) = \begin{cases} \frac{4}{T^2}\rho_i \sum_{l=0}^{N} (D_{tt})_{jl} u_{il} - 4 \sum_{k=0}^{N} (D_{xx})_{ik} u_{kj}, & m = (j-1)(N-1) + i, \\ & i, j = 1, 2, \cdots, N-1, \\ u_{Nj} - f_j, & m = (N-1)^2 + j, \\ & j = 0, 1, \cdots, N, \\ u_{0j} - g_j, & m = (N^2 - N + 2) + j, \\ & j = 0, 1, \cdots, N, \\ u_{iN} - \alpha_i, & m = N^2 + 3 + i, \\ & i = 1, 2, \cdots, N-1, \\ \frac{2}{T} \sum_{l=0}^{N} (D_t)_{Nl} u_{il} - \beta_i, & m = N^2 + N + 2 + j, \\ & j = 1, 2, \cdots, N-1, \\ 2 \sum_{k=0}^{N} (D_x)_{Nk} u_{kj} - h_j, & m = (N+1)^2 + j, \\ & j = 1, 2, \cdots, N-1, \end{cases}$$

where $M = N(N+3)$. $F'(v)$ is an $M \times M$ square matrix. In Newton's method, we need an initial guess v^0.

4.4 Numerical results

First we consider three simple examples involving the 5×5 matrices A and B. A is a given upper triangular matrix, so the diagonal elements of A are the eigenvalues of A and the number of nonzero diagonal elements is the rank of A. B is obtained from A by applying the elementary operations of row addition, and row and column switching. These operations are the same in all examples. It is well known that elementary operations do not change the rank. SVD, GECP, and GECPE are applied to B. They are performed in both double precision and multiple precision[13].

Example 4.1.

$$A = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ 0 & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ 0 & 0 & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ 0 & 0 & 0 & 10^{-20} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \Longrightarrow \quad B = \begin{pmatrix} \frac{1}{3} & \frac{1}{5} & \frac{1}{4} & 0 & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{5} + \frac{1}{6} & \frac{1}{4} + \frac{1}{5} & 0 & \frac{1}{6} + \frac{1}{7} \\ 0 & 10^{-20} & 0 & 0 & 0 \\ 0 & \frac{1}{6} & \frac{1}{5} & 0 & \frac{1}{7} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{3} & 1 & \frac{1}{5} \end{pmatrix}$$

Obviously, $RD(A) = RD(B) = 1$. Numerical results for Example 4.1 are shown in Tables 4.1.1–4.1.3. From Tables 4.1.1–4.1.3, double precision is not enough. 50 digits or 100 digits are enough. $NRD(B;50) = NRD(B;100) = 1$ in either of GECP, GECPE and SVD. Therefore we may determine $RD(B) = 1$.

Table 4.1.1 Diagonal elements (DE) and singular values (SV) (double precision(L=16)).

	DE in GECPE	DE in GECP	SV in SVD
1	1.0000$\cdots$e+000	1.0000$\cdots$e+000	1.3492$\cdots$e+000
2	1.0000$\cdots$e+000	4.5000$\cdots$e−001	6.9520$\cdots$e−001
3	1.0000$\cdots$e+000	-1.4815$\cdots$e−001	1.8291$\cdots$e−001
4	1.0000$\cdots$e+000	-5.5511$\cdots$e−017	1.6588$\cdots$e−017
5	1.1102$\cdots$e−016	1.0000$\cdots$e−020	5.8905$\cdots$e−019
	$NRD(B;16) = 1$	$NRD(B;16) = 2$	$NRD(B;16) = 2$

Table 4.1.2. Diagonal elements (DE) and singular values (SV) (50 digits).

	DE in GECPE	DE in GECP	SV in SVD
1	1.0000$\cdots$e+000	1.0000$\cdots$e+000	1.3492$\cdots$e+000
2	1.0000$\cdots$e+000	4.5000$\cdots$e−001	6.9520$\cdots$e−001
3	1.0000$\cdots$e+000	-1.4815$\cdots$e−001	1.8291$\cdots$e−001
4	1.0000$\cdots$e+000	1.2500$\cdots$e−001	8.2760$\cdots$e−021
5	-3.8234$\cdots$e−078	-1.9914$\cdots$e−059	1.1746$\cdots$e−059
	$NRD(B;50) = 1$	$NRD(B;50) = 1$	$NRD(B;50) = 1$

Table 4.1.3. Diagonal elements (DE) and singular values (SV) (100 digits).

	DE in GECPE	DE in GECP	SV in SVD
1	1.0000$\cdots$e+000	1.0000$\cdots$e+000	1.3492$\cdots$e+000
2	1.0000$\cdots$e+000	4.5000$\cdots$e−001	6.9520$\cdots$e−001
3	1.0000$\cdots$e+000	-1.4815$\cdots$e−001	1.8290$\cdots$e−001
4	1.0000$\cdots$e+000	1.2500$\cdots$e−001	8.2760$\cdots$e−021
5	-1.2690$\cdots$e−116	-3.1724$\cdots$e−117	6.9997$\cdots$e−117
	$NRD(B;100) = 1$	$NRD(B;100) = 1$	$NRD(B;100) = 1$

Example 4.2.

$$A = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ 0 & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ 0 & 0 & 2\cdot 10^{-20} & 0 & 0 \\ 0 & 0 & 0 & 10^{-20} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \Longrightarrow \quad B = \begin{pmatrix} \frac{1}{3} & \frac{1}{5} & \frac{1}{4} & 0 & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{5} & \frac{1}{4}+2\cdot 10^{-20} & 0 & \frac{1}{6} \\ 0 & 10^{-20} & 0 & 0 & 0 \\ 0 & 0 & 2\cdot 10^{-20} & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{3} & 1 & \frac{1}{5} \end{pmatrix}$$

Obviously, $RD(A) = RD(B) = 1$. Numerical results for Example 4.2 are shown in Tables 4.2.1–4.2.3. From Tables 4.2.1–4.2.3, double precision is not enough. 50 digits or 100 digits are enough. $NRD(B;50) = NRD(B;100) = 1$ in either of GECP, GECPE and SVD. Therefore we may determine $RD(B) = 1$.

Table 4.2.1. Diagonal elements (DE) and singular values (SV) (double precision(L=16)).

	DE in GECPE	DE in GECP	SV in SVD
1	1.0000$\cdots$e+000	1.0000$\cdots$e+000	1.2859$\cdots$e+000
2	1.0000$\cdots$e+000	3.3333$\cdots$e$-$001	5.4125$\cdots$e$-$001
3	1.0000$\cdots$e+000	2.0000$\cdots$e$-$020	1.6080$\cdots$e$-$018
4	1.0000$\cdots$e+000	1.0000$\cdots$e$-$020	9.0566$\cdots$e$-$020
5	0.0000$\cdots$e+000	0.0000$\cdots$e+000	1.1919$\cdots$e$-$020
	$NRD(B;16)=1$	$NRD(B;16)=3$	$NRD(B;16)=3$

Table 4.2.2. Diagonal elements (DE) and singular values (SV) (50 digits).

	DE in GECPE	DE in GECP	SV in SVD
1	1.0000$\cdots$e+000	1.0000$\cdots$e+000	1.2859$\cdots$e+000
2	1.0000$\cdots$e+000	3.3333$\cdots$e$-$001	5.4125$\cdots$e$-$001
3	1.0000$\cdots$e+000	2.0000$\cdots$e$-$020	2.1250$\cdots$e$-$020
4	1.0000$\cdots$e+000	1.0000$\cdots$e$-$020	8.7378$\cdots$e$-$021
5	0.0000$\cdots$e+000	0.0000$\cdots$e+000	2.3958$\cdots$e$-$059
	$NRD(B;50)=1$	$NRD(B;50)=1$	$NRD(B;50)=1$

Table 4.2.3. Diagonal elements (DE) and singular values (SV) (100 digits).

	DE in GECPE	DE in GECP	SV in SVD
1	1.0000$\cdots$e+000	1.0000$\cdots$e+000	1.2859$\cdots$e+000
2	1.0000$\cdots$e+000	3.3333$\cdots$e$-$001	5.4125$\cdots$e$-$001
3	1.0000$\cdots$e+000	2.0000$\cdots$e$-$020	2.1250$\cdots$e$-$020
4	1.0000$\cdots$e+000	1.0000$\cdots$e$-$020	8.7378$\cdots$e$-$021
5	0.0000$\cdots$e+000	0.0000$\cdots$e+000	3.1746$\cdots$e$-$118
	$NRD(B;100)=1$	$NRD(B;100)=1$	$NRD(B;100)=1$

Example 4.3.

$$A=\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ 0 & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ 0 & 0 & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \Longrightarrow \quad B=\begin{pmatrix} \frac{1}{3} & \frac{1}{5} & \frac{1}{4} & 0 & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{5}+\frac{1}{6} & \frac{1}{4}+\frac{1}{5} & 0 & \frac{1}{6}+\frac{1}{7} \\ \frac{1}{2}+\frac{1}{3} & \frac{1}{4}+\frac{1}{5} & \frac{1}{3}+\frac{1}{4} & 1 & \frac{1}{5}+\frac{1}{6} \\ 0 & \frac{1}{6} & \frac{1}{5} & 0 & \frac{1}{7} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{3} & 1 & \frac{1}{5} \end{pmatrix}$$

Obviously, $RD(A)=RD(B)=2$. Numerical results for Example 4.3 are shown in Tables 4.3.1–4.3.3. From Tables 4.3.1–4.3.3, double precision or 50 digits or 100 digits are enough. $NRD(B;16)=NRD(B;50)=NRD(B;100)=2$ in either of GECP, GECPE and SVD. Therefore we may determine $RD(B)=2$.

Table 4.3.1. Diagonal elements (DE) and singular values (SV) (double precision(L=16)).

	DE in GECPE	DE in GECP	SV in SVD
1	1.0000$\cdots$e+000	1.0000$\cdots$e+000	2.0457$\cdots$e+000
2	1.0000$\cdots$e+000	4.5000$\cdots$e−001	6.9640$\cdots$e−001
3	1.0000$\cdots$e+000	1.4815$\cdots$e−001	1.9664$\cdots$e−001
4	-1.1102$\cdots$e−016	-5.5511$\cdots$e−017	7.4868$\cdots$e−017
5	-5.5511$\cdots$e−017	4.3368$\cdots$e−019	3.0880$\cdots$e−017
	$NRD(B;16) = 2$	$NRD(B;16) = 2$	$NRD(B;16) = 2$

Table 4.3.2. Diagonal elements (DE) and singular values (SV) (50 digits).

	DE in GECPE	DE in GECP	SV in SVD
1	1.0000$\cdots$e+000	1.0000$\cdots$e+000	2.0457$\cdots$e+000
2	1.0000$\cdots$e+000	4.5000$\cdots$e−001	6.9640$\cdots$e−001
3	1.0000$\cdots$e+000	1.4815$\cdots$e−001	1.9664$\cdots$e−001
4	-7.9655$\cdots$e−059	-3.9827$\cdots$e−059	8.1417$\cdots$e−059
5	-3.9827$\cdots$e−059	-1.9914$\cdots$e−059	4.1058$\cdots$e−059
	$NRD(B;50) = 2$	$NRD(B;50) = 2$	$NRD(B;50) = 2$

Table 4.3.3. Diagonal elements (DE) and singular values (SV) (100 digits).

	DE in GECPE	DE in GECP	SV in SVD
1	1.0000$\cdots$e+000	1.0000$\cdots$e+000	2.0457$\cdots$e+000
2	1.0000$\cdots$e+000	4.5000$\cdots$e−001	6.9640$\cdots$e−001
3	1.0000$\cdots$e+000	1.4815$\cdots$e−001	1.9664$\cdots$e−001
4	1.2098$\cdots$e−116	-9.4181$\cdots$e−117	1.1770$\cdots$e−116
5	6.3449$\cdots$e−117	-3.1724$\cdots$e−117	4.0701$\cdots$e−117
	$NRD(B;100) = 2$	$NRD(B;100) = 2$	$NRD(B;100) = 2$

Example 4.4 is the case in Remark 4.1, drived from Problem 2. We consider it for the nonuniqueness of the solution. Problem 2 is an inverse problem, so numerical computations in double precision are omitted.

Example 4.4.
$F'(v^0)$, $u^0_{ij} = x_i + t_j$ $(0 \leqq i, j \leqq N)$, $\rho^0_i = x_i$ $(1 \leqq i \leqq N-1), \alpha_i = x_i, \beta_i = 1$ $(1 \leqq i \leqq N-1)$, $f_j = t_j, g_j = t_j + 1$ $(0 \leqq j \leqq N)$, $h_j = 1$ $(1 \leqq j \leqq N-1)$. SVD, GECP, and GECPE are applied to $F'(v^0)$, and they are performed in multiple precision.

Theoretically, $\sum_{l=0}^{N} (D_{tt})_{jl} u^0_{il}$ is zero in $F'(v^0)$. This is because $u_{tt}(x,t) = 0$ for $u(x,t) = x + t$ in Remark 4.1. So, $RD(F'(v^0)) \geq N - 1$. The results are shown in Tables 4.4.1 and 4.4.2. $NRD(F'(v^0); 50) = NRD(F'(v^0); 100) = 9 (= N - 1)$ in GECP and SVD. On the other hand, $NRD(F'(v^0); 50) = NRD(F'(v^0); 100) = 0$ in GECPE. These results show that GECPE does not work well, which is due to the equilibration in GECPE. $\sum_{l=0}^{N} (D_{tt})_{jl} u^0_{il}$ in $F'(v^0)$, which is an approximation of $u_{tt}(x_i, t_j) (= 0)$, is as small as a rounding error. However, equilibration enlarges it to $O(1)$. This spoils the results of GECPE.

Table 4.4.1. Diagonal elements (DE) and singular values (SV) (N=10, 50 digits).

	DE in GECPE	DE in GECP	SV in SVD
1	1.0000$\cdots$e+000	-1.6850$\cdots$e+003	2.4093$\cdots$e+003
$\vdots$	$\vdots$	$\vdots$	$\vdots$
120	-3.6703$\cdots$e−001	6.7964$\cdots$e−001	1.2595$\cdots$e−001
121	3.6091$\cdots$e−001	-3.1883$\cdots$e−002	2.0038$\cdots$e−002
122	3.1520$\cdots$e−001	8.1338$\cdots$e−057	8.0292$\cdots$e−057
$\vdots$	$\vdots$	$\vdots$	$\vdots$
130	-1.7248$\cdots$e−003	-6.8415$\cdots$e−060	6.7414$\cdots$e−060
	$NRD(F'(v^0);50) = 0$	$NRD(F'(v^0);50) = 9$	$NRD(F'(v^0);50) = 9$

Table 4.4.2. Diagonal elements (DE) and singular values (SV) (N=10, 100 digits).

	DE in GECPE	DE in GECP	SV in SVD
1	1.0000$\cdots$e+000	-1.6850$\cdots$e+003	2.4093$\cdots$e+003
$\vdots$	$\vdots$	$\vdots$	$\vdots$
120	1.8482$\cdots$e−001	-6.1686$\cdots$e−001	1.2595$\cdots$e−001
121	-1.8942$\cdots$e−001	-2.4825$\cdots$e−002	2.0038$\cdots$e−002
122	-1.4152$\cdots$e−001	-9.8656$\cdots$e−115	1.7578$\cdots$e−114
$\vdots$	$\vdots$	$\vdots$	$\vdots$
130	-5.1602$\cdots$e−004	-6.8917$\cdots$e−118	4.4327$\cdots$e−117
	$NRD(F'(v^0);100) = 0$	$NRD(F'(v^0);100) = 9$	$NRD(F'(v^0);100) = 9$

In examples 4.1–4.3 rank deficit of matrices can be known by using our numerical rank deficit. Thus our criterion works well. They also showed that both SVD and GECP work well, GECPE does not(we found this).

We can expect the rank deficit of the coefficient matrix in the discretized problem when the nonuniqueness holds. In Example 4.4, numerical results show that our expectation is right.

5 Conclusion

In this dissertation an inverse problem governed by a one-dimensional hyperbolic equation is focused on. Some related topics are considered and investigated numerically.

First, we consider direct problems. The existence of the unique solution to a direct problem governed by a one-dimensional hyperbolic equation on the semi-axis is discussed. Then, a direct problem defined on the finite interval(Problem 1) is solved numerically. IPNS is used to obtain accurate data of Problem 1. IPNS can reduce the truncation error and the rounding error arbitrarily if the solution is smooth. Accurate data is important in the development of methods for solving inverse problems. It is used to check the validity of the method. Several examples concerned with Problem 1 are considered and solved numerically. Numerical results are satisfactory in accuracy. They are also successful for showing the smoothness of the solution. This means that the smoothness of the solution can be determined numerically by the convergence order of errors. Moreover, the smoothness or analyticity of the solution can be determined numerically by means of behaviors of errors. Behaviors of $\widetilde{Err}$ or Err in Example 3.3(the exact solution is smooth but not analytic) and in Example 3.4(the exact solution may be smooth) are almost same(zigzag). On the other hand, those in Example 3.5(the exact solution is analytic) are quite different from those in Example 3.3(or 3.4). To check the compatibility condition for direct problems is difficult if $\rho(x)$ is not constant. This means that theoretical investigation about the regularity (the smoothness) of the solution is difficult. However, by using our results it is easy to know the regularity(the smoothness) of the solution.

Secondly, we focus on an inverse problem(Problem 2). Inverse problems are usually ill-posed and difficult to solve numerically. Moreover, the uniqueness of the solution does not always hold. This makes them even more difficult. This is because when the nonuniqueness holds, the coefficient matrix in the discretized problem may be rank deficit. To overcome this difficulty we considere the numerical computation of rank deficit of matrices in multiple precision and propose a simple criterion for determining rank deficit. Three famous solvers GECP, GECPE, and SVD for a linear system are compared. These solvers and our criterion are applied to some examples. Numerical results for Examples 4.1–4.3 show that our criterion works well. This means that rank deficit of matrices can be known by using our numerical rank deficit. Numerical results for Example 4.4 show that our expectation is right. Thus, the nonuniqueness of the solution to inverse problems may be determined numerically by using our criterion. Our criterion is also useful for the investigation of the suitable choice of the regularization method. Numerical results for Examples 4.1–4.4 show that both SVD and GECP work well, GECPE does not.

Numerical methods developed here are useful to obtain basic and important information for solving inverse problems.

6 Acknowledgement

I would like to express my deepest gratitude to my supervisor, Prof. Hitoshi IMAI for enabling me to study for a doctoral degree at Graduate School of Advanced Technology and Science, The University of Tokushima. His rigoruos guidance has served me well and has helped to keep me focusing on the task at hand. I would also like to express my grateful appreciation to Assi. Prof. Hideo SAKAGUCHI and Dr. Naoki WADA for their kind coorperation and constructive suggestions during the study.

I would like to thank Prof. Toshiki TAKEUCHI for his suggestions. I would like to thank to Ms. Yuko FUJII for her help and encouragement.

Finally, I would like to thank the Mongolian State Universtity of Education and the Government of Mongolia for the scholarship throughout my study in Japan.

References

[1] S.A. Avdonin, B.P. Belinsky and J.V. Matthews, Inverse problem on the semi-axis: Local approach, Tamkang Journal of Mathematics, 42(3) (2011), pp. 275-293.

[2] S.A. Avdonin and V. Mikhaylov, Boundary control approach to inverse spectral theory, Inverse Problems, 26 (2010), pp. 1-19.

[3] S.A. Avdonin, V. Mikhaylov and A. Rybkin, The boundary control approach to the Titchmarsh-Weyl m-function, Comm. Math. Phys., 275 (2007), pp. 791-803.

[4] E. Azjargal, N. Wada, H. Imai and H. Sakaguchi, Numerical computation for smoothness of the solution of a one-dimensional hyperbolic equation, "Proceeding of the sixth international conference on science and mathematics education in developing countries", The Mathematical Society of Myanmar, Myingyan offset, Yangon, 2013, pp. 78-87.

[5] E. Azjargal, H. Imai and H. Sakaguchi, On numerical computation of rank deficit of matrices in multiple precision, Advances in Mathematical Sciences and Applications, 23(2) (2013), to appear.

[6] A. S. Blagoveshchenskii, The local nonstationary inverse problem method for an inhomogeneous string, Tr. Mat. Inst. Akad. Nauk SSSR, 115, 28-33(1971).

[7] M. Belishev, On an approach to multidimensional inverse problems for the wave equation. English translation: J. Soviet Math., 36(3) (1988), pp. 481-484.

[8] M. Belishev and A. Kachalov, The Boundary control method in the spectral inverse problem for inhomogeneous string, English translation: J. Soviet Math., 57(3) (1991), pp. 3072-3077.

[9] M. Belishev, Boundary control and inverse problems: The one-dimansional variant of the BC-method, J. Math. Sci, 155(3) (2008), pp. 343-378.

[10] M. Belishev and T. L. Sheronova, The boundary control method in dynamical inverse problem for inhomogeneous string, J. Math. Sci.,73(3) (1995), pp. 320-329.

[11] C. Canuto, M. Y. Hussaini, A. Quarteroni and T.A. Zang, "Spectral Methods in Fluid Dynamics", Springer-Verlag, New York, 1988.

[12] N. Flyer and N. Swarztrauber, The convergence of spectral and finite difference method for initial-boundary value problems, SIAM J. Sci. Comput., 23(5) (1992), pp. 1731-1751.

[13] H. Fujiwara and Y. Iso, Design of a multiple-precision arithmetic package for a 64-bit computing environment and its application to numerical computation of ill-posed problems, IPSJJ., 44(3) (2003), pp. 925-931.

[14] H. Fujiwara and Y. Iso, Some remarks on the choice of regularization parameters under multiple-precision arithmetic, Theoretical and Applied Mechanics Japan, 51 (2002), pp. 387–393.

[15] G. H. Golub and C. F. V. Loan, "Matrix Computations, 4th edition", The Johns Hopkins University Press, Baltimore, 2013.

[16] M. Ikawa, Mixed problems for hyperbolic equations of second order, J. Math. Soc. Japan, 20 (1968), pp. 580-608.

[17] H. Imai, T. Takeuchi and M. Kushida, On Numerical Simulation of Partial Differential Equations in Infinite Precision, Advances in Mathematical Sciences and Applications, 9(2) (1999), pp. 1007-1019.

[18] H. Imai, T. Takeuchi, M. Nakamura and N. Ishimura, A direct approach to an inverse problem, *GAKUTO Internat. Ser. Math. Sci. Appl.*, 12, pp. 223-232, 1999.

[19] H. Imai, H. Sakaguchi and Y. Iso, Numerical simulation on non-existence and non-uniqueness of solutions for the Tricomi equation, *GAKUTO Internat. Ser. Math. Sci. Appl.*, 34, pp. 39-58, 2011.

[20] Imai, H. and Sakaguchi, H., Numerical continuation for the Laplace equation with higher order regularization, *GAKUTO Internat. Ser. Math. Sci. Appl.*, 32, pp.131–144, 2010.

[21] S. Kabanikhin, Definitions and examples of inverse and ill-posed problems, J. Inv. Ill-Posed problems, 61 (2008), pp. 317-357.

[22] D. R. Kincaid and E. W. Cheney, Numerical Analysis: Mathematics of Scientific Computing, 3rd edition, American Mathematical Society, 2001.

[23] D. E. Knuth, The Art of Computer Programming, Addison-Wesley, Boston, 1997.

[24] M. G. Krein, A method of effective solution of the inverse boundary-value problem, Dokl. Akad. Nauk SSSR, 94(6) (1954), 13-16.

[25] H. Lewy, An example of a smooth linear partial differential equation without solution, Annals of Math., 66 (1957), pp. 155-158.

[26] S.McKEE, T.TANG and T.DIOGO, An Euler-type method for two-dimensional Volterra integral equations of the first kind, IMA Jour. of Num. Ana., 20(3) (2000), pp. 423-440.

[27] J. Rauchi and F. Massey, Differentiability of solution to hyperbolic initial-boundary value problems, Trans. Amer. Math. Soc., 189 (1974), pp. 303-318.

[28] Rainer Kress, "Numerical Analysis", Graduate Texts in Mathematics, 181 (1997).

[29] J. Stoer and R. E.Bulirsch, "Introduction to numerical analysis", Springer-Verlag, New York Heidelberg Berlin (1980).

[30] Tarmizi, Numerical Simulation of Free Boundary Problems in Infinite Precision. Doctoral thesis, The University of Tokushima, 2000.

[31] A. N. Tikhonov and V. Y. Arsenin, "Solutions of Ill-Posed Problems", John Wiley, New York, 1977.

[32] V.S.Vladimirov, "Equations of Mathematical physics, English transl. of 1st ed.", Marcel Dekker, New York, 1971