Discrete Artificial Boundary Conditions for the Korteweg-de Vries Equation

Christophe Besse, Matthias Ehrhardt, Ingrid Lacroix-Violet

Discrete Artificial Boundary Conditions for the Korteweg-de Vries Equation

C. Besse
Institut de Mathématiques de Toulouse UMR5219,
Université de Toulouse; CNRS,
UPS IMT, F-31062 Toulouse Cedex 9, France.
Christophe.Besse@math.univ-toulouse.fr

M. Ehrhardt
Bergische Universität Wuppertal,
Fachbereich Mathematik und Naturwissenschaften,
Angewandte Mathematik - Numerische Analysis,
Gaußstrasse 20, 42119 Wuppertal, Germany.
ehrhardt@math.uni-wuppertal.de

I. Lacroix-Violet
Laboratoire Paul Painlevé, Université Lille Nord de France,
CNRS UMR 8524, INRIA MEPHYSTO Team,
Université Lille 1 Sciences et Technologies, Cité Scientifique,
59655 Villeneuve d'Ascq Cedex, France.
Ingrid.Violet@math.univ-lille1.fr

April 2, 2015

1 Introduction

Korteweg-de Vries (KdV) equations are typical dispersive nonlinear partial differential equations (PDEs). Zabusky and Kruskal [24] observed that KdV equation owns wave-like solutions which can retain their initial forms after collision with another wave. This led them to name these solitary wave solutions "solitons".

These special solutions were observed and investigated for the first time in 1834 by Scott Russell [11, 18]. Later in 1895, Korteweg and de Vries [14] showed that the soliton could be expressed as a solution of a rather simple one-dimensional nonlinear PDE describing small amplitude waves in a narrow and shallow channel of water [1]:

$$\frac{\partial \eta}{\partial \tau} = \frac{3}{2}\frac{\partial}{\partial \xi}\left(\frac{1}{2}\eta^2 + \frac{2}{3}\alpha\eta + \frac{1}{3}\sigma\frac{\partial^2 \eta}{\partial \xi^2}\right), \qquad \sigma = \frac{1}{3}h^3 - \frac{Th}{\rho g}, \qquad \tau \in \mathbb{R}^+, \quad \xi \in \mathbb{R}, \quad (1.1)$$

where α is some constant, g denotes the gravitational constant, ρ is the density, T the surface tension and $\eta = \eta(\xi, \tau)$ denotes the surface displacement of the wave above the

undisturbed water level h. The equation (1.1) can be written in non-dimensional, simplified form by the transformation [1]:

$$t = \frac{1}{2}\sqrt{\frac{gh\sigma}{\tau}}, \qquad x = -\frac{\xi}{\sqrt{\sigma}}, \qquad u = \frac{1}{2}\eta + \frac{1}{3}\alpha \tag{1.2}$$

to obtain the usual KdV equation

$$u_t + 6uu_x + u_{xxx} = 0, \qquad t \in \mathbb{R}^+, \quad x \in \mathbb{R}, \tag{1.3}$$

(subscripts x and t denoting partial differentiations), with the soliton solution is given by

$$u(x,t) = \gamma \operatorname{sech}^2\big(\beta(x - ct)\big), \qquad t \in \mathbb{R}^+, \quad x \in \mathbb{R}. \tag{1.4}$$

The KdV equation (1.3) has a broad range of applications [12]: description of the asymptotic behaviour of small- but finite-amplitude shallow-water waves [14], hydromagnetic waves in a cold plasma, ion-acoustic waves [21], interfacial electrohydrodynamics [13], internal wave in the coastal ocean [17], water wave power stations [7], acoustic waves in an anharmonic crystal [25], or pressure pulse propagation in blood vessels [15].

In this paper, we focus on the *linearized KdV equation* in one space dimension

$$u_t + U_1 u_x + U_2 u_{xxx} = h(t,x), \qquad t \in \mathbb{R}^+, \quad x \in \mathbb{R}, \tag{1.5}$$

where h stands for a source term and U_1 and U_2 are real constants such that $U_1 \geq 0$ and $U_2 > 0$. Recall that for $U_1 = 0$ and $U_2 = 1$ we recover the case considered by Zheng, Wen & Han [28]. Although the PDE (1.5) looks very simple, it has a lot of applications, e.g. Whitham [22] used it for the modelling of the propagation of long waves in the shallow water equations, see also [23].

We emphasize the fact that the restriction of the solution to equation (1.5) to a finite interval is not periodic. Thus concerning the numerical simulation, we cannot use the FFT method and we consider instead the equation set on an interval and supplemented with specially designed boundary conditions. Since the linear PDE (1.5) is defined on an unbounded domain, one has to confine the unbounded domain in a numerical finite computational domain for simulation. A common used method in such situation consists in reducing the computational domain by introducing *artificial boundary conditions*. Such artificial boundary conditions are constructed with the goal to approximate the exact solution on the whole domain restricted to the computational one. They are called *absorbing boundary conditions (ABCs)* if they lead to a well-posed initial boundary value problem where some energy is absorbed at the boundary. If the approximate solution coincides on the computational domain with the exact solution on the whole domain, they are called *transparent boundary conditions (TBCs)*. See [2] for a review on the techniques used to construct such transparent or artificial boundary conditions for the Schrödinger equation.

The linearity property of equation (1.5) allows to use many analytical tools such as the Laplace transform. Using this tool, Zheng, Wen & Han [28] derived the exact TBCs for equation (1.5) at fixed boundary points and then obtained an initial boundary value problem "equivalent" to the problem in the whole space domain. Moreover, using a dual Petrov-Galerkin scheme [20] the authors proposed a numerical approximation of this initial boundary value problem. Thus the derivation in [28] of the adapted boundary conditions is carried out at the continuous level and then discretized afterwards. Recently, Zhang,

Li and Wu [26] revisited the approach of Zheng, Wen & Han [28] and proposed a fast approximation of the exact TBCs based on Padé approximation of the Laplace-transformed TBCs.

In this paper we will follow a different strategy: we first discretize the equation (1.5) with respect to time and space and then derive the suitable artificial boundary conditions for the fully discrete problem using the $\mathcal{Z}$-transformation. The goal of this paper is therefore to derive analogous conditions of the transparent boundary conditions obtained by the authors in [28] but in the fully discrete case. These discrete artificial boundary conditions are superior since they are by construction perfectly adapted to the used interior scheme and thus retain the stability properties of the underlying discretization method and theoretically do not produce any reflections when compared to the discrete whole space solution. However, there will be some small errors induced by the numerical root finding routine and the numerical inverse $\mathcal{Z}$-transformation and also later due to the fast sum-of-exponentials approximation. Let us finally remark that there exists also an alternative approach in this "discrete spirit", namely to use discrete multiple scales, following the work of Schoombie [19].

The paper is organized as follow. In Section 2 we use the ideas of Zheng, Wen & Han [28] to obtain the TBCs for the linearized KdV equation (1.5) and we briefly recall the results given in [28] for the special case $U_1 = 0$ and $U_2 = 1$. In Section 3 we present an appropriate space and time discretization and explain the procedure to derive the artificial boundary conditions for the purely discrete problem mimicking the ideas presented in Section 2. Since exact ABCs are too time-consuming, especially for higher dimensional problems, we propose in Section 4 to use a sum-of-exponentials approach [4], to speed up the (approximate) computation of the discrete convolutions at the boundaries. Finally, in Section 5 we present some numerical benchmark examples from the literature to illustrate our findings.

2 Transparent boundary conditions for the continuous case

The motivation for this section is twofold. First, we briefly recall from the literature the construction of TBCs for the 1D linearized KdV equation (1.5) for the special case $U_1 = 0$ and $U_2 = 1$ and the well-posedness of the resulting initial boundary value problem [28].

Secondly, we extend the derivation of TBCs to the generalized case $U_1 \geq 0$ and $U_2 > 0$; these results will serve us as a guideline for the completely discrete case in Section 3.

To do so, we consider the Cauchy problem

$$u_t + U_1 u_x + U_2 u_{xxx} = h(t,x), \qquad t \in \mathbb{R}^+, \quad x \in \mathbb{R}, \tag{2.1}$$

$$u(0,x) = u_0(x), \qquad x \in \mathbb{R}, \tag{2.2}$$

$$u \to 0, \qquad x \to \pm\infty, \tag{2.3}$$

where (for simplicity) the initial function u_0 and the source term h are assumed to be compactly supported in a finite computational interval $[a,b]$, with $a < b$ and where $U_1 \geq 0$ and $U_2 > 0$ are given constants. For the construction of TBCs in the case of non-compactly supported initial data we refer the interested reader to [10].

The construction of (continuous) artificial boundary conditions associated to problem (2.1)–(2.3) is established by considering the problem on the complementary of $[a,b]$, *i.e.*

$$u_t + U_1 u_x + U_2 u_{xxx} = 0, \qquad t \in \mathbb{R}^+, \quad x < a \quad \text{or} \quad x > b, \tag{2.4}$$

$$u(0,x) = 0, \qquad x < a \quad \text{or} \quad x > b, \tag{2.5}$$

$$u \to 0, \quad x \to \pm\infty. \tag{2.6}$$

Denoting by $\widehat{u} = \widehat{u}(s,x)$ the Laplace transform in time of the function $u = u(t,x)$, we obtain from (2.4) the *transformed exterior problems*

$$s\widehat{u} + U_1 \widehat{u}_x + U_2 \widehat{u}_{xxx} = 0, \qquad x < a \quad \text{or} \quad x > b, \tag{2.7}$$

$$\widehat{u} \to 0, \quad x \to \pm\infty, \tag{2.8}$$

where $s \in \mathbb{C}$, with $\mathrm{Re}(s) > 0$, stands for the argument of the transformation, i.e. the dual time variable. The general solutions of the ODE (2.7) are given explicitly by

$$\widehat{u}(s,x) = c_1(s)\, e^{\lambda_1(s)x} + c_2(s)\, e^{\lambda_2(s)x} + c_3(s)\, e^{\lambda_3(s)x}, \qquad x < a \quad \text{or} \quad x > b, \tag{2.9}$$

where $\lambda_1(s)$, $\lambda_2(s)$, $\lambda_3(s)$ denote the roots of the (depressed) cubic equation

$$s + U_1 \lambda + U_2 \lambda^3 = 0. \tag{2.10}$$

The three solutions are given by

$$\lambda_1(s) = \zeta(s) - \frac{1}{2}\frac{U_1}{U_2}\frac{1}{\zeta(s)}, \quad \lambda_2(s) = \omega\zeta(s) - \frac{1}{2}\frac{U_1}{U_2}\frac{1}{\omega\zeta(s)}, \quad \lambda_3(s) = \omega^2\zeta(s) - \frac{1}{2}\frac{U_1}{U_2}\frac{1}{\omega^2\zeta(s)}, \tag{2.11}$$

where $\omega = \exp(2i\pi/3)$ and

$$\zeta(s) = -\frac{1}{2^{1/3}}\left(\frac{s}{U_2} + \sqrt{\frac{s}{U_2} + \frac{4}{27}\left(\frac{U_1}{U_2}\right)^3}\right)^{1/3}.$$

We have strong numerical evidences (see Figure 1) that the roots of the cubic equation (2.10) possess the following *separation property*

$$\mathrm{Re}(\lambda_1(s)) < 0, \qquad \mathrm{Re}(\lambda_2(s)) > 0, \qquad \mathrm{Re}(\lambda_3(s)) > 0, \tag{2.12}$$

which is crucial for defining later the TBCs; the *separation* property allows to separate the fundamental solutions into outgoing and incoming waves.

Remark 2.1. *Considering, as in [28], the case $U_1 = 0$ and $U_2 = 1$ we have*

$$\lambda_1(s) = -\sqrt[3]{s}, \qquad \lambda_2(s) = -\sqrt[3]{s}\omega, \qquad \lambda_3(s) = -\sqrt[3]{s}\omega^2, \quad \textit{with} \quad \omega = e^{2i\pi/3}$$

and then it can be easily verified that the separation property (2.12) *is satisfied.*

Now using the decay condition (2.8), the general solution (2.9), the separation property (2.12) and since solutions of (2.7) have to belong to $L^2(\mathbb{R})$, we obtain

$$c_1(s) = 0 \quad \text{for} \quad x \le a, \qquad c_2(s) = c_3(s) = 0 \quad \text{for} \quad x \ge b, \tag{2.13}$$

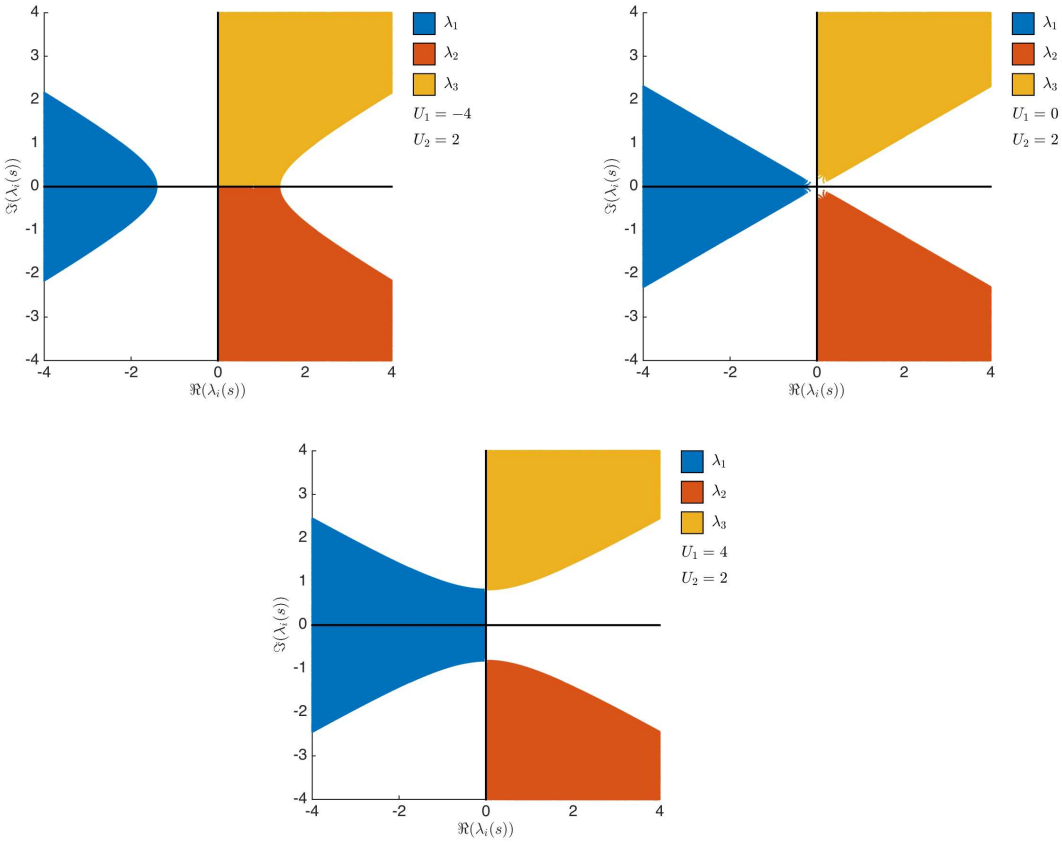

Figure 1: Roots $\lambda_1(s)$, $\lambda_2(s)$, $\lambda_3(s)$ to the cubic equation (2.10) for $U_1 = -4,\ 0,\ 4$ and $U_2 = 2$, where $s = r\,e^{i\phi}$, $r \in]0, 500]$, $\phi \in]-\pi/2, \pi/2[$.

which yields the following TBCs in the Laplace-transformed space

$$\widehat{u}_{xx}(s,a) - \big(\lambda_2(s) + \lambda_3(s)\big)\,\widehat{u}_x(s,a) + \lambda_2(s)\lambda_3(s)\,\widehat{u}(s,a) = 0, \tag{2.14}$$

$$\widehat{u}(s,b) - \frac{1}{\lambda_1^2(s)}\,\widehat{u}_{xx}(s,b) = 0, \qquad \widehat{u}_x(s,b) - \frac{1}{\lambda_1(s)}\,\widehat{u}_{xx}(s,b) = 0. \tag{2.15}$$

Since λ_1, λ_2 and λ_3 are roots of the cubic equation (2.10) we obtain immediately

$$\lambda_2(s)\lambda_3(s) = -\frac{s}{U_2\lambda_1(s)} \qquad \text{and} \qquad \lambda_2(s) + \lambda_3(s) = -\lambda_1(s),$$

and hence the transformed left TBC (2.14) can be rewritten solely in terms of $\lambda_1(s)$

$$\widehat{u}(s,a) - \frac{U_2\lambda_1(s)^2}{s}\,\widehat{u}_x(s,a) - \frac{U_2\lambda_1(s)}{s}\,\widehat{u}(s,a) = 0. \tag{2.16}$$

Now applying the inverse Laplace transform to equations (2.16) and (2.15) we get

$$u(t,a) - U_2\,\mathcal{L}^{-1}\left(\frac{\lambda_1(s)^2}{s}\right) * u_x(t,a) - U_2\,\mathcal{L}^{-1}\left(\frac{\lambda_1(s)}{s}\right) * u_{xx}(t,a) = 0, \tag{2.17}$$

$$u(t,b) - \mathcal{L}^{-1}\left(\frac{1}{\lambda_1(s)^2}\right) * u_{xx}(t,b) = 0, \qquad u_x(t,b) - \mathcal{L}^{-1}\left(\frac{1}{\lambda_1(s)}\right) * u_{xx}(t,b) = 0, \tag{2.18}$$

where $\mathcal{L}^{-1}(f(s))$ stands for the inverse Laplace transform of f and $*$ denotes the convolution operator. We emphasize that those boundary conditions strongly depend on U_1 and U_2 through the root $\lambda_1(s)$.

Remark 2.2. *Considering, as in [28], the special case $U_1 = 0$ and $U_2 = 1$ we easily obtain from* (2.17)–(2.18)

$$u(t,a) - I_t^{1/3} u_x(t,a) + I_t^{2/3} u_{xx}(t,a) = 0, \tag{2.19}$$

$$u(t,b) - I_t^{2/3} u_{xx}(t,b) = 0, \qquad u_x(t,b) + I_t^{1/3} u_{xx}(t,b) = 0, \tag{2.20}$$

where I_t^p with $p > 0$ is the nonlocal-in-time fractional integral operator given by the Riemann-Liouville formula

$$I_t^p f(t) = \frac{1}{\Gamma(p)} \int_0^t (t-\tau)^{p-1} f(\tau)\, d\tau,$$

where $\Gamma(z) = \int_0^{+\infty} e^{-t} t^{z-1}\, dt$ is the Gamma function. We refer to [28] for more details.

To summarize our findings so far, the derived initial boundary value problem reads

$$u_t + U_1 u_x + U_2 u_{xxx} = 0, \qquad t \in \mathbb{R}^+,\ x \in [a,b], \tag{2.21}$$

$$u(0,x) = u_0(x), \qquad x \in [a,b], \tag{2.22}$$

$$u(t,a) - U_2\, \mathcal{L}^{-1}\Big(\frac{\lambda_1(s)^2}{s}\Big) * u_x(t,a) - U_2\, \mathcal{L}^{-1}\Big(\frac{\lambda_1(s)}{s}\Big) * u_{xx}(t,a) = 0, \tag{2.23}$$

$$u(t,b) - \mathcal{L}^{-1}\Big(\frac{1}{\lambda_1(s)^2}\Big) * u_{xx}(t,b) = 0, \tag{2.24}$$

$$u_x(t,b) - \mathcal{L}^{-1}\Big(\frac{1}{\lambda_1(s)}\Big) * u_{xx}(t,b) = 0. \tag{2.25}$$

Note that a solution of (2.21)–(2.25) can be regarded as the restriction on $[a,b]$ of the solution on the whole space domain.

For the special case $U_1 = 0$ and $U_2 = 1$ (cf. Remark 2.2) the following stability theorem is shown in [28].

Theorem 2.1. *[28] The initial boundary value problem* (2.21)–(2.25) *for $U_1 = 0$ and $U_2 = 1$ is L^2-stable. More precisely, for any $t > 0$, there is a constant positive number $c(t)$ such that*

$$\int_a^b u^2(t,x)\, dx \le c(t) \Big(\int_a^b u_0^2(x)\, dx + \int_0^t \int_a^b h^2(t,x)\, dxdt \Big). \tag{2.26}$$

In the sequel we use the same procedure to obtain the artificial boundary conditions for the fully discrete case, i.e. for the discretized version of the equation (1.5). These so-called discrete artificial boundary conditions are better adapted to the numerical scheme and thus do not alter the stability properties. Also, they do not suffer from discretization errors of convolution integrals and theoretically do not produce any unphysical reflections. However, since some steps in the calculation of the convolution coefficients, like the root finding and the inverse $\mathcal{Z}$-transformation have to be done numerically, this procedure will lead to some small errors.

3 Discrete transparent boundary conditions

In this section we present how to obtain the artificial boundary conditions in the fully discrete case for the problem (2.1)–(2.3). For simplicity we focus here on the case without source term, *i.e.* we assume $h(t,x) = 0$ for all $t > 0$ and $x \in \mathbb{R}$. Moreover we consider the problem restricted to the computational interval $[a,b]$ for the finite time $t \in [0,T]$, *i.e.*

$$u_t + U_1 u_x + U_2 u_{xxx} = 0, \qquad t \in [0,T], \quad x \in [a,b], \tag{3.1}$$

$$u(0,x) = u_0(x), \qquad x \in [a,b]. \tag{3.2}$$

Let us denote by $(t_n)_{0\le n\le N}$ a uniform subdivision of the time interval $[0,T]$ given by $t_n = n\Delta t$ with the temporal step size $\Delta t = T/N$:

$$0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T.$$

We also define $(x_j)_{0\le j\le J}$ a uniform subdivision of $[a,b]$ given by $x_j = a + j\Delta x$ with the spatial step size $\Delta x = (b-a)/J$:

$$a = x_0 < x_1 < \cdots < x_{J-1} < x_J = b.$$

We emphasize here that the temporal discretization must remain uniform due to the usage of the $\mathcal{Z}$-transform to derive the discrete TBCs. On the other hand, the space discretization in the interior domain could have been non uniform. In the following, we denote by $u_j^{(n)}$ the pointwise approximation of the solution $u(t_n, x_j)$.

We will consider in the sequel two different numerical schemes based on trapezoidal rule in time (semi discrete Crank-Nicolson approximation). The first one is the *Rightside Crank-Nicolson* (proposed by Mengzhao [16]) (R-CN) scheme defined for $U_1 = 0$ and $U_2 > 0$. It reads

$$\frac{u_j^{(n+1)} - u_j^{(n)}}{\Delta t} + \frac{U_2}{2(\Delta x)^3}\left(u_{j+2}^{(n+1)} - 3u_{j+1}^{(n+1)} + 3u_j^{(n+1)} - u_{j-1}^{(n+1)}\right) + \frac{U_2}{2(\Delta x)^3}\left(u_{j+2}^{(n)} - 3u_{j+1}^{(n)} + 3u_j^{(n)} - u_{j-1}^{(n)}\right) = 0. \tag{3.3}$$

The second one is the *Centered Crank-Nicolson* (C-CN) scheme [16] which is used for the generalized linear Korteweg-de Vries equation (1.5) where $U_1 \ge 0$ and $U_2 \ge 0$. It reads

$$\begin{aligned}\frac{u_j^{(n+1)} - u_j^{(n)}}{\Delta t} &+ \frac{U_1}{4\Delta x}\left(u_{j+1}^{(n+1)} - u_{j-1}^{(n+1)}\right) + \frac{U_1}{4\Delta x}\left(u_{j+1}^{(n)} - u_{j-1}^{(n)}\right) \\ &+ \frac{U_2}{4(\Delta x)^3}\left(u_{j+2}^{(n+1)} - 2u_{j+1}^{(n+1)} + 2u_{j-1}^{(n+1)} - u_{j-2}^{(n+1)}\right) \\ &+ \frac{U_2}{4(\Delta x)^3}\left(u_{j+2}^{(n)} - 2u_{j+1}^{(n)} + 2u_{j-1}^{(n)} - u_{j-2}^{(n)}\right) = 0.\end{aligned} \tag{3.4}$$

Here, the convection term is discretized in a centered way. Indeed, using simply an *upwind Crank-Nicholson* scheme for the first order term and (R-CN) scheme for the third order term leads to a strongly dissipative scheme.

Both schemes are absolutely stable and their truncation errors are respectively

$$\begin{aligned} E_{\mathrm{R-CN}} &= O(\Delta x + \Delta t^2), \\ E_{\mathrm{C-CN}} &= O(\Delta x^2 + \Delta t^2). \end{aligned} \tag{3.5}$$

The stencil of the different scheme involves respectively 4 nodes for (R-CN) and 5 nodes for (C-CN) schemes. This structure will have a strong influence on the computation of the roots for the corresponding equation to (2.10) at the discrete level. For the (R-CN) scheme, we will recover as in the continuous case a cubic equation, but a quartic equation for the (C-CN) scheme. The later case will turn out to be more difficult.

3.1 Discrete artificial boundary conditions for (R-CN) scheme

Let us first consider the (R-CN) scheme (3.3) for the interior problem, i.e. with a spatial index j such that $1 \leq j \leq J-2$. Let us recall that this scheme is only valid in the case $U_1 = 0$ and $U_2 > 0$. For this scheme, as in the continuous case, we will obtain one boundary condition at point $x_0 = a$ and two boundary conditions at the right side which will involve the two nodes x_{J-1} and x_J, cf. the continuous ABCs (2.23)–(2.25).

In order to derive appropriate artificial boundary conditions, we follow the same procedure as in Section 2, but on a purely discrete level. First we apply the $\mathcal{Z}$-transform with respect to the time index n, which is the discrete analogue of the Laplace transform in time, to the partial difference equation (3.3). We refer the reader to the appendix of [2, 8, 9] for a proper definition of the $\mathcal{Z}$-transform and its basic properties. The standard definition reads

$$\hat{u}(z) = \mathcal{Z}\{(u^n)_n\}(z) = \sum_{k=0}^{\infty} u^k z^{-k}, \qquad |z| > R \geq 1, \tag{3.6}$$

where R is the convergence radius of the Laurent series and $z \in \mathbb{C}$.

Denoting by $\widehat{u}_j = \widehat{u}_j(z)$ the $\mathcal{Z}$-transform of the sequence $(u_j^{(n)})_{n\in\mathbb{N}_0}$ we obtain from (3.3) the homogeneous *third order difference equation*

$$\widehat{u}_{j+2} - 3\widehat{u}_{j+1} + \left(3 + \frac{2(\Delta x)^3}{U_2 \Delta t}\frac{z-1}{z+1}\right)\widehat{u}_j - \widehat{u}_{j-1} = 0, \qquad 1 \leq j \leq J-2. \tag{3.7}$$

It is well-known that homogeneous difference equations with constant coefficients possess solutions of the power form $\widehat{u}_j = \sum_k c_k(z)\,\ell_k^j(z)$, where $\ell = \ell(z)$ solves the cubic equation

$$\ell^3 - 3\ell^2 + \left(3 + \frac{2(\Delta x)^3}{U_2 \Delta t}\frac{z-1}{z+1}\right)\ell - 1 = 0. \tag{3.8}$$

Equation (3.8) admits three fundamental solutions denoted here by ℓ_1, ℓ_2 and ℓ_3 that can be computed analytically or numerically up to a very high precision. Thus the general solution of (3.7) on the exterior domains is of the form (cf. (2.9))

$$\widehat{u}_j(z) = c_1(z)\,\ell_1^j(z) + c_2(z)\,\ell_2^j(z) + c_3(z)\,\ell_3^j(z), \qquad j \leq 1 \text{ or } j \geq J-2.$$

Let

$$\alpha(z) = 3 + \frac{2(\Delta x)^3}{U_2 \Delta t}\frac{z-1}{z+1}, \quad \zeta(z) = \left(\frac{(3-\alpha(z))}{3}\left(1 + \sqrt{1 - \frac{4}{27}(3-\alpha(z))}\right)\right)^{1/3}.$$

The three solutions of (3.8) are

$$\ell_j(z) = \omega^{j-1}\zeta(z) + \frac{3-\alpha(z)}{3}\frac{1}{\omega^{j-1}\zeta(z)}, \quad j = 1, 2, 3.$$

We have strong numerical evidences (see Figure 2), that the roots are well separated according to

$$|\ell_1(z)| < 1, \qquad |\ell_2(z)| > 1, \qquad |\ell_3(z)| > 1, \qquad \text{for all } z, \tag{3.9}$$

which is the *discrete separation property* (cf. (2.12)). Like in the continuous case, the plots of Figure 2 are obtained using a polar representation $z = r\,e^{i\Phi}$ with the radius $r > 1$ and the phase $\Phi \in [0, 2\pi)$. The choice for the value of r is linked to the convergence radius for the $\mathcal{Z}$-transform (3.6). We present in Figure 2 the plots obtained for the case $r = 1.001$ (figure on the left) and for the case $r = 1.02$ (figure on the right). We clearly observe the discrete separation property (3.9): two solutions have a modulus strictly greater than one and one solution has a modulus strictly less than one.

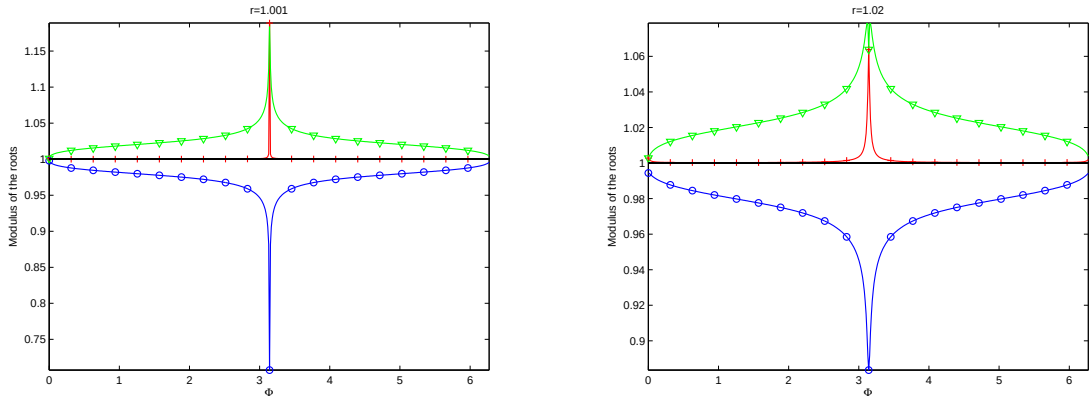

Figure 2: Numerical evidence for the discrete separation property (3.9): Modulus of the three solutions of the cubic equation (3.8) with respect to $\Phi \in [0, 2\pi)$ and with $\Delta t = 4/2560$, $\Delta x = 12/5000$ and $r = 1.001$ (figure on the left) or $r = 1.02$ (figure on the right).

Then using the decay condition we obtain, cf. (2.13)

- for the left exterior domain $c_1(z) = 0$, $j \leq 1$ and thus $\widehat{u}_j(z) = c_2(z)\,\ell_2^j(z) + c_3(z)\,\ell_3^j(z)$, $j \leq 1$
- for the right exterior domain $c_2(z) = c_3(z) = 0$, $j \geq J-2$ and thus $\widehat{u}_j(z) = c_1(z)\,\ell_1^j(z)$, $j \geq J-2$

Let us now derive the boundary conditions for (R-CN) scheme.

Left boundary. On the left boundary we only need one relation. It is easy to see that

$$\widehat{u}_{j+1}(z) - \big(\ell_2(z) + \ell_3(z)\big)\,\widehat{u}_j(z) + \ell_2(z)\ell_3(z)\,\widehat{u}_{j-1}(z) = 0. \tag{3.10}$$

Applying it for $j = 1$ and denoting by $\mathcal{Z}^{-1}\{f(z)\}$ the inverse $\mathcal{Z}$-transform of $f(z)$, we obtain

$$\mathcal{Z}^{-1}\{\ell_2(z)\ell_3(z))\} *_d u_0^{(n)} - \mathcal{Z}^{-1}\{\ell_2(z) + \ell_3(z))\} *_d u_1^{(n)} + u_2^{(n)} = 0, \quad n = 0, 1, 2 \ldots, \tag{3.11}$$

where $*_d$ stands for the discrete convolution with respect to the temporal index n:

$$P *_d u_i^{(n)} = \sum_{k=0}^{n} P^{(k)} u_i^{(n-k)},$$

for $P = (P^k)$ a sequence and i an integer. Let us denote the convolution kernels by $k_{1,R}(z) = \ell_2(z) + \ell_3(z)$ and $k_{2,R}(z) = \ell_2(z)\ell_3(z)$ and by $Y_{i,R}$ the sequences of the inverse $\mathcal{Z}$-transform of kernel $k_{i,R}$ *i.e.* $Y_{i,R} = \mathcal{Z}^{-1}\{k_{i,R}(z)\}$. Then (3.11) can be written

$$Y_{2,R} *_d u_0^{(n)} - Y_{1,R} *_d u_1^{(n)} + u_2^{(n)} = 0, \quad n = 0, 1, 2 \ldots. \tag{3.12}$$

Right boundary. On the right boundary we need two relations since the fully discrete scheme involved four grid points. It is easy to see that

$$\widehat{u}_{j+2}(z) = \ell_1(z)^2\, \widehat{u}_j(z), \qquad \text{and} \qquad \widehat{u}_{j+1}(z) = \ell_1(z)\, \widehat{u}_j(z), \tag{3.13}$$

Applying them for $j = J - 2$ and using inverse $\mathcal{Z}$-transformation, we obtain

$$u_J^{(n)} - Y_{4,R} *_d u_{J-2}^{(n)} = 0, \quad u_{J-1}^{(n)} - Y_{3,R} *_d u_{J-2}^{(n)} = 0, \quad n = 0, 1, 2 \ldots, \tag{3.14}$$

where $k_{3,R}(z) = \ell_1(z)$ and $k_{4,R}(z) = \ell_1^2(z)$.

Remark 3.1. *We draw on Figure 3 (top figures) the behaviour of the inverse $\mathcal{Z}$-transform only for the two kernels $k_{1,R}(z)$ and $k_{3,R}(z)$ (the behavior of $k_{2,R}(z)$ and $k_{4,R}(z)$ being analogous). We clearly see that the signs of the coefficients alternate. This will possibly create subtractive cancellation errors when we will use them in the boundary convolutions. Following [4, 5], we modify the convolution kernels. The idea is to multiply a $\mathcal{Z}$-transformed kernel $k_{i,R}(z)$ by $\xi(z) = 1 + z^{-1}$ which corresponds to add two neighboured values in the series $\mathcal{Z}^{-1}\{k_{i,R}(z)\}$. We draw on Figure 3 (bottom figures) the behavior of the inverse $\mathcal{Z}$-transform for $\xi(z)k_{1,R}(z)$ and $\xi(z)k_{3,R}(z)$. We can clearly see that the signs of the coefficients do not alternate anymore. In the sequel we introduce the notations $Y_{i,R}^{\xi} = \mathcal{Z}^{-1}\{\xi(z)k_{i,R}(z)\}$. We refer to section 5 for more details on the numerical procedure used to compute the inverse $\mathcal{Z}$-transform for a kernel.*

Remark 3.1 yields the algorithm used in Section 5 to solve numerically the problem. Assuming that the solution on the previous time level $(u_j^{(n)})_{0\le j\le J}$ is known, $(u_j^{(n+1)})_{0\le j\le J}$ is given for $n \ge 0$ by

$$\begin{cases} Y_{2,R}^{\xi} *_d u_0^{(n+1)} - Y_{1,R}^{\xi} *_d u_1^{(n+1)} + u_2^{(n+1)} = -u_2^n, \\ -\alpha u_{j-1}^{(n+1)} + (3\alpha + 1)u_j^{(n+1)} - 3\alpha u_{j+1}^{(n+1)} + \alpha u_{j+2}^{n+1} \\ \qquad\qquad = \alpha u_{j-1}^{(n)} - (3\alpha - 1)u_j^{(n)} + 3\alpha u_{j+1}^{(n)} - \alpha u_{j+2}^{(n)}, \qquad 1 \le j \le J-2, \\ u_{J-1}^{(n+1)} - Y_{3,R}^{\xi} *_d u_{J-2}^{(n+1)} = -u_{J-1}^n, \\ u_J^{(n+1)} - Y_{4,R}^{\xi} *_d u_{J-2}^{(n+1)} = -u_J^n, \end{cases} \tag{3.15}$$

with the mesh ratio $\alpha = U_2\Delta t/(2(\Delta x)^3)$.

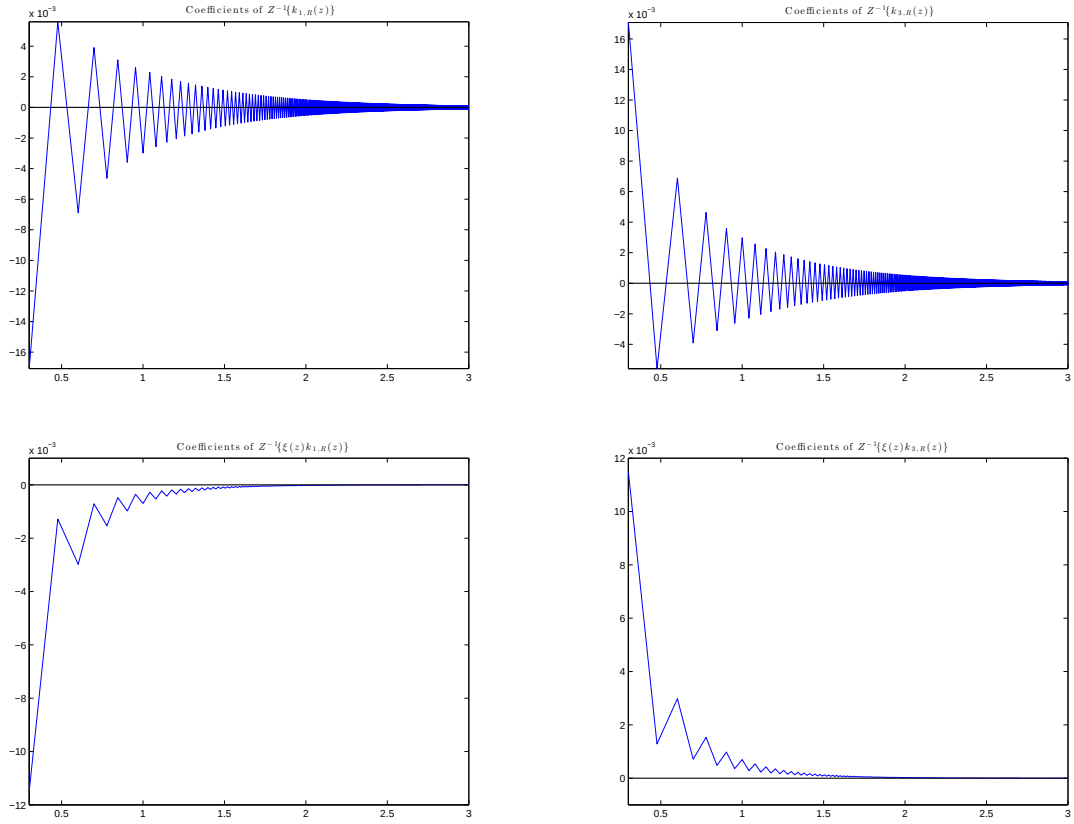

Figure 3: Coefficients of the inverse $\mathcal{Z}$-transform for the kernels $k_{i,R}(z)$, $i = 1, 3$ and kernels $\xi(z)k_{i,R}(z)$, $i = 1, 3$ with $\Delta t = 4/2560$, $\Delta x = 12/5000$ and $r = 1.001$.

3.2 Discrete artificial boundary conditions for (C-CN) scheme

We treat here the case of the (C-CN) scheme for the interior nodes with a spatial index j such that $2 \leq j \leq J-2$. Since we consider a difference scheme with a five points stencil, we need two artificial boundary conditions on each side of the computational interval $[a, b]$.

In order to derive suitable artificial boundary conditions for the (C-CN) scheme (3.4) we follow the same procedure as in Section 3.1 for (R-CN) scheme. First we apply the $\mathcal{Z}$-transform with respect to the time index n, denoting by $\widehat{u}_j$ the $\mathcal{Z}$-transform of the sequence $(u_j^{(n)})_{n\in\mathbb{N}_0}$ we obtain from (3.4) the homogeneous *fourth order difference equation*:

$$\widehat{u}_{j+2} - \left(2 - \frac{U_1(\Delta x)^2}{U_2}\right)\widehat{u}_{j+1} + \frac{4(\Delta x)^3}{U_2 \Delta t}\frac{z-1}{z+1}\widehat{u}_j + \left(2 - \frac{U_1(\Delta x)^2}{U_2}\right)\widehat{u}_{j-1} - \widehat{u}_{j-2} = 0, \quad (3.16)$$

for the spatial index range $2 \leq j \leq J-2$.

The solutions of this difference equation are again of the power form $\widehat{u}_j(z) = \sum_k c_k(z)\,\ell_k^j(z)$, where $\ell = \ell(z)$ solves now the *quartic* equation

$$\ell^4 - \left(2 - \frac{U_1(\Delta x)^2}{U_2}\right)\ell^3 + \frac{4(\Delta x)^3}{U_2 \Delta t}\frac{z-1}{z+1}\ell^2 + \left(2 - \frac{U_1(\Delta x)^2}{U_2}\right)\ell - 1 = 0. \quad (3.17)$$

Equation (3.17) admits four roots denoted here by ℓ_1, ℓ_2, ℓ_3, ℓ_4 (that can be computed numerically or analytically by the well-known Ferrari's solution formula) and thus the general solution of (3.16) is of the form

$$\widehat{u}_j(z) = c_1(z)\,\ell_1^j(z) + c_2(z)\,\ell_2^j(z) + c_3(z)\,\ell_3^j(z) + c_4(z)\,\ell_4^j(z).$$

Like in the previous sections, we have strong numerical evidence (cf. Figure 4), that we can define these four roots of (3.17) such that they *separate* appropriately, i.e.

$$|\ell_1(z)| < 1, \qquad |\ell_2(z)| < 1, \qquad |\ell_3(z)| > 1, \qquad |\ell_4(z)| > 1, \qquad \text{for all } z. \tag{3.18}$$

Let us mention that we obtain Figure 4 as Figure 2 using the polar representation $z = r\,e^{i\Phi}$ with the radius $r > 1$ and the phase $\Phi \in [0, 2\pi)$. We present in Figure 4 the plots obtained in the case $r = 1.001$ (figures on the left) and for the case $r = 1.02$ (figures on the right) with $U_1 = 1$ and $U_2 = 1$ (top figures) or $U_1 = 0$ and $U_2 = 1$ (bottom figures).

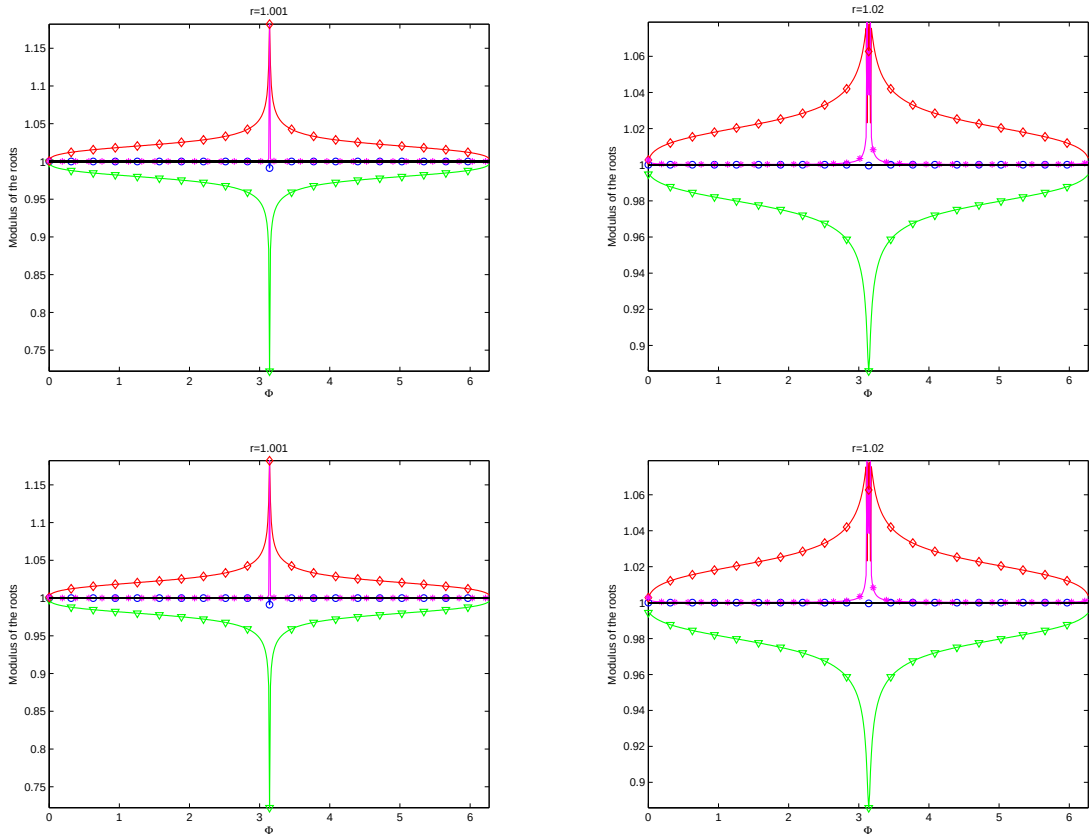

Figure 4: Numerical evidence for the discrete separation property (3.18): Modulus of the four solutions of the quartic equation (3.17) for $\Delta t = 4/2560$, $\Delta x = 12/5000$ and $r = 1.001$ or $r = 1.02$ in the two cases: $U_1 = U_2 = 1$ (top figures), and $U_1 = 0$, $U_2 = 1$ (bottom figures).

Then using the decay condition of the solution we obtain analogously

- for the left exterior domain $c_1(z) = c_2(z) = 0$, $j \le 2$ and thus $\widehat{u}_j(z) = c_3(z)\,\ell_3^j(z) + c_4(z)\,\ell_4^j(z)$, $j \le 2$

- for the right exterior domain $c_3(z) = c_4(z) = 0$, $j \geq J-2$ and thus $\widehat{u}_j(z) = c_1(z)\,\ell_1^j(z) + c_2(z)\,\ell_2^j(z)$, $j \geq J-2$

Right boundary. On the right boundary we need two relations. It is easy to see that

$$\widehat{u}_{j+2}(z) - \big(\ell_1(z)+\ell_2(z)\big)\widehat{u}_{j+1}(z) + \ell_1(z)\ell_2(z)\widehat{u}_j(z) = 0, \tag{3.19}$$

$$\widehat{u}_{j+2}(z) - 2(\ell_1(z)+\ell_2(z))\widehat{u}_{j+1}(z) + (\ell_1(z)+\ell_2(z))^2\widehat{u}_j(z) - (\ell_1(z)\ell_2(z))^2\widehat{u}_{j-2}(z) = 0, \tag{3.20}$$

which will give a link between $u_J^{(n)}$, $u_{J-1}^{(n)}$, $u_{J-2}^{(n)}$ and $u_{J-4}^{(n)}$ with $j = J-2$.

For brevity of the notation we introduce $Y_{i,C} = \mathcal{Z}^{-1}\{k_{i,C}(z)\}, i = 1,2,3,4$ with

$$k_{1,C}(z) = \ell_1(z) + \ell_2(z), \qquad k_{2,C}(z) = \big(\ell_1(z) + \ell_2(z)\big)^2,$$
$$k_{3,C}(z) = \ell_1(z)\ell_2(z), \qquad k_{4,C}(z) = \big(\ell_1(z)\ell_2(z)\big)^2,$$

and we obtain from (3.22)–(3.23)

$$\begin{aligned} u_J^{(n)} - Y_{1,C} *_d u_{J-1}^{(n)} + Y_{3,C} *_d u_{J-2}^{(n)} = 0,\\ u_J^{(n)} - 2Y_{1,C} *_d u_{J-1}^{(n)} + Y_{2,C} *_d u_{J-2}^{(n)} - Y_{4,C} *_d u_{J-4}^{(n)} = 0. \end{aligned} \tag{3.21}$$

Left boundary. On the left boundary we need now two relations. We can easily verify

$$\widehat{u}_j(z) - \big(\ell_3(z)+\ell_4(z)\big)\,\widehat{u}_{j-1}(z) + \ell_3(z)\ell_4(z)\,\widehat{u}_{j-2}(z) = 0, \tag{3.22}$$

$$\widehat{u}_{j+2}(z) - 2(\ell_3(z)+\ell_4(z))\widehat{u}_{j+1}(z) + (\ell_3(z)+\ell_4(z))^2\widehat{u}_j(z) - (\ell_3(z)\ell_4(z))^2\widehat{u}_{j-2}(z) = 0, \tag{3.23}$$

which give a link between $u_0^{(n)}$, $u_1^{(n)}$, $u_2^{(n)}$, $u_3^{(n)}$ and $u_4^{(n)}$ setting $j = 2$. Indeed, denoting by $Y_{i,C} = \mathcal{Z}^{-1}\{k_{i,C}(z)\}$, $i = 5,6,7,8$ with

$$k_{5,C}(z) = \ell_3(z) + \ell_4(z), \qquad k_{6,C}(z) = (\ell_3(z) + \ell_4(z))^2,$$
$$k_{7,C}(z) = \ell_3(z)\ell_4(z), \qquad k_{8,C}(z) = (\ell_3(z)\ell_4(z))^2,$$

we obtain from (3.22)-(3.23)

$$\begin{aligned} Y_{7,C} *_d u_0^{(n)} - Y_{5,C} *_d u_1^{(n)} + u_2^{(n)} = 0,\\ -Y_{8,C} *_d u_0^{(n)} + Y_{6,C} *_d u_2^{(n)} - 2Y_{5,C} *_d u_3^{(n)} + u_4^{(n)} = 0. \end{aligned} \tag{3.24}$$

Following Remark 3.1, we finally obtain the algorithm used in Section 5 to solve numerically the problem. Assuming that the solution $(u_j^{(n)})_{0\leq j\leq J}$ on the previous time level is known, then $(u_j^{(n+1)})_{0\leq j\leq J}$ is given for $n \geq 0$ by

$$\begin{cases} Y_{7,C}^{\xi} *_d u_0^{(n)} - Y_{5,C}^{\xi} *_d u_1^{(n)} + u_2^{(n)} = -u_2^{(n-1)},\\ -Y_{8,C} *_d u_0^{(n)} + Y_{6,C} *_d u_2^{(n)} - 2Y_{5,C} *_d u_3^{(n)} + u_4^{(n)} = -u_4^{(n-1)},\\ -\dfrac{\alpha}{2}u_{j-2}^{(n+1)} + (\alpha-\beta)u_{j-1}^{(n+1)} + u_j^{(n+1)} + (-\alpha+\beta)u_{j+1}^{(n+1)} + \dfrac{\alpha}{2}u_{j+2}^{(n+1)}\\ \quad = \dfrac{\alpha}{2}u_{j-2}^{(n)} + (-\alpha+\beta)u_{j-1}^{(n)} + u_j^{(n)} + (\alpha-\beta)u_{j+1}^{(n+1)} - \dfrac{\alpha}{2}u_{j+2}^{(n+1)}, \quad 1 \leq j \leq J-2,\\ u_J^{(n)} - Y_{1,C}^{\xi} *_d u_{J-1}^{(n)} + Y_{3,C}^{\xi} *_d u_{J-2}^{(n)} = -u_J^{(n-1)},\\ u_J^{(n)} - 2Y_{1,C}^{\xi} *_d u_{J-1}^{(n)} + Y_{2,C}^{\xi} *_d u_{J-2}^{(n)} - Y_{4,C}^{\xi} *_d u_{J-4}^{(n)} = -u_J^{(n-1)}. \end{cases} \tag{3.25}$$

with the mesh ratios $\alpha = U_2\Delta t/(2(\Delta x)^3)$ and $\beta = U_1\Delta t/(4\Delta x)$.

Remark 3.2. *Concerning the implementation, it is usual to define the midpoint unknown* $v_j^{(n+1/2)} = (u_j^{(n+1)} + u_j^{(n)})/2$, *with* $v_j^{(-1/2)} = u_j^{(0)}$. *In this case,* (3.15) *for the (R-CN) scheme reads*

$$\begin{cases} Y_{2,R}^{\xi,(0)} v_0^{(n+1/2)} - Y_{1,R}^{\xi,(0)} v_1^{(n+1/2)} + v_2^{(n+1/2)} = -v_2^{(n-1/2)} - \displaystyle\sum_{k=1}^{n} Y_{2,R}^{\xi,(k)} v_0^{(n+1/2-k)} - \\ \qquad\qquad Y_{2,R}^{\xi,(n+1)} u_0^{(0)}/2 + \displaystyle\sum_{k=1}^{n} Y_{1,R}^{\xi,(k)} v_1^{(n+1/2-k)} + Y_{1,R}^{\xi,(n+1)} u_1^{(0)}/2, \\ v_j^{(n+1/2)} + \alpha \left(v_{j+2}^{(n+1/2)} - 3v_{j+1}^{(n+1/2)} + 3v_j^{(n+1/2)} - v_{j-1}^{(n+1/2)} \right) = u_j^n, \ 1 \le j \le J-2. \\ -Y_{3,R}^{\xi,(0)} v_{J-2}^{(n+1/2)} + v_{J-1}^{(n+1/2)} = -v_{J-1}^{(n-1/2)} + \displaystyle\sum_{k=1}^{n} Y_{3,R}^{\xi,(k)} v_{J-2}^{(n+1/2-k)} + Y_{3,R}^{\xi,(n+1)} u_{J-2}^{(0)}/2, \\ -Y_{4,R}^{\xi,(0)} v_{J-2}^{(n+1/2)} + v_J^{(n+1/2)} = -v_J^{(n-1/2)} + \displaystyle\sum_{k=1}^{n} Y_{4,R}^{\xi,(k)} v_{J-2}^{(n+1/2-k)} + Y_{4,R}^{\xi,(n+1)} u_{J-2}^{(0)}/2, \end{cases} \tag{3.26}$$

For the (C-CN) scheme, (3.25) *reads for* $1 \le j \le J-2$.

$$\begin{cases} Y_{7,C}^{\xi,(0)} v_0^{(n+1/2)} - Y_{5,C}^{\xi,(0)} v_1^{(n+1/2)} + v_2^{(n+1/2)} = -v_2^{(n-1/2)} - \displaystyle\sum_{k=1}^{n} Y_{7,C}^{\xi,(k)} v_0^{(n+1/2-k)} - \\ \qquad\qquad Y_{7,C}^{\xi,(n+1)} u_0^{(0)}/2 + \displaystyle\sum_{k=1}^{n} Y_{5,C}^{\xi,(k)} v_1^{(n+1/2-k)} + Y_{5,C}^{\xi,(n+1)} u_1^{(0)}/2, \\ -Y_{8,C}^{\xi,(0)} v_0^{(n+1/2)} + Y_{6,C}^{\xi,(0)} v_2^{(n+1/2)} - 2Y_{5,C}^{\xi,(0)} v_3^{(n+1/2)} + v_4^{(n+1/2)} \\ \quad = -v_4^{(n-1/2)} + \displaystyle\sum_{k=1}^{n} Y_{8,C}^{\xi,(k)} v_0^{(n+1/2-k)} + Y_{8,C}^{\xi,(n+1)} u_0^{(0)}/2 - \sum_{k=1}^{n} Y_{6,C}^{\xi,(k)} v_2^{(n+1/2-k)} \\ \qquad -Y_{6,C}^{\xi,(n+1)} u_2^{(0)}/2 + 2\displaystyle\sum_{k=1}^{n} Y_{5,C}^{\xi,(k)} v_3^{(n+1/2-k)} + Z_5^{(n+1)} u_3^{(0)}, \\ -\dfrac{\alpha}{2} v_{j-2}^{(n+1/2)} + (\alpha-\beta) v_{j-1}^{(n+1/2)} + v_j^{(n+1/2)} + (-\alpha+\beta) v_{j+1}^{(n+1/2)} + \dfrac{\alpha}{2} v_{j+2}^{(n+1/2)} = u_j^n, \\ Y_{3,C}^{\xi,(0)} v_{J-2}^{(n+1/2)} - Y_{1,C}^{\xi,(0)} v_{J-1}^{(n+1/2)} + v_J^{(n+1/2)} = -v_J^{(n-1/2)} - \displaystyle\sum_{k=1}^{n} Y_{3,C}^{\xi,(k)} v_{J-2}^{(n+1/2-k)} - \\ \qquad\qquad Y_{3,C}^{\xi,(n+1)} u_{J-2}^{(0)}/2 + \displaystyle\sum_{k=1}^{n} Y_{1,C}^{\xi,(k)} v_{J-1}^{(n+1/2-k)} + Y_{1,C}^{\xi,(n+1)} u_{J-1}^{(0)}/2, \\ -Y_{4,C}^{\xi,(0)} v_{J-4}^{(n+1/2)} + Y_{2,C}^{\xi,(0)} v_{J-2}^{(n+1/2)} - 2Y_{1,C}^{\xi,(0)} v_{J-1}^{(n+1/2)} + v_J^{(n+1/2)} \\ \quad = -v_J^{(n-1/2)} + \displaystyle\sum_{k=1}^{n} Y_{4,C}^{\xi,(k)} v_{J-4}^{(n+1/2-k)} + Y_{4,C}^{\xi,(n+1)} u_{J-4}^{(0)}/2 - \sum_{k=1}^{n} Y_{2,C}^{\xi,(k)} v_{J-2}^{(n+1/2-k)} \\ \qquad -Y_{2,C}^{\xi,(n+1)} u_{J-2}^{(0)}/2 + 2\displaystyle\sum_{k=1}^{n} Y_{1,C}^{\xi,(k)} v_{J-1}^{(n+1/2-k)} + Y_{1,C}^{\xi,(n+1)} u_{J-1}^{(0)}. \end{cases} \tag{3.27}$$

Solving (3.26) *or* (3.27), *we recover* $u_j^{(n+1)}$ *by* $u_j^{(n+1)} = 2v_j^{(n+1/2)} - u_j^{(n)}$.

4 The Sum-of-Exponentials Approach

An ad-hoc implementation of the discrete convolutions of the form

$$\sum_{k=1}^{n} X_m^{(k)} u_j^{(n-k)}$$

with convolution coefficients $X_m^{(n)}$ has still one disadvantage. The boundary conditions are non–local in time (and space for higher dimensions) and therefore computations are too expensive. As a remedy, to get rid of the time non-locality, we use as in [4] the sum of exponentials ansatz, i.e. to approximate the convolution coefficients $X_m^{(n)}$ by a finite sum (say L_m terms) of exponentials that *decay* with respect to time. This approach allows for a fast (approximate) evaluation of the discrete convolution since the convolution can now be evaluated with a simple recurrence formula for L_m auxiliary terms and the numerical effort per time step now stays constant.

4.1 The Exponential Approximation

To do so we will follow the ideas of [4] and approximate the coefficients of a sequence $X_m^{(n)}$ by the following *sum-of-exponentials ansatz*

$$X_m^{(n)} \approx \widetilde{X}_m^{(n)} := \begin{cases} X_m^{(n)}, & n = 0, \dots, \nu_m - 1, \\ \sum_{l=1}^{L_m} b_{m,l} q_{m,l}^{-n}, & n = \nu_m, \nu_m + 1, \dots, \end{cases} \tag{4.1}$$

where $L_m \in \mathbb{Z}$, $\nu_m \geq 0$ are given integer parameters, e.g. $L_m = 20$, $\nu_m = 2$, that have to be chosen appropriately to guarantee good approximation properties of $\widetilde{X}_m^{(n)}$. In the following, $X_m^{(n)}$ has to be seen as $Y_{m,R}^{\xi,(n)}$ or $Y_{m,C}^{\xi,(n)}$ respectively for (R-CN) and (C-CN) schemes.

In [4] the authors presented a deterministic method of choosing such an optimal approximation, i.e. finding the set $\{b_{m,l}, q_{m,l}\}$ for fixed L_m and ν_m.

The "split" definition of $\widetilde{X}_m^{(n)}$ in (4.1) is motivated by the fact that the implementation of the discrete TBCs involves a convolution sum with k ranging only from 1 to $k = n$. Since the first coefficient $X_m^{(0)}$ does not appear in this convolution, it makes no sense to include it in our sum-of-exponential approximation, which aims at simplifying the evaluation of the convolution. Hence, one may choose $\nu_m = 1$ in (4.1). We observe numerically that the two first coefficients have a larger magnitude compared to the other ones. This suggests even to exclude $Z_m^{(1)}$ from this approximation and to choose $\nu_m = 2$ in (4.1). We use this choice in our numerical implementation.

Let us fix L_m and consider the formal power series:

$$g_m(x) := s^{(\nu_m)} + s^{(\nu_m+1)} x + s^{(\nu_m+2)} x^2 + \dots, \quad |x| \leq 1. \tag{4.2}$$

If there exists the $[L_m - 1 | L_m]$ *Padé approximation*

$$\tilde{g}_m(x) := \frac{P_{L_m-1}(x)}{Q_{L_m}(x)}$$

of (4.2), then its Taylor series

$$\tilde{g}_m(x) = \widetilde{X}_m^{(\nu_m)} + \widetilde{X}_m^{(\nu_m+1)} x + \widetilde{X}_m^{(\nu_m+2)} x^2 + \dots$$

satisfies the conditions

$$\widetilde{X}_m^{(n)} = X_m^{(n)}, \qquad n = \nu_m, \nu_m + 1, \dots, 2L_m + \nu_m - 1, \tag{4.3}$$

due to the definition of the Padé approximation rule.

Theorem 4.1 ([4]). *Let $Q_{L_m}(x)$ have L_m simple roots $q_{m,l}$ with $|q_{m,l}| > 1, \quad l = 1, \ldots, L_m$. Then*

$$\widetilde{X}_m^{(n)} = \sum_{l=1}^{L_m} b_{m,l} q_{m,l}^{-n}, \qquad n = \nu_m, \nu_m + 1, \ldots, \tag{4.4}$$

where

$$b_{m,l} := -\frac{P_{L_m-1}(q_{m,l})}{Q'_{L_m}(q_{m,l})}\, q_{m,l} \neq 0, \qquad l = 1, \ldots, L_m. \tag{4.5}$$

Evidently, the approximation of the convolution coefficients $X_m^{(n)}$ by the representation (4.1) using a $[L_m - 1|L_m]$ Padé approximant to (4.2) behaves as follows: the first $2L_m$ coefficients are reproduced exactly, see (4.3). However, the asymptotics of $X_m^{(n)}$ and $\widetilde{X}_n^{(n)}$ (as $n \to \infty$) differ strongly – algebraic versus exponential decay.

4.2 Fast Evaluation of the Discrete Convolution

Let us consider the approximation (4.1) with $\nu_m = 2$ for the discrete convolution kernel appearing in the discrete TBCs. With these "exponential" coefficients the *approximated convolution*

$$\tilde{C}_{m,j}^{(n)} := \sum_{k=2}^{n} \widetilde{X}_m^{(k)} u_j^{(n-k)}, \qquad \widetilde{X}_m^{(n)} = \sum_{l=1}^{L_m} b_{m,l} q_{m,l}^{-n}, \qquad |q_l| > 1, \tag{4.6}$$

of the discrete function $u_j^{(n-k)}$, $k = 1, 2, \ldots$ with the coefficients $\widetilde{X}_m^{(n)}$ can be calculated by recurrence formulas, and this will reduce the numerical effort significantly.

A straightforward calculation ([4]) yields (for $\nu_m = 2$):

$$\tilde{C}_{m,j}^{(n)}(\{u_j^{(n)})_n\} = \sum_{l=1}^{L_m} \tilde{C}_{m,j,l}^{(n)}, \qquad n \geq 2, \tag{4.7}$$

where

$$\tilde{C}_{m,j,l}^{(2)} \equiv 0,$$
$$\tilde{C}_{m,j,l}^{(n)} = q_{m,l}^{-1} \tilde{C}_{m,j,l}^{(n-1)} + b_{m,l} q_{m,l}^{-1} u_j^{(n-2)}, \tag{4.8}$$

$n = 2, 3, \ldots, \; l = 1, \ldots, L_m$.

In order to use this fast evaluation procedure in our implementation point of view, we must transform it before to use it for midpoint $v_j^{(n+1/2)}$ unknown. It is easy to see that the second relation of (4.8) can be transformed as

$$\tilde{C}_{m,j,l}^{(n)}\{(u_j^{(n)})_n\} = b_{m,l} \sum_{k=2}^{n-1} q_{m,l}^{-k} u_j^{(n-k)}. \tag{4.9}$$

For example, let us consider the last TBC of (3.15)

$$u_J^{(n+1)} - Y_{4,R}^{\xi} *_d u_{J-2}^{(n+1)} = -u_J^n.$$

We have to see $Y_{4,R}^{\xi} *_d u_{J-2}^{(n+1)}$ as

$$Y_{4,R}^{\xi,(0)} u_{J-2}^{(n+1)} + Y_{4,R}^{\xi,(1)} u_{J-2}^{(n)} + \tilde{C}_{4,J-2}^{(n+1)}\{(u_{J-2}^{(n+1)})_n\}.$$

In order to get an equation for $v_{J-2}^{(n+1/2)}$, we write the previous relation at discrete time level n and average the equations. We therefore get

$$v_J^{(n+1/2)} - Y_{4,R}^{\xi,(0)} v_{J-2}^{(n+1/2)} - Y_{4,R}^{\xi,(1)} v_{J-2}^{(n-1/2)} \\ = -v_J^{(n-1/2)} + \frac{1}{2}\left(\tilde{C}_{4,J-2}^{(n+1)}\{(u_{J-2}^{(n+1)})_n\} + \tilde{C}_{4,J-2}^{(n)}\{(u_{J-2}^{(n)})_n\}\right). \quad (4.10)$$

Thanks to (4.7) and (4.9), we obtain

$$v_J^{(n+1/2)} - Y_{4,R}^{\xi,(0)} v_{J-2}^{(n+1/2)} - Y_{4,R}^{\xi,(1)} v_{J-2}^{(n-1/2)} \\ = -v_J^{(n-1/2)} + \tilde{C}_{4,J-2}^{(n)}\{(v_{J-2}^{(n+1/2)})_n\} + \sum_{l=1}^{L_m} b_{4,l} q_{4,l}^{-n} v_0^{(1/2)}. \quad (4.11)$$

These equation then replaces the last one of (3.26).

These computations can be easily transferred to other convolutions appearing in other TBCs.

5 Numerical Results

In this section we first present the numerical procedure used to compute the inverse $\mathcal{Z}$-transforms required for the discrete absorbing boundary conditions. Then we consider two different examples for which we give some numerical results. The first example can be considered for either (R-CN) or (C-CN) scheme since $U_1 = 0$, $U_2 = 1$. The second one can only be considered for the (C-CN) scheme since in this ones $U_1 = U_2 = 1$.

5.1 Numerical procedure for the inverse $\mathcal{Z}$-transform

In this section, we recall for a self-contained presentation the numerical procedure presented in [29] to compute the inverse $\mathcal{Z}$-transform.

Let us write the $\mathcal{Z}$-transform of a finite sequence $(u_n)_{n\geq 1}^N$ as

$$U(z) = \sum_{n=1}^{N} u_n z^{-(n-1)},$$

cf. (3.6). Next, we define the sample points $z_k = re^{i\phi_k}$ with $0 \leq \phi_k < 2\pi$, $\phi_k = \frac{2\pi}{K}(k-1)$, $1 \leq k \leq K$ for a given K (which should be really bigger than N in practice). Then

$$U_k = U(z_k) = U(re^{i\phi_k}) = \sum_{n=1}^{N} u_n r^{-(n-1)} e^{-(n-1)i\phi_k} = \sum_{n=1}^{N} u_n r^{-(n-1)} \omega_k^{(n-1)(k-1)},$$

with $\omega_k = e^{-\frac{2i\pi}{K}}$. Applying the discrete inverse Fourier transform to U_k, we obtain

$$A = \sum_{n=1}^{N} u_n r^{-(n-1)} \frac{1}{K} \sum_{k=1}^{K} \omega_k^{-[(j-n+1)-1](k-1)}.$$

Since the inverse Fourier transform of 1 is the Dirac function in zero we obtain

$$A = u_j r^{-(j-1)},$$

and then to obtain u_j it remains to multiply A by $r^{(j-1)}$. Note that the choice of the inversion radius r is crucial to guarantee the good approximation of the inverse $\mathcal{Z}$-transform and then the convergence of the numerical scheme as presented in Figure 10. Note that in [29] the best choice of r seems to be 1.02 while in our case it seems to be 1.001. For a concise discussion on the choice of r we refer the reader to [29].

5.2 Numerical Example 1

Let us first consider the example from Zheng, Wen and Han [28] which is concerned with the following equation ($U_1 = 0,\ U_2 = 1$):

$$u_t + u_{xxx} = 0, \qquad x \in \mathbb{R}, \tag{5.1}$$

$$u(0,x) = e^{-x^2}, \qquad x \in \mathbb{R}, \tag{5.2}$$

$$u \to 0, \qquad |x| \to \infty. \tag{5.3}$$

The fundamental solution of equation (5.1) is [28]

$$E(t,x) = \frac{1}{\sqrt[3]{3t}}\,\mathrm{Ai}\left(\frac{x}{\sqrt[3]{3t}}\right),$$

where $\mathrm{Ai}(\cdot)$ is the Airy function. The exact solution of (5.1)-(5.3) can be written in terms of $E(t,x)$ as

$$u_{\text{exact}}(t,x) = E(t,x) * e^{-x^2},$$

where $*$ denotes the convolution product on the whole real axis.

We present in Figure 5 the exact solution and the approximate solution obtained with (R-CN) scheme for $\Delta t = 4/2560$, $\Delta x = 12/5000$ and $r = 1.001$ at different times $t = 1, 2, 3, 4$. We see that the (R-CN) solution is a very good approximation of the exact solution all along the time. No unphysical reflections can be seen at the boundaries. The same can be obtained by using the (C-CN) scheme.

We present in Figure 6 a comparison at time $T = 1$ between the exact solution and the approximate solution obtained with the sum of exponential approach either for various values of N ($N = 640, 1280, 2560$) and a fixed value of L_m ($L_m = 20$) or for a fixed value of N ($N = 2560$) and various values of L_m ($L_m = 10, 20$). In each case $\Delta x = 12/5000$ and $r = 1.001$. We observe that the accuracy of the approximate solution depend on N for a fixed L_m and on L_m for a fixed N.

Let us define as $e^{(n)}$ the *relative ℓ^2-error* at time $t = n\Delta t$ given by:

$$e^{(n)} = \left\| u^{(n)}_{\text{exact}} - u^{(n)}_{\text{num}} \right\|_2 / \left\| u^{(n)}_{\text{exact}} \right\|_2,$$

where we use trapezoidal rule to compute the ℓ^2-norm. Note that here u_{num} stands for the numerical solution computed with either (R-CN) or (C-CN) scheme. We decided to

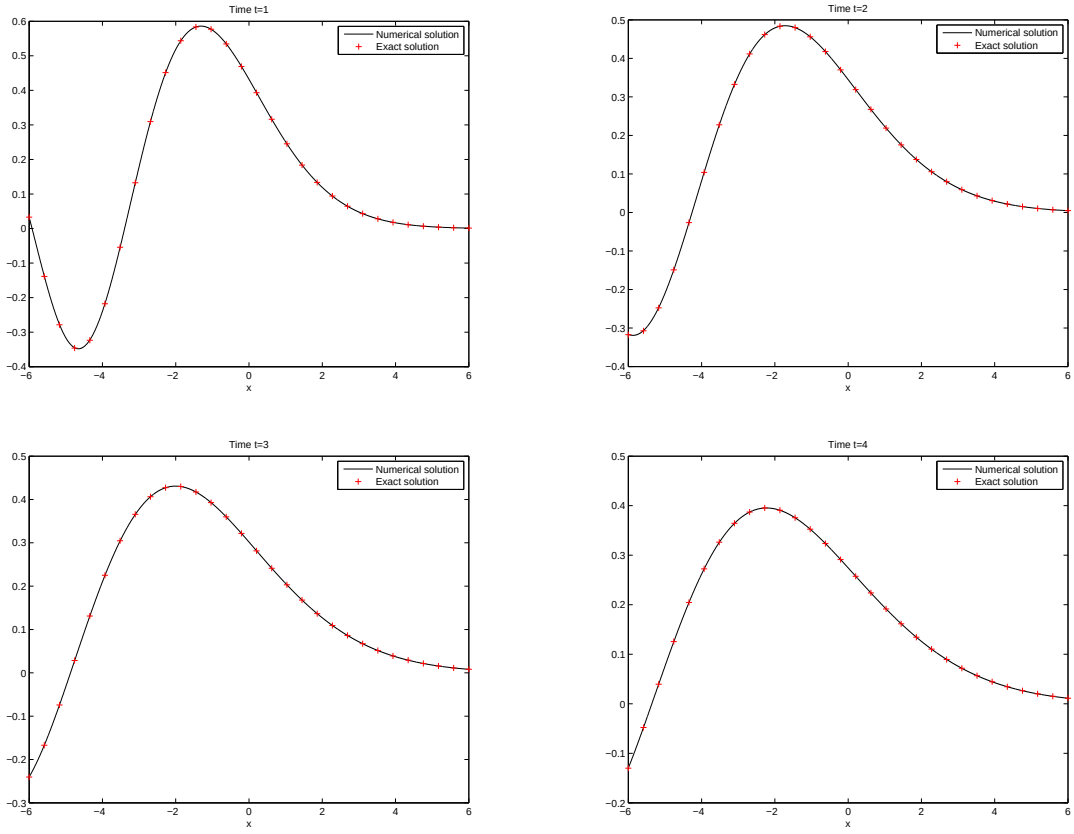

Figure 5: Numerical and exact solutions at times $t = 1$, $t = 2$, $t = 3$ and final time $T = 4$ for the first example with $\Delta t = 4/2560$, $\Delta x = 12/5000$ and $r = 1.001$.

compute from $e^{(n)}$ two error functions; first the maximum in time and secondly the ℓ^2-error in time:

$$rel.ErrTm = \max_{0<n<N} \left(e^{(n)} \right), \qquad rel.ErrL2 = \left(\Delta t \sum_{n=1}^{N} (e^{(n)})^2 \right)^{1/2}.$$

The behaviour of these two errors with respect to Δx are presented in Figure 7. We observe that we obtained numerically the expected order of accuracy for each scheme: the (R-CN) scheme has a convergence order of one and the (C-CN) scheme is of order two. We can also see that for each value of N there is a saturation phenomena for the error, for very small Δx the round-off errors balances with the errors in the solution. Also, changing Δx also modifies the roots of the cubic/quartic equations needed in the boundary convolution and the numerical inverse $\mathcal{Z}$-transform of the convolution kernels may degrade the overall accuracy (at least for the selected inversion radius).

We present in Figure 8 the $rel.ErrTm$ and $rel.ErrL2$ with respect to Δt for $J = 20000$ and $r = 1.001$. Again we obtain for each scheme a numerical rate of convergence of order two in time. Surprisingly, the saturation effects from the previous figure do not show up, although with smaller Δt the size of the boundary convolutions is increasing and this often leads to additional errors.

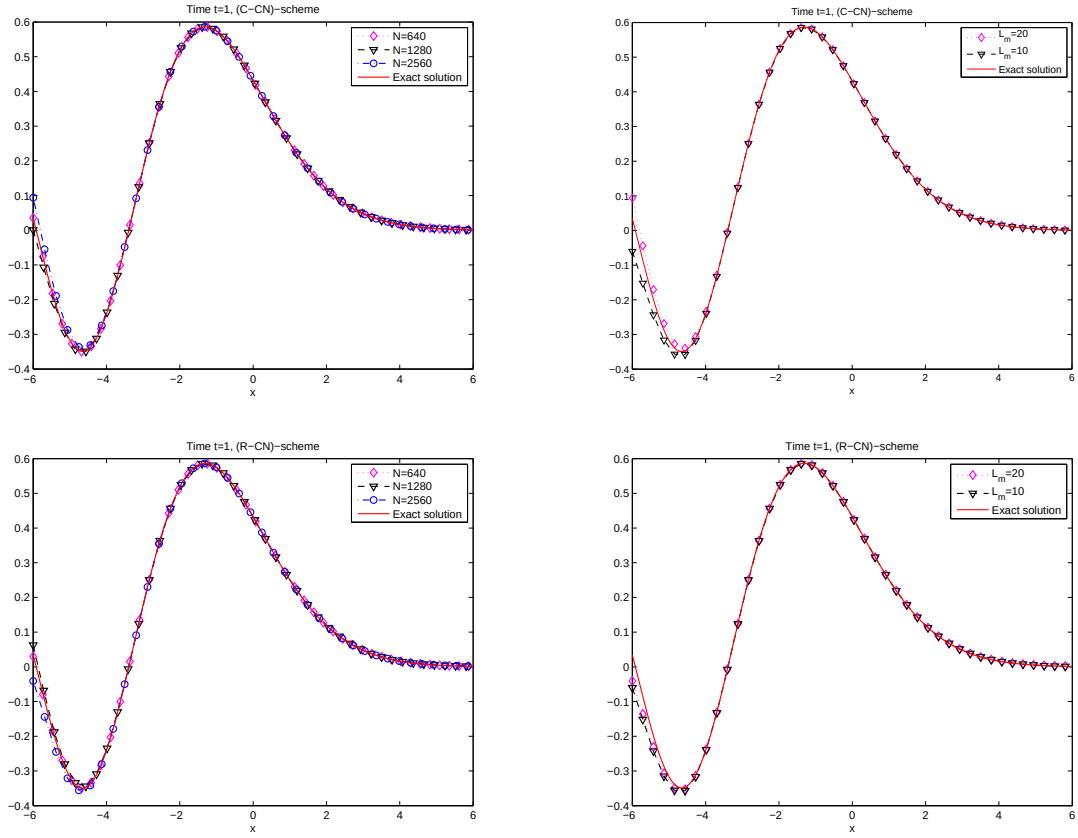

Figure 6: Comparison at time $T = 1$ between the exact solution and the approximate solution obtained with the sum of exponential approach for the (C-CN) scheme (top figures) and the (R-CN)-scheme (bottom figures). The left figures are obtained with a fixed $L_m = 20$ and various $N = 640, 1280, 2560$ and the right figures with a fixed $N = 2560$ and various $L_m = 10, 20$ for (C-CN) and (R-CN)-schemes.

We present in Figure 9 the evolution of the ℓ^2-error with respect to time for various values of N, $J = 20000$ and $r = 1.001$. As expected, the error decreases for increasing N, i.e. finer mesh size. In any case, the error remains moderately bounded over the whole simulation time which shows the usefulness of the proposed method. At the beginning the first increase is due to the interaction with the artificial boundaries and the second long term growth is due to an accumulation effect of errors, e.g. due to the increasing time convolution at the boundaries.

We present in Figure 10 the $rel.ErrL2$ with respect to r for each scheme and either with $N = 2560$ and various J or $J = 5000$ and various N. The choice of r in the inverse $\mathcal{Z}$-transform procedure is clearly impacting the error and depend on the values of J and N. It seems that our choice, $r = 1.001$, is a good choice for a large set of values of N and J.

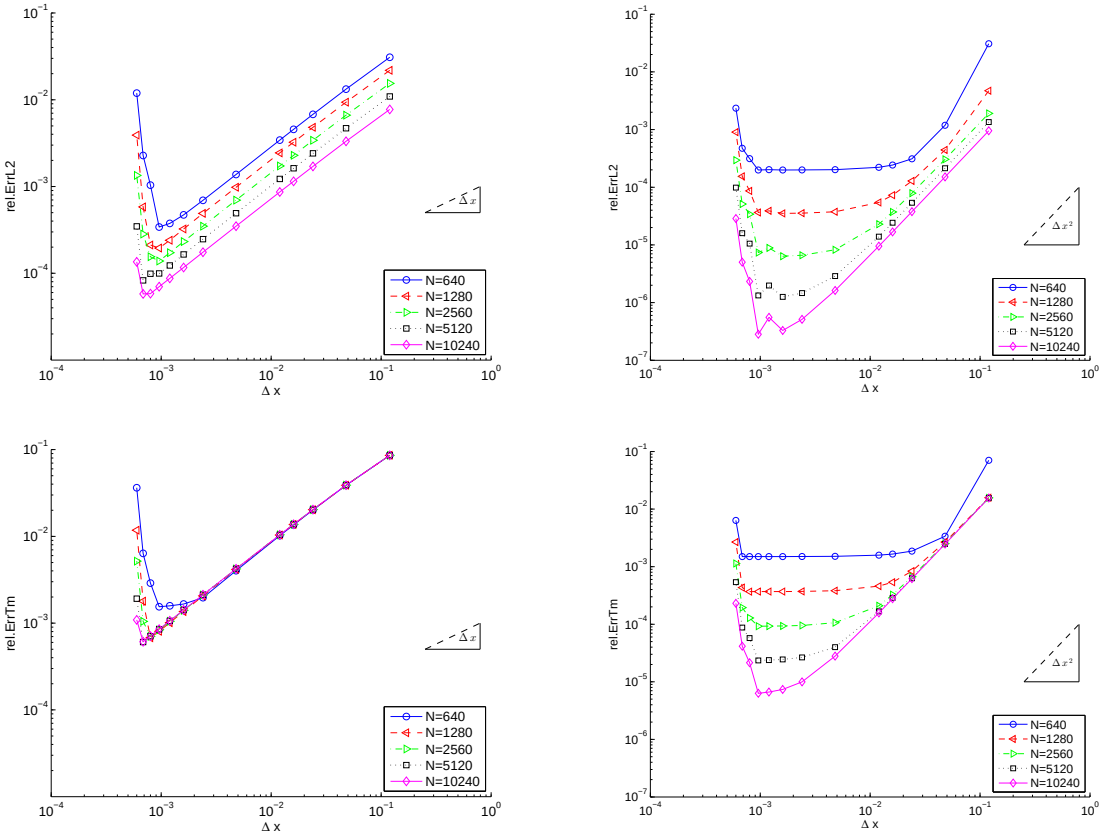

Figure 7: Relative errors with respect to Δx at time $T = 4$ for the (R-CN) scheme (figures on left) and the (C-CN)-scheme (figures on right) and for different values of N.

5.3 Numerical Example 2

Let us now consider a second example. We consider the dispersive equation (1.5) with $U_1 = U_2 = 1$ and we choose as initial condition

$$u_0(x) = \exp(-8(x-5)^2)\sin\left(\frac{50\pi}{4}\right),$$

for $0 \leq x \leq 10$ and for a final time $T = 4.8 \times 10^{-4}$. This example was already considered in [6]. Note that using the Fourier transform, the problem being a linear and periodic problem, we can compute the exact solution $u_{\text{exact}}(t,x)$. Indeed, applying the Fourier transformation in the space variable to the equation (1.5) we obtain

$$\widehat{u}_t + i\xi\widehat{u} - i\xi^3\widehat{u} = 0,$$

where ξ stands for the Fourier variable. Then it is easy to see that the transformed exact solution reads

$$\widehat{u}_{\text{exact}}(t,\xi) = \widehat{u}_0 \exp\left(-\left(i\xi - i\xi^3\right)t\right).$$

Using the inverse Fourier transform we have the exact solution of the problem. A reference solution is computed using 50000 points in space and 2560 iterations in time and used for Figures 11 and 12.

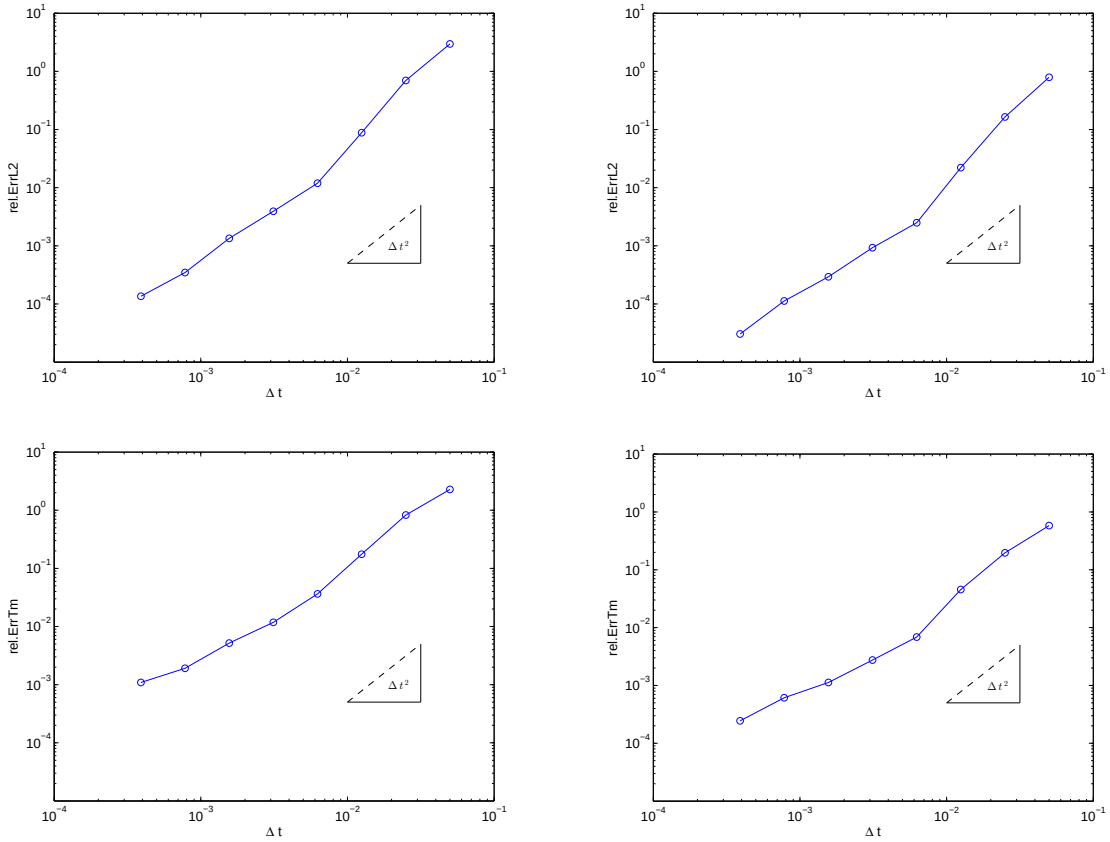

Figure 8: Relative errors with respect to Δt at time $T = 4$ for the (R-CN) scheme (figures on left) and the (C-CN)-scheme (figures on right) and for $J = 20000$.

We present in Figure 11 the exact solution and the approximate solution obtained with (C-CN) scheme for $\Delta t = T/2560$, $\Delta x = 10/5000$ and $r = 1.001$ at final time. We see that the (C-CN) solution is a very good approximation of the exact oscillatory solution, the two solutions are nearly indistinguishable.

We present in Figure 12 the relative ℓ^2-error $e^{(n)}$ computing at final time (*i.e.* for n=N) with respect to Δx and for various values of N and $r = 1.001$. Again, we see that we obtain the order two as predicted. As for the first example there is a saturation phenomena.

Conclusion and Outlook

In this work we presented some new discrete absorbing boundary conditions adapted to two different numerical schemes for the linearized KdV equation (1.5). The orders of each scheme in time and space are shown numerically and given evidence that they are not perturbed by the discrete absorbing boundary conditions. To speed up the calculations of the costly boundary convolutions, especially in higher-dimensional cases, we proposed to use the sum-of-exponentials ansatz. We gave finally two numerical examples that supported our theoretical findings.

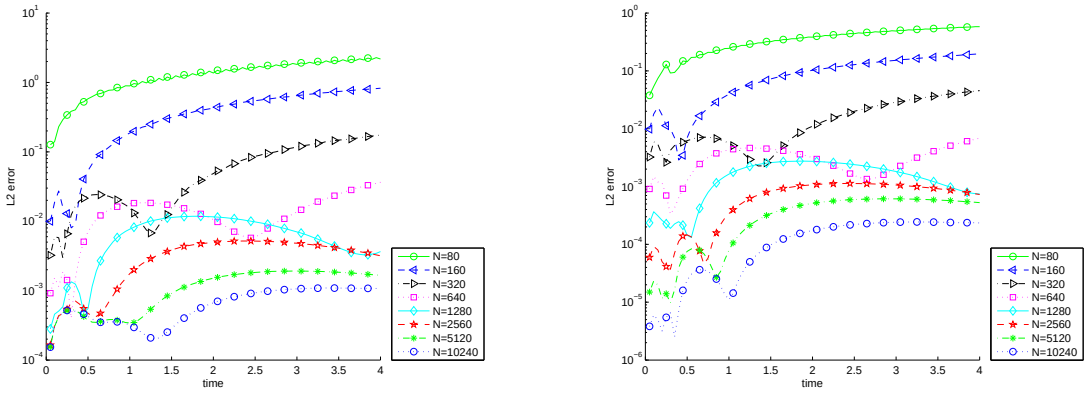

Figure 9: Evolution of the ℓ^2 error between $T = 0$ and $T = 4$ for the (R-CN) scheme (figure on left) and the (C-CN)-scheme (figure on right) and for $J = 20000$ and various values of N.

Future work will be to design an automatic good choice of the inversion radius, establish a transformation rule in the spirit of [4] for the KdV equation, treat the 2D and the nonlinear case.

Acknowledgements

This work was partially supported by the French ANR grant MicroWave NT09 460489 ("Programme Blanc" call) and Université Paul Sabatier Toulouse 3. The first author also acknowledges support from the French ANR grant BonD ANR-13-BS01-0009-01.

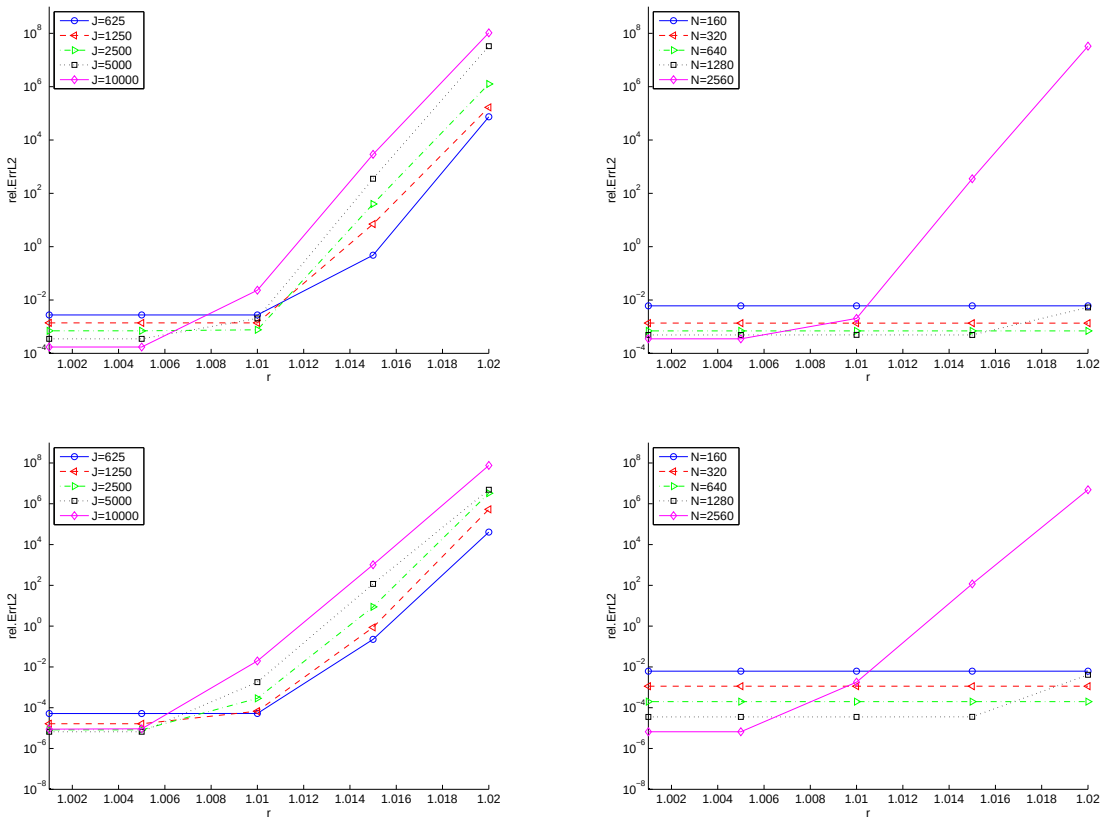

Figure 10: Error with respect to r at time $T = 4$ for the (R-CN) scheme (top figures) and the (C-CN)-scheme (bottom figures) with either $N = 2560$ and various J (figures on the left) or $J = 5000$ and various N (figures on the right).

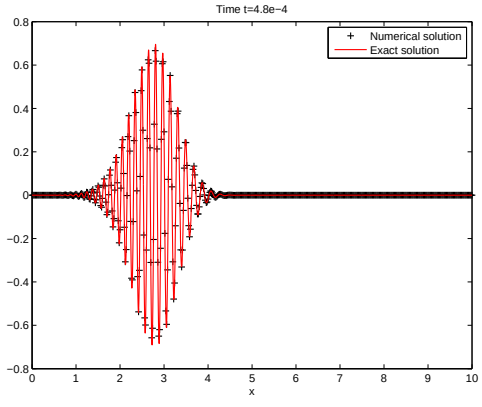

Figure 11: Numerical and exact solutions at final time for the second example with $\Delta t = T/2560$, $\Delta x = 10/5000$ and $r = 1.001$.

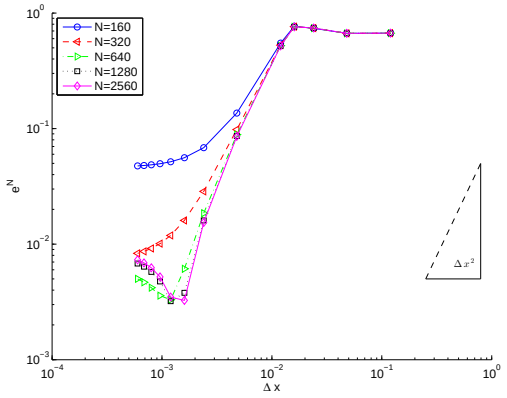

Figure 12: Relative error $e^{(N)}$ with respect to Δx for various values of N using (C-CN) scheme for the second example

References

[1] M.J. Ablowitz, and P.A. Clarkson, *Solitons, nonlinear evolution equations and inverse scattering*, Cambridge University Press, New York (1991).

[2] X. Antoine, A. Arnold, C. Besse, M. Ehrhardt, and A. Schädle, *A review of transparent and artificial boundary conditions techniques for linear and nonlinear Schrödinger equations*, Commun. Comput. Phys., 4 (2008), 729-796.

[3] X. Antoine, C. Besse, and S. Descombes, *Artificial boundary conditions for one-dimensional cubic nonlinear Schrödinger equations*, SIAM J. Numer. Anal. 43 (2006), 2272-2293.

[4] A. Arnold, M. Ehrhardt, and I. Sofronov, *Discrete transparent boundary conditions for the Schrödinger equation: fast calculation, approximation, and stability*, Commun. Math. Sci. 1 (2003), 501-556.

[5] A. Arnold, M. Ehrhardt, M. Schulte, and I. Sofronov, *Discrete transparent boundary conditions for the Schrödinger equation on circular domains*, Commun. Math. Sci. 10 (2012), 889-916.

[6] W.L. Briggs, and T. Sarie, *Finite difference solutions of dispersive partial differential equations*, Math. Comput. Simul. 25 (1983), 268-278.

[7] J. Cruz, *Ocean Wave Energy - Current Status and Future Prospects*, Springer, 2008.

[8] M. Ehrhardt, *Discrete artificial boundary conditions*, PhD, Technische Universität Berlin, 2001.

[9] M. Ehrhardt, and A. Arnold, *Discrete transparent boundary conditions for the Schrödinger equation*, Riv. Math. Univ. Parma 6 (2001), 57-108.

[10] M. Ehrhardt, *Discrete Transparent Boundary Conditions for Schrödinger-type equations for non-compactly supported initial data*, Appl. Numer. Math. 58 (2008), 660-673.

[11] C. Eilbeck, (1998) `http://www.ma.hw.ac.uk/~chris/scott_russell.html`

[12] C.S. Gardner, J.M. Greene, M.D. Kruskal, and R.M. Miura, *Method for solving the Korteweg-de Vries equation*, Phys. Lett. 19 (1967), 1095-1097.

[13] H. Gleeson, P. Hammerton, D.T. Papageorgiou, and J.-M. Vanden-Broeck, *A new application of the Korteweg-de Vries Benjamin-Ono equation in interfacial electrohydrodynamics*, Phys. Fluids 19 (2007), 031703

[14] D.J. Korteweg, and G. de Vries, *On the Change of Form of Long Waves Advancing in a Rectangular Canal, and on a New Type of Long Stationary Waves*, Philosophical Magazine 39 (1895), 422-443.

[15] N.A. Kudryashov, and I.L. Chernyavskii, *Nonlinear Waves in Fluid Flow through a Viscoelastic Tube*, Fluid Dynamics 41 (2006), 49-62.

[16] Q. Mengzhao, *Difference schemes for the dispersive equation*, Computing, 31 (1983), 261–267.

[17] L.A. Ostrovsky, and Y.A. Stepanyants, *Do internal solitons exist in the ocean?*, Reviews of Geophysics, 27 (1989), 293-310.

[18] J.S. Russell, *Report of the committee on waves*, Rep. Meet. Brit. Assoc. Adv. Sci. 7th Liverpool (1837) 417, London, John Murray.

[19] S.W. Schoombie, *A discrete multiple scales analysis of a discrete version of the Korteweg-de Vries equation*, J. Comp. Phys. 101 (1992), 55-70.

[20] J. Shen, *A new dual-Petrov-Galerkin method for third and higher odd-order differential equations: application to the KdV equation*, SIAM J. Numer. Anal. 41 (2003), 1595-1619.

[21] H. Washimi, and T. Taniuti, *Propagation of ion-acoustic solitary waves of small amplitude*, Phys. Rev. Lett. 17 (1966), 966.

[22] G.B. Whitham, *Linear and nonlinear waves*, Wiley, New York, 1974.

[23] J.D. Wright, *Corrections to the KdV Approximation for Water Waves*, SIAM J. Math. Anal. 37 (2005), 1161-1206.

[24] N.J. Zabusky, and M.D. Kruskal, *Interactions of Solitons in a Collisionless Plasma and the Recurrence of Initial States*, Phys. Rev. Lett. 15 (1965), 240-243.

[25] N.J. Zabusky, *Phenomena Associated with the Oscillations of a Nonlinear Model String*, in Mathematical Models in Physical Sciences, S. Drobot (ed.) Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

[26] W. Zhang, H. Li, and X. Wu, *Local absorbing boundary conditions for a linearized Korteweg-de Vries equation*, Phys. Rev. E 89 (2014), 053305.

[27] C. Zheng, *Numerical simulation of a modified KdV equation on the whole real axis*, Numer. Math. 105 (2006), 315-335.

[28] C. Zheng, X. Wen, and H. Han, *Numerical Solution to a Linearized KdV Equation on Unbounded Domain*, Numer. Meth. Part. Diff. Eqs. 24 (2008), 383-399.

[29] A. Zisowsky, *Discrete transparent boundary conditions for systems of evolution equations*, PhD, Technische Universität Berlin, 2003.