Journal of Computational Mathematics
Vol.33, No.1, 2015, 17–32.

http://www.global-sci.org/jcm
doi:10.4208/jcm.1406-m4443

FINITE DIFFERENCE METHODS FOR THE HEAT EQUATION WITH A NONLOCAL BOUNDARY CONDITION*

V. Thomée
Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE-412 96 Göteborg, Sweden
E-mail: thomee@chalmers.se
A.S. Vasudeva Murthy
TIFR Centre for Appl. Math, Yelahanka New Town, Bangalore, India
E-mail: vasu@math.tifrbng.res.in

Abstract

We consider the numerical solution by finite difference methods of the heat equation in one space dimension, with a nonlocal integral boundary condition, resulting from the truncation to a finite interval of the problem on a semi-infinite interval. We first analyze the forward Euler method, and then the $\theta-$method for $0 < \theta \le 1$, in both cases in maximum-norm, showing $O(h^2 + k)$ error bounds, where h is the mesh-width and k the time step. We then give an alternative analysis for the case $\theta = 1/2$, the Crank-Nicolson method, using energy arguments, yielding a $O(h^2 + k^{3/2})$ error bound. Special attention is given the approximation of the boundary integral operator. Our results are illustrated by numerical examples.

Mathematics subject classification: 65M06, 65M12, 65M15
Key words: Heat equation, Artificial boundary conditions, unbounded domains, product quadrature.

1. Introduction

We are concerned with the numerical solution of the parabolic problem on a semi-infinite interval,

$$u_t = u_{xx} + f(x,t), \qquad \text{for } x \ge 0, \quad t > 0, \tag{1.1a}$$
$$u(0,t) = b(t), \qquad \text{for } t > 0, \tag{1.1b}$$
$$u(x,0) = v(x), \qquad \text{for } x \ge 0, \tag{1.1c}$$
$$u \to 0, \qquad \text{for } x \to +\infty, \tag{1.1d}$$

where $f(x,t)$ and $v(x)$ vanish outside a finite interval in x, which in the sequel we normalize to be $[0,1)$. To be able to use finite difference or finite element methods for this problem, it is useful to truncate it to this finite spatial interval. This necessitates setting up a boundary condition at the right hand endpoint of the interval, $x = 1$, usually referred to as an artificial boundary condition (*abc*). Han and Huang [3] have recently proposed such an *abc* for (1.1)

* Received October 17, 2013 / Revised version received May 19, 2014 / Accepted June 26, 2014 / Published online December 1, 2014 /

resulting in the initial-boundary value problem

$$u_t = u_{xx} + f(x,t), \qquad \text{for } x \in (0,1), \quad t > 0 \tag{1.2a}$$
$$u(0,t) = b(t), \qquad \text{for } t > 0, \tag{1.2b}$$
$$u_x(1,t) + Gu(1,t) = g(t), \qquad \text{for } t > 0, \tag{1.2c}$$
$$u(x,0) = v(x), \qquad \text{for } x \in (0,1), \tag{1.2d}$$

with $g(t) = 0$, where Gu may be thought of as a fractional derivative of order $\frac{1}{2}$ of u, cf. [8], or

$$Gu(t) = Ju_t(t), \quad \text{where } Jw(t) = \frac{1}{\sqrt{\pi}} \int_0^t \frac{w(s)}{\sqrt{t-s}} ds. \tag{1.3}$$

The function $g(t)$ will be included below for the purpose of our analysis.

To derive this *abc* at $x = 1$, we set $b_1(t) = u(1,t)$, with u the solution of (1.1), and note that u also solves

$$u_t = u_{xx}, \qquad \text{for } x \geq 1, \quad t > 0,$$
$$u(1,t) = b_1(t), \qquad \text{for } t > 0, \quad \text{and} \quad u(x,0) = 0, \quad \text{for } x \geq 1.$$

Using Laplace transformation one shows that the solution of this problem is

$$u(x,t) = \frac{x-1}{2\sqrt{\pi}} \int_0^t (t-s)^{-3/2} b_1(s) e^{-(x-1)^2/(4(t-s))}\, ds, \quad \text{for } x > 1,\ t > 0.$$

From this one finds, after some calculation, that

$$u_x(1,t) = -\frac{1}{\sqrt{\pi}} \int_0^t (t-s)^{-1/2} b_1'(s)\, ds = -Ju_t(1,t), \quad \text{for } t > 0,$$

and hence that the boundary condition at $x = 1$ in (1.2) holds. Although [3] does not conatin any error analysis, the authors demonstrated the effectiveness of this *abc* by numerical computation. Recently Wu and Sun [7] have analyzed this *abc* for a slightly more complicated difference scheme than the Crank-Nicolson one, and Zheng [8] employs the same condition for the time discretized heat equation using the $\mathcal{Z}$ transform. For a technique that does not truncate the domain, see Li and Greengard [4]. Tsynkov [6] contains a survey of numerical solution on infinite domains.

Our purpose here is to analyze the solution of the truncated problem (1.2) by finite differences, using the θ-method, for $0 \leq \theta \leq 1$. For $\theta = 0$ this reduces to the explicit forward Euler method, and for $\theta > 0$ the method is implicit, with the backward Euler method corresponding to $\theta = 1$, and the Crank-Nicolson method to $\theta = \frac{1}{2}$.

We use the spatial grid $x_m = mh, m = 0, 1, \ldots M+1$, with $h = 1/M'$, where M is a positive integer and $M' = M + \frac{1}{2}$, thus also using the grid point $x_{M+1} = 1 + \frac{1}{2}h$ to the right of the right hand boundary, but with no gridpoint at $x = 1$. The step size in time is denoted by k, with the corresponding time levels $t_n = nk$. We denote by U_m^n the difference approximation of $u(x_m, t_n)$ and introduce the forward and backward difference quotients in space and time by

$$\partial_x U_m^n = \frac{U_{m+1}^n - U_m^n}{h}, \qquad \bar{\partial}_x U_m^n = \frac{U_m^n - U_{m-1}^n}{h},$$
$$\partial_t U_m^n = \frac{U_m^{n+1} - U_m^n}{k}, \qquad \bar{\partial}_t U_m^n = \frac{U_m^n - U_m^{n-1}}{k}.$$

We consider first, in Section 3 below, the forward Euler approximation

$$\partial_t U_m^n - \partial_x \overline{\partial}_x U_m^n = f_m^n, \quad \text{for } m = 1, \dots, M,\ n \ge 0, \quad f_m^n = f(x_m, t_n), \tag{1.4}$$

with the left side boundary values and initial values given by

$$\begin{aligned} &U_0^n = b^n, \quad \text{for } n \ge 1, \quad b^n = b(t_n), \\ &U_m^0 = v_m, \quad \text{for } m = 0. \dots, M+1, \quad v_m = v(x_m). \end{aligned} \tag{1.5}$$

To approximate the boundary condition at $x = 1$ in (1.2) we use a second order symmetric finite difference approximation to the spatial derivative at $x = 1$, and then need to find a suitable approximation to the integral operator

$$Gw(t) = Jw_t(t) = \int_0^t \beta(t-s)\, w_t(s)\, ds, \quad \text{where } \ \beta(t) = (\pi t)^{-1/2}. \tag{1.6}$$

Our starting point is the product integration rule for the convolution operator J, setting $w^j = w(t_j)$ and $I_j = (t_{j-1}, t_j)$, for $j = 1, \dots, n$,

$$Jw(t_n) \approx J_k w^n = \sum_{j=1}^n \int_{I_j} \beta(t_n - s)\, ds\, w^j = k^{1/2} \sum_{j=1}^n \omega_{n-j} w^j, \quad n \ge 1, \tag{1.7}$$

where

$$\sqrt{\pi}\omega_j = k^{-1/2} \int_{t_j}^{t_{j+1}} y^{-1/2}\, dy = 2(\sqrt{j+1} - \sqrt{j}) = \frac{2}{\sqrt{j+1} + \sqrt{j}}, \quad j \ge 0.$$

We now set

$$G_k U^n = J_k \bar{\partial}_t U^n = k^{-1/2} \sum_{j=1}^n \omega_{n-j} (U^j - U^{j-1}), \quad \text{for } n \ge 1. \tag{1.8}$$

This operator will be discussed in detail in Section 2 below.

For the boundary condition at $x = 1$ we then apply this quadrature rule to the average of the values at $m = M, M+1$, thus prescribing

$$\partial_x U_M^n + G_k U_{M'}^n = g^n, \quad \text{for } n \ge 1, \quad \text{where } U_{M'}^n = \tfrac{1}{2}(U_M^n + U_{M+1}^n). \tag{1.9}$$

We begin our analysis with an abstract stability result which we use to show the maximum-norm stability of the forward Euler scheme, in the case $b(t) = 0$, setting

$$\|v\| = \max_{0 \le m \le M+1} |v_m| \quad \text{and} \quad \|v\|_0 = \max_{1 \le m \le M} |v_m|.$$

We demonstrate that, provided the standard mesh-ratio condition $\lambda = k/h^2 \le 1/2$ is satisfied, and in addition a bound for λ from below, $\lambda \ge \lambda_0 > 1/\pi$, then, for the solutions of (1.4), (1.5), with $b^n = 0$, and (1.9), we have

$$\|U^n\| \le C(t_n) \left(\|U^0\| + k \sum_{j=0}^{n-1} \|f^j\|_0 + \max_{j \le n} |g^j| \right), \quad \text{for } n \ge 1. \tag{1.10}$$

Together with estimates for the truncation errors, based on an analysis of the quadrature error in (1.8), this will show an error bound of the form

$$\|U^n - u(t_n)\| \le C(u, t_n)\, h^2, \quad \text{for } n \ge 1. \tag{1.11}$$

In Section 4 we then extend our analysis to the θ-method with $\theta > 0$,

$$\partial_t U_m^n - \partial_x \overline{\partial}_x(\theta U_m^{n+1} + (1-\theta)U_m^n) = f_m^{n+\theta}, \qquad m = 1, \dots, M,\ n \ge 1, \tag{1.12a}$$
$$U_0^n = b^n, \qquad n \ge 0, \tag{1.12b}$$
$$\partial_x U_M^n + G_k U_{M'}^n = g^n, \qquad n \ge 1, \tag{1.12c}$$
$$U_m^0 = v_m, \qquad m = 0. \dots, M+1, \tag{1.12d}$$

where $f_m^{n+\theta} = f(x_m, t_n + \theta k)$. In this case, the condition for stability is reduced to $2\lambda(1-\theta) \le 1$, which is satisfied for any $\lambda > 0$ in the backward Euler case $\theta = 1$, with the requirement $\lambda > 1/\pi$ remaining. The error estimate now reads

$$\|U^n - u(t_n)\| \le C(u, t_n)(h^2 + k), \quad \text{for } n \ge 1.$$

In the particular case $\theta = \frac{1}{2}$, the approximation of the heat equation in (1.12a) is the Crank-Nicolson method. In this case our stability condition becomes $\lambda \le 1$, or $k \le h^2$, and the error bound will be of order $O(h^2)$. For the standard boundary value problem for the heat equation, the Crank-Nicolson method is unconditionally stable in L_2, and one is able to show a $O(h^2 + k^2)$ order error bound. In Section 5 we therefore give an alternative analysis of this case, based on energy arguments. In treating the boundary condition in (1.12a) we will then have reason to take advantage of a discrete version of the fact that the kernel $\beta(t)$ in (1.6) is positive definite, i.e.

$$\int_0^t Jw(s)\, w(s)\, ds = \int_0^t \int_0^s \beta(s-y)\, w(y)\, dy\, w(s)\, ds \ge 0, \quad \forall w. \tag{1.13}$$

Such discrete analogues of J are also discussed in Section 2, cf. [5]. For the discrete Crank-Nicolson method we show an error bound of nonoptimal order $O(h^2 + k^{3/2})$, where we have lost a factor $k^{-1/2}$ in the quadrature error in (1.6) because of the weak singularity in $\beta(t)$.

In Section 6 we illustrate our findings by some numerical examples.

2. The Quadrature Formula

In this section we discuss some properties of the quadrature operator G_k defined in (1.8), (1.7), approximating the operator G in (1.3). We begin by showing that this operator is accurate of order $O(k^{3/2})$, see also Wu and Sun [7], Lemma 1. We shall use the notation

$$\|w\|_{\mathbb{C}_t^p} = \max_{l \le p} \sup_{s \le t} |w^{(l)}(s)|.$$

Lemma 2.1. *For the quadrature operators G_k and G, defined in* (1.8), (1.7), *and* (1.3), *respectively, we have,*

$$|\bar\partial_t^q (G_k w^n - Gw(t_n))| \le Ck^{3/2} \|w\|_{\mathbb{C}_{t_n}^{2+q}}, \quad n \ge 1+q,\ q = 0, 1.$$

where in the case $q = 1$ we assume that $w^{(l)}(0) = 0$ for $l = 0, 1, 2, 3$.

Proof. By our definitions we have

$$G_k w^n - Gw(t_n) = \sum_{j=1}^n \int_{I_j} (\bar\partial_t w^j - w'(s))\beta(t_n - s)\, ds.$$

We find, by Taylor expansion with remainder term, for $t \in I_j$,

$$\bar\partial_t w^j - w'(t) = k^{-1}\Big(\int_t^{t_j} (t_j - s)\, w''(s)\, ds - \int_{t_{j-1}}^{t} (s - t_{j-1})\, w''(s)\, ds\Big).$$

Hence, after changing the order of integration in two double integrals,

$$\int_{I_j} (\bar\partial_t w^j - w'(t))\beta(t_n - t)dt = \int_{I_j} w''(s)\chi_{nj}(s)ds,$$

where

$$\chi_{nj}(s) = k^{-1}\Big((t_j - s)\int_{t_{j-1}}^{s} \beta(t_n - t)dt - (s - t_{j-1})\int_s^{t_j} \beta(t_n - t)dt\Big).$$

To show our claim for $q = 0$, we need to prove that

$$\sum_{j=1}^{n} \int_{I_j} |\chi_{nj}(s)|ds \le Ck^{3/2}.$$

We note that, for $s \in I_j$,

$$\chi'_{nj}(s) = -k^{-1}\int_{I_j} \beta(t_n - t)dt + \beta(t_n - s) \quad \text{and} \quad \chi''_{nj}(s) = -\beta'(t_n - s).$$

Since $\chi_{nj}(t_{j-1}) = \chi_{nj}(t_j) = 0$ we have

$$\chi_{nj}(s) = \tfrac{1}{2}(s - t_{j-1})(s - t_j)\chi''_{nj}(\xi_j) \quad \text{with} \quad \xi_j \in I_j,$$

and hence

$$|\chi_{nj}(s)| \le \tfrac{1}{8}k^2|\beta'(t_n - t_j)| \quad \text{for} \quad s \in I_j.$$

Thus

$$\sum_{j=1}^{n-1} \int_{I_j} |\chi_{nj}(s)|\, ds \le Ck^3 \sum_{j=1}^{n-1} |\beta'(t_n - t_j)| \le Ck^{3/2} \sum_{j=1}^{n-1} (n - j)^{-3/2} \le Ck^{3/2}.$$

Finally, for $s \in I_n$,

$$|\chi_{nj}(s)| \le \int_s^{t_n} |\chi'_{nj}(t)|dt \le k^{-1}\int_{I_n}\int_{I_n} \beta(t_n - s)dsdt + \int_{I_n} \beta(t_n - s)\, ds \le Ck^{1/2},$$

so that

$$\int_{I_n} |\chi_{nj}(s)|ds \le Ck^{3/2}.$$

Together these estimates complete the proof for $q = 0$.

For $q = 1$ we find easily, if $w(t)$ is extended by 0 for $t < 0$,

$$\bar\partial_t G_k w^n = G_k \bar\partial_t w^n \quad \text{and} \quad \bar\partial_t G w(t_n) = G\bar\partial_t w(t_n), \quad \text{for } n \ge 1. \tag{2.1}$$

By the result for $q = 0$, and the assumed regularity of w, in particular, that on I_1, e.g., $|D_t^3 \bar\partial_t w(t)| = k^{-1}|D_t^3 w(t)| \le C\|w\|_{\mathbb{C}^4_{t_1}}$, since $D_t^3 w(0) = 0$,

$$\begin{aligned} &|\bar\partial_t(G_k w^n - G w(t_n))| \\ &= |G_k \bar\partial_t w^n - G\bar\partial_t w(t_n)| \le Ck^{3/2}\|\bar\partial_t w\|_{\mathbb{C}^2_{t_n}} \le Ck^{3/2}\|w\|_{\mathbb{C}^3_{t_n}}, \end{aligned}$$

which completes the proof. □

We note that although the quadrature operator J_k is only first order accurate, because of the nonsymmetric approximation w^j of $w(t)$ on I_j, the operator $G_k = J_k\bar\partial_t$ is of higher order $O(k^{3/2})$, where the loss of $k^{1/2}$ over second order accuracy results from the weak singularity of $\beta(t)$.

We shall also use the following boundedness property of the operator G_k.

Lemma 2.2. *For the quadrature operator G_k of* (1.8), (1.7), *we have*

$$|\bar\partial_t^q G_k w^n| \le Ct_n^{1/2}\|w\|_{\mathbb{C}^{1+q}_{t_n}}, \quad \text{for } n \ge 1+q,\ q = 0, 1,$$

where for $q = 1$ we assume $w^{(l)}(0) = 0$ for $l = 0, 1, 2$.

Proof. We have for the operator J_k in (1.7)

$$|J_k w^n| \le \int_0^{t_n} \beta(t_n - s)\, ds \max_{1\le j\le n} |w^j| = Ct_n^{1/2}\|w\|_{\mathbb{C}^0_{t_n}},$$

and the result for $q = 0$ now follows at once from $G_k w^n = J_k\bar\partial_t w^n$, $n \ge 1$. For $q = 1$ one uses the first part of (2.1), so that $\bar\partial_t G_k w^n = J_k\bar\partial_t^2 w^n$.

In Section 5 we shall also need to use the positive definiteness of the quadrature formula defined by, with J_k and ω_j as in (1.7),

$$q_n(W) = k^{-1/2}\tfrac12 J_k(W^n + W^{n-1}) = \sum_{j=1}^n \sigma_{n-j}W^j, \quad \text{with } J_k W^0 = 0, \tag{2.2}$$
$$\text{where } \sigma_n = \omega_n + \omega_{n-1} \quad \text{for } n \ge 1, \quad b_0 = \omega_0.$$

We say that such a quadrature operator of convolution type is positive definite for $n \le N$ if, for the quadratic form analogous to the double integral in (1.13), we have, cf. [5], for all $W = \{W^j\}_1^N$,

$$B_N(W) = k\sum_{n=1}^N q_n(W)W^n = k\sum_{n=1}^N\sum_{j=1}^n \sigma_{n-j}W^nW^j \ge 0. \tag{2.3}$$

Writing $B_N(W) = \mathcal{B}_N W\cdot W$, where $\mathcal{B}_N$ is a lower triangular $N\times N$ Toeplitz matrix, this condition may also be expressed in terms of the positivity of the symmetrized quadratic form, or

$$\widetilde B_N(W) = \widetilde{\mathcal{B}}_N W\cdot W \ge 0, \quad \forall\, W, \quad \text{where } \widetilde{\mathcal{B}}_N = \mathcal{B}_N + \mathcal{B}_N^T. \tag{2.4}$$

For (2.3) to hold it is therefore sufficient to show that the symmetric Toeplitz matrix $\widetilde{\mathcal{B}}_N$ is positive definite, or that $\underline\lambda_N = \min_j \lambda_j(\widetilde{\mathcal{B}}_N) \ge 0$. It is clear that $\underline\lambda_N$ is a nonincreasing function of N. Hence if $\widetilde{\mathcal{B}}_N$ is positive definite, then we may conclude that $\widetilde{\mathcal{B}}_n$ is positive definite for $n \le N$.

By computation, using Jacobi's method by means of the code written by John Burkardt [1] we could compile Table 1 and hence have the following proposition.

Proposition 2.1. *With $\widetilde{\mathcal{B}}_N$ defined in* (2.4) *we have $\underline\lambda_{10^4} > 0$, and thus $q_n(W)$, as defined in* (2.2), *is positive definite for $n \le 10^4$.*

Table 1: Computed smallest eigenvalues of $\widetilde{\mathcal{B}}_N$.

N	$\underline{\lambda}_N$
10	2.0094×10^{-2}
10^2	2.3066×10^{-4}
10^3	2.3390×10^{-6}
10^4	2.3423×10^{-8}

3. The Forward Euler Method

We begin this section with an abstract stability lemma which will be used for the maximum-norm estimates in this and the next section.

Lemma 3.1. *Let $\{B_n\}_{n\geq 1}$, and $\{F_n\}_{n\geq 0}$ be given sequences of nonnegative real numbers, and assume that the nonnegative sequence $\{\varphi_n\}_{n\geq 0}$ satisfies*

$$\varphi_{n+1} \leq (1+\nu k)\alpha_n \max_{j\leq n} \varphi_j + (1-\alpha_n)B_{n+1} + kF_n, \quad \text{for } n \geq 0,$$

with $\nu \geq 0$ and $0 \leq \alpha_n \leq 1$. Then

$$\varphi_n \leq \max\left(e^{\nu t_n}\varphi_0, \max_{j\leq n}(e^{\nu t_{n-j}}B_j)\right) + k\sum_{j=0}^{n-1} e^{\nu t_{n-1-j}}F_j, \quad \text{for } n \geq 0.$$

Proof. The proof is by induction over n. The result clearly holds for $n = 0$. Assume it holds for some $n \geq 0$. Then

$$\begin{aligned}\varphi_{n+1} &\leq e^{\nu k}\Big(\alpha_n \max\left(e^{\nu t_n}\varphi_0, \max_{j\leq n}(e^{\nu t_{n-j}}B_j)\right) + k\sum_{j=0}^{n-1} e^{\nu t_{n-1-j}}F_j\Big) + (1-\alpha_n)B_{n+1} + kF_n \\ &\leq \max\left(e^{\nu t_{n+1}}\varphi_0, \max_{j\leq n+1}(e^{\nu t_{n+1-j}}B_j)\right) + k\sum_{j=0}^{n} e^{\nu t_{n-j}}F_j.\end{aligned}$$

We now use this lemma to show the following stability result for the forward Euler method.

Theorem 3.1. *Let U^n be the solution of (1.4), (1.5), (1.9), with $b^n = 0$, for $n \geq 1$. Assume that $1/\pi < \lambda_0 \leq \lambda \leq 1/2$ and $\nu > 1$. Then we have, with $C = C(\lambda_0)$,*

$$\|U^n\| \leq Ce^{\nu t_n}\|U^0\| + Ck\sum_{j=0}^{n-1} e^{\nu t_{n-1-j}}\|f^j\|_0 + C\max_{j\leq n}(e^{\nu t_{n-j}}|g^j|).$$

Proof. Given U^n we set $U_0^{n+1} = 0$ and for the interior mesh-points

$$U_m^{n+1} = \lambda U_{m-1}^n + (1-2\lambda)U_m^n + \lambda U_{m+1}^n + kf_m^n, \quad 1 \leq m \leq M.$$

At the right hand boundary the definition (1.9) leads to

$$U_{M+1}^{n+1} = U_M^{n+1} - hG_kU_{M'}^{n+1} + hg^{n+1}. \tag{3.1}$$

We now note that the approximate integral operator may be written as

$$G_kU^{n+1} = k^{-1/2}\Big(\omega_0 U^{n+1} - \sum_{j=1}^{n} d_{n+1-j}U^j - \omega_n U^0\Big), \tag{3.2}$$

where

$$d_j = \omega_{j-1} - \omega_j > 0, \quad \text{for } j \ge 1 \quad \text{and} \quad \sum_{j=1}^{n} d_j + \omega_n = \omega_0 = 2/\sqrt{\pi}. \tag{3.3}$$

Hence we have, with $\gamma = 1/\sqrt{\pi\lambda}$,

$$hG_kU^{n+1} = 2\gamma U^{n+1} - H_kU^{n+1}, \text{ where } H_kU^n = \lambda^{-1/2}\Big(\sum_{j=1}^{n-1} d_{n-j}U^j + \omega_{n-1}U^0\Big).$$

Solving (3.1) for U_{M+1}^{n+1} we obtain, for $\gamma \le 1/\sqrt{\pi\lambda_0} < 1$,

$$(1+\gamma)U_{M+1}^{n+1} = (1-\gamma)U_M^{n+1} + H_kU_{M'}^{n+1} + hg^{n+1}. \tag{3.4}$$

Setting $\mu = (1-\gamma)/(1+\gamma) < 1$, this takes the form

$$U_{M+1}^{n+1} = \mu U_M^{n+1} + W^{n+1}, \quad \text{where } W^n = (1+\gamma)^{-1}\big(H_kU_{M'}^n + hg^n\big). \tag{3.5}$$

To show our stability result we shall make a change of variables by setting

$$V_m^n = e^{-x_m}U_m^n, \quad \widetilde{f}_m^n = e^{-x_m}f_m^n, \quad \text{and} \quad \widetilde{g}^n = e^{-1-h/2}g^n. \tag{3.6}$$

Note that, for small h,

$$\|V^n\| \le \|U^n\| \le e^{1+h/2}\|V^n\| \le 3\|V^n\|, \tag{3.7}$$

and similarly for f^n, $\widetilde{f}^n$ and g^n, $\widetilde{g}^n$. For the interior mesh-points we then have

$$V_m^{n+1} = \lambda e^{-h}V_{m-1}^n + (1-2\lambda)V_m^n + \lambda e^h V_{m+1}^n + k\widetilde{f}_m^n, \quad m = 1, \dots, M,$$

and hence, since $\lambda \le 1/2$, $\cosh h \le 1 + \frac{1}{2}\nu h^2$, for small h, and $\lambda h^2 = k$,

$$|V_m^{n+1}| \le (1 - 2\lambda + 2\lambda\cosh h)\|V^n\| + k\|\widetilde{f}^n\| \le (1+\nu k)\|V^n\| + k\|\widetilde{f}^n\|_0. \tag{3.8}$$

For V_{M+1}^{n+1} (3.5) yields

$$V_{M+1}^{n+1} = \mu e^{-h}V_M^{n+1} + (1+\gamma)^{-1}\Big(\tfrac{1}{2}H_k(V_{M+1}^{n+1} + e^{-h}V_M^{n+1}) + h|\widetilde{g}^{n+1}|\Big).$$

Hence, using (3.3), we have, since $|H_kV^n| \le \lambda^{-1/2}\omega_0 \max_{j\le n-1}|V^j|$ and $\omega_0/(\lambda^{1/2}(1+\gamma)) = 1-\mu$,

$$|V_{M+1}^{n+1}| \le \mu e^{-h}|V_M^{n+1}| + (1-\mu)\max_{j\le n}\|V^j\| + h(1+\gamma)^{-1}|\widetilde{g}^{n+1}|. \tag{3.9}$$

Using (3.8) to bound V_M^{n+1}, and $e^{-h} \le 1 - \frac{1}{2}h$, we find

$$\begin{aligned}|V_{M+1}^{n+1}| &\le (1+\nu k)(\mu e^{-h} + 1 - \mu)\max_{j\le n}\|V^j\| + k\mu e^{-h}\|\widetilde{f}^n\|_0 + h(1+\gamma)^{-1}|\widetilde{g}^{n+1}| \\ &\le (1+\nu k)(1 - \tfrac{1}{2}\mu h)\max_{j\le n}\|V^j\| + \tfrac{1}{2}\mu h\Big(2(1-\gamma)^{-1}|\widetilde{g}^{n+1}|\Big) + k\mu\|\widetilde{f}^n\|_0.\end{aligned}$$

Hence, with $\alpha_n = 1$ if the maximum is taken in the interior, and $\alpha_n = 1 - \frac{1}{2}\mu h$ if it is taken at the right hand boundary point,

$$\|V^{n+1}\| \le (1+\nu k)\alpha_n \max_{j\le n}\|V^j\| + (1-\alpha_n)\Big(2(1-\gamma)^{-1}|\widetilde{g}^{n+1}|\Big) + k\mu\|\widetilde{f}^n\|_0. \tag{3.10}$$

Thus with $\varphi_n = \|V^n\|,\ B_n = 2(1-\gamma)^{-1}|\widetilde{g}^n|,\ F_n = \mu\|\widetilde{f}^n\|_0$, we have

$$\varphi_{n+1} \le (1+\nu k)\alpha_n \max_{j\le n}\varphi_j + (1-\alpha_n)B_{n+1} + kF_n,$$

and by Lemma 3.1 we conclude

$$\|V^n\| \le \max\left(e^{\nu t_n}\|V_0\|, 2(1-\gamma)^{-1}\max_{j\le n}(e^{\nu t_{n-j}}\widetilde{g}^j)\right) + k\sum_{j=0}^{n-1} e^{\nu t_{n-1-j}}\mu\|\widetilde{f}^j\|_0.$$

Using (3.7) and $2(1-\gamma)^{-1} \le C(\lambda_0)$, $\mu \le C(\lambda_0)$, this shows the theorem. □

We remark that without making the transformation of variables (3.6), and taking $f = 0$ for simplicity, the above argument would yield

$$\|U^{n+1}\| \le \max_{j\le n}\|U^j\| + h|g^{n+1}|,$$

or, after repeated application,

$$\begin{aligned}\|U^n\| &\le \|U^0\| + h\sum_{j=1}^{n}|g^j| \le \|U^0\| + nh\max_{j\le n}|g^j| \\ &= \|U^0\| + \lambda^{-1}h^{-1}t_n\max_{j\le n}|g^j|,\end{aligned}$$

which does not show stability in g^n.

We now state and prove our error estimate. Setting $D_x = \partial/\partial x,\ D_t = \partial/\partial t$, and letting $\delta > \frac{1}{2}h$, we use the notation

$$\|u\|_{\mathbb{C}^p(\Omega_t)} = \max_{j+2l\le p}\ \sup_{(x,s)\in\widetilde{\Omega}_\tau}\left|D_x^j D_t^l\, u(x,s)\right|, \quad \text{with } \widetilde{\Omega}_t = (0, 1+\delta)\times(0,t).$$

Theorem 3.2. *Let U^n be the solution of* (1.4), (1.5), (1.9), *and u that of* (1.1). *Assume that $1/\pi < \lambda_0 \le \lambda \le 1/2$. Then we have*

$$\|U^n - u(t_n)\| \le C(t_n,\lambda_0)h^2\|u\|_{\mathbb{C}^4(\widetilde{\Omega}_{t_n})}, \quad \textit{for } n \ge 0.$$

Proof. Setting $\varepsilon_m^n = U_m^n - u_m^n$, where $u_m^n = u(x_m, t_n)$, we have

$$\begin{aligned}
&\partial_t\varepsilon_m^n - \partial_x\overline{\partial}_x\varepsilon_m^n = \tau_m^n, && \text{for } m = 1,\dots,M,\ n \ge 1,\\
&\varepsilon_0^n = 0, && \text{for } n \ge 1,\\
&\partial_x\varepsilon_M^n + G_k\varepsilon_{M'}^n = \psi^n, && \text{for } n \ge 1,\\
&\varepsilon_m^0 = 0, && \text{for } m = 0,\dots,M+1.
\end{aligned}$$

Here

$$\tau_m^n = f_m^n - \partial_t u_m^n + \partial_x\overline{\partial}_x u_m^n = ((u_t)_m^n - \partial_t u_m^n) - ((u_{xx})_m^n - \partial_x\overline{\partial}_x u_m^n),$$

and hence, since $\lambda \le 1/2$,

$$\|\tau^n\|_0 \le C(k+h^2)\|u\|_{\mathbb{C}^4(\Omega_{t_n})} \le Ch^2\|u\|_{\mathbb{C}^4(\widetilde{\Omega}_{t_n})}, \quad \text{for } n \ge 0.$$

Further,

$$\begin{aligned}\psi^n &= g^n - \partial_x u_M^n - G_k u_{M'}^n \\ &= (D_x(1,t_n) - \partial_x u_M(t_n)) + (Gu(1,t_n) - G_k u(1,t_n)) + G_k(u(1,t_n) - u_{M'}(t_n)) \\ &= \psi_1 + \psi_2 + \psi_3. \end{aligned} \tag{3.11}$$

Here, by the symmetry of $[x_M, x_{M+1}]$ around $x = 1$, $|\psi_1| \le Ch^2\|u\|_{\mathbb{C}^3(\widetilde{\Omega}_{t_n})}$. For ψ_2 we have, by Lemma 2.1,

$$|\psi_2| \le Ck^{3/2}\|u(1,\cdot)\|_{\mathbb{C}^2_{t_n}} \le Ch^2\|u\|_{\mathbb{C}^4(\widetilde{\Omega}_{t_n})},$$

and Lemma 2.2, with $q = 0$, shows

$$|\psi_3| = |G_k(u_{M'}(t_n) - u(1,t_n))| \le Ch^2 t_n^{1/2}\|u\|_{\mathbb{C}^4(\widetilde{\Omega}_{t_n})}.$$

As a result we find

$$|\psi^n| \le C(t_n)h^2\|u\|_{\mathbb{C}^4(\widetilde{\Omega}_{t_n})}. \tag{3.12}$$

Applying the stability estimate of Theorem 3.1 to ε_m^n we thus find

$$\begin{aligned}\|U^n - u^n\| &\le C(t_n,\lambda_0)\Big(k\sum_{j\le n-1}\|\tau^j\|_0 + \max_{j\le n}|\psi^j|\Big) \\ &\le C(t_n,\lambda_0)h^2\|u\|_{\mathbb{C}^4(\widetilde{\Omega}_{t_n})},\end{aligned}$$

which completes the proof. □

4. The θ-method with $\theta > 0$

In the θ method (1.12a) with $\theta > 0$, given U^n, U^{n+1} is determined from the system

$$\begin{aligned}-\lambda\theta U_{m-1}^{n+1} + (1+2\lambda\theta)U_m^{n+1} - \lambda\theta U_{m+1}^{n+1} &= \lambda(1-\theta)U_{m-1}^n \\ + (1-2\lambda(1-\theta))U_m^n + \lambda(1-\theta)U_{m+1}^n + kf_m^{n+\theta}, &\quad m = 1,\dots,M.\end{aligned}$$

For $m = 1$ we use $U_0^{n+1} = b^{n+1}$, and for U_{M+1}^{n+1} we still have (3.1), and hence (3.5). After elimination of U_{M+1}^{n+1}, the equation for $m = M$ is therefore

$$\begin{aligned}-\lambda\theta U_{M-1}^{n+1} + (1+(2-\mu)\lambda\theta)U_M^{n+1} &= \lambda(1-\theta)(U_{M-1}^n + U_{M+1}^n) \\ + (1-2\lambda(1-\theta)U_M^n + kf_M^{n+\theta} + \lambda\theta W^{n+1},&\end{aligned}$$

where W^n was defined in (3.5). The system of equations in $(U_1^{n+1},\cdots,U_M^{n+1})$ thus has a tridiagonal matrix, with diagonal elements $1 + 2\lambda\theta$ except the bottom element which is $1 + (2-\mu)\lambda\theta$. Since $\mu < 1$ we see that the matrix is diagonally dominant. In particular, it is nonsingular, and thus the solution U^{n+1} exists for given U^n. Here the stability result reads as follows.

Theorem 4.1. *Let $0 < \theta \le 1$ and let U^n be the solution of* (1.12a)*, with $b^n = 0$ for $n \ge 1$. Assume that $2\lambda(1-\theta) \le 1$, $\lambda \ge \lambda_0 > 1/\pi$, and $\nu > 1$. Then we have, with $C = C(\theta,\lambda_0)$,*

$$\|U^n\| \le Ce^{\nu t_n}\|U^0\| + Ck\sum_{j=0}^{n-1} e^{\nu t_{n-1-j}}\|f^{j+\theta}\|_0 + C\max_{j\le n}(e^{\nu t_{n-j}}|g^j|), \quad \text{for } n \ge 1.$$

Proof. Introducing again the new variables V_m^n and f_m^n as in (3.6), we find this time, for $m = 1, \dots, M$,

$$\begin{aligned} &- \lambda\theta e^{-h} V_{m-1}^{n+1} + (1 + 2\lambda\theta) V_m^{n+1} - \lambda\theta e^{h} V_{m+1}^{n+1} \\ &= \lambda(1-\theta) e^{-h} V_{m-1}^{n} + (1 - 2\lambda(1-\theta)) V_m^n + \lambda(1-\theta) e^{h} V_{m+1}^n + k\widetilde{f}_m^{n+1}. \end{aligned}$$

and hence, if the maximum is attained for an interior point m_0, $1 \le m_0 \le M$, using that $2\lambda(1-\theta) \le 1$,

$$\begin{aligned} (1 + 2\lambda\theta)\|V^{n+1}\| = (1 + 2\lambda\theta)|V_{m_0}^{n+1}| &\le 2\lambda\theta \cosh h \|V^{n+1}\| \\ &\quad + (1 + 2\lambda(1-\theta)(\cosh h - 1)\|V^n\| + k\|\widetilde{f}^{n+\theta}\|_0. \end{aligned}$$

Since $\cosh h \le 1 + \frac{1}{2}\nu' h^2$, with $1 < \nu' < \nu$, for small h, we conclude

$$(1 - \theta\nu' k)\|V^{n+1}\| \le (1 + (1-\theta)\nu' k)\|V^n\| + k\|\widetilde{f}^{n+\theta}\|_0.$$

Since, for small k, $(1 - \theta\nu' k)^{-1}(1 + (1-\theta)\nu' k) \le 1 + \nu k$, we have then

$$\|V^{n+1}\| \le (1 + \nu k)(\|V^n\| + 2k\|\widetilde{f}^{n+\theta}\|_0). \tag{4.1}$$

If the maximum of V_m^{n+1} is instead taken for $m = M+1$, (3.9) remains valid, and using (4.1) to bound the first term on the right in (3.9) as before, that (3.10) holds, and the proof is concluded as in Theorem 3.1, using Lemma 3.1. □

The error estimate in this case is the following.

Theorem 4.2. *Let $0 < \theta \le 1$ and let U^n be the solution of* (1.12a)*, and u that of* (1.1)*. Assume that $2\lambda(1-\theta) \le 1$ and $\lambda \ge \lambda_0 > 1/\pi$. Then we have*

$$\|U^n - u(t_n)\| \le C(t_n, \theta, \lambda_0)(h^2 + k)\|u\|_{\mathbb{C}^4(\widetilde{\Omega}_{t_n})}.$$

Proof. Here the truncation error in the difference equation is

$$\left|\tau_m^{n+\theta}\right| = \left|f_m^{n+\theta} - \partial_t u_m^n + \partial_x\bar{\partial}_x(\theta u_m^{n+1} + (1-\theta)u_m^n)\right| \le C(h^2 + k)\|u\|_{\mathbb{C}^4(\widetilde{\Omega}_{t_{n+1}})},$$

and ψ_j^n is again defined by (3.11) so that (3.12) holds. Applying Theorem 4.1 to $\varepsilon_m^n = U_m^n - u_m^n$ therefore shows that

$$\|U^n - u^n\| \le C(t_n)\left(k\sum_{j=1}^{n-1}\|\tau^{j+\theta}\|_0 + \max_{j\le n}|\psi^j|\right) \le C(t_n)(h^2 + k)\|u\|_{\mathbb{C}^4(\widetilde{\Omega}_{t_n})},$$

which shows our claim. □

5. The Crank-Nicolson Method

In this section we shall be concerned with the Crank-Nicolson method

$$\partial_t U_m^n - \partial_x\bar{\partial}_x\widehat{U}_m^n = f_m^{n+\frac{1}{2}}, \quad m = 1, \dots, M,\ n \ge 0, \tag{5.1}$$

where $\widehat{U}^n = \frac{1}{2}(U^n + U^{n+1})$, i.e., (1.12a) with $\theta = \frac{1}{2}$. We again use the initial and boundary conditions in (1.12a), in particular, on the right hand side

$$\partial_x U_M^n + G_k U_{M'}^n = g^n, \quad \text{for } n \ge 1, \quad \text{where } G_k = J_k \bar{\partial}_t. \tag{5.2}$$

In our numerical approximation of (1.2), $g^n = 0$, but the inhomogeneity is included in (5.2) for the purpose of our error analysis.

To illustrate our approach we begin by proving a stability estimate for the continuous problem (1.2), with $b(t) = v(x) = 0$. We recall that the kernel of J is positive definite so that (1.13) holds. Let

$$(v, w) = \int_0^1 vw\, dx, \quad \|v\|_{L_2} = (v, v)^{1/2}, \quad \Omega_t = (0, 1) \times (0, t).$$

Proposition 5.1. *Let u be the solution of* (1.2) *with $b(t) = v(x) = g(0) = 0$. Then*

$$\|u_t\|_{L_2(\Omega_t)} + \|u_x(\cdot, t)\|_{L_2} \le C(t)(\|f\|_{L_2(\Omega_t)} + \|g_t\|_{L_2(0,t)}), \quad \textit{for } t \ge 0.$$

Proof. Multiplying the differential equation by u_t and integrating in x we have

$$\|u_t\|_{L_2}^2 - (u_t, u_{xx}) = (u_t, f),$$

or, integrating by parts and using the boundary condition at $x = 1$,

$$\begin{aligned} \|u_t\|_{L_2}^2 + \tfrac{1}{2}\frac{d}{dt}\|u_x\|_{L_2}^2 &= u_t(1, t)\, u_x(1, t) + (u_t, f) \\ &= -u_t(1, t)\, Ju_t(1, t) + u_t(1, t)\, g(t) + (u_t, f). \end{aligned}$$

Our claim now easily follows, after integration in t, and using that by (1.13), with $w(t) = u_t(1, t)$, the integral of the third to last term is nonpositive, so that

$$\int_0^t \|u_t(s)\|_{L_2}^2 ds + \tfrac{1}{2}\|u_x(t)\|_{L_2}^2 \le \int_0^t u_t(1, s)\, g(s) ds + \int_0^t (u_t(s), f(s)) ds.$$

Here

$$\begin{aligned} \int_0^t u_t(1, s)\, g(s) ds &= u(1, t)\, g(t) - \int_0^t u(1, s)\, g_t(s) d \\ &\le \tfrac{1}{4}u(1, t)^2 + g(t)^2 + \int_0^t g_t(s)^2 ds + \int_0^t u(1, s)^2 ds. \end{aligned}$$

Using $|u(1, t)| \le \|u_x(t)\|_{L_2}$, we conclude

$$\|u_t\|_{L_2(\Omega_t)}^2 + \|u_x(t)\|_{L_2}^2 \le C\bigg(\|f\|_{L_2(\Omega_t)}^2 + (1 + t)\|g_t\|_{L_2(0,t)}^2 + \int_0^t \|u_x(s)\|^2 ds\bigg),$$

and the proposition now follows by Gronwall's lemma. □

We begin our analysis of the discrete problem with the following discrete version of Green's formula. We use the notation

$$(V, W)_{m_0, m_1} = h \sum_{m=m_0}^{m_1} V_m W_m \quad \text{and} \quad \|V\|_{m_0, m_1} = (V, V)_{m_0, m_1}^{1/2}.$$

Lemma 5.1. *If $U_0 = V_0 = 0$, we have*

$$(\bar\partial_x \partial_x U, V)_{1,M} = -(\partial_x U, \partial_x V)_{0,M} + \partial_x U_M \, V_{M+1}.$$

Proof. We have, if $V_0 = 0$,

$$\begin{aligned}(\bar\partial_x U, V)_{1,M} &= h \sum_{m=1}^{M} \bar\partial_x U_m V_m = -\sum_{m=0}^{M} U_m (V_{m+1} - V_m) + U_M V_{M+1} \\ &= -(U, \partial_x V)_{0,M} + U_M V_{M+1}.\end{aligned}$$

Hence the lemma follows by replacing U by $\partial_x U$.

We now turn to a stability result for (5.1), which can be though of as a discrete analogue of Proposition 5.1. We write

$$\begin{aligned}(V, W)_n &= k \sum_{j=0}^{n} V^j \, W^j, \\ (V, W)_{m_0, m_1; n} &= \sum_{m=m_0}^{m_1} (V_m, W_m)_n = kh \sum_{m=m_0}^{m_1} \sum_{j=0}^{n} V_m^j W_m^j,\end{aligned}$$

and correspondingly for the norms $\|\cdot\|_n$ and $\|\cdot\|_{m_0,m_1;n}$.

Theorem 5.1. *Let U be the solution of* (5.1) *and* (5.2)*, with $b(t) = v(x) = 0$, and set $g^0 = g^{-1} = 0$. Assume that $q_n(W)$ as defined in* (2.2) *is positive definite for $n \le N$. Then we have*

$$\begin{aligned}&\|\partial_t U\|_{1,M+1;n-1} + \|\partial_x U^n\|_{0,M} \\ &\le C(t_n) \Big(\|f^{\cdot + \frac12}\|_{1,M;n-1} + \|\bar\partial_t g\|_n \Big), \qquad 1 \le n \le N.\end{aligned}$$

Proof. Multiplying (5.1) by V_m^n, $m = 1, \dots, M$, and using Lemma 5.1 we find, with $\widehat U^n = \frac12 (U^n + U^{n+1})$, for $n \ge 0$,

$$(\partial_t U^n, V^n)_{1,M} + (\partial_x \widehat U^n, \partial_x V^n)_{0,M} = \partial_x \widehat U_M^n \, V_{M+1}^n + (f^{n+\frac12}, V^n)_{1,M}.$$

Choosing $V_m^n = \partial_t U_m^n, m = 1, \dots, M$, and $V_{M+1}^n = \partial_t U_{M'}^n$, we get

$$\begin{aligned}&\|\partial_t U^n\|_{1,M}^2 + \tfrac12 \partial_t \|\partial_x U^n\|_{0,M-1}^2 + \tfrac14 \, h \, \partial_t (\partial_x U_M^n)^2 \\ &= \partial_x \widehat U_M^n \, \partial_t U_{M'}^n + (f^{n+\frac12}, \partial_t U^n)_{1,M}.\end{aligned}$$

After multiplication by k and summation in n, from 0 to $n-1$, we obtain

$$\tfrac12 \|\partial_t U\|_{1,M;n-1}^2 + \tfrac14 \|\partial_x U^n\|_{0,M}^2 \le (\partial_x \widehat U_M, \, \partial_t U_{M'})_{n-1} + C \|f^{\cdot+\frac12}\|_{1,M;n-1}^2. \tag{5.3}$$

The boundary condition (5.2) for $x = 1$ now yields, with $J_k W^0 = 0$,

$$(\partial_x \widehat U_M, \partial_t U_{M'})_{n-1} = -(J_k \bar\partial_t \widehat U_{M'}, \partial_t U_{M'})_{n-1} + (\widehat g, \partial_t U_{M'})_{n-1}. \tag{5.4}$$

Here, by (2.2), for $j \ge 1$ and $U^j = U_{M'}^j$,

$$J_k \bar\partial_t \widehat U^j = J_k \partial_t \widehat U^{j-1} = \tfrac12 J_k \partial_t (U^j + U^{j-1}) = k^{1/2} q_j(\partial_t U),$$

and hence, since $J_k W^0 = 0$, by (2.3),

$$(J_k \bar\partial_t \widehat U, \partial_t U)_{n-1} = k^{1/2} \sum_{j=1}^{n-1} q_j(\partial_t U)\, \partial_t U^j = k^{1/2} B_{n-1}(\partial_t U) \ge 0.$$

Thus, the first term on the right in (5.4) is nonpositive,

We now turn to the last term in (5.4) and note that with $\bar\partial_t g^0 = 0$ and arbitrary $\varepsilon > 0$,

$$\begin{aligned}|(\widehat g,\, \partial_t U_{M'})_{n-1}| &= |\widehat g^{n-1}\, U^n_{M'} - (\bar\partial_t \widehat g, U_{M'})_{n-1}| \\ &\le C_\varepsilon\big(|\widehat g^{n-1}|^2 + \|\bar\partial_t \widehat g\|^2_{n-1} + \|U_{M'}\|^2_{n-1}\big) + \tfrac18 |U^n_{M'}|^2.\end{aligned}$$

We now observe that $|U^j_{M'}| \le \|\partial_x U^j\|_{0,M}$ and hence that (5.3) yields

$$\|\partial_t U\|^2_{1,M;n-1} + \|\partial_x U^n\|^2_{0,M} \le C\bigg(\|f^{\cdot+\frac12}\|^2_{1,M;n-1} + \|\bar\partial_t g\|^2_n\bigg) + C\|\partial_x U\|^2_{0,M;n}.$$

An application of the discrete Gronwall's lemma bounds the last term and thus completes the proof. □

We are now ready for our maximum-norm error estimate.

Theorem 5.2. *Let U be the solution of* (5.1) *and* (5.2) *and u that of* (1.2), *both with $g(t) = 0$. Assume that $q_N(w)$ as defined in* (2.2) *is positive definite for $n \le N$. We then have*

$$\|U^n - u(t_n)\| \le C(u, t_n)\,(h^2 + k^{3/2}), \quad \textit{for } 0 \le n \le N.$$

Proof. Let $\varepsilon^n_m = U^n_m - u(x_m, t_n)$ be the error, and note that $\varepsilon^n_m = \varepsilon^n_0 = 0$ for $m = 0, \dots, M+1$, $n \ge 0$. We have for the truncation errors, for $n \ge 0$,

$$\big|\tau^{n+\frac12}_m\big| = \big|\partial_t \varepsilon^n_m - \partial_x \bar\partial_x \widehat\varepsilon^n_m)\big| = \big|f^{n+\frac12}_m - (\partial_t u^n_m - \partial_x \bar\partial_x \widehat u^n_m)\big| \le C(u)(h^2 + k^2),$$

and, for $n \ge 1$,

$$\begin{aligned}\psi^n &= \partial_x \varepsilon^n_M + G_k \varepsilon^n_{M'} = g^n - \partial_x u^n_M - G_k u^n_{M'} \\ &= (u_x(1,t_n) - \partial_x u^n_M) + (Gu(1,t_n) - G_k u(1,t_n)) + (G_k(u(1,t_n) - u_{M'}(t_n))) \\ &= \psi^n_1 + \psi^n_2 + \psi^n_3.\end{aligned}$$

Setting $\psi^0_l = \psi^{-1}_l = 0$ for $l = 1, 2, 3$, we have $\bar\partial_t \psi^n_1 = \bar\partial_t u_x(1, t_n) - (\partial_x \bar\partial_t u)^n_M$ for $n \ge 1$, and hence

$$\|\bar\partial_t \psi_1\|_n \le Ch^2 \|\bar\partial_t u\|_{\mathbb{C}^2(\Omega_{t_n})} \le Ch^2 \|u\|_{\mathbb{C}^4(\Omega_{t_n})}, \quad \text{for } n \ge 1.$$

Further, using Lemma 2.1, with $q = 1$, noting that $D^l_t u(1,0) = 0$ for any $l \ge 0$, we find

$$\|\bar\partial_t \psi_2\|_n \le Ck^{3/2} \|u(1)\|_{\mathbb{C}^3_{t_n}}, \quad \text{for } n \ge 1,$$

and, using Lemma 2.2, with $q = 1$,

$$\|\bar\partial_t \psi_3\|_n \le Ct_n^{1/2} \|u(1) - u_{M'}\|_{\mathbb{C}^2_{t_n}} \le Ct_n^{1/2} h^2 \|u\|_{\mathbb{C}^4(\Omega_{t_n})}, \quad \text{for } n \ge 1.$$

Theorem 5.1, applied to ε^n_m, therefore shows, since $\|\varepsilon^n\| \le \|\partial_x \varepsilon^n\|_{0,M}$, that

$$\begin{aligned}\|U^n - u^n\| \le \|\partial_x \varepsilon^n\|_{0,M} &\le C\bigg(\|\tau^{\cdot+\frac12}\|_{1,M;n-1} + \|\bar\partial_t \psi\|_n\bigg) \\ &\le C(u, t_n)(h^2 + k^{3/2}),\end{aligned}$$

which completes the proof. □

6. Numerical Examples

In this section we shall illustrate our convergence results in Theorems 3.2, 4.2 and 5.2. For our test problem we consider (1.2) with $f(x,t) = v(x) = g(t) = 0$, and

$$b(t) = 1 - erf(\frac{1}{2\sqrt{t}}), \quad \text{where } \; erf(y) = \frac{2}{\sqrt{\pi}} \int_0^y e^{-s^2} ds.$$

By Laplace transformation the exact solution is then easily seen to be

$$u(x,t) = 1 - erf\left(\frac{1+x}{2\sqrt{t}}\right), \quad \text{for } x \in (0,1), \; t > 0, \tag{6.1}$$

which, modulo a trival translation, is the exact solution used for the same problem by Han and Huang [3] and Wu and Sun [7]. In the tables below the errors $\varepsilon = \{\varepsilon_m\}_{m=0}^{M+1}$ are measured in the norms $\|\varepsilon\|_{\ell^\infty} = \max_{0\le m\le M+1} |\varepsilon_m|$ and $\|\varepsilon\|_{\ell^2} = (h\sum_{m=0}^{M+1} |\varepsilon_m|^2)^{1/2}$.

Since we have chosen $h = 1/M'$, where $M' = M + \frac{1}{2}$ with M the number of interior spatial mesh-points, we shall consider refinements of the mesh which essentially treble the number of mesh-points, by replacing M by $M_1 = 3M + 1$, corresponding to choosing the new mesh width $h_1 = 1/((3M+1) + \frac{1}{2}) = h/3$, and then check if the error scales with factors of $3^2, 3^{3/2}, 3$, for orders of convergence 2, 3/2 and 1. For the forward Euler (FE) method we choose the time step $k = \frac{1}{2}(M+1)^{-2}$, so that λ is just below 1/2. Table 1 lists the results for this case. The maximum-norm error between the exact solution and the numerical solution is evaluated at $t = 1$ and displayed in the third column, while the last column shows the ratio of the errors. Note that they are approaching the value of 3^2 confirming the second order convergence of Theorem 3.2.

For the backward Euler (BE) and Crank-Nicolson (CN) methods we choose $k = 1/M \approx h$ and check that the error behaves as the predicted $O(h)$ and $O(h^{3/2})$, or that it scales as 3 and $3^{3/2} \approx 5.2$, respectively. As in the FE case the numerical solution is computed at $t = 1$, and the errors in the ℓ^∞ and ℓ^2 norms are evaluated. These are provided in Table 3, along with the ratio of errors. Modulo the first entry in the CN case the ratios clearly confirm the orders of convergence of Theorems 4.2 and 5.2 above. It is surprising to note that the CN error is large

Table 2: Errors for FE with $h = 1/M' \;\; k = \frac{1}{2}(M+1)^{-2}$.

M	λ	FE ℓ^∞ error	ratio
10	0.4556	8.20×10^{-5}	***
31	0.4845	8.86×10^{-6}	9.26
94	0.4948	9.74×10^{-7}	9.10
283	0.4982	1.08×10^{-7}	9.02
850	0.4994	1.20×10^{-8}	9.00

Table 3: Errors for BE and CN with $h = 1/M', \;\; k = 1/M$.

M	BE ℓ^∞ error	ratio	CN ℓ^∞ error	ratio	CN ℓ^2 error	ratio
10	8.88×10^{-4}	***	1.73×10^{-3}	***	6.10×10^{-4}	***
31	2.99×10^{-4}	2.97	4.08×10^{-5}	42.37	3.20×10^{-5}	18.78
94	1.05×10^{-4}	2.85	8.36×10^{-6}	4.88	6.42×10^{-6}	4.98
283	3.61×10^{-5}	2.91	1.67×10^{-6}	5.01	1.27×10^{-6}	5.06
850	1.22×10^{-5}	2.96	3.29×10^{-7}	5.08	2.50×10^{-7}	5.08

Table 4: Computing (CPU) times in seconds.

M	FE	BE	CN
10	3.37×10^{-3}	3.61×10^{-3}	3.54×10^{-3}
31	1.19×10^{-2}	3.52×10^{-3}	3.71×10^{-3}
94	6.20×10^{-1}	4.05×10^{-3}	3.94×10^{-3}
283	4.78×10^{1}	7.56×10^{-3}	8.11×10^{-3}
850	4.86×10^{3}	3.85×10^{-2}	4.36×10^{-2}

compared to the BE error for $M = 10$. The CN errors behave slightly better in the ℓ^2 norm as shown in Table 3 but the first entry continues to be large. Such a phenomenon has also been reported by Faragó and Kovács [2] for the heat equation with smooth initial condition and zero Dirichlet boundary conditions; the first line in their Table 16 shows an error in the maximum-norm for CN which is much greater than that for BE. As in our case it is only their first entry that plays the spoilsport. Han and Huang [3] used the CN method with $k = h$ and found errors indicating $O(h)$ convergence in the energy norm, and, for their scheme, Wu and Sun [7] show maximum-norm errors consistent with their $O(h^2 + k^{3/2})$ error bounds. Table 4 shows the computational effort required in terms of CPU time, and emphasizes that the stability requirements makes the FE method inadequately slow compared to the BE and CN methods.

References

[1] J. Burkardt, *http://people.sc.fsu.edu/ jburkardt.*

[2] I. Faragó and M. Kovács, On maximum norm contractivity of second order damped single step methods, *Calcolo*, **40** (2003), 91-108.

[3] H.D. Han and Z.Y. Huang, A class of artificial boundary conditions for heat equation is unbounded domains, *Comput. Math. Appl.*, **43** (2002), 889–900

[4] J -R. Li and L. Greengard, On the numerical solution of the heat equation I: fast solvers in free space, *J. Comput. Physics.*, **226** (2007), 1891-1901.

[5] W. McLean and V. Thomée, Numerical solution of an evolution equation with a positive type memory term, *J. Austral. Math. Soc. Ser. B.*, **35** (1993), 23-70.

[6] S.V. Tsynkov, Numerical solution of problems on unbounded domains, A review, *Appl. Numer. Math.*, **27** (1998), 465–532.

[7] X. Wu and Z.Z. Sun, Convergence of difference scheme for heat equation is unbounded domains using artificial boundary conditions, *Appl. Numer. Math.*, **50** (2004), 261–277

[8] C. Zheng, Approximation, stability and fast evaluation of exact artificial boundary condition for the one-dimensional heat equation. *J. Comput. Math.* **25** (2007), 730-745.