

On the maximum norm contractivity of second order damped single step methods

István Faragó *

Eötvös Loránd University,
Department of Applied Analysis,
Pázmány Péter sétány 1/c, H-1117 Budapest, Hungary
email address: faragois@cs.elte.hu

Mihály Kovács

Louisiana State University,
Department of Mathematics,
Baton Rouge, LA 70803
email address: kmisi@math.lsu.edu

*Corresponding author

Abstract

In this paper we consider $A(\theta)$ -stable finite difference methods for numerical solutions of dissipative partial differential equations of parabolic type. Combining two rational approximation methods with different orders of accuracy, where the lower order method is applied n_0 times (n_0 fixed) at each time step, we prove the existence of a second order method which is contractive for all time steps. Moreover, we shed light on the conditions on the lower order method which are sufficient (and sometimes necessary) to obtain the optimal order of accuracy. For the one-dimensional heat equation we construct a family of numerical methods which are contractive in the maximum norm for all values of the discretization parameters. We also present numerical examples to illustrate our results.

1 Introduction

In this paper we consider $A(\theta)$ -stable finite difference methods for numerical solutions of dissipative partial differential equations of parabolic type. Under homogeneous boundary conditions the classical solutions satisfy the inequality $\max_x |u(t, x)| \leq \max_x |u_0(x)|$ for all $t > 0$, where $u_0(\cdot)$ denotes the initial function. It is natural to require from a numerical approximation method to preserve this property. Namely, denoting by y^j the approximation to the solution at the j th time level, we say that an approximation scheme is contractive if $\|y^j\|_\infty \leq \|y^0\|_\infty$ for all j , where y^0 is an approximation to the initial function in \mathbf{R}^s for some $s \in \mathbf{N}$, and $\|y\|_\infty := \max\{|y_i|, i = 1, \dots, s\}$. Clearly, if an approximation scheme is strongly contractive, i.e., $\|y^{j+1}\|_\infty \leq \|y^j\|_\infty$ for all j , then it is also contractive.

Let $\mu := \frac{\tau}{h^2}$, where τ is the time step and h the space discretization parameter. For one-step finite difference methods which are based on some rational approximation of the exponential function e^{-z} , Spijker has shown in [12] that there is an order barrier: only methods with first order of accuracy can be contractive in the maximum norm for all $\mu > 0$. (Such a method is the backward Euler method with approximating functions $r_{be}(z) = (1 + z)^{-1}$.) For higher order methods it is necessary to restrict the choice of μ (some upper bound) in order to preserve the maximum norm contractivity. For example, if the second order Crank-Nicolson method $r_{cn}(z) = (1 - \frac{z}{2})(1 + \frac{z}{2})^{-1}$ is applied to the one-dimensional heat equation, then the sharp restriction is $\mu \leq 1.5$ (see [5],[7],[8]). Therefore, the use of the Crank-Nicolson method

requires the choice of a very small time step τ in case of a small space discretization parameter h if we want to preserve the maximum norm contractivity. However, as we will see later, the choice $\mu \leq 1.5$ requires not only considerable computational efforts but also results in an essential loss of accuracy.

Our aim is to construct, for all fixed $h > 0$, a second order method for the one-dimensional heat equation which is contractive in the maximum norm for all $\tau > 0$. Clearly, due to the result of Spijker, such a method cannot be based solely on one rational approximation of the exponential function.

For construction of higher order unconditional contractive Runge-Kutta type methods we refer to [1] and [2].

Our approach follows the one of Luskin and Rannacher who introduced a second order stable approximation method with optimal convergence properties by combining the robust stability and approximation property of the backward Euler method with the second order accuracy of the Crank-Nicolson method (see [9], [11]). This result was generalized by Hansbo in [6] to Banach spaces (see below Theorem 1.1). In these works, qualitative properties (such as the maximum norm contractivity) were not considered. This will be done in Section 3 and Section 4 of this paper which is organized as follows.

In Section 2 we investigate pairings of $A(\theta)$ -stable approximation methods of the exponential function, where the lower order method is applied n_0 times before the higher order method takes over. We recall that a rational function $r(z)$ is $A(\theta)$ -stable (see [13]) if, for some $\theta \in (0, \frac{\pi}{2}]$,

$$|r(z)| \leq 1, \text{ for } |\arg(z)| \leq \theta.$$

We say that $r(z)$ approximates e^{-z} of order $q \geq 1$ if

$$r(z) = e^{-z} + O(z^{q+1}), \text{ as } z \rightarrow 0.$$

We will denote $A(\theta)$ -stable rational functions of approximation order q by $r_q(z)$. We also recall the definition of a sectorial operator. Let X be a complex Banach space and $A : X \rightarrow X$ a linear operator with domain $\text{dom}(A) \subset X$. We denote by $\mathcal{B}(X)$ the space of bounded linear operators on X . If the resolvent set

$$\rho(A) := \{z \in \mathbf{C} : (zI - A)^{-1} \text{ exists in } \mathcal{B}(X)\}$$

is such that

$$\rho(A) \supset \Sigma_\delta := \{z \in \mathbf{C} : \frac{\pi}{2} - \delta \leq |\arg(z)| \leq \pi, z \neq 0\} \cup \{0\} \quad (1)$$

for some $\delta \in (0, \frac{\pi}{2})$ and the resolvent, $R(z, A) := (zI - A)^{-1}$, satisfies

$$\|R(z, A)\| \leq \frac{M}{|z|}, \text{ for all } z \in \Sigma_\delta, z \neq 0 \quad (2)$$

with some $M \geq 1$, then the operator A is called a *sectorial operator of angle δ* . Let us denote the semigroup generated by $-A$ by $T(t)$. The basis for our investigation is a result of Luskin and Rannacher which was proven in the following general form by Hansbo in ([6]).

Theorem 1.1 *Let X be a complex Banach space, $A : X \rightarrow X$ a sectorial operator of angle δ , and let $r_p(z)$ and $r_{p-1}(z)$ be $A(\theta)$ -stable ($\theta \in (\frac{\pi}{2} - \delta, \frac{\pi}{2})$) rational approximations of order p and $p - 1$ respectively. If $r_{p-1}(\infty) = 0$, then*

$$\|r_p^{n-p}(\tau A)r_{p-1}^p(\tau A)x - T(n\tau)x\| \leq \frac{M}{n^p}\|x\|$$

for all $x \in X$, $\tau > 0$ and $p \leq n \in \mathbf{N}$.

We will show in Section 2 that the condition $r_{p-1}(\infty) = 0$ is more than solely a technical condition. It means that the lower order approximation method is a smoothing starting method in the sense that it maps every initial data into the domain of the operator. In a special case we prove that this is also a necessary condition to obtain the optimal accuracy (which is p).

In Section 3 we focus our attention on constructing a second order scheme, which is contractive for all time steps, by using the backward Euler scheme as the smoothing starting scheme.

In Section 4, for the numerical solution of the one-dimensional heat equation, we analyze the combination of the Crank-Nicolson and the backward Euler methods. We examine the relation between the number of the backward Euler steps, the space discretization parameter and the possible choices of the time step. We show that for any fixed h , there exists a suitable combined method which is contractive in the maximum norm for all $\tau > 0$. In Section 5 we also provide numerical examples to illustrate the advantages of the combined method.

2 Range preservation

Let us consider the initial value problem

$$\dot{u}(t) + Au(t) = 0, \quad t \geq 0, \quad u(0) = u_0 \in X. \quad (3)$$

If A is sectorial with dense domain $\text{dom}(A)$, then the operator $-A$ generates a uniformly bounded analytic semigroup $T(t) = e^{-tA}$ and the solution to (3) can be obtained as $u(t) = T(t)u_0$ (see [4]).

In the following we investigate a class of approximation schemes for the semigroup $T(t)$ which are based on the combination of two different rational approximation methods. Let n_0, p, q be positive integers, $p \geq q$, and

$$V_n(z) := r_p^{n-n_0}(z)r_q^{n_0}(z), \quad n \geq n_0.$$

The operators $V_n(\tau A)$ can be viewed as numerical schemes for the solution of (3). It is known that for some higher order schemes (e.g. Crank-Nicolson scheme) optimal error estimates are only available for sufficiently smooth initial data. If $|r_p(\infty)| = 1$, then we have q -th order error estimate only if the initial data belongs to $\text{dom}(A^q)$ [13]. This deficiency can be resolved by considering combined schemes of the form $V_n(z)$, where error estimations are available for all initial data (see Theorem 1.1). We need the following variant of Theorem 1.1 whose proof is essentially the same as the one of Theorem 1.1 given in [6].

Proposition 2.1 *If we replace $r_p^{n-p}(z)r_{p-1}^p(z)$ by $r_p^{n-n_0}(z)r_{p-1}^{n_0}(z)$ with $n_0 \geq p$, then the same estimate holds as in Theorem 1.1.*

The condition $r_{p-1}(\infty) = 0$ in Theorem 1.1 seems to be solely technical. To show that this is not the case, we first recall the Dunford functional calculus for closed operators (see eg. [3]). Let $A : X \rightarrow X$ with $\text{dom}(A) \subset X$ be a closed operator with $\rho(A) \neq \emptyset$. Let $f(z)$ be analytic on $V \cup \partial V$, where V is an open neighborhood of the spectrum $\sigma(A) := \mathbf{C} - \rho(A)$ of A , and at infinity. Assume further that ∂V consists of a finite number of Jordan arcs and has a positive orientation with respect to the (possibly unbounded) set V . Then

$$f(A) := f(\infty)I + \frac{1}{2\pi i} \int_{\partial V} f(z)R(z; A)dz$$

is well defined, $f(A) \in \mathcal{B}(X)$, and we obtain an algebra homeomorphism $\phi : \mathcal{F}(A) \rightarrow \mathcal{B}(X)$ by setting $\phi(f) = f(A)$, where $\mathcal{F}(A)$ is the set of all functions that are analytic on some open neighborhood of $\sigma(A)$ and at infinity. We note that if $r(z)$ is an $A(\theta)$ -stable rational approximation to the exponential function, then $r^n(\tau z) \in \mathcal{F}(A)$ for all $n \in \mathbf{N}$ and $\tau > 0$.

Definition 2.1 *Let A be a sectorial operator. We say that an approximation $r(\tau A)$ is range preserving if for every $\tau > 0$ the inclusion $\text{ran}(r(\tau A)) \subset \text{dom}(A)$ holds.*

If $-A$ generates an analytic C_0 semigroup $T(t)$, then $\text{ran}(T(t)) \subset \text{dom}(A)$, for all $t > 0$. It is our point of view that "good" numerical schemes should preserve "good" qualitative properties of the semigroup they approximate. In the following we discuss range preserving schemes.

Proposition 2.2 *Let A be an unbounded closed operator with a nonempty resolvent set and $f(z) \in \mathcal{F}(A)$. Then the following statements are equivalent.*

- (i) $\lim_{|z| \rightarrow \infty} f(z) = 0$.
- (ii) $\text{ran}(f(A)) \subset \text{dom}(A)$.

Proof. (i) \Rightarrow (ii). From assumption (i) it follows that

$$f(A) = \frac{1}{2\pi i} \int_{\partial V} f(z)R(z, A)dz.$$

The function $z \rightarrow Af(z)R(z, A) = f(z)(zR(z, A) - I)$ is continuous along the rectifiable path ∂V . Since A is closed we have that

$$\frac{1}{2\pi i} \int_{\partial V} f(z)R(z, A)dz \in \text{dom}(A),$$

and $A \frac{1}{2\pi i} \int_{\partial V} f(z)R(z, A)dz = \frac{1}{2\pi i} \int_{\partial V} Af(z)R(z, A)dz$. This implies $\text{ran}(f(A)) \subset \text{dom}(A)$.

(ii) \Rightarrow (i). It follows from assumption (ii) that

$$\begin{aligned} Af(A) &= f(\infty)A + A \frac{1}{2\pi i} \int_{\partial V} f(z)R(z, A)dz \\ &= f(\infty)A + \frac{1}{2\pi i} \int_{\partial V} f(z)(zR(z, A) - I)dz. \end{aligned} \tag{4}$$

Since A is closed, $f(A) \in \mathcal{B}(X)$, and $f(A)X \subset \text{dom}(A)$ we have that $Af(A)$ is closed and everywhere defined. By the closed graph theorem, $Af(A) \in \mathcal{B}(X)$. But

$$\frac{1}{2\pi i} \int_{\partial V} f(z)(zR(z, A) - I)dz \in \mathcal{B}(X),$$

and thus it follows from (4) that $f(\infty)A \in \mathcal{B}(X)$. Thus, $f(\infty) = 0$

Corollary 2.1 *If $r(z)$ is a rational approximation to the exponential function, then $r(\tau A)$ is range preserving if and only if the degree of the denominator of $r(z)$ is greater than the degree of the numerator.*

For example, the backward Euler scheme is range preserving and the Crank-Nicolson scheme is not. Thus, the condition $r_{p-1}(\infty) = 0$ expresses a nice qualitative property of the starting method. In view of Proposition 2.2 we call a rational approximation method *smoothing* or *range preserving* if $r(\infty) = 0$. We also remark that a method with the latter property is sometimes also called *L-stable*. The question arises naturally if the range preservation property of the starting scheme is also necessary to obtain an optimal error estimation for the damped method. In the following special case we show that the answer is yes.

Theorem 2.1 *Let X be a Hilbert space and $T(t) = e^{-tA}$, where A an unbounded densely defined positive definite operator with a compact resolvent. If*

$$\|(r_p^{n-p}(\tau A)r_{p-1}^p(\tau A) - T(n\tau))\| \leq \frac{M}{n^p} \quad (5)$$

and $|r_p(\infty)| = 1$, then $r_{p-1}(\tau A)$ is range preserving.

Remark 2.1 *If $|r_p(\infty)| < 1$, then we do not need any steps with the lower order scheme (see [13]).*

Proof. By the spectral theorem and Parseval's identity, the estimate in (5) is equivalent to the condition

$$|n^p(r_p^{n-p}(\lambda)r_{p-1}^p(\lambda) - e^{-n\lambda})| \leq M \quad (6)$$

for all $\lambda \in \sigma(A)$. Let $\lambda_n \in \sigma(A)$ be such that $\lim_{n \rightarrow \infty} \lambda_n = +\infty$, and $|r_p^{n-p}(\lambda_n)| > \frac{1}{2}$ (since $|\lim_{|z| \rightarrow \infty} r_p(z)| = 1$, such λ_n exist). Using the fact that $e^{-n\lambda_n}$ is bounded, by substituting λ_n into (6) we have

$$|n^p r_p^{n-p}(\lambda_n) r_{p-1}^p(\lambda_n)| \leq C$$

for some $C > 0$. By the properties of λ_n we obtain

$$|n^p r_p^{n-p}(\lambda_n) r_{p-1}^p(\lambda_n)| \geq \frac{1}{2} |n^p r_{p-1}^p(\lambda_n)|.$$

Therefore, the sequence $n^p r_{p-1}^p(\lambda_n)$ needs to be bounded, which yields

$$\lim_{|z| \rightarrow \infty} r_{p-1}(z) = 0.$$

Thus, by Proposition 2.2, the operator $r_{p-1}(\tau A)$ is range preserving.

3 Bound preservation

Let $T(t)$ be a strongly continuous semigroup generated by $-A$ satisfying

$$\|T(t)\| \leq M \tag{7}$$

for some $M \geq 1$ and all $t \geq 0$. We say that an approximating operator family $\{V_n(\tau A)\}_{n=1}^\infty$ is *unconditionally bound preserving* (in the Banach space X), if

$$\|V_n(\tau A)\| \leq M \tag{8}$$

for all $\tau > 0$ and $n \in \mathbf{N}$ and for all M with the property (7). If (8) holds only for $\tau \leq \tau^*$ with some $\tau^* > 0$, we say that the approximating operator family $\{V_n(\tau A)\}_{n=1}^\infty$ is *conditionally bound preserving*. By the Hille-Yosida theorem, the backward Euler scheme

$$V_n(\tau A) = r_{be}^n(\tau A)$$

is unconditionally bound preserving, whereas the Crank-Nicolson scheme

$$V_n(\tau A) = r_{cn}^n(\tau A)$$

is not bound preserving in an arbitrary Banach space. In fact in [5] it is shown, that for $X = l_\infty$ and $A = \text{tridiag}[1, -2, 1]$, the family $\{r_{cn}^n(\tau A)\}_{n=1}^\infty$ is conditionally bound preserving but not bound preserving.

In the next theorem we show how to construct second order unconditionally bound preserving schemes for exponentially decaying contraction semigroups.

Theorem 3.1 *Let A be a sectorial operator and $T(t)$ be the strongly continuous semigroup, generated by $-A$. Assume that $\|T(t)\| \leq e^{-\omega t}$ for some $\omega > 0$ and all $t \geq 0$. Let $r_2(\tau A)$ be a conditionally bound preserving scheme for $0 < \tau \leq \tau^*$. Then there exists $n_0 \in \mathbf{N}$ such that $r_2^{n-n_0}(-\tau A)r_{be}^{n_0}(-\tau A)$ is unconditionally bound preserving with the optimal second order error estimation.*

Proof. By Proposition 2.1, the scheme $r_2^{n-n_0}(-\tau A)r_{be}^{n_0}(-\tau A)$ has the optimal second order error estimation. Since $r_2(z)$ is $A(\theta)$ -stable it satisfies

$$\|r_2^m(-\tau A)\| \leq M_1 \quad (9)$$

for all $m \in \mathbf{N}$ with some $M_1 > 0$ [13]. By the Hille-Yosida theorem

$$\|r_{be}^n(\tau A)\| \leq \frac{1}{(1 + \tau\omega)^n} \quad (10)$$

for all $n \in \mathbf{N}$ and $\tau > 0$. Now, let n_0 be such that

$$\|r_{be}^{n_0}(-\tau^* A)\| \leq \frac{1}{M_1}. \quad (11)$$

For $\tau > \tau^*$ by (10),(11) and (9) we have:

$$\|r_2^{n-n_0}(-\tau A)r_{be}^{n_0}(-\tau A)\| \leq \|r_2^{n-n_0}(-\tau A)\| \|r_{be}^{n_0}(-\tau A)\| \leq M_1 \frac{1}{M_1} = 1.$$

If $0 < \tau \leq \tau^*$, then $\|r_2^{n-n_0}(-\tau A)\| \leq 1$ since $r_2(z)$ is conditionally bound preserving. Also, $\|r_{be}^{n_0}(-\tau A)\| \leq \frac{1}{(1+\tau\omega)^{n_0}} \leq 1$ for all $\tau > 0$. Thus, $\|r_2^{n-n_0}(-\tau A)r_{be}^{n_0}(-\tau A)\| \leq 1$.

In the next section we construct, for all $h > 0$, a second order method for the one-dimensional heat equation which is contractive in the maximum norm for all $\tau > 0$.

4 The one-dimensional heat equation

We consider the Cauchy problem

$$\begin{aligned} \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} &= 0 \text{ for } t \geq 0, x \in (0, 1), \\ u(t, 0) = u(t, 1) &= 0 \text{ for } t \geq 0, \\ u(0, x) &= u_0(x) \text{ for } x \in [0, 1], \end{aligned} \quad (12)$$

and use the finite difference method for its numerical solution. In order to apply the results of the previous sections, we discretize first the space variable. We denote by $y_i(t)$, ($i = 0, 1, \dots, s$) the approximation of $u(t, ih)$, where $h := \frac{1}{s}$ and s is the dimension of the space discretization. Let the Banach space be $X := (\mathbf{R}^{s+1}, \|\cdot\|_\infty)$. Then the equation for the semidiscrete solution can be written as

$$\dot{y}(t) + Qy(t) = 0 \text{ for } t \geq 0, \quad y(0) = y^0, \quad (13)$$

where $Q : X \rightarrow X$ is defined as $(Qy(t))_i = -\frac{1}{h^2}(y_{i+1}(t) - 2y_i(t) + y_{i-1}(t))$, ($i = 1, \dots, s-1$); $(Qy(t))_0 := (Qy(t))_s = 0$, and $(y^0)_i = u_0(ih)$, ($i = 0, 1, \dots, s$). It is easy to see that Q is sectorial and generates an analytic contraction semigroup on X with a growth bound less than zero.

Theorem 4.1 *The combined scheme $r_{cn}^{n-n_0}(-\tau Q)r_{be}^{n_0}(-\tau Q)$ has second order accuracy. Moreover, for a suitable n_0 , it is unconditionally bound preserving, i.e., contractive in the maximum norm for all $\tau > 0$ and $n \geq n_0$.*

Proof. It is shown in [13] that $\|r_{cn}^n(-\tau Q)\| \leq M$ for all $n, s \in \mathbf{N}$ and $\tau > 0$. In particular, for $\mu := \frac{\tau}{h^2} \leq \frac{3}{2}$ we have $\|r_{cn}^n(-\tau Q)\| \leq 1$ for all $n \in \mathbf{N}$ (see [5]). Therefore, the statement follows from Theorem 3.1.

In order to determine n_0 , note that the backward Euler scheme satisfies

$$\|r_{be}(-\tau Q)\|_\infty = 1 - \frac{1}{\text{coshyp}[\frac{s+1}{2}\text{arcoshyp}(1 + \frac{h^2}{2\tau})]} := g(\tau).$$

for all $\tau > 0$ (see [7]). (We denote the hyperbolic cosine and sine functions by $\text{coshyp}(\cdot)$ and $\text{sinhyp}(\cdot)$, the inverse hyperbolic cosine function by $\text{arcoshyp}(\cdot)$.) It is shown in [5] that the Crank-Nicolson scheme satisfies $\|r_{cn}^n(-\tau Q)\|_\infty \leq 4.325$ for all $\tau > 0$. Therefore, if we fix the dimension of the space discretization s (or, equivalently, fix h) we seek n_0 such that $g(\tau^*)^{-n_0} \leq 4.325$. Then $g(\tau) \leq 4.325^{-\frac{1}{n_0}} =: \beta_1$ for all $\tau > \tau^*$. We note that for the Crank-Nicolson scheme $\tau^* = 1.5s^2$. Using the notation $\beta := (1 - \beta_1)^{-1}$, the inequality $g(\tau) \leq \beta_1$ is equivalent to

$$\beta \geq \text{coshyp}[\frac{s+1}{2}\text{arcoshyp}(1 + \frac{h^2}{2\tau})],$$

or

$$\frac{2\operatorname{arccoshyp}\beta}{s+1} \geq \operatorname{arccoshyp}\left(1 + \frac{h^2}{2\tau}\right).$$

This yields

$$\tau \geq \frac{h^2}{2[\operatorname{coshyp}(2\frac{\operatorname{arccoshyp}\beta}{s+1}) - 1]}.$$

Finally, using the identity $\operatorname{coshyp}2\gamma - 1 = 2\operatorname{sinhyp}^2\gamma$, we have

$$\tau \geq \frac{h^2}{4\operatorname{sinhyp}^2\frac{\operatorname{arccoshyp}\beta}{s+1}}, \quad (14)$$

or, equivalently,

$$\mu \geq \frac{1}{4\operatorname{sinhyp}^2\frac{\operatorname{arccoshyp}\beta}{s+1}}. \quad (15)$$

Since in case $\mu \leq 1.5$ for all $n \in \mathbf{N}$ $\|r_{cn}^n(-\tau Q)\| \leq 1$, we have to consider only the case $\mu > 1.5$. Thus, we end up with the condition

$$\frac{3}{2} \geq \frac{1}{4\operatorname{sinhyp}^2\frac{\operatorname{arccoshyp}\beta}{s+1}}. \quad (16)$$

It is easy to see that for fixed s the sequence

$$\frac{1}{4\operatorname{sinhyp}^2\frac{\operatorname{arccoshyp}\beta}{s+1}} = \frac{1}{4\operatorname{sinhyp}^2\frac{\operatorname{arccoshyp}[(1-(\frac{1}{4.325})^{\frac{1}{n_0}})^{-1}]}{s+1}}$$

tends to zero as n_0 tends to infinity. Therefore, there exists n_0 such that (16) holds. Using this formula n_0 can be determined.

Although (16) allows us to define n_0 such that the method is contractive in the maximum norm for all $\tau > 0$, if we choose the dimension of the space discretization s large, the value of n_0 becomes extremely big. Therefore, from a practical point of view it is reasonable to choose some smaller n_0^* . In this case we cannot use arbitrary $\tau > 0$, but only those $\tau > \tau_s^*$, where τ_s^* can be computed via (14). Then we have a choice. Either we use $\tau \leq 1.5h^2$ (the well-known uniform contractivity condition for the Crank-Nicolson scheme) or $\tau > \tau_s^*$, (the behavior of the method between the two bounds is still unknown). The method has the optimal accuracy (i.e., second order), if

$\tau \sim \frac{1}{s}$. Therefore, for fixed s , the optimal choice of n_0 can be determined by the dimension of the space discretization. Otherwise the choice of bigger τ (i.e., smaller n_0) means that the order of the error is defined only by τ .

One of the main problems when using the Crank-Nicolson scheme that it preserves the maximum norm contractivity only for $\mu \leq 1.5$. That means if we have a fine mesh for the space variable we must choose the time step $\tau \leq 1.5h^2$ in order to obtain the maximum norm contractivity. For large values of s it would mean very small (even useless) τ . The computational error, as the result of the large number of iteration steps, might cause an essential loss in the accuracy, i.e., the scheme will loose its second order accuracy. Moreover the computational time could be enormous (see eg. [10]). Now we make a few steps with the backward Euler scheme before starting the Crank-Nicolson method. For fixed s and fixed n_0 , using (14) we obtain a lower bound

$$\hat{\tau}(s, n_0) = \frac{1}{4s^2 \operatorname{sinhyp}^2 \frac{\operatorname{arcoshyp}\beta}{s+1}} \quad (17)$$

for τ . This means, if we take any time step larger than $\hat{\tau}$, the damped method is contractive in the maximum norm. If we look at

$$\begin{aligned} \frac{1}{4 \operatorname{arcoshyp}^2 \beta} \left(\frac{s+1}{s}\right)^2 &> \frac{1}{4s^2 \operatorname{sinhyp}^2 \frac{\operatorname{arcoshyp}\beta}{s+1}} \\ &> \frac{1}{4 \operatorname{arcoshyp}^2 \beta} \frac{1}{\left(\frac{s+1}{\operatorname{arcoshyp}\beta}\right)^2 \operatorname{sinhyp}^2 \frac{\operatorname{arcoshyp}\beta}{s+1}}, \end{aligned} \quad (18)$$

then it is easy to see what is going to happen to this lower bound $\hat{\tau}(s, n_0)$ if we increase s by fixed n_0 ; i.e.,

$$\lim_{s \rightarrow \infty} \hat{\tau}(s, n_0) = \lim_{s \rightarrow \infty} \frac{1}{4s^2 \operatorname{sinhyp}^2 \frac{\operatorname{arcoshyp}\beta}{s+1}} = \frac{1}{4 \operatorname{arcoshyp}^2 \beta} \quad (19)$$

We can also see that the sequence of the upper bounds decreases monotonically towards $\frac{1}{4 \operatorname{arcoshyp}^2 \beta}$. Therefore, we can give an upper bound, uniform in s , by taking the value $s = 1$. Then (17) turns into the condition

$$\hat{\tau}(n_0) = \frac{1 - \beta_1}{2\beta_1}. \quad (20)$$

Table 1 shows the uniform lower bound for $\hat{\tau}$ defined by (17). As we can see, it could still happen that with the computed uniform $\hat{\tau}$ the local error becomes very large, (especially for small values n_0). To avoid this problem, we can use the sharper condition (17). For $s = 10000$ we calculate $\hat{\tau}$ for different values of n_0 (see Table 2). We note that for the Crank-Nicolson scheme's maximum norm contractivity we must choose $\tau \leq 1.5 \times 10^{-8}$!

As the example shows, we have the following conclusion. If we aim to have a maximum norm-contractive, second order scheme, then we have two possibilities. Either we use the Crank-Nicolson method with an extremely small time step or we use the damped method with a reasonable small time step. If we look at our example we can conclude that if we make four steps with the backward Euler method and then apply the Crank-Nicolson method, the local error will be of $O(10^{-3})$, which looks quite reasonable.

Finally, we note that the convergence in (19) is very fast; i.e., even for not very large values of s we have

$$\frac{1}{4s^2 \operatorname{sinhyp}^2 \frac{\operatorname{arcoshyp}\beta}{s+1}} \approx \lim_{s \rightarrow \infty} \frac{1}{4s^2 \operatorname{sinhyp}^2 \frac{\operatorname{arcoshyp}\beta}{s+1}} = \frac{1}{4 \operatorname{arcoshyp}^2 \beta}.$$

Table 3 shows the result for $\hat{\tau}$ with $s = 50$. If we compare Tables 2 and 3 we can observe that the values of $\hat{\tau}$ in Table 2 are representative.

5 Numerical examples

In this section we summarize some numerical investigations for the model problem (12) with both smooth and nonsmooth initial functions using different numerical methods discussed in the previous section at the time level $T = 1$.

The first model problem with the initial function $u_0(x) = \sin(\pi x)$ has the exact solution $u(x, t) = \sin(\pi x) \exp(-\pi^2 t)$.

The following tables show the maximum norm error for the Crank-Nicolson (CN) and the backward Euler (BE) methods on different meshes.

As Tables 4 and 5 show with refining the mesh under the maximum norm contractivity condition, the Crank-Nicolson scheme loses its higher accuracy with respect to the backward Euler method and in the limit they result in the same accuracy. The following Tables 6 - 8 serve to demonstrate the

behavior of the maximum norm error of the Crank-Nicolson method with further different discretization parameters.

These results show that the optimal accuracy of the Crank-Nicolson method is attained at some value $\mu_{opt} = \mu_{opt}(h)$ which is greater than 1.5. Moreover, by decreasing h (that is, by refining the mesh), the values μ_{opt} are increasing. Table 9 shows the loss in the accuracy. The fifth column in this table shows how much more CPU-time is used to obtain the less accurate result (with $\mu = 1.5$). The accuracy with the choice $\mu = 1.5$ is attained with some $\mu_{big} > 1.5$, too. The approximate values of these parameters and the corresponding CPU ratios are included in the last two columns.

The second model problem is the problem (12) with the initial function

$$u_0(x) = \begin{cases} 1 & \text{if } x \in [0.25, 0.75] \\ 0 & \text{otherwise} \end{cases}$$

which has the exact solution

$$u(x, t) = \frac{2}{\pi} \sum_{m=1}^{\infty} \frac{1}{m} \left(\cos \frac{m\pi}{4} - \cos \frac{3m\pi}{4} \right) \sin(m\pi x) \exp(-m^2\pi^2 t).$$

Table 10 summarizes the error in maximum norm for the CN and BE methods. The behavior of the Crank-Nicolson method is similar as for the smooth initial function. However, the smoothing property of the backward Euler method is considerable. We remark that the same conclusions can be made for the other choices of h .

In the following we give numerical results for the damped method.

First, we analyze the behavior of the damped method on the problem with smooth initial function. Tables 11 - 13 show the numerical results for the damped method with different space discretization steps. Each table contains the maximum norm error for different numbers of smoothing steps n_0 and time discretization steps, (the values in the columns $n_0 = 0$ correspond to the Crank-Nicolson method and the errors with bold numbers are the result of the backward Euler method). We observe that with the increase in n_0 the damped method loses its accuracy and the "almost best" choice is $n_0 = 2$. Table 14 shows the result for this fixed choice with the small space discretization stepsize $h = 0.002$.

For the nonsmooth initial function the behavior of the damped method for $n_0 = 1, 2, 3$ damping steps is given in Table 15.

Finally, we compare the damped method, the Crank-Nicolson method, and the backward Euler method on the mesh where the maximum norm is preserved for the damped method, that is, the mesh is chosen according to the condition (15). Clearly, on such a mesh the BE method is also maximum norm contractive while the CN method is usually not.

Tables 16-18 contain the results for the smooth initial function and Tables 19-21 contain the results for the nonsmooth problem. Especially remarkable is the advantage of the damped method on the nonsmooth problem.

6 Conclusions and further problems

The main results of the paper are the following.

For combined schemes we showed that the condition $r(\infty) = 0$ on the starting scheme is equivalent to saying that the lower order approximation scheme maps all initial data into the domain of the operator A . In a special case, we proved that this is also a necessary condition to obtain the optimal accuracy (which is p). For the numerical solution of the one-dimensional heat equation we analyzed the damped Crank-Nicolson method using the backward Euler scheme for the first n_0 steps. We examined the relation between the number of the backward Euler steps, the space discretization parameter and the possible choices of the time step. We showed, that for arbitrary fixed h , there exists a suitable damped Crank-Nicolson method, having second order of accuracy, which is contractive in the maximum norm for all $\tau > 0$. For the damped Crank-Nicolson method we showed, that for any fixed n_0 , there exists $\hat{\tau}$ (depending on n_0 and the space discretization parameter) such that for any time step bigger than $\hat{\tau}$, the damped Crank-Nicolson method is contractive in the maximum norm. (Of course a small time step can be chosen also, according to the well-known maximum norm contractivity upper bound for the Crank-Nicolson method). We have also provided numerical examples to illustrate the advantages of the damped method.

We list some further open problems.

- a. The behavior of the damped Crank-Nicolson method with the choice of the time step between the contractivity bound of the Crank-Nicolson method and $\hat{\tau}$, in the one-dimensional heat equation problem.

- b. The investigation of some other qualitative properties of the damped Crank-Nicolson method, for instance, nonnegativity preservation, shape preservation, etc. in the one-dimensional heat equation problem.
- c. Analysis of the damped methods for more general problems, i.e., problems in several space variables, non-autonomous problems, nonlinear problems.

Acknowledgement We wish to thank Frank Neubrander and Anita Hansbo for helpful comments and suggestions.

References

- [1] Bellen, A., Jackiewicz, Z., Zennaro, M.: Contractivity of waveform relaxation Runge-Kutta iterations and related limit methods for dissipative systems in the maximum norm. *SIAM J. Numer. Anal.* **31**, no. 2, 499-523 (1994).
- [2] Bellen, A., Torelli, L.: Unconditional contractivity in the maximum norm of diagonally split Runge-Kutta methods. *SIAM J. Numer. Anal.* **34** no. 2, 528-543 (1997).
- [3] Dunford, N., Schwartz, J.T.: *Linear Operators: Part 1. General Theory.* New York: Interscience Publ. 1958.
- [4] Engel, K.J., Nagel, R.: *One-parameter Semigroups for Linear Evolution Equations.* New York: Springer-Verlag 2000.
- [5] Farago, I., Palencia, C.: Sharpening the estimate of the stability constant in the maximum-norm of the Crank-Nicolson scheme for the one-dimensional heat equation. *Appl. Numer. Math.* **42**, 133-140 (2002).
- [6] Hansbo, A.: Nonsmooth error estimates for damped single step methods for parabolic equations in Banach spaces. *Calcolo* **36**, 75-101 (1999).
- [7] Horváth, R.: Maximum norm contractivity in the numerical solution of the one-dimensional heat equation. *Appl. Numer. Math.* **31**, 451-462 (1999).
- [8] Kraaijevanger, J.F.B.M.: Maximum norm contractivity of the discretization schemes for the heat equation. *Appl. Numer. Math.* **9**, 475-492 (1992).
- [9] Luskin, M., Rannacher, R.: On the smoothing property of the Crank-Nicolson scheme. *Appl. Anal.* **14**, 117-135 (1982).
- [10] Ortega, J.M., Poole, W.G.: *An Introduction to Numerical Methods for Differential Equations.* Pittman Publ. 1981.
- [11] Rannacher, R.: Finite element solution of diffusion problems with irregular data. *Numer. Math.* **43**, 309-327 (1984).
- [12] Spijker, M.N.: Contractivity in the numerical solution of initial value problems. *Numer. Math.* **42**, 271-290 (1983).

- [13] Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems.
Berlin: Springer-Verlag 1997.

TABLES

n_0	1	2	4	10	50
$\hat{\tau}$	1.625	0.531	0.218	0.078	0.015

Table 1: The values of $\hat{\tau}$ uniform for all h .

n_0	1	2	4	10	50
$\hat{\tau}$	0.437	0.155	0.073	0.035	0.014

Table 2: The values of $\hat{\tau}$ with $h = 0.0001$.

n_0	1	2	4	10	50
$\hat{\tau}$	0.453	0.160	0.076	0.036	0.014

Table 3: The upper bounds for $\hat{\tau}$ with $h = 0.02$.

μ	100	50	40	20	10	8
CN-error	$6.35E-5$	$4.14E-5$	$3.10E-5$	$8.79E-6$	$1.53E-6$	$6.04E-7$
BE-error	$6.90E-3$	$1.60E-3$	$1.00E-3$	$2.81E-4$	$9.83E-5$	$7.29E-5$
μ	7	6	5	4	2	1.5
CN-error	$2.22E-7$	$1.147E-7$	$4.05E-7$	$6.40E-7$	$9.54E-7$	$9.89E-7$
BE-error	$6.27E-5$	$4.86E-5$	$4.07E-5$	$3.15E-5$	$1.50E-5$	$1.20E-5$
μ	1	0.4	0.1	0.05	0.01	0.005
CN-error	$1.03E-6$	$1.06E-6$	$1.06E-6$	$1.06E-6$	$1.06E-6$	$1.06E-6$
BE-error	$7.75E-6$	$3.67E-6$	$1.70E-6$	$1.38E-6$	$1.12E-6$	$1.09E-6$

Table 4: The maximum norm error for $h = 0.05$ for the CN and BE methods

μ	4000	2000	1500	1000	500
CN-error	$3.16E-5$	$9.70E-6$	$5.12E-6$	$2.54E-6$	$6.35E-7$
BE-error	$9.91E-4$	$2.76E-4$	$1.58E-4$	$9.59E-5$	$3.91E-5$
μ	400	200	100	40	20
CN-error	$4.03E-7$	$9.30E-8$	$1.54E-8$	$6.35E-9$	$9.46E-9$
BE-error	$3.00E-5$	$1.38E-5$	$6.59E-6$	$2.58E-6$	$1.28E-6$
μ	10	4	2	1.5	1
CN-error	$1.02E-8$	$1.05E-8$	$1.05E-8$	$1.05E-8$	$1.05E-8$
BE-error	$6.43E-7$	$2.63E-7$	$1.37E-7$	$1.05E-7$	$7.35E-8$

Table 5: The maximum norm error for $h = 0.005$ for the CN and BE methods

μ	10	5	2	1.5	1
error	$2.93E-5$	$5.93E-6$	$2.62E-6$	$3.23E-6$	$3.92E-6$
μ	0.5	0.1	0.05	0.01	0.005
error	$4.25E-6$	$4.35E-6$	$4.36E-6$	$4.36E-6$	$4.36E-6$

Table 6: The maximum norm error of the Crank-Nicolson method for $h = 0.1$

μ	1000	500	250	160	80	40	20
error	0.245	0.014	$9.54E - 5$	$3.14E - 5$	$9.48E - 6$	$2.29E - 6$	$3.84E - 7$
μ	16	4.8	1.6	1.5	1	0.8	0.16
error	$1.52E - 7$	$1.57E - 7$	$2.59E - 7$	$2.59E - 7$	$2.61E - 7$	$1.85E - 7$	$1.86E - 7$

Table 7: The maximum norm error of the Crank-Nicolson method for $h = 0.025$

μ	1000	500	100	50	40	30
error	$3.16E - 5$	$9.67E - 6$	$3.72E - 7$	$6.16E - 8$	$2.43E - 8$	$4.61E - 9$
μ	20	10	5	1.5	1	0.5
error	$2.54E - 8$	$3.79E - 8$	$4.10E - 8$	$2.96E - 8$	$2.97E - 8$	$2.97E - 8$

Table 8: The maximum norm error of the Crank-Nicolson method for $h = 0.01$

h	μ_{opt}	error	error for $\mu = 1.5$	CPU ratio	μ_{big}	CPU ratio
0.1	2	$2.62E - 6$	$3.23E - 6$	1.266	4.6	3.07
0.05	4	$6.40E - 7$	$9.87E - 7$	2.71	8.7	5.80
0.025	16	$1.52E - 7$	$2.59E - 7$	10.81	18	12.00
0.01	30	$4.61E - 9$	$4.19E - 8$	20.5	45	30.00
0.005	40	$6.35E - 9$	$1.05E - 8$	27.6	94	62.67
0.004	62.5	$2.57E - 9$	$6.71E - 9$	42.9	112.5	75.00

Table 9: Comparison of the accuracy and consumed CPU time

μ	4000	2000	400	200	100
CN-error	0.4765	0.4438	0.2397	$7.86E - 2$	$2.30E - 3$
BE-error	$8.99E - 4$	$2.51E - 4$	$2.76E - 5$	$1.29E - 5$	$6.35E - 6$
μ	75	60	50	45	40
CN-error	$8.18E - 5$	$1.48E - 6$	$3.68E - 7$	$3.69E - 7$	$3.71E - 7$
BE-error	$4.83E - 6$	$3.88E - 6$	$3.30E - 6$	$2.99E - 6$	$2.70E - 6$
μ	20	10	5	1.5	1
CN-error	$3.73E - 7$	$3.74E - 7$	$3.74E - 7$	$3.74E - 7$	$3.74E - 7$
BE-error	$1.53E - 6$	$9.48E - 7$	$6.61E - 7$	$4.60E - 7$	$4.31E - 7$

Table 10: The maximum norm error for $h = 0.005$ for the CN and BE methods for nonsmooth initial function

n_0	0	1	2	3	5
$\mu = 5$	$4.92E - 5$	$4.91E - 5$	$4.78E - 5$	$2.87E - 5$	$4.51E - 3$
$\mu = 1$	$1.01E - 5$	$1.49E - 5$	$1.91E - 5$	$2.36E - 5$	$3.34E - 5$
$\mu = 0.1$	$1.85E - 5$	$1.86E - 5$	$1.86E - 5$	$1.86E - 5$	$1.87E - 5$
n_0	10	25	50	100	250
$\mu = 5$	—	—	—	—	—
$\mu = 1$	$6.41E - 5$	—	—	—	—
$\mu = 0.1$	$1.90E - 5$	$1.97E - 5$	$2.10E - 5$	$2.35E - 5$	$3.18E - 5$

Table 11: The maximum norm error for $h = 0.2$ for the damped method for smooth initial function

n_0	0	1	2	3	5
$\mu = 125$	$9.57E - 6$	$5.01E - 6$	$4.79E - 8$	$5.65E - 6$	$1.87E - 5$
$\mu = 50$	$1.48E - 6$	$5.70E - 7$	$3.53E - 7$	$1.29E - 6$	$3.22E - 6$
$\mu = 5$	$1.52E - 7$	$1.62E - 7$	$1.72E - 7$	$1.82E - 7$	$2.02E - 7$
n_0	10	25	50	100	250
$\mu = 125$	$6.60E - 5$	$2.77E - 4$	—	—	—
$\mu = 50$	$8.36E - 6$	$2.01E - 5$	$7.11E - 5$	—	—
$\mu = 5$	$2.52E - 7$	$3.52E - 7$	$6.54E - 7$	$2.71E - 6$	$5.40E - 6$

Table 12: The maximum norm error for $h = 0.02$ for the damped method for smooth initial function

n_0	0	1	2	3	5
$\mu = 50000$	$5.17E - 5$	$5.17E - 5$	$5.01E - 5$	$3.22E - 5$	$4.20E - 3$
$\mu = 5000$	$1.64E - 6$	$7.35E - 7$	$1.85E - 7$	$1.12E - 6$	$3.05E - 6$
$\mu = 125$	$6.43E - 10$	$1.27E - 9$	$1.90E - 9$	$2.53E - 9$	$3.78E - 9$
n_0	10	50	100	500	2000
$\mu = 50000$	—	—	—	—	—
$\mu = 5000$	$8.17E - 6$	$7.07E - 5$	—	—	—
$\mu = 125$	$6.93E - 9$	$3.21E - 8$	$6.53E - 8$	$3.16E - 7$	$1.27E - 6$

Table 13: The maximum norm error for $h = 0.002$ for the damped method for smooth initial function

μ	50000	37500	25000	15000	10000
max. error	$5.17E - 5$	$1.95E - 5$	$7.38E - 6$	$8.39E - 7$	$1.74E - 7$
μ	5000	4000	3000	2000	1500
max. error	$1.85E - 7$	$1.60E - 7$	$9.72E - 8$	$4.69E - 8$	$2.79E - 8$
μ	1000	500	375	250	125
max. error	$1.44E - 8$	$5.06E - 9$	$3.59E - 9$	$2.55E - 9$	$1.90E - 9$

Table 14: The maximum norm error for $h = 0.002$ and $n_0 = 2$ for the damped method for smooth initial function

μ	4000	2000	400	200	100
$n_0 = 1$	$3.10E - 3$	$1.40E - 3$	$2.76E - 3$	$1.31E - 4$	$6.77E - 6$
$n_0 = 2$	$1.42E - 4$	$2.43E - 5$	$1.20E - 6$	$5.24E - 7$	$3.79E - 7$
$n_0 = 3$	$2.10E - 5$	$5.76E - 6$	$6.54E - 7$	$4.48E - 7$	$3.93E - 7$
μ	75	50	40	10	5
$n_0 = 1$	$5.55E - 7$	$3.72E - 7$	$3.72E - 7$	$3.74E - 7$	$3.74E - 7$
$n_0 = 2$	$3.79E - 7$	$3.76E - 7$	$3.75E - 7$	$3.75E - 7$	$3.74E - 7$
$n_0 = 3$	$3.87E - 7$	$3.79E - 7$	$3.77E - 7$	$3.75E - 7$	$3.75E - 7$

Table 15: The maximum norm error for $h = 0.005$ for the damped method for nonsmooth initial function

μ	10.58	3.72	2.32	1.722	1.395
n_0	1	2	3	4	5
DM	0.064	$1.01E - 5$	$2.97E - 5$	$3.32E - 5$	$4.58E - 5$
CN	0.3216	$2.88E - 5$	$1.32E - 5$	$1.72E - 6$	$4.17E - 6$
BE	0.037	0.0019	$8.54E - 4$	$4.49E - 4$	$3.89E - 4$

Table 16: The maximum norm error for $h = 0.2$ for the different methods on maximum norm contractive mesh for smooth initial function

μ	1066	380	240	86	35
n_0	1	2	3	10	50
DM	0.0685	$1.79E - 5$	$2.38E - 5$	$2.70E - 5$	$2.99E - 5$
CN	0.1262	$2.63E - 5$	$4.29E - 5$	$4.70E - 6$	$6.77E - 7$
BE	0.037	0.0016	0.0012	$1.56E - 4$	$4.75E - 5$

Table 17: The maximum norm error for $h = 0.02$ for the different methods on maximum norm contractive mesh for smooth initial function

μ	106000	38000	24000	8620	3460
n_0	1	2	3	10	50
DM	0.0684	$1.80E - 5$	$2.34E - 5$	$2.63E - 5$	$2.82E - 5$
CN	0.1245	$2.63E - 5$	$4.30E - 5$	$4.78E - 6$	$8.13E - 7$
BE	0.037	0.0016	0.0012	$1.53E - 4$	$4.56E - 5$

Table 18: The maximum norm error for $h = 0.002$ for the different methods on maximum norm contractive mesh for smooth initial function

μ	10.58	3.72	2.32	1.722	1.395
n_0	1	2	3	4	5
DM	0.057	$4.75E - 4$	$2.03E - 5$	$2.03E - 5$	$2.83E - 5$
CN	0.2894	0.0242	$2.32E - 4$	$6.42E - 6$	$3.40E - 6$
BE	0.0286	0.0015	$6.37E - 4$	$3.37E - 4$	$2.89E - 4$

Table 19: The maximum norm error for $h = 0.2$ for the different methods on maximum norm contractive mesh for nonsmooth initial function

μ	1066	380	240	86	35
n_0	1	2	3	10	50
DM	0.0667	$3.66E - 4$	$2.54E - 5$	$2.43E - 5$	$2.70E - 5$
CN	0.4970	0.4256	0.3905	0.2057	0.0220
BE	0.0332	0.0015	0.0011	$1.40E - 4$	$4.28E - 5$

Table 20: The maximum norm error for $h = 0.02$ for the different methods on maximum norm contractive mesh for nonsmooth initial function

μ	106000	38000	24000	8620	3460
n_0	1	2	3	10	50
DM	0.0667	$3.63E - 4$	$2.53E - 5$	$2.39E - 5$	$2.56E - 5$
CN	0.5210	0.4941	0.4927	0.4673	0.4192
BE	0.0336	0.0015	0.0011	$1.38E - 4$	$4.14E - 5$

Table 21: The maximum norm error for $h = 0.002$ for the different methods on maximum norm contractive mesh for nonsmooth initial function