NON-NEGATIVITY PRESERVATION
OF THE DISCRETE NONSTATIONARY HEAT EQUATION
IN 1D AND 2D

FARAGÓ István, (HU), KOROTOV Sergey, (FI), SZABÓ Tamás, (HU)

Abstract. In this paper we analyse the preservation of the non-negativity property for
the semidiscrete and fully discretized (by finite differences and finite elements) numerical
solutions of the linear parabolic problem in one and two space dimensions. In particular,
we derive the exact (necessary and sufficient) conditions for the 1D problem, and also give
sufficient conditions for the 2D case, under which the non-negativity preservation property
of the fully discretized problem is valid.

Key words and phrases. parabolic problems, semidiscretization, fully discretized prob-
lem, linear and bilinear finite elements, finite differences, non-negativity preservation,
tridiagonal matrix, block-tridiagonal matrix.

Mathematics Subject Classification. 15A06, 65M06, 65M60

1 Introduction

Besides the convergence, another natural requirement in the process of numerical solution of
partial differential equations, is a preservation of basic qualitative properties of the original
(physical) solution, assuming that they are inherent to the continuous mathematical model.
As an example, consider the advection-diffusion-reaction equation arising, e.g., in the large
air-pollution modeling [22]:

$$\frac{\partial c}{\partial t} = -\text{div} \ (vc) + \text{div} \ (D \ \text{grad} \ c) + R(c), \quad t \in (0, T], \ c(0, x) = c_0(x), \quad (1)$$

where the vector-valued function $c(t, x)$ denotes the concentration of compounds, $v = v(t, x)$
presents the current velocity of the medium, $D$ is the so-called diffusion coefficient matrix,
and the function $R$ describes the chemical reactions between the compounds and includes the parametrized deposition and emission. (As a rule, we use equation (1) in the componentwise sense.) We solve the above problem by suitably chosen numerical method. Since $c$ denotes the concentration, which is always non-negative, it is natural to require the non-negativity from the numerical approximations of $c$ as well.

The technique, often used to solve problems like (1) (based on several physical processes) is the so-called operator splitting method, see e.g. [4]: we choose a suitable time-step $\Delta t > 0$, and for $j = 1, 2, \ldots$ solve the sequence of subproblems

\[
\begin{align*}
\frac{\partial u_1^{(j)}}{\partial t} &= -\text{div} (vu_1^{(j)}), \quad u_1^{(j)}((j - 1)\Delta t, x) = u_3^{(j-1)}((j - 1)\Delta t, x), \\
\frac{\partial u_2^{(j)}}{\partial t} &= \text{div} (D \text{grad} u_2^{(j)}), \quad u_2^{(j)}((j - 1)\Delta t, x) = u_1^{(j)}(j\Delta t, x), \\
\frac{\partial u_3^{(j)}}{\partial t} &= R(u_3^{(j)}), \quad u_3^{(j)}((j - 1)\Delta t, x) = u_2^{(j)}(j\Delta t, x),
\end{align*}
\]

on the intervals $[(j - 1)\Delta t, j\Delta t]$, where $u_3^{(0)}(0, x) = c_0(x)$. In this procedure, $u_3^{(j)}$ yields an approximation to $c(j\Delta t, x)$.

For numerical solution of each of the above subproblems, we can choose a certain suitable numerical method (naturally, all three may be completely different each from other). Obviously, if all numerical techniques used for solving (2)–(4) are non-negativity preserving, then the whole computational scheme, used for solving (1), is non-negativity preserving.

In the above sequence of subproblems (2)–(4), the central role belongs to subproblem (3), which, in the simplest setting, has the following form

\[
\begin{align*}
\frac{\partial u}{\partial t} &= \Delta u \quad \text{in}\; \Omega_T = (0, T) \times \Omega, \\
u &= 0 \quad \text{on}\; \Gamma_T = (0, T) \times \partial \Omega, \\
u|_{t=0} &= u_0 \quad \text{on}\; \Omega,
\end{align*}
\]

where $\Omega \subset \mathbb{R}^d$ is a segment (for $d = 1$, and the problem is referred to as the one-dimensional), or a rectangle (for $d = 2$, and the problem is called two-dimensional), the symbol $\Delta$ denotes the Laplace operator, and $u_0$ is the initial function defined from the splitting procedure.

This equation has an importance of its own. For instance, it describes the heat conduction process, therefore hereafter we refer to it as the heat conduction equation. For this equation, the non-negativity preservation principle reads as follows: for any non-negative initial function $u_0$, the solution $u$ has to be non-negative in $\Omega_T$ as well, see, e.g. [19].

A typical numerical technique for solving (5)–(7) presents a combination of separate discretizations in space and time. For the first one, we can employ the finite element method (based e.g. on the linear elements in the one-dimensional case, and on bilinear elements - for the two-dimensional case), or the standard finite difference method with the mesh-size $h$. As a result, we get the following Cauchy problem for the semidiscrete solution $u_h$

\[
\frac{du_h}{dt}(t) = \Delta_h u_h(t), \quad t \in (0, T),
\]

volume 3 (2010), number 2
where the initial value $u_h(0)$ is given and $\Delta_h$ denotes the corresponding discrete Laplace operator (represented by a matrix). Applying some suitable time discretization method with the time-step $\tau$ to problem (8), we finally arrive at the fully discretized problem, presenting the following algebraic iterative procedure

$$X_1 y^{j+1} = X_2 y^j,$$

where $X_1$ and $X_2$ are given matrices, and the vector $y^j$ represents the approximation to $u_h(j\tau)$.

Our aim is to formulate, for a fixed standard parameter $\theta \in [0, 1]$, such conditions on the discretization parameters $h$ and $\tau$, under which the corresponding semidiscrete and fully discretized solutions preserve the non-negativity property.

In this paper we study these problems. We show that in the one-dimensional case, the exact (necessary and sufficient) conditions (with respect to $\tau$ and $h$, for any fixed $\theta$) can be obtained. The results are based on finding the exact representation of the matrix $X = X_1^{-1}X_2$, where the crucial point consists of computing the inverse $X_1^{-1}$.

In the two-dimensional case, a similar problem is more difficult, due to more complicated structures of the corresponding matrices. A certain sufficient condition is given in [8] and it is based on the following requirements: $X_2 \geq 0$ and $X_1$ is a monotone matrix. However, finding the necessary and sufficient condition is still an open problem.

In our paper, we analyse first the non-negativity preservation for the semidiscrete solutions of (8), and, additionally, we establish direct connection between the non-negativity preservation of the semidiscrete solutions (8) for the one and two-dimensional cases. In Section 3 we give the exact condition for the non-negativity preservation in 1D and we give “the bounds” for the finite difference and linear finite element methods. In Section 4 we formulate the conditions under which the discrete problems are non-negativity preserving in 2D. In the final section we illustrate numerically the theoretically derived results.

We note that the non-negativity preservation property has a close relation to the validity of the discrete maximum principle. This topic is addressed, e.g., in [10], [5], [6] for the parabolic problem, and in [12], [13] - for the elliptic problems.

2 Non-negativity preservation for the semidiscrete solutions

In this section, we consider the non-negativity preservation for the Cauchy problem (8), where $\Delta_h$ arises from the space uniform discretization of the Laplace operator on rectangular mesh. The solution has a form

$$u_h(t) = \exp(\Delta_h t)u_0, \quad t \in (0, T),$$

where $\exp(\Delta_h t)$ denotes the exponent of the matrix $\Delta_h t$. Therefore, the problem of the non-negativity preservation is equivalent to finding conditions on discretization methods for which the matrix exponential $\exp(\Delta_h t)$ is non-negative.

The following useful lemma holds (cf. [1, p. 172]).
Lemma 2.1 Let \( A \) be an arbitrary square matrix with the entries \( a_{ij} \). Then \( \exp(At) \) is non-negative for any \( t \geq 0 \) if and only if the condition
\[
a_{ij} \geq 0 \text{ for all } i \neq j
\] holds.

Proof: By the definition of the exponential, we have
\[
\exp(At) = I + At + \ldots,
\]
where \( I \) denotes the identity matrix. This series immediately shows the necessity of condition (11). Let now \( s \) be a scalar such that \( A + sI \) is a non-negative matrix. Then, obviously, \( \exp((A + sI)t) \) is non-negative if \( t \geq 0 \). Moreover, \( \exp(-sIt) \) is also non-negative, and the matrices \( (A + sI)t \) and \( -sIt \) commute. Therefore, due to the identity
\[
\exp(At) = \exp((A + sI)t - sIt) = \exp((A + sI)t) \cdot \exp(-sIt),
\]
the sufficiency of condition (11) is also proven.

In the following we consider the application of this lemma to the one-dimensional case.

2.1 One-dimensional case

If problem (5)–(7) is considered in the one-dimensional case, the structure of the matrix \( \Delta_h \) is well-known for the both finite difference and finite element (space) discretizations. Namely, using the standard denotations
\[
Q = \text{tridiag}(1, -2, 1), \quad M = \frac{1}{6} \text{tridiag}(1, 4, 1),
\]
we have:
for the finite difference method
\[
\Delta_h = \frac{1}{h^2}Q,
\]
for the linear finite element method
\[
\Delta_h = \frac{1}{h^2}M^{-1}Q.
\]

The matrix \( \Delta_h \) from (13) obviously satisfies the condition (11), at the same time, the matrix \( \Delta_h \) from (14) (which can be computed explicitly) is known to have its entries changing the sign chessboard-likely [7], i.e., it does not satisfy the condition (11).

Thus, using Lemma 2.1, we obtain the following result.

Theorem 2.1 For the one-dimensional problem (8), the semidiscrete numerical solutions, obtained by the finite difference discretization, preserve the non-negativity property. However, this property is not preserved, in general, for the numerical solutions resulting from the linear finite element discretization.
Remark 2.1 The lumped mass method for the linear finite element method results in the semidiscrete problem which coincides with the finite difference semidiscrete problem. Therefore, this approach makes possible to improve the qualitative property of the finite element discretization.

Remark 2.2 In [2] there is given the condition of the positivity of the time derivative of the semidiscrete solutions. Clearly, this condition can be regarded as a sufficient condition of the non-negativity preservation of the semidiscrete solutions.

Remark 2.3 Let us introduce the following denotation

\[ T_1(p) = \text{tridiag}(1, p, 1), \]  

where \( p \in \mathbb{R} \), i.e., \( Q = T_1(-2) \). Then, the semidiscretization in the form

\[ \frac{du_h}{dt}(t) = \frac{1}{h^2} T_1(p) u_h(t), \quad t \in (0, L), \]  

is non-negativity preserving for any value of the parameter \( p \). Therefore, instead of (5), we can consider a more general equation

\[ \frac{\partial u}{\partial t} = \Delta u + ku \quad \text{in} \; \Omega_L = (0, L) \times \Omega, \]  

where \( k \) is any constant, and prove that the finite difference semi-discretization for such an equation is non-negativity preserving, because the approximation of the new term effects only the diagonal elements of the matrix in (16).

Remark 2.4 However, the non-negativity property for the above problem is known to hold only for some \( k \geq k_0 \) for the original continuous problem (see, e.g., [19]), and, of course, the preservation of this property in the numerical realization is somewhat meaningless for certain values of \( k \).

2.2 Two-dimensional case

We now consider the discretization on the uniform mesh (of the step-size \( h \)) of problem (5)–(7) in the two-dimensional case. We introduce the following denotation

\[ T_2(p) = \text{tridiag}(I, T_1(p), I), \]  

for a block tridiagonal matrix (from \( \mathbb{R}^{n^2 \times n^2} \)), where \( p \in \mathbb{R} \).

We consider the both semidiscretization methods. Clearly, if we apply the usual finite difference method, then in the Cauchy-problem (8) the matrix \( \Delta_h \) has the form

\[ \Delta_h = \frac{1}{h^2} T_2(-4), \]
that is, all its off-diagonal elements are non-negative. If we use the linear finite element method, then we get \( \Delta \triangleq (1/h^2)M^{-1}T_2(-4) \).

In what follows, we establish the relation between the exponentials of the matrices \( T_1(p) \) from \( \mathbb{R}^{n \times n} \) and \( T_2(p) \) from \( \mathbb{R}^{n^2 \times n^2} \), i.e., the matrix exponential between the 1D and 2D discrete Laplacians. Using the denotation

\[
A(p) = \text{tridiag}(1, p + 2, 1) = T_1(p) + 2I, \tag{19}
\]

we have

\[
T_2(p) = \text{tridiag}(I, A(p) - 2I, I) = \text{tridiag}(0, A(p), 0) + \text{tridiag}(I, -2I, I) = I \otimes A(p) + Q \otimes I, \tag{20}
\]

where \( \otimes \) denotes the Kronecker product of matrices, see e.g. [9, 17]. In order to attribute the matrix exponential of the matrix \( T_2(p) \in \mathbb{R}^{n^2 \times n^2} \) to the matrix exponentials of the matrices \( A(p) \) and \( Q \) from \( \mathbb{R}^{n \times n} \), i.e., the two-dimensional problem to the one-dimensional one, we prove the following lemma.

**Lemma 2.2** For the matrices \( T_2(p), A(p), \) and \( Q \), the relation

\[
\exp(T_2(p)) = \exp(Q) \otimes \exp(A(p)) \tag{21}
\]

holds.

**Proof:** For any matrices \( A, B, C, D \) of the same size, we have [9, p. 228],

\[
(A \otimes B)(C \otimes D) = AC \otimes BD.
\]

Therefore

\[
(I \otimes A(p))(Q \otimes I) = Q \otimes A(p), \\
(Q \otimes I)(I \otimes A(p)) = Q \otimes A(p). \tag{22}
\]

Consequently, the corresponding exponential can be written by use of the binomial rule as follows

\[
\exp(T_2(p)) = \exp((I \otimes A(p)) + (Q \otimes I)) = \\
= \sum_{n=0}^{\infty} \frac{1}{n!} (I \otimes A(p) + Q \otimes I)^n = \\
= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{1}{k!(n-k)!} (I \otimes A(p))^k (Q \otimes I)^{n-k} = \\
= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{1}{k!(n-k)!} Q^k \otimes A(p)^{n-k}. \tag{23}
\]
On the other hand, by the definition of the tensor product, we have

\[
\exp(Q) \otimes \exp(A(p)) = \sum_{i=0}^{\infty} \frac{1}{i!} Q^i \otimes \sum_{j=0}^{\infty} \frac{1}{j!} A(p)^j = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{1}{i! j!} Q^i \otimes A(p)^j. \tag{24}
\]

Since the right-hand sides in (23) and (24) are equal, we obtain the relation (21). ■

Obviously, the tensor product of two matrices is non-negative if and only if both involved matrices are non-negative. Moreover, \(\exp(Q)\) is non-negative and \(\exp(A(p))\) is non-negative if and only if \(\exp(T_1(p))\) is non-negative. Based on Lemma 2.1, we obtained the following statement.

**Theorem 2.2** For the two-dimensional problem on rectangular mesh, the semidiscrete numerical solution, obtained by the regular finite difference discretization, preserves the non-negativity property. However, this property is not preserved, in general, for the linear finite element discretization.

**Remark 2.5** Clearly, the statement of Theorem 2.2 is also valid for the more general equation (17).

### 3 The non-negativity preservation in 1D case

We consider the non-negativity preservation property of the discretization of the one-dimensional heat conduction problem with first homogeneous boundary conditions. (For the simplicity, the constant coefficient is assumed to be equal one.) Then we get problem (9) with the matrices from \(\mathbb{R}^{N \times N}\) in the form:

\[
X_1 = \frac{1}{\Delta t} M - \theta Q, \quad X_2 = \frac{1}{\Delta t} M + (1 - \theta) Q, \tag{25}
\]

where \(M = I\) for the finite difference method and it has the form from (12) for the linear finite element method. Hence, these matrices have the following entries:

- for the finite difference method

\[
X_1 = \text{tridiag} \left[ -\frac{\theta}{h^2}, \frac{1}{\Delta t} + 2\frac{\theta}{h^2}, -\frac{\theta}{h^2} \right],
\]

\[
X_2 = \text{tridiag} \left[ \frac{1-\theta}{h^2}, \frac{1}{\Delta t} - 2\frac{1-\theta}{h^2}, \frac{1-\theta}{h^2} \right], \tag{26}
\]

- for the linear finite element method the corresponding matrices are

\[
X_1 = \text{tridiag} \left[ \frac{1}{6\Delta t} - \frac{\theta}{h^2}, \frac{2}{3\Delta t} + 2\frac{\theta}{h^2}, \frac{1}{6\Delta t} - \frac{\theta}{h^2} \right],
\]

\[
X_2 = \text{tridiag} \left[ \frac{1}{6\Delta t} + \frac{1-\theta}{h^2}, \frac{2}{3\Delta t} - 2\frac{1-\theta}{h^2}, \frac{1}{6\Delta t} + \frac{1-\theta}{h^2} \right]. \tag{27}
\]
For the non-negativity preservation property we require the condition

$$X = X_1^{-1}X_2 \geq 0.$$  \hspace{1cm} (28)

Let us notice that the matrices in (26) and (27) have special structure: only the entries of the main-, super- and sub-diagonals differ from zero and the elements standing on the same diagonal are equal. Moreover, these matrices are symmetric, too. Such kind of matrix is called uniformly continuant, symmetrical tridiagonal matrix and they have some special qualitative properties, which will be considered in the sequel.

3.1 Non-negativity of the iteration matrix in general form

We consider the real, uniformly continuant, symmetrical tridiagonal matrices

$$X_1 = z \cdot \text{tridiag}[-1, 2w, -1]; \quad X_2 = s \cdot \text{tridiag}[1, p, 1]$$  \hspace{1cm} (29)

with the assumptions

$$z > 0, \quad s > 0; \quad w > 1.$$  \hspace{1cm} (30)

Our aim is to define for this case those conditions under which the iteration matrix $X$ is non-negative.

We introduce the following one-pair matrix $G = (G_{ij})$, depending on the parameter $w$:

$$G_{i,j} = \begin{cases} \gamma_{i,j}, & \text{if } i \leq j \\ \gamma_{j,i}, & \text{if } j \leq i \end{cases} \quad (i, j = 1, 2, \ldots, N),$$  \hspace{1cm} (31)

where

$$\gamma_{i,j} = \frac{\text{sh}(i\vartheta)\text{sh}(N+1-j)\vartheta}{\text{sh}(N+1)\vartheta}, \quad \vartheta = \text{arch}(w), \quad \text{with } w > 1.$$  \hspace{1cm} (32)

We have the relation $X_1^{-1} = (1/z)G$ (see [17]), thus a direct computation verifies the validity of the following

**Lemma 3.1** For the matrices $X_1$ and $X_2$ of the form (29) the iteration matrix $X = X_1^{-1}X_2$ can be expressed as

$$X = \frac{s}{z}[(2w + p)G - I].$$  \hspace{1cm} (33)

Hence, taking into the account the conditions (30), we get the following statement.

**Lemma 3.2** Under the condition (30) the iteration matrix $X \in \mathbb{R}^{N \times N}$ for arbitrary fixed $N$ is non-negative if and only if the conditions

$$2w + p > 0$$  \hspace{1cm} (34)

and

$$\gamma_{i,i} \geq \frac{1}{2w + p}, \quad i = 1, 2, \ldots, N$$  \hspace{1cm} (35)

are fulfilled.
Now we analyze the expression on left hand side in condition (35).

**Lemma 3.3** For the diagonal elements of the matrix \(X\) the relation

\[
\min \{ \gamma_{i,i}, \; i = 1, 2, \ldots, N \} = \gamma_{1,1} = \gamma_{N,N}
\]  

holds.

**Proof.** Introducing the functions

\[
h_1(y) = K_1 \text{sh}(Cy) \; \text{sh}(C(N + 1 - y)) \quad \text{and} \quad h_2(y) = K_2y(N + 1 - y)
\]

on the interval \([1, N]\), (where \(K_1, \; K_2, \; C\) are some positive constants), one can check that both functions take their maxima at the same point \(y=(N+1)/2\). Moreover, on the interval \([1,(N+1)/2]\) they are monotonically increasing, while on the interval \(((N+1)/2,N]\) they are monotonically decreasing. Using this fact and the expressions for \(\gamma_{i,i}\), we get the statement. □

Combining Lemma 3.2 and Lemma 3.3, we obtain

**Theorem 3.1** Under the conditions (30), for arbitrary fixed \(N\) the iteration matrix \(X \in \mathbb{R}^{N \times N}\) is non-negative if and only if the conditions (34) and

\[
a(N) := \frac{\text{sh}(N \vartheta)}{\text{sh}((N + 1)\vartheta)} \geq \frac{1}{2w + p}
\]  

are satisfied.

Obviously, (34) and (37) are necessary and sufficient conditions of the non-negativity for some fixed dimension \(N\). Let us turn to the examination of the varying \(N\). Due the relations

\[
\frac{\text{sh}(N \vartheta)}{\text{sh}((N + 1)\vartheta)} = \text{ch}(\vartheta) - \text{coth}((N + 1)\vartheta)\text{sh}(\vartheta),
\]

we have

\[
\sup \left\{ \frac{\text{sh}(N \vartheta)}{\text{sh}((N + 1)\vartheta)} : \; N \in \mathbb{N} \right\} = \text{ch}(\vartheta) - \text{sh}(\vartheta) = \exp(-\vartheta).
\]  

Since the sequence \(a(N)\) is monotonically increasing, it converges to its limit (which is its superior) monotonically. Thus, the conditions (34) and (37), that is, the necessary and sufficient conditions for some fixed \(N\), serve as sufficient condition of the non-negativity of the matrices \(X \in \mathbb{R}^{N_1 \times N_1}\) for all \(N_1 \geq N\).

Let us observe that

\[
\exp(-\vartheta) = \exp(-\text{arch}(w)) = \exp \left( \ln \left[ w + \sqrt{w^2 - 1} \right]^{-1} \right) = \left[ w + \sqrt{w^2 - 1} \right]^{-1}.
\]

Therefore, from some sufficiently large \(N_0 \in \mathbb{N}\) the relation \(X \geq 0\) may be true only if the condition

\[
\left[ w + \sqrt{w^2 - 1} \right]^{-1} > \frac{1}{2w + p},
\]

volume 3 (2010), number 2

69
i.e., the condition
\[ p > -w + \sqrt{w^2 - 1} \]  
(42)
is fulfilled. This proves the following

**Theorem 3.2** Assume that the conditions in (30) are satisfied. If, for some number \( N_0 \in \mathbb{N} \), the conditions (34) and (37) are satisfied, then, all matrices \( X \in \mathbb{R}^{N \times N} \) with \( N \geq N_0 \), are non-negative. Moreover, there exists such a number \( N_0 \), if and only if the condition (42) holds.

**Remark 3.1** Since
\[ a(1) = \frac{\text{sh} \vartheta}{\text{sh}(2 \vartheta)} = \frac{1}{2 \text{ch} \vartheta} = \frac{1}{2w}, \]
therefore, (37) results in the condition
\[ p \geq 0. \]  
(43)

**Remark 3.2** Due to the relation
\[ a(2) = \frac{\text{sh}(2 \vartheta)}{\text{sh}(3 \vartheta)} = \frac{2 \text{ch}(\vartheta)}{4 \text{ch}^2(\vartheta) - 1} = \frac{2w}{4w^2 - 1}, \]
condition (37) results in the assumption
\[ p \geq -\frac{1}{2w}. \]  
(44)

That is, \( X \in \mathbb{R}^{N \times N} \) is non-negative for all \( N = 2, 3, \ldots \), if and only if \( X_1 \) is an M-matrix and (44) is valid.

**Remark 3.3** The conditions (43) and (44) (corresponding to the cases \( N = 1 \) and \( N = 2 \), respectively) are sufficient conditions for the non-negativity of the matrix \( X \) in any larger dimension. For increasing \( N \), the new conditions, which we obtain, are approaching to the necessary condition of non-negativity. Using (38) and (39) we can characterize the rate of the convergence: it is equal to the rate of convergence of the sequence \( \{\coth(N\vartheta), n = 1, 2, \ldots\} \) to one. Clearly,

\[ \coth(N\vartheta) = 1 + \frac{2}{[\exp(\vartheta)]^{2N} - 1}. \]

Using (40),
\[ \exp(\vartheta) = w + \sqrt{w^2 - 1} =: \beta. \]  
(45)

Hence, the sequence of the bounds of the sufficient conditions converges linearly with the ratio \( 1/\beta^2 \) to the bound of the necessary condition.
3.2 Non-negativity of difference schemes in 1D

The results of the previous part can be used in the qualitative analysis of the finite difference and linear finite element mesh operators in 1D, given by the formula (26) and (27), respectively. First we investigate the finite difference method. According to (26), the corresponding matrices are uniformly continuant and, using the notation $q = \Delta t / h^2$, they can be written in the form (29) with the choice

$$z = \frac{\theta q}{\Delta t}, \quad s = \frac{(1 - \theta)q}{\Delta t}, \quad w = \frac{1 + 2\theta q}{2\theta q}, \quad p = \frac{1 - 2(1 - \theta)q}{(1 - \theta)q}. \quad (46)$$

First we consider two special choices for the parameter $\theta$. For the case $\theta = 0$, according to (26), we have $X_1 = \frac{1}{\Delta t} I$ and $X_2 = \frac{1}{\Delta t} I - (1 - \theta)Q$. Hence, $X$ is non-negative if and only if the condition

$$q \leq \frac{1}{2}, \quad (47)$$

is satisfied. For the case $\theta = 1$ we get $X_1 = \frac{1}{\Delta t} I + Q$ and $X_2 = \frac{1}{\Delta t} I$. Because such $X_1$ is monotone matrix, therefore we do not have any condition for the choice of the parameters $h$ and $\Delta t$.

In what follows we pass to the analysis of the case $\theta \in (0, 1)$. For this case, the conditions of (30) clearly are satisfied. Moreover, let us notice, that, under the choice (46) we have

$$2w + p = 1/\theta (1 - \theta)q,$$

hence the condition (34) is always satisfied.

Using (43), we directly get that the condition of the non-negativity preservation for all $N = 1, 2, \ldots$ is the condition

$$q \leq \frac{1}{2(1 - \theta)}. \quad (48)$$

However, the non-negativity preservation for all $N = 2, 3, \ldots$ should be guaranteed by the weaker condition (44), which, in our case yields the inequality

$$\frac{1 - 2(1 - \theta)q}{(1 - \theta)q} \geq -\frac{\theta q}{1 + 2\theta q}. \quad (49)$$

Solving this problem, we get the upper bound

$$q \leq \frac{-1 + 2\theta + \sqrt{1 - \theta(1 - \theta)}}{3\theta(1 - \theta)}, \quad (50)$$

which is larger than the bound in (47).

Our aim is to get the largest value for $q$ under which the non-negativity preservation for sufficiently large values $N$ still holds. Therefore we put the values $w$ and $p$ from (46) into the necessary condition (42). Then we should solve the inequality

$$\frac{1 - 2(1 - \theta)q}{(1 - \theta)q} \geq -\frac{1 + 2\theta q}{2\theta q} + \frac{\sqrt{1 + 4\theta q}}{2\theta q}. \quad (51)$$
The solution of (51) gives the bound

$$q \leq \frac{1 - \sqrt{1 - \theta}}{\theta(1 - \theta)}.$$ (52)

We can summarize our results in the following

**Theorem 3.3** The finite difference method is non-negativity preserving for each $N \geq 1$ if and only if the condition (48) holds. It is non-negativity preserving for each $N \geq 2$ only under the condition (50). There exists a number $N_0 \in \mathbb{N}$ such that the method is non-negativity preserving for each $N \geq N_0$, if and only if the condition (52) is satisfied.

We demonstrate our results on some special choice of $\theta$. Namely, we define the upper bounds for
- explicit Euler method ($\theta = 0$);
- fourth order method $\theta = 1/2 - 1/(12q)$, $q > 1/6$;
- Crank-Nicolson second order method ($\theta = 0.5$);
- implicit Euler method ($\theta = 1$).

The results are shown in Table 1.

We pass to the investigation of the linear finite element mesh operator. According to (27), the corresponding matrices are also symmetric, uniformly continuant, tridiagonal.

First we consider the special choices $\theta = 0$ and $\theta = 1$.

For $\theta = 0$ we get $X_1 = (1/6\Delta t)\text{tridiag}[1, 4, 1]$, i.e., we are not able to guarantee the monotonicity of $X_1$, which is required in the Remark 3.2. When $\theta = 1$, then $X_2 = (1/6\Delta t)\text{tridiag}[1, 4, 1]$, hence, the monotonicity of $X_1$ is the necessary and sufficient condition of the non-negativity preservation of the of the method. Hence, we get the condition $q \geq 1/6$.

Now we assume that $\theta \in (0, 1)$.

When $q = 1/(6\theta)$ then $X_1 = (1/\Delta t)I$, hence the only condition of the non-negativity preservation is $X_2 \geq 0$. This can be guaranteed only by the condition $q \leq (3(1 - \theta))^{-1}$. When $q = (3(1 - \theta))^{-1}$ then $X_2 = (1/6\Delta t)\text{tridiag}[1, 4, 1]$, hence the only condition is the monotonicity of $X_1$. As we can see, for this case this matrix is M-matrix, therefore, there is no additional condition for the non-negativity preservation.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>$N = 1$</th>
<th>$N = 2$</th>
<th>$N = \infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$0.5 - (12q)^{-1}$</td>
<td>0.8333</td>
<td>0.9574</td>
<td>0.9661</td>
</tr>
<tr>
<td>1</td>
<td>no bound</td>
<td>no bound</td>
<td>no bound</td>
</tr>
</tbody>
</table>

Table 1: Upper bounds for $q$ in several finite difference methods providing the non-negativity.
In what follows we may assume that $\theta \in (1/3, 1)$ and
\[
\frac{1}{6\theta} < q < \frac{1}{3(1-\theta)}.
\] (53)

Then we can use the form (29) with the choice
\[
z = \frac{1}{6\Delta t} - \frac{\theta}{h^2}, \quad s = \frac{1}{6\Delta t} + \frac{1-\theta}{h^2},
\]
\[
w = \frac{1}{3} + \theta q - \frac{1}{6}, \quad p = \frac{2}{3} - 2(1-\theta)q + \frac{1}{6}.
\] (54)

For this choice the assumption (30) is valid and $2w + p = [(\theta q - 1/6)((1-\theta)q + 1/6)]^{-1} > 0$. Therefore (34) is always satisfied. Let us notice that under the condition (53) the condition $z > 0$ is also satisfied.

Using (43), we get that the condition of the non-negativity preservation for all $N = 1, 2, \ldots$ is (43), which results in the upper bound
\[
q \leq \frac{1}{3(1-\theta)}.
\] (55)

The non-negativity preservation for all $N = 2, 3, \ldots$ should be guaranteed by the weaker condition (44), which, in our case yields the upper bound
\[
q \leq \frac{3(-1+2\theta) + \sqrt{9 - 16\theta(1-\theta)}}{12\theta(1-\theta)},
\] (56)

which is larger than the bound in (55).

Our aim is to get the largest value for $q$ under which the non-negativity preservation for sufficiently large values $N$ is still valid. Therefore we put the values $w$ and $p$ from (54) into the necessary condition (42). Hence, we obtain that for any fixed $\theta \in (0, 1)$ the suitable $q$ are the solution of the inequality
\[
\theta(1-\theta)q^2 - 1/6(\theta + 4)q + A \leq 0; \quad A = \sqrt{q\theta + 1/12[1/6 + (1-\theta)q]}.
\] (57)

We can summarize our results in the following

**Theorem 3.4** The linear finite element method is non-negativity preserving for any $\theta \in [1/3, 1]$,

- for each $N \geq 1$ if and only if the condition
\[
\frac{1}{6\theta} \leq q \leq \frac{1}{3(1-\theta)};
\] (58)
<table>
<thead>
<tr>
<th>$\theta$</th>
<th>$N = 1$</th>
<th>$N = 2$</th>
<th>$N = \infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>not allowed</td>
<td>not allowed</td>
<td>not allowed</td>
</tr>
<tr>
<td>0.5</td>
<td>$1/3 \leq q \leq 2/3$</td>
<td>$1/3 \leq q \leq \sqrt{5}/3$</td>
<td>$1/3 \leq q \leq 0.748$</td>
</tr>
<tr>
<td>1</td>
<td>$1/6 \leq q$</td>
<td>$1/6 \leq q$</td>
<td>$1/6 \leq q$</td>
</tr>
</tbody>
</table>

Table 2: Upper and lower bounds for $q$ in several linear finite element methods providing the non-negativity.

- for each $N \geq 2$ if and only if the condition

$$\frac{1}{6\theta} \leq q \leq \frac{3(-1 + 2\theta) + \sqrt{9 - 16\theta(1 - \theta)}}{12\theta(1 - \theta)}$$

holds. There exists a number $N_0 \in \mathbb{N}$ such that the method is non-negativity preserving for each $N \geq N_0$ if and only if the condition (57) is satisfied.

We demonstrate our results again on some special choice of $\theta$. The results are shown in Table 2.

4 The non-negativity preservation in 2D FEM case

We consider the non-negativity preservation property of the discretization of the two-dimensional heat conduction problem with pure homogenous Dirichlet boundary conditions. For the simplicity, the constant coefficient is assumed to be equal one, however, a more general analysis was done in a previous work [20].

The general form of the two-dimensional heat conduction equation on $\Omega \times (0, T)$, where $\Omega := (0, L_x) \times (0, L_y)$, is

$$\frac{\partial u}{\partial t} = \Delta u, \quad (x, y) \in \Omega, \ t \in (0, T),$$
$$u|_{\Gamma_0} = 0, \ t \in [0, T)$$
$$u(x, y, 0) = u_0(x, y), \ (x, y) \in \Omega, \quad (60)$$

where $u$ is the temperature of the analyzed domain, $t$ and $x, y$ denote the time and space variables, respectively.

In the course of the analysis of the problem the space was divided into $2 \cdot (n_x + 1) \cdot (n_y + 1)$ triangle elements. Then we get the problem (9) with the matrices from $\mathbb{R}^{n_x^2 \times n_y^2}$ in the form for the linear finite element method the corresponding matrices are

$$X_1 = \frac{1}{\Delta t} M - \theta Q, \quad (61)$$
$$X_2 = \frac{1}{\Delta t} M + (1 - \theta)Q, \quad (62)$$
where, for bilinear shape functions \[15\]

\[
Q = \text{tridiag}(Q_I, Q_A, Q_I)
\] (63)

and

\[
M = h_x h_y \text{tridiag}(M^T_D, M_A, M_D)
\] (64)

respectively, where

\[
Q_A = \frac{h_y}{h_x} \text{tridiag} \left(1, -2 \left[1 + \frac{h_x^2}{h_y^2}\right], 1\right),
\] (65)

\[
Q_I = \frac{h_x}{h_y} \text{tridiag}(0, -1, 0),
\] (66)

\[
M_A = \frac{1}{12} \text{tridiag}(1, 6, 1), ~ M_D = \frac{1}{12} \text{tridiag}(0, 1, 1),
\] (67)

moreover, \(h_x\) and \(h_y\) are the lengths of the spatial approximations. For one dimensional linear spline functions see \[11\].

It is clear that for non-negativity preservation property we require the condition

\[
X = X_1^{-1} X_2 \geq 0.
\] (68)

The sufficient conditions of the non-negativity of \(X\) are the following:

\[
X_1^{-1} \geq 0 \text{ and } X_2 \geq 0.
\] (69)

**Remark 4.1** The decomposition of \(X_1 - X_2 = \Delta t Q\) with the property (69), is called a regular matrix splitting \[16\].

For \(X_2\) it is easy to give a condition that guarantees its non-negativity by analyzing the elements of the matrix. By a direct computation we get the condition

\[
\frac{h_y h_x}{2 \Delta t} - 2 \left(\frac{h_x}{h_y} + \frac{h_y}{h_x}\right)(1 - \theta) \geq 0,
\] (70)

which yields the upper bound

\[
\frac{h_x h_y}{4 \left(\frac{h_x}{h_y} + \frac{h_y}{h_x}\right)(1 - \theta)} \geq \Delta t.
\] (71)

It is not possible to obtain a sufficient condition for the non-negativity of the matrix \(X_1^{-1}\) by the so-called M-matrix method \[21\]. This also follows from the fact that \(X_1\) contains positive elements in its off-diagonal. Therefore, a sufficient condition for the inverse-positivity of matrix \(X_1\) will be obtained by some other criteria.
Lemma 4.1 \cite{14} Let $A$ be an $n$-by-$n$ matrix, denote $A_d$ and $A^-$ the diagonal and the negative off-diagonal part of the matrix $A$, respectively.

Let $A^- = A^z + A^s = (a^z_{ij}) + (a^s_{ij})$. If

$$a_{ij} \leq \sum_{k=1}^{n} a^z_{ik} a^s_{kj}, \text{ for all } a_{ij}, i \neq j,$$

then $A$ is a product of two $M$-matrices, i.e., $A$ is monotone.

We will analyze the monotonicity of $X_1$ with the help of this lemma. We can do it because it is a square matrix and it can be decomposed into the diagonal part, the positive off-diagonal part, the upper triangular and lower triangular negative parts. All the conditions of the lemma are satisfied if

$$\frac{1}{12} \leq \left( \frac{1}{12} - \frac{\Delta t \theta}{h_x^2} \right) \left( \frac{1}{12} - \frac{\Delta t \theta}{h_y^2} \right),$$

which implies the lower bound

$$\frac{h_y^2}{12 \theta} \left( \frac{3}{2} \left( \frac{h_x^2}{h_y^2} + 1 \right) + \sqrt{\frac{9}{4} \left( \frac{h_x^4}{h_y^4} + 1 \right) + \frac{19}{2} \left( \frac{h_x^2}{h_y^2} \right)} \right) \leq \Delta t.$$

Hence, the next statement is proven.

Theorem 4.3 Let us assume that the conditions (71) and (74) hold. Then for the problem (60) on a rectangular domain with an arbitrary non-negative initial condition the linear finite element method results in a non-negative solution on any time level.

Remark 4.2 If $\theta = 1$, there is no upper bound for the time-step size, nor any condition for the ratio of the lengths of the spatial approximations.

Remark 4.3 If the conditions of the theorem hold, then the following complementary conditions are also satisfied:

- For the ratio of the lengths of the spatial approximations

$$\sqrt{\omega - \sqrt{\omega^2 - 1}} \leq \frac{h_x}{h_y} \leq \sqrt{\omega + \sqrt{\omega^2 - 1}},$$

where

$$\omega = \frac{10 T^2 + 2T - 1}{-\frac{10}{9} T^2 - 2T}.$$
This yields a geometrical restriction for the shape of the partition of the space domain in the linear FEM.

• For \( \theta \), which is the parameter of the applied numerical method, we have the bound

\[
\theta \geq \frac{1}{\frac{63}{50} + \frac{1}{10}}.
\]  

Since the right-hand side is greater than 0.818, this implies that for the Crank-Nicolson method (\( \theta = 0.5 \)) we cannot guarantee the non-negativity by this principle [3].

5 Numerical experiments

In the following figures we illustrate the possible choice of the discretization step-sizes \( q = \Delta t/h^2 \) in the one-dimensional case for finite difference and for linear finite element methods, for different values of \( \theta \).

In the course of the numerical experiments in 2D \((n_x = 20, n_y = 25, h_x = 0.1, h_y = 0.04)\) for the homogenous initial condition \( u_0(x, y) = 300K \) was considered. For the numerical experiments, the tridiagonal matrix algorithm (TDMA) was used for the inversion of the sparse tridiagonal matrices [18]. The following figures are in three dimensions, in Fig. 3 the first two dimensions are the spatial ones \((x, y)\) and the third is the temperature at the nodes. First, we apply the Crank-Nicolson method and a relatively long time step \((\theta = 0.5, \Delta t = 10^{-2}, \text{Timesteps} = 1)\), which results in a negative \( X_2 \).

For the sake of completeness, in Fig. 4 we applied the time-step size from the interval (71) and (74) \((\theta = 1, \Delta t = 0.02, \text{Timesteps} = 10)\), and it can be seen we have got a more realistic solution.
Figure 2: Condition for the choice of the discretization step-sizes for linear FEM in 1D.

Figure 3: The solution obtained by the Crank-Nicolson method and relatively long time step.
6 Conclusions

In this paper, we have considered the non-negativity preservation property for the linear parabolic PDE’s. We established the direct connection of the non-negativity preservation for the semidiscrete solutions between the one and two-dimensional cases. Namely, we proved that the two-dimensional problem has this property if the corresponding one-dimensional problem is non-negativity preserving. The conditions posed are satisfied for all arbitrary (linear FEM and FDM) methods. Moreover, we gave the explicit formula for the inversion of the block-tridiagonal matrices with scalar tridiagonal matrices placed along their diagonals.

Acknowledgement

The first author was supported by Hungarian National Research Fund OTKA No. K67819. The second author was supported by project no. 124619 from the Academy of Finland. The first and the third authors were supported by Jedlik project ”ReCoMend” 2008–2011.

References


Current address

István Faragó
Department of Applied Analysis and Computational Mathematics
Eötvös Loránd University
H-1117, Budapest, Pázmány P. s. 1/c. Hungary
e-mail: faragois@cs.elte.hu

Sergey Korotov
Department of Applied Analysis and Computational Mathematics
Eötvös Loránd University
H-1117, Budapest, Pázmány P. s. 1/c. Hungary
e-mail: smkorotov@gmail.com

Department of Mathematics
Tampere University of Technology
P.O. Box 553, FIN–33101 Tampere, Finland
e-mail: sergey.korotov@tut.fi

Tamás Szabó
Department of Applied Analysis and Computational Mathematics
Eötvös Loránd University
H-1117, Budapest, Pázmány P. s. 1/c. Hungary
e-mail: szabot@cs.elte.hu