On the monotonicity conservation in numerical solutions of the heat equation

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computing Science

RANA 00-15
December 2000

On the Monotonicity Conservation in Numerical Solutions of the Heat Equation

by

R. Horvath

Reports on Applied and Numerical Analysis
Department of Mathematics and Computing Science
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven, The Netherlands
ISSN: 0926-4507

On the Monotonicity Conservation in Numerical Solutions of the Heat Equation

Róbert Horváth

University of Technology Eindhoven
Eindhoven, The Netherlands
e-mail: rhorvath@natlab.research.philips.com

Abstract

It is important to choose such numerical methods in practice that mirror the characteristic properties of the described procces beyond the stability and convergence. The investigated qualitative property in this paper is the conservation of the monotonicity in space of the initial heat distribution. We prove some statements about the monotonicity conservation of one-step vector-iterations. Then, applying these results, we consider the numerical solutions of the one-dimensional heat equation where the approximations of the exact solution are generated by the so-called (σ, θ)-method ([1]). Our main theorem formulates the necessary and sufficient condition of the uniform monotonicity conservation.

1 Introduction

The temperature changes of a homogeneous, isotropic rod can be described by the one-dimensional heat equation, which has the form

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad x \in (0,1), \quad t > 0, \tag{1.1}$$

$$u(0,t) = \mu_1, \quad t \geq 0, \tag{1.2}$$

$$u(1,t) = \mu_2, \quad t \geq 0, \tag{1.3}$$

$$u(x,0) = u_0(x), \quad x \in (0,1), \tag{1.4}$$

where we have chosen constant boundary conditions and have used non-dimensional variables ([2]). The sufficiently smooth function $u_0 : (0,1) \to \mathbb{R}$ describes the initial temperature of the rod and the function $u = u(x,t)$ denotes the temperature in the point $x \in [0,1]$ and at the point of time $t \geq 0$.

To solve the above problem numerically we define a uniform mesh on the set $\Omega = [0,1] \times [0,\infty)$ with the parameters $\tau > 0$ and $h = 1/(n+1)$ $(n \in \mathbb{N})$ as follows

$$\Omega_{h,\tau} := \{(x_i, t_j) \in \Omega \quad | \quad x_i = ih \quad (i = 0, \ldots, n+1), \quad t_j = j\tau \quad (j \in \mathbb{N})\}. \tag{1.5}$$

We denote the approximation to the exact value $u(ih, j\tau)$ by $y_i^{(j)}$ and we set $\mathbf{y}^{(j)} = (y_1^{(j)}, \ldots, y_n^{(j)})^\top \in \mathbb{R}^n$. In this paper the approximating values are generated by the so-called (σ, θ)-method ([1]), that is we generate the values $y_i^{(j)}$ by the iteration

$$\mathbf{M}\frac{\mathbf{y}^{(j+1)} - \mathbf{y}^{(j)}}{\tau} = -\theta\frac{1}{h^2}\mathbf{Q}\mathbf{y}^{(j+1)} - (1-\theta)\frac{1}{h^2}\mathbf{Q}\mathbf{y}^{(j)} + \frac{1}{h^2}(\mu_1\mathbf{e}^1 + \mu_2\mathbf{e}^n)\ , \quad j = 0, 1, \ldots\ . \tag{1.6}$$

Here $\mathbf{e}^1 = (1, 0, \ldots, 0)^\top$, $\mathbf{e}^n = (0, \ldots, 1)^\top \in \mathbb{R}^n$, $\mathbf{Q} = tridiag[-1, 2, -1] \in \mathbb{R}^{n\times n}$ and $\mathbf{M} = \mathbf{I} - \sigma\mathbf{Q} \in \mathbb{R}^{n\times n}$ are tridiagonal matrices ($\mathbf{I}$ is the unit matrix), $\sigma \in [0, 1/4)$ and $\theta \in [0, 1]$ are given parameters. The vector $\mathbf{y}^{(0)}$ is a suitable approximation to the initial function u_0.

The (σ, θ)-method unites a few remarkable numerical methods. For example, the $(0, \theta)$-method gives the classical finite difference θ-method (the $(0, 1/2)$-method is the well-known Crank-Nicolson method) and the $(1/6, \theta)$-method results in the finite element method with linear elements. In this sense the (σ, θ)-method can be considered as a generalization of the classical methods.

The condition of the convergence of the method can be found in [1]. Moreover, it is known, that it is not enough to construct a convergent numerical method in practice, the method must be qualitatively adequate, too. This means that we have to require among others the nonnegativity and shape conservation of the initial function, the sign-stability and maximum norm contractivity of the numerical solution (see e.g. [3],[4],[5],[6],[7],[8],[9]). Let us execute the (σ, θ)-method with the parameters $\sigma = 0$, $\theta = 1/2$ (this is the Crank-Nicolson method, which is unconditionally stable), $\mu_1 = 0$, $\mu_2 = 1$, $q = 20$ and with the initial vector $\mathbf{y}^{(0)} = (0, 1/2, 1/2, 1/2, 1)^\top$. This method is stable, but it does not suit the qualitative requirements. Namely, the initial vector describes a monotonically increasing heat distribution in space while the first iterate $\mathbf{y}^{(1)} = (0.32, 0.17, 0.5, 0.82, 0.68)$ does not. In this case we say that the numerical method does not conserve the monotonicity. If the monotonicity is not conserved, then the considered method does not describe a real physical process, because this phenomenon contradicts the second law of Thermodynamics.

In this paper we determine the conditions of the monotonicity conservation of the (σ, θ)-method. Introducing the notations $q = \tau/h^2$, $z = \theta q - \sigma$, $\mathbf{T}_1 = \mathbf{I} + z\mathbf{Q}$, $\mathbf{T}_2 = \mathbf{I} - (q - z)\mathbf{Q}$ and $\mathbf{T} = \mathbf{T}_1^{-1}\mathbf{T}_2$ the iteration (1.6) can be written in the one-step vector-iteration form

$$\widehat{\mathbf{y}}^{(j+1)} = \begin{bmatrix} 1 & \mathbf{0}^\top & 0 \\ q\mathbf{T}_1^{-1}\mathbf{e}^1 & \mathbf{T} & q\mathbf{T}_1^{-1}\mathbf{e}^n \\ 0 & \mathbf{0}^\top & 1 \end{bmatrix} \widehat{\mathbf{y}}^{(j)} = \widehat{\mathbf{T}}\widehat{\mathbf{y}}^{(j)}, \quad j = 0, 1, \ldots, \tag{1.7}$$

where $\widehat{\mathbf{y}}^{(j)} = (\mu_1, (\mathbf{y}^{(j)})^\top, \mu_2)^\top \in \mathbb{R}^{n+2}$ and $\widehat{\mathbf{T}} \in \mathbb{R}^{(n+2)\times(n+2)}$. That is why we investigate the one-step iterations from monotonicity conservation point of view in Section 2. In Section 3 we apply the linear algebraic results of Section 2 for the special iteration (1.7).

2 Monotonicity conservation of one-step vector-iterations

A vector $\mathbf{x} \in \mathbb{R}^n$ is said to be monotonically increasing (decreasing) if the relations $x_1 \leq x_2 \leq \ldots \leq x_n$ ($x_1 \geq x_2 \geq \ldots \geq x_n$) hold. Let us consider the one-step vector-iteration process

$$\mathbf{x}^{(j+1)} = \mathbf{A}\mathbf{x}^{(j)}, \quad j = 0, 1, \ldots, \tag{2.1}$$

where $\mathbf{x}^{(0)}$ is an arbitrary initial vector and $\mathbf{A} \in \mathbb{R}^{n \times n}$ is an arbitrary matrix. We say that the iteration (2.1) is monotonicity conserving if the monotone increase (monotone decrease) of any initial vector $\mathbf{x}^{(0)}$ implies the same for the vector $\mathbf{x}^{(1)}$. These notions of the monotonicity of a vector and the monotonicity conservation of an iteration can be handled only with difficulties, therefore we introduce an other monotonicity notion, where we can use a matrix representation form. Let us denote the matrix *tridiag* $[-1, 1, 0]$ by $\mathbf{D}$. Since $D^{-1}_{i,j} = 1$ if $i \geq j$ and $D^{-1}_{i,j} = 0$ if $i < j$, therefore $\mathbf{D}$ is a so-called inverse-positive matrix.

DEFINITION 2.1. The iteration (2.1) is called totally monotone if for any vector $\mathbf{x}^{(0)}$ such that $\mathbf{D}\mathbf{x}^{(0)} \geq 0$ (or $\mathbf{D}^\top\mathbf{x}^{(0)} \geq 0$) the relation $\mathbf{D}\mathbf{x}^{(1)} \geq 0$ (or $\mathbf{D}^\top\mathbf{x}^{(1)} \geq 0$) holds.

REMARK 2.2. Since the relation $\mathbf{D}\mathbf{x}^{(0)} \geq 0$ can be satisfied only for nonnegative vectors, therefore, a totally monotone iteration produces nonnegative monotonically increasing (decreasing) vectors from all nonnegative monotonically increasing (decreasing) vectors. However, a monotonicity conserving iteration produces monotonically increasing (decreasing) vectors from all monotonically increasing (decreasing) vectors. Later we will prove (Theorem 2.10) that in case of $\mathbf{A} \geq 0$ these properties are equivalent.

Theorem 2.3. *The iteration (2.1) is totally monotone if and only if both of the conditions* $\mathbf{D}\mathbf{A}\mathbf{D}^{-1} \geq 0$ *and* $\mathbf{D}^\top\mathbf{A}\mathbf{D}^{-\top} \geq 0$ *are fulfilled.*

Proof. Let $\mathbf{x} \in \mathbb{R}^n$ be an arbitrary nonnegative vector. In this case for the vector $\mathbf{D}^{-1}\mathbf{x}$ the condition $\mathbf{D}(\mathbf{D}^{-1}\mathbf{x}) \geq 0$ is valid. Thus, if the iteration is totally monotone, then the relation $\mathbf{D}\mathbf{A}\mathbf{D}^{-1}\mathbf{x} \geq 0$ holds for all nonnegative vectors $\mathbf{x}$. So we get the condition that the matrix $\mathbf{D}\mathbf{A}\mathbf{D}^{-1}$ must be nonnegative. Conversely, if $\mathbf{D}\mathbf{A}\mathbf{D}^{-1} \geq 0$, then assuming $\mathbf{D}\mathbf{x}^{(0)} \geq 0$ follows $(\mathbf{D}\mathbf{A}\mathbf{D}^{-1})(\mathbf{D}\mathbf{x}^{(0)}) = \mathbf{D}\mathbf{x}^{(1)} \geq 0$. The condition $\mathbf{D}^\top\mathbf{A}\mathbf{D}^{-\top} \geq 0$ can be obtained considering the vectors $\mathbf{D}^{-\top}\mathbf{x}$. ■

REMARK 2.4. We would like to shed light on the conditions of Theorem 2.3. Let us introduce the notations $S^r_{i,k} = \sum_{j=k}^{n} A_{i,j}$ and $S^l_{i,k} = \sum_{j=1}^{k} A_{i,j}$. Then the iteration (2.1) is totally monotone if and only if the relations

$$S^r_{i+1,k} \geq S^r_{i,k}, \quad S^l_{i+1,k} \leq S^l_{i,k} \quad (k = 1, \ldots, n;\ i = 1, \ldots, n-1) \tag{2.2}$$

hold and the numbers $S^r_{1,k}, S^l_{n,k}$ are nonnegative. Using the above consequences for $k = 1$ and $k = n$, respectively, we get a necessary condition of the total monotonicity: the sum of the elements in the rows of the matrix $\mathbf{A}$ is some fixed nonnegative constant.

REMARK 2.5. Multiplying the inequalities $\mathbf{DAD}^{-1} \geq 0$ and $\mathbf{D}^{\top}\mathbf{AD}^{-\top} \geq 0$ by the nonnegative matrices $\mathbf{D}^{-1}$ and $\mathbf{D}^{-\top}$, respectively, we get that a necessary condition of the total monotonicity is $\mathbf{AD}^{-1} \geq 0$ and $\mathbf{AD}^{-\top} \geq 0$.

REMARK 2.6. We notice that the nonnegativity of the matrix $\mathbf{A}$ is not necessary to the total monotonicity of the iteration. As one can see, with the not nonnegative matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & -1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.3}$$

the iteration is a totally monotone one.

REMARK 2.7. Let $\tilde{\mathbf{I}}$ denote the matrix with the elements $\tilde{\mathbf{I}}_{i,(n-i+1)} = 1$ if $i = 1, \ldots, n$ and $\tilde{\mathbf{I}}_{i,j} = 0$ otherwise. We call the matrix $\mathbf{A}$ doubly symmetric if the relation $\tilde{\mathbf{I}}\mathbf{A}\tilde{\mathbf{I}} = \mathbf{A}$ holds. If the matrix $\mathbf{A}$ is symmetric and it is also symmetric for the secondary diagonal, then the matrix $\mathbf{A}$ is doubly symmetric. If the matrix $\mathbf{A}$ is doubly symmetric, then the iteration (2.1) is totally monotone if and only if $\mathbf{DAD}^{-1} \geq 0$. To see this we have to show that the condition $\mathbf{DAD}^{-1} \geq 0$ yields the condition $\mathbf{D}^{\top}\mathbf{AD}^{-\top} \geq 0$. If $\mathbf{DAD}^{-1} \geq 0$, then $\tilde{\mathbf{I}}\mathbf{DAD}^{-1}\tilde{\mathbf{I}} \geq 0$. Hence

$$0 \leq \tilde{\mathbf{I}}\mathbf{DAD}^{-1}\tilde{\mathbf{I}} = \mathbf{D}^{\top}\tilde{\mathbf{I}}\mathbf{A}\tilde{\mathbf{I}}\mathbf{D}^{-\top} = \mathbf{D}^{\top}\mathbf{AD}^{-\top}. \tag{2.4}$$

REMARK 2.8. If an iteration is totally monotone with the matrices $\mathbf{A}_1$ and $\mathbf{A}_2$, then it is that with the matrix $\mathbf{A}_1\mathbf{A}_2$, too.

Theorem 2.9. *If the iteration (2.1) is totally monotone, then it conserves the monotonicity, too.*

Proof. For simplicity we can assume that $\mathbf{x}^{(0)}$ is a monotonically increasing vector. Then $\mathbf{x}^{(0)}$ can be written in the form $\mathbf{x}^{(0)} = \mathbf{v} - c\mathbf{e}$, where $\mathbf{v} \geq 0$ is a monotonically increasing vector, $\mathbf{e} = (1, \ldots, 1)^{\top}$ and c is a suitable real number. Then $\mathbf{Ax}^{(0)} = \mathbf{A}(\mathbf{v} - c\mathbf{e}) = \mathbf{Av} - c\mathbf{Ae}$. Since the sum of the elements in the rows of the matrix $\mathbf{A}$ is constant therefore the vector $c\mathbf{Ae}$ is a constant vector. On the other

hand, the vector $\mathbf{Av}$ is monotonically increasing because of the total monotonicity. Thus the vector $\mathbf{Ax}^{(0)}$ is also monotonically increasing. ■

Theorem 2.10. *Assume that* $\mathbf{A} \geq 0$. *Then the iteration (2.1) is totally monotone if and only if it conserves the monotonicity.*

Proof. The necessity follows from the previous theorem. To see that the condition is also sufficient let us suppose that for a nonnegative monotonically increasing vector $\mathbf{x}^{(0)}$ the condition $\mathbf{Dx}^{(0)} \geq 0$ is valid. Obviously, $\mathbf{Ax}^{(0)}$ is a nonnegative and monotonically increasing vector. So the relation $\mathbf{DAx}^{(0)} \geq 0$ is valid. For the case of $\mathbf{D}^{\top}\mathbf{x}^{(0)} \geq 0$ the proof is similar. ■

3 Monotonicity conservation of the (σ, θ)-method

DEFINITION 3.1. We say that the (σ, θ)-method is monotonicity conserving (resp. totally monotone) if the iteration (1.7) is the same. The (σ, θ)-method is said to be uniformly monotonicity conserving (resp. uniformly totally monotone) for a fixed value q if the iteration (1.7) is monotonicity conserving (resp. totally monotone) for all step-sizes $h = 1/(n+1)$.

In this section we give the necessary and sufficient condition of the uniform monotonicity conservation (resp. uniform total monotonicity) of the numerical solution. Our results follow from the application of the theorems of the previous section for the special iteration (1.7). Moreover, we apply the fact (see [8]) that the matrix $\mathbf{T}$ can be expressed in the form

$$\mathbf{T} = \frac{1}{z}\left[\frac{q}{z}\mathbf{G} - (q-z)\mathbf{I}\right] \text{ if } z \neq 0, \ \mathbf{T} = \mathbf{I} - q\mathbf{Q} \text{ if } z = 0, \tag{3.1}$$

where the matrix $\mathbf{G} \in \mathbb{R}^{n\times n}$ is a symmetric, one-pair matrix defined as follows: $G_{i,j} = \Gamma_{i,j}$ if $z > 0$ and $G_{i,j} = (-1)^{i+j-1}\Gamma_{i,j}$ if $z < 0$, where $\alpha = \operatorname{arch}|1 + 1/(2z)|$ and

$$\Gamma_{i,j} = \begin{cases} \gamma_{i,j}, & \text{if } i \leq j \\ \gamma_{j,i}, & \text{if } i > j \end{cases}, \quad \gamma_{i,j} = \frac{\operatorname{sh}(i\alpha)\operatorname{sh}((n+1-j)\alpha)}{\operatorname{sh}\alpha \operatorname{sh}((n+1)\alpha)}. \tag{3.2}$$

Because of Theorem 2.9, if the iteration (1.7) is totally monotone then it conserves the monotonicity, too. If (1.7) conserves the monotonicity, then it is totally monotone, because $\widehat{y}_1^{(j)} \geq 0$ (resp. $\widehat{y}_{n+2}^{(j)} \geq 0$) implies $\widehat{y}_1^{(j+1)} \geq 0$ (resp. $\widehat{y}_{n+2}^{(j+1)} \geq 0$) with the matrix $\widehat{\mathbf{T}}$. Taking into consideration Remark 2.7 and Theorem 2.3 the iteration (1.7) is totally monotone if and only if the condition

$$\widehat{\mathbf{D}}\widehat{\mathbf{T}}\widehat{\mathbf{D}}^{-1} \geq 0 \tag{3.3}$$

holds, where

$$\widehat{\mathbf{D}} = \begin{bmatrix} 1 & \mathbf{0}^{\top} & 0 \\ -\mathbf{e}^1 & \mathbf{D} & \mathbf{0} \\ 0 & -(\mathbf{e}^n)^{\top} & 1 \end{bmatrix} \in \mathbb{R}^{(n+2)\times(n+2)} \tag{3.4}$$

and $\mathbf{D} \in \mathbb{R}^{n\times n}$ is the matrix introduced in the previous section. Our next aim is to express the condition (3.3) with the parameters σ, θ and q.

If $z = 0$, then $\widehat{\mathbf{D}}\widehat{\mathbf{T}}\widehat{\mathbf{D}}^{-1}$ is the matrix

$$\widehat{\mathbf{D}}\widehat{\mathbf{T}}\widehat{\mathbf{D}}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & & \dots & & 0 \\ 0 & 1-q & q & 0 & & \dots & & 0 \\ 0 & q & 1-2q & q & 0 & \dots & & 0 \\ 0 & 0 & q & 1-2q & q & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & & 0 & q & 1-2q & q & 0 \\ 0 & & \dots & & 0 & q & 1-2q & q \\ 0 & & & \dots & & 0 & q & 1-q \end{bmatrix}. \tag{3.5}$$

This matrix is nonnegative if and only if $q \leq 1/2$ (if $n = 1$, then $q \leq 1$). Let us suppose now that $z \neq 0$. Then $\widehat{\mathbf{D}}\widehat{\mathbf{T}}\widehat{\mathbf{D}}^{-1}$ can be written in the form

$$\widehat{\mathbf{D}}\widehat{\mathbf{T}}\widehat{\mathbf{D}}^{-1} = \begin{bmatrix} 1 & \mathbf{0}^\top & 0 \\ \mathbf{0} & \mathbf{D}\mathbf{T}\mathbf{D}^{-1} + q\mathbf{D}\mathbf{T}_1^{-1}\mathbf{e}^n\mathbf{e}^\top & q\mathbf{D}\mathbf{T}_1^{-1}\mathbf{e}^n \\ 0 & -(\mathbf{e}^n)^\top(\mathbf{T}\mathbf{D}^{-1} + q\mathbf{T}_1^{-1}\mathbf{e}^n\mathbf{e}^\top) + \mathbf{e}^\top & 1 - q(\mathbf{e}^n)^\top\mathbf{T}_1^{-1}\mathbf{e}^n \end{bmatrix} \tag{3.6}$$

($\mathbf{e} = (1, \dots, 1)^\top$). Applying that $z\mathbf{T}_1^{-1} = \mathbf{G}$ the matrix is nonnegative if and only if the conditions

$$\begin{array}{ll} (A) & (q/z)\mathbf{D}\mathbf{G}\mathbf{e}^n \geq 0 \\ (B) & 1 - (q/z)(\mathbf{e}^n)^\top\mathbf{G}\mathbf{e}^n \geq 0 \\ (C) & -(\mathbf{e}^n)^\top(\mathbf{T}\mathbf{D}^{-1} + (q/z)\mathbf{G}\mathbf{e}^n\mathbf{e}^\top) + \mathbf{e}^\top \geq 0 \\ (D) & \mathbf{D}\mathbf{T}\mathbf{D}^{-1} + (q/z)\mathbf{D}\mathbf{G}\mathbf{e}^n\mathbf{e}^\top \geq 0 \end{array} \tag{3.7}$$

are fulfilled. One can see that z must be positive (in this case $\mathbf{G} = \mathbf{\Gamma}$), because for negative values the matrix $\mathbf{G}$ would have "chess-board like" sign distribution and condition (A) could not be satisfied. For the term in condition (A) the estimation

$$(q/z)\mathbf{D}\mathbf{G}\mathbf{e}^n = (q/z)\mathbf{D}(\gamma_{1,n}, \dots, \gamma_{n,n})^\top = \tag{3.8}$$

$$= \frac{q}{z \cdot \mathrm{sh}((n+1)\alpha)}\mathbf{D}(\mathrm{sh}(\alpha), \mathrm{sh}(2\alpha), \dots, \mathrm{sh}(n\alpha))^\top \geq 0$$

is valid because $0 < \mathrm{sh}(\alpha) < \mathrm{sh}(2\alpha) < \dots < \mathrm{sh}(n\alpha)$.

Condition (B) results in the relation

$$1 - \frac{q}{z}\gamma_{n,n} \geq 0. \tag{3.9}$$

This is fulfilled for all values of n if and only if $1 - (q/z)e^{-\alpha} \geq 0$, because

$$\lim_{n\to\infty} \gamma_{n,n} = \lim_{n\to\infty} \frac{\mathrm{sh}(n\alpha)}{\mathrm{sh}((n+1)\alpha)} = e^{-\alpha} \tag{3.10}$$

and the sequence is monotonically increasing. Due to the equalities $e^{-\alpha} = \mathrm{ch}\alpha - \mathrm{sh}\alpha$ and $\mathrm{ch}\alpha = 1 + 1/(2z)$ the condition (3.9) is valid for all values of n if and only if $\sqrt{1+4z} \geq 2(q-z) - 1$. Considering the relation $z = \theta q - \sigma$ we get the necessary and sufficient condition

$$q \leq \frac{1 - 2(1-\theta)\sigma + \sqrt{1 - 4(1-\theta)\sigma}}{2(1-\theta)^2} \tag{3.11}$$

(if $\theta = 1$, then there is no restriction for q).

The term in the condition (C) can be rewritten in the following manner

$$-(\mathbf{e}^n)^\top(\mathbf{T}\mathbf{D}^{-1} + (q/z)\mathbf{G}\mathbf{e}^n\mathbf{e}^\top) + \mathbf{e}^\top = \tag{3.12}$$

$$= \mathbf{e}^\top - (\frac{q}{z^2}\Gamma_{n,1}, \ldots, \frac{q}{z^2}\Gamma_{n,n-1}, \frac{q}{z^2}\Gamma_{n,n} - \frac{q-z}{z})\mathbf{D}^{-1} - (q/z)(\gamma_{n,n}, \ldots, \gamma_{n,n}).$$

Considering the equality (see [8])

$$\sum_{j=1}^{n} |G_{i,j}| = \sum_{j=1}^{n} \Gamma_{i,j} = \frac{1}{|(1/z) + 2| - 2}(1 - \gamma_{1,i} - \gamma_{i,n}) \quad , \quad (i = 1, \ldots, n) \tag{3.13}$$

this vector is nonnegative if and only if the condition

$$1 - \frac{q}{z^2} \cdot z(1 - \gamma_{1,n} - \gamma_{n,n}) + \frac{q-z}{z} - \frac{q}{z}\gamma_{n,n} = \frac{q}{z}\gamma_{1,n} \geq 0 \tag{3.14}$$

is valid, which in turn means that the condition (C) is always satisfied.

For the left-hand side of the inequality in the condition (D) are true the following transformations

$$\mathbf{D}\mathbf{T}\mathbf{D}^{-1} + (q/z)\mathbf{D}\mathbf{G}\mathbf{e}^n\mathbf{e}^\top = \frac{1}{z}\mathbf{D}\mathbf{G}(\mathbf{T}_2\mathbf{D}^{-1} + q\mathbf{e}^n\mathbf{e}^\top) =$$

$$= \frac{1}{z}\mathbf{D}\mathbf{G}((\mathbf{T}_1 - q\mathbf{Q})\mathbf{D}^{-1} + q\mathbf{e}^n\mathbf{e}^\top) = \mathbf{I} - \frac{q}{z}\mathbf{D}\mathbf{G}\mathbf{Q}\mathbf{D}^{-1} + \frac{q}{z}\mathbf{D}\mathbf{G}\mathbf{e}^n\mathbf{e}^\top = \tag{3.15}$$

$$= \mathbf{I} - \frac{q}{z}\mathbf{D}\mathbf{G}(\mathbf{Q}\mathbf{D}^{-1} - \mathbf{e}^n\mathbf{e}^\top) = \mathbf{I} - \frac{q}{z}\mathbf{D}\mathbf{G}\mathbf{D}^\top.$$

We investigate the validity of the condition $\mathbf{I} - (q/z)\mathbf{D}\mathbf{G}\mathbf{D}^\top \geq 0$. The offdiagonal elements of the matrix are nonnegative, because the offdiagonal elements of the matrix $\mathbf{D}\mathbf{G}\mathbf{D}^\top$ are negative. To see this let us consider the following relations for the indices $i < j$ setting $\mathbf{G}_{i,j} = 0$ if $i = 0$ or $j = 0$ (because of the symmetricity it is sufficient to consider only these index choices)

$$(\mathbf{D}\mathbf{G}\mathbf{D}^\top)_{i,j} = \mathbf{G}_{i,j} + \mathbf{G}_{i-1,j-1} - \mathbf{G}_{i-1,j} - \mathbf{G}_{i,j-1} = \gamma_{i,j} + \gamma_{i-1,j-1} - \gamma_{i-1,j} - \gamma_{i,j-1} =$$

$$= \frac{\mathrm{sh}(i\alpha)\,\mathrm{sh}((n+1-j)\alpha) + \mathrm{sh}((i-1)\alpha)\,\mathrm{sh}((n+1-j+1)\alpha)}{\mathrm{sh}(\alpha)\,\mathrm{sh}((n+1)\alpha)} + \tag{3.16}$$

$$+ \frac{-\mathrm{sh}((i-1)\alpha)\,\mathrm{sh}((n+1-j)\alpha) - \mathrm{sh}(i\alpha)\,\mathrm{sh}((n+1-j+1)\alpha)}{\mathrm{sh}(\alpha)\,\mathrm{sh}((n+1)\alpha)} =$$

$$= \frac{(\mathrm{sh}((n+1-j)\alpha) - \mathrm{sh}((n+1-j+1)\alpha)) \cdot (\mathrm{sh}(i\alpha) - \mathrm{sh}((i-1)\alpha))}{\mathrm{sh}(\alpha)\,\mathrm{sh}((n+1)\alpha)} < 0.$$

The diagonal elements of the matrix $\mathbf{DGD}^\top$ are

$$(\mathbf{DGD}^\top)_{i,i} = \gamma_{i,i} + \gamma_{i-1,i-1} - 2\gamma_{i-1,i}, \quad i = 1, 2, \ldots, n. \tag{3.17}$$

One can show with a tedious calculation that

$$(\mathbf{DGD}^\top)_{i,i} \le \gamma_{1+n/2,1+n/2} + \gamma_{n/2,n/2} - 2\gamma_{n/2,1+n/2}, \quad i = 1, \ldots, n \tag{3.18}$$

and if n is even, then this estimation cannot be improved. Thus the following relations are true

$$\begin{aligned}
&(\mathbf{I} - (q/z)\mathbf{DGD}^\top)_{i,i} = 1 - (q/z)(\mathbf{DGD}^\top)_{i,i} \ge \\
&\ge 1 - (q/z)(\gamma_{1+n/2,1+n/2} + \gamma_{n/2,n/2} - 2\gamma_{n/2,1+n/2}) = \\
&= 1 - (q/z)\frac{2\mathrm{sh}(n\alpha/2)(\mathrm{sh}((1+n/2)\alpha) - \mathrm{sh}(n\alpha/2))}{\mathrm{sh}\alpha\,\mathrm{sh}((n+1)\alpha)} = \\
&= 1 - (q/z)\frac{\mathrm{sh}(n\alpha/2)}{\mathrm{sh}((n+1)\alpha/2)\,\mathrm{ch}(\alpha/2)} \ge \\
&\ge 1 - (q/z)\frac{e^{-\alpha/2}}{\mathrm{ch}(\alpha/2)} = 1 - (q/z)(1 - \mathrm{th}(\alpha/2)).
\end{aligned} \tag{3.19}$$

Here we applied the fact that the numbers $\mathrm{sh}(n\alpha/2)/\mathrm{sh}((n+1)\alpha/2)$ tend to $e^{-\alpha/2}$ if n approaches infinity and the convergence is monotonically increasing. It follows from the estimations (3.19) that the condition (D) is fulfilled for all values of n if and only if $1-(q/z)(1-\mathrm{th}(\alpha/2)) \ge 0$. From this condition we obtain the inequality $1/\sqrt{1+4z} \ge 1 - z/q$, which is fulfilled if and only if

$$\frac{\sigma}{\theta} < q \le \frac{8\sigma(\theta-1) + 2 - \theta + \sqrt{(8\sigma(1-\theta) - 2 + \theta)^2 - 16(1-\theta)^2\sigma(4\sigma-1)}}{8(1-\theta)^2} \tag{3.20}$$

(in case of $\theta = 1$ the condition is $q \ge \sigma$). We can easily verify the relation

$$\begin{aligned}
&\frac{8\sigma(\theta-1) + 2 - \theta + \sqrt{(8\sigma(1-\theta) - 2 + \theta)^2 - 16(1-\theta)^2\sigma(4\sigma-1)}}{8(1-\theta)^2} \le \\
&\le \frac{1 - 2(1-\theta)\sigma + \sqrt{1 - 4(1-\theta)\sigma}}{2(1-\theta)^2}.
\end{aligned} \tag{3.21}$$

Comparing the conditions in the cases $(A), (B), (C)$ and (D) we can formulate our result in

Theorem 3.2. *The (σ,θ)-method is uniformly totally monotone and at the same time it is uniformly monotonicity conserving if and only if the condition*

$$\frac{\sigma}{\theta} \le q \le \frac{8\sigma(\theta-1)+2-\theta+\sqrt{(8\sigma(1-\theta)-2+\theta)^2-16(1-\theta)^2\sigma(4\sigma-1)}}{8(1-\theta)^2} \tag{3.22}$$

(in case of $\theta = 1$ the condition is $q \ge \sigma$ and if $\theta = 0$, then $q \le 1/2$) holds.

REMARK 3.3. Let us observe that the above condition corresponds with the condition of the uniform maximum norm contractivity ([8]). That is the (σ,θ)-method is uniformly totally monotone and at the same time it is uniformly monotonicity conserving if and only if it is uniformly contractive in maximum norm.

4 Summary

Summarizing our result we can establish, that the requirement of the monotonicity conservation entails stricter conditions for the step-size choice than the stability bounds. If we would like to use a qualitatively adequate numerical method to solve the heat equation, then we have to choose the mesh-parameters according to the bounds (3.22). Let us calculate these bounds for two well-known methods. In the case of $\sigma = 0$ the finite difference method is uniformly monotonicity conserving if and only if

$$0 \le q \le \frac{2-\theta}{4(1-\theta)^2}. \tag{4.1}$$

The choice $\theta = 0$ corresponds to the explicit Euler method (in this case the condition is $q \le 0.5$), $\theta = 0.5$ corresponds to the Crank-Nicolson method (in this case the condition is $q \le 1.5$) and the choice $\theta = 1$ corresponds to the implicit Euler method (in this case q is optional).
The other special choice is $\sigma = 1/6$. Then the finite element method with linear elements is uniformly monotonicity conserving if and only if $\theta \ge 1/3$ and the condition

$$\frac{1}{6\theta} \le q \le \frac{\theta+2+\sqrt{9\theta^2-12\theta+12}}{24(1-\theta)^2} \tag{4.2}$$

holds (if $\theta = 1$, then $q \ge 1/6$).

In the numerical example in Introduction the parameter q was too large. For the Crank-Nicolson method the necessary and sufficient condition of the uniform monotonicity conservation is $0 \le q \le 1.5$ (see condition (4.1)). This yields a sufficient condition for the case $n = 5$. With the choice $q = 1.5$ we obtain the iterates

$$\begin{aligned}
\mathbf{y}^{(1)} &= (0.2308, 0.2692, 0.5000, 0.7308, 0.7692)^\top, \\
\mathbf{y}^{(2)} &= (0.1420, 0.3580, 0.5000, 0.6420, 0.8580)^\top, \\
\mathbf{y}^{(3)} &= (0.1761, 0.3239, 0.5000, 0.6761, 0.8239)^\top,
\end{aligned}$$

$$\begin{array}{c} \vdots \\ \mathbf{y}^{(20)} = (0.1666, 0.3333, 0.5000, 0.6666, 0.8333)^{\top}, \\ \vdots \end{array}$$

which are monotonically increasing vectors, indeed.

Acknowledgement. The author is very thankfull to István Faragó for his fruitful suggestions.

References

[1] I. FARAGÓ, *One-step Methods of Solving a Parabolic Problem and their Qualitative Properties*, Publications on Applied Analysis, Eötvös L. University, Department of Applied Analysis, Budapest, 1996/3.

[2] J.M. HILL, J.N. DEWYNNE, *Heat Conduction*, Applied Mathematics and Engineering Science Texts, Blackwell Scientific Publications 1987.

[3] I. FARAGÓ, *Nonnegativity of the Difference Schemes*, Pure Math. Appl. 6 (1996), 38-50.

[4] I. FARAGÓ, T. PFEIL, *Preserving Concavity in Initial-Boundary Value Problems of Parabolic Type and its Numerical Solution*, Periodica Math. Hung. 30 (1995), 135-139.

[5] K. GLASHOFF, H. KRETH, *Vorzeichenstabile Differenzenverfahren für parabolische Anfangsrandwertaufgaben*, Numer. Math. 35 (1980), 343-354. (in German)

[6] R. HORVÁTH, *On the Sign-Stability of the Numerical Solution of the Heat Equation*, Pure Math. Appl. (to appear)

[7] J.F.B.M. KRAAIJEVANGER, *Maximum Norm Contractivity of Discretization Schemes for the Heat Equation*, Applied Numerical Mathematics 9 (1992), 475-492.

[8] R. HORVÁTH, *Maximum Norm Contractivity in the Numerical Solution of the One-Dimensional Heat Equation*, Appl. Num. Math. 31 (1999), 451-462.

[9] I. FARAGÓ, R. HORVÁTH, *Qualitative Linear Algebra and its Application to the Numerical Solution of the Heat Equation*, Publications on Applied Analysis, Eötvös Loránd University, Department of Applied Analysis, 1999/1.