On Positivity Preservation in Some Finite Element Methods for the Heat Equation

V. Thomée(✉)

Mathematical Sciences, Chalmers University of Technology
and the University of Gothenburg, 412 96 Gothenburg, Sweden
thomee@chalmers.se

Abstract. We consider the initial boundary value problem for the homogeneous heat equation, with homogeneous Dirichlet boundary conditions. By the maximum principle the solution is nonnegative for positive time if the initial data are nonnegative. We study to what extent this property carries over to some piecewise linear finite element discretizations, namely the Standard Galerkin method, the Lumped Mass method, and the Finite Volume Element method. We address both spatially semidiscrete and fully discrete methods.

Keywords: Heat equation · Finite element method · Positivity preservation

1 Introduction

We consider the following model problem for the homogeneous heat equation, to find $u = u(x,t)$ for $x \in \Omega,\ t \geq 0$, satisfying

$$u_t = \Delta u \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega, \quad \text{for } t \geq 0, \quad \text{with } u(\cdot,0) = v \quad \text{in } \Omega, \tag{1}$$

where Ω is a polygonal domain in $\mathbb{R}^2$. The initial values v are thus the only data of the problem, and the solution of (1) may be written $u(t) = E(t)v$ for $t \geq 0$, where $E(t) = e^{\Delta t}$ is the solution operator. By the maximum principle, $E(t)$ is a nonnegative operator, so that

$$v \geq 0 \quad \text{in } \Omega \quad \text{implies} \quad E(t)v \geq 0 \quad \text{in } \Omega, \text{ for } t \geq 0. \tag{2}$$

Our purpose here is to discuss analogues of this property for some finite element methods, based on piecewise linear finite elements, including, in particular, the Standard Galerkin (SG), the Lumped Mass (LM), and the Finite Volume Element (FVE) method. For general information about these methods, and especially error estimates, see Thomée [7], Chou and Li [3], and Chatzipantelidis, Lazarov and Thomée [1,2]. We consider both spatially semidiscrete and fully discrete approximations.

I. Dimov et al. (Eds.): NMA 2014, LNCS 8962, pp. 13–24, 2015.
DOI: 10.1007/978-3-319-15585-2_2

The basis for the methods studied is the variational formulation of the model problem, to find $u = u(\cdot, t) \in H_0^1 = H_0^1(\Omega)$ for $t \geq 0$, such that

$$(u_t, \varphi) + A(u, \varphi) = 0, \quad \forall \varphi \in H_0^1, \quad \text{for } t \geq 0, \quad \text{with } u(0) = v, \tag{3}$$

where $(v, w) = (v, w)_{L_2(\Omega)}$ and $A(v, w) = (\nabla v, \nabla w)$. The finite element methods are based on regular triangulations $\mathcal{T}_h = \{K\}$ of Ω, with $h = \max_{\mathcal{T}_h} \operatorname{diam}(K)$, using the finite element spaces

$$S_h = \{\chi \in \mathcal{C}(\overline{\Omega}) : \chi \text{ linear on each } K \in \mathcal{T}_h;\ \chi = 0 \quad \text{on } \partial\Omega\}.$$

The spatially semidiscrete SG method consists in using (3) restricted to S_h, and the corresponding LM and FVE methods on variational formulations in which the first term (u_t, φ) has been modified, or to find $u_h(t) \in S_h$ for $t \geq 0$, such that

$$[u_{h,t}, \chi] + A(u_h, \chi) = 0, \quad \forall \chi \in S_h, \quad \text{for } t \geq 0, \quad \text{with } u(0) = v_h, \tag{4}$$

where $[\cdot, \cdot]$ is an inner product in S_h, approximating $(\cdot, \cdot)$. The specific choices of $[\cdot, \cdot]$ in the LM and FVE cases will be given in Sect. 2 below.

We now formulate (4) in matrix form. Let $Z_h = \{P_j\}_{j=1}^N$ be the interior nodes of $\mathcal{T}_h$, and $\{\Phi_j\}_{j=1}^N \subset S_h$ the corresponding nodal basis, with $\Phi_j(P_i) = \delta_{ij}$. Writing

$$u_h(t) = \sum_{j=1}^N \alpha_j(t)\Phi_j, \quad \text{with } v_h = \sum_{j=1}^N \widetilde{v}_j \Phi_j,$$

the semidiscrete problem (4) may then be formulated, with $\alpha = (\alpha_1, \ldots, \alpha_N)^T$,

$$\mathcal{M}\alpha' + \mathcal{S}\alpha = 0, \text{ for } t \geq 0, \quad \text{with } \alpha(0) = \widetilde{v}, \tag{5}$$

where $\mathcal{M} = (m_{ij})$, $m_{ij} = [\Phi_i, \Phi_j]$, $\mathcal{S} = (s_{ij})$, $s_{ij} = A(\Phi_i, \Phi_j)$, and $\widetilde{v} = (\widetilde{v}_1, \ldots, \widetilde{v}_N)^T$. The mass matrix $\mathcal{M}$ and the stiffness matrix $\mathcal{S}$ are both symmetric, positive definite. The solution of (5) can be written, with $\mathcal{E}(t)$ the solution matrix,

$$\alpha(t) = \mathcal{E}(t)\,\widetilde{v}, \quad \text{where } \mathcal{E}(t) = e^{-\mathcal{H}t}, \quad \mathcal{H} = \mathcal{M}^{-1}\mathcal{S}. \tag{6}$$

We note that the semidiscrete solution $u_h(t) \in S_h$ is ≥ 0 (> 0) if and only if, elementwise, $\alpha(t) \geq 0$ (> 0).

It was shown in Thomée and Wahlbin [8] that, for the semidiscrete SG method, the discrete analogue of (2) does not hold for small $t > 0$. However, in the case of the LM method, it is valid if and only if the triangulation is of Delaunay type; it had been shown already in Fujii [5] that nonnegativity holds for triangulations with all angles $\leq \frac{1}{2}\pi$. For the FVE method we will show here that the situation is the same as for the SG method, i.e., that $\mathcal{E}(t) \geq 0$ does not hold for small $t > 0$.

In cases where the solution operator is not nonnegative for all positive times, we shall also discuss if it becomes nonnegative for larger time, or if $\mathcal{E}(t) \geq 0$ for $t \geq t_0 > 0$; the smallest such t_0, if it exists, will be referred to as the *threshold of positivity*. Clearly, this is particularly interesting if t_0 is relatively small.

We also study fully discrete schemes based on time stepping in the spatially semidiscrete methods. With k a time step, we consider approximations of the solution matrix $\mathcal{E}(t) = e^{-t\mathcal{H}}$ in (6) at $t_n = nk$ of the form $\mathcal{E}_k^n$, where $\mathcal{E}_k = r(k\mathcal{H})$, with $r(\xi)$ a rational function satisfying certain conditions. We will be particularly concerned here with the Backward Euler and $(0, 2)$ Padé time stepping methods, corresponding to $r(\xi) = 1/(1 + \xi)$ and $r(\xi) = 1/(1 + \xi + \frac{1}{2}\xi^2)$, respectively.

In Schatz, Thomée and Wahlbin [6] some positivity results were obtained for fully discrete schemes related to those for the spatially semidiscrete SG and LM methods, and some of these are extended here to include also the FVE method.

After the introductory Sects. 1 and 2, the positivity properties of the spatially semidiscrete methods are analyzed in Sect. 3, and then, in Sect. 4, of the fully discrete methods. In Sect. 5 we give a concrete example, with Ω the unit square, using the most basic uniform triangulation $\mathcal{T}_h$, with the stiffness matrix corresponding to the 5-point finite difference Laplacian. Computations in MATLAB are used to elucidate our theoretical results, and to determine actual positivity thresholds.

The author gratefully acknowledges the help of Panagiotis Chatzipantelidis with the computer experiments and the figures.

2 The Spatially Semidiscrete Methods

We begin our discussion of the semidiscrete problem (4), or (5), by observing that for the stiffness matrix $\mathcal{S} = (s_{ij})$, which is common to all cases of (4), simple calculations show, see, e.g., [4],

$$s_{ij} = (\nabla\Phi_i, \nabla\Phi_j) = \begin{cases} \sum_{K\subset \mathrm{supp}(\Phi_i)} h_i^{-2}|K|, & \text{if } i = j, \\ -\frac{1}{2}\sin(\alpha+\beta)/(\sin\alpha\sin\beta), & \text{if } P_i, P_j \text{ neighbors}, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Here h_i is the height of K with respect to the edge opposite the vertex P_i, and α and β are the angles opposite the edge P_iP_j, see Fig. 1. We assume throughout that the triangulations $\mathcal{T}_h$ are such that the corresponding $\mathcal{S}$ are irreducible matrices.

We now turn to the three different semidiscrete versions of (4) mentioned above, and specify the corresponding discrete inner products $[\cdot, \cdot]$ on S_h.

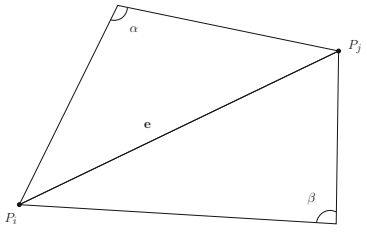

Fig. 1. An interior edge $\mathbf{e} = P_iP_j$.

The *Standard Galerkin* (SG) method is defined by (4) with $[\cdot,\cdot] = (\cdot,\cdot) = (\cdot,\cdot)_{L_2(\Omega)}$, and we find for the mass matrix, with $|V| = \text{area}(V)$,

$$m_{ij} = m_{ij}^{SG} = (\Phi_i, \Phi_j) = \begin{cases} \frac{1}{6}|\text{supp}(\Phi_i)|, & \text{if } i = j, \\ \frac{1}{12}|\text{supp}(\Phi_i\Phi_j)|, & \text{if } P_i, P_j \text{ neighbors}, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The *Lumped Mass* (LM) method uses (4) with $[\cdot,\cdot] = (\cdot,\cdot)_h$, where the latter is defined by quadrature: with $\{P_{K,j}\}_{j=1}^3$ the vertices of the triangle K, we set

$$(\psi,\chi)_h = \sum_{K\in\mathcal{T}_h} Q_{K,h}(\psi\,\chi), \quad \text{with } Q_{K,h}(f) = \tfrac{1}{3}\,|K| \sum_{j=1}^{3} f(P_{K,j}) \approx \int_K f\,dx.$$

In the matrix formulation (5), this means that $\mathcal{M} = \mathcal{D} = (d_{ij})$, with $d_{ij} = (\Phi_i, \Phi_j)_h = 0$ for $j \neq i$, so that $\mathcal{D}$ is a diagonal matrix.

To define the spatially semidiscrete *Finite Volume Element* (FVE) method, following [2], we note that a solution of the differential equation $u_t = \Delta u$ in (1) satisfies the local conservation law

$$\int_V u_t\,dx - \int_{\partial V} \frac{\partial u}{\partial n}\,ds = 0, \quad \text{for } t \geq 0, \tag{9}$$

for any $V \subset \Omega$, with n the unit exterior normal to ∂V. The semidiscrete FVE method is then to find $u_h(t) \in S_h$, for $t \geq 0$, satisfying

$$\int_{V_j} u_{h,t}\,dx - \int_{\partial V_j} \frac{\partial u_h}{\partial n}\,ds = 0, \ \text{for } j = 1,\dots,N, \ t \geq 0, \quad \text{with } u_h(0) = v_h, \tag{10}$$

where the V_j are the so called control volumes, defined as follows, see Fig. 2. For $K \in \mathcal{T}_h$, let b_K be its barycenter, and connect b_K with the midpoints of the edges of K, thus partitioning K into three quadrilaterals K_l, $l = j, m, n$, where P_j, P_m, P_n are the vertices of K. The control volume V_j is then the union of the quadrilaterals K_j, sharing the vertex P_j. The equations (10) thus preserves (9) for any union of control volumes.

To write (10) in weak form, we introduce the finite dimensional space

$$Y_h = \{\eta \in L_2 : \ \eta|_{V_j} = \text{constant}, \ j = 1,\dots,N; \ \eta = 0 \text{ outside } \cup_{j=1}^N V_j\}.$$

For $\eta \in Y_h$, we multiply (10) by $\eta(P_j)$, and sum over j, to obtain the Petrov–Galerkin formulation

$$(u_{h,t},\eta) + a_h(u_h,\eta) = 0, \quad \forall \eta \in Y_h, \ \ t \geq 0, \quad \text{with } u_h(0) = v_h, \tag{11}$$

where

$$a_h(\chi,\eta) = -\sum_{j=1}^{N} \eta(P_j) \int_{\partial V_j} \frac{\partial \chi}{\partial n}\,ds, \quad \forall \chi \in S_h, \ \eta \in Y_h. \tag{12}$$

In order to rephrase this as a pure Galerkin method, we shall introduce a new inner product on S_h. Let $J_h : \mathcal{C}(\Omega) \to Y_h$ be the interpolant defined by $(J_h v)(P_j) = v(P_j)$, $j = 1,\dots,N$. The following lemma then holds, see [3].

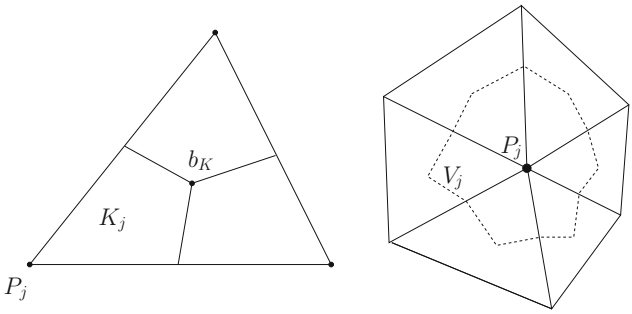

Fig. 2. A triangle $K \in \mathcal{T}_h$ and a patch Π_j around a vertex P_j

Lemma 1. *The bilinear form* $(\chi, J_h\psi)$ *is symmetric, positive definite on* S_h*, and*

$$a_h(\chi, J_h\psi) = (\nabla\chi, \nabla\psi) = A(\chi, \psi), \quad \forall \chi, \psi \in S_h. \tag{13}$$

We now define the inner product $\langle \chi, \psi \rangle = (\chi, J_h\psi)$, for $\chi, \psi \in S_h$. By (13), the Petrov-Galerkin equation (11), (12) may then be written in the Galerkin formulation (4), with $[\cdot, \cdot] = \langle \cdot, \cdot \rangle$, and the mass matrix $\mathcal{M} = (m_{ij})$ in (5) is

$$m_{ij} = m_{ij}^{FVE} = \langle \Phi_i, \Phi_j \rangle = \begin{cases} \frac{11}{54}|\text{supp}(\Phi_i)|, & \text{if } i = j, \\ \frac{7}{108}|\text{supp}(\Phi_i\Phi_j)|, & \text{if } P_i, P_j \text{ neighbors}, \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

We note that the FVE mass matrix is more concentrated on the diagonal than that of SG. In fact, with $\mathcal{D}$ the diagonal mass matrix of LM, we have

$$\mathcal{M}^{FVE} = \tfrac{2}{9}\mathcal{D} + \tfrac{7}{9}\mathcal{M}^{SG}. \tag{15}$$

3 Positivity Preservation in the Spatially Semidiscrete Methods

In this section we shall consider the general spatially semidiscrete problem (4), in matrix form (5), where $\mathcal{S}$ is the stiffness matrix, and $\mathcal{M} = (m_{ij})$, $m_{ij} = [\Phi_i, \Phi_j]$, is the mass matrix. We assume that $[\cdot, \cdot]$ is such that either $m_{ij} > 0$ for all neighbors P_i, P_j, or such that $m_{ij} = 0$ for all neighbors P_i, P_j. In the former case $\mathcal{M}$ is a nondiagonal matrix, and in the latter diagonal. We shall make the technical assumption that $\mathcal{T}_h$ has a strictly interior node, P_j say, such that any neighbor of P_j has a neighbor which is not a neighbor of P_j; we shall refer to such a triangulation as *normal*. Note that $\mathcal{T}_h$ is normal if it has a strictly interior node P_j, with all its neighbors strictly interior, such that the associated patch Π_j is convex. In the case of a nondiagonal mass matrix we have the following negative result, which was shown in [8] for the SG method.

Theorem 1. *Assume that $\mathcal{T}_h$ is normal and that $\mathcal{M}$ is nondiagonal. Then the solution matrix $\mathcal{E}(t) = e^{-\mathcal{H}t}$ for (5) cannot be nonnegative for small $t > 0$.*

Proof. Assume that $\mathcal{E}(t) \ge 0$ for small $t > 0$. Then $h_{ij} \le 0$ for $i \ne j$ since

$$\mathcal{E}(t) = e^{-\mathcal{H}t} = \mathcal{I} - \mathcal{H}\,t + O(t^2) \ge 0, \quad \text{as } t \to 0.$$

Let P_j be the strictly interior node in the definition of a normal $\mathcal{T}_h$. We shall show that $h_{ij} = 0$ for $i \ne j$. If this has been proven, then

$$s_{ij} = \sum_{l=1}^{N} m_{il}h_{lj} = h_{jj}m_{ij}, \quad i = 1, \dots, N, \tag{16}$$

with $h_{jj} \ne 0$, and hence the j^{th} columns of $\mathcal{S}$ and $\mathcal{M}$ are proportional. Since P_j is strictly interior, we have $\sum_{i=1}^{N} \Phi_i = 1$ on $\operatorname{supp}(\Phi_j)$ and hence $\sum_{i=1}^{N} s_{ij} = \sum_{i=1}^{N}(\nabla\Phi_i, \nabla\Phi_j) = (\nabla 1, \nabla\Phi_j) = 0$. Together with $\sum_{i=1}^{N} m_{ij} > 0$, this contradicts (16) and thus shows our claim.

It remains to show that $h_{ij} = 0$ for $i \ne j$. Consider first the case that P_i is not a neighbor of P_j, so that $m_{ij} = s_{ij} = 0$. Since $\mathcal{S} = \mathcal{M}\mathcal{H}$, we find $s_{ij} = \sum_{l\ne j} m_{il}h_{lj} = 0$, and since $h_{lj} \le 0$ for $l \ne j$, we have $m_{il}h_{lj} \le 0$ and hence $m_{il}h_{lj} = 0$ for $l \ne j$. In particular, $h_{ij} = 0$. When P_i is a neighbor of P_j, it has a neighbor P_q which is not a neighbor of P_j and hence $s_{qj} = \sum_{l\ne j} m_{ql}h_{lj} = 0$, now implying $h_{ij} = 0$ since $m_{qi} > 0$ (where we have used that $\mathcal{M}$ is nondiagonal). This completes the proof.

This result thus covers the SG and FVE methods, but not the LM method, which has a diagonal mass matrix. We recall that an edge $\mathbf{e}$ of $\mathcal{T}_h$ is a Delaunay edge if the sum of the angles α and β opposite $\mathbf{e}$ is $\le \pi$ (see Fig. 1), and that $\mathcal{T}_h$ is a Delaunay triangulation if all *interior* edges are Delaunay. Using (7) this shows that $\mathcal{T}_h$ is Delaunay if and only if $s_{ij} \le 0$ for all $i \ne j$. But this is equivalent to $\mathcal{S}$ being a Stieltjes matrix, i.e., a symmetric, positive definite matrix with nonpositive off-diagonal entries. The following result was shown in [8].

Theorem 2. *The LM solution matrix $\mathcal{E}(t) = e^{\mathcal{H}t}$, $\mathcal{H} = \mathcal{D}^{-1}\mathcal{S}$, is nonnegative for all $t \ge 0$ if and only if $\mathcal{T}_h$ is Delaunay.*

Proof. As in the proof of Theorem 1 we find that $\mathcal{E}(t) \ge 0$ for $t \ge 0$ implies $h_{ij} \le 0$ for $i \ne j$, and hence, since $\mathcal{S} = \mathcal{D}\mathcal{H}$, that $s_{ij} \le 0$ for $i \ne j$, so that $\mathcal{T}_h$ Delaunay.

On the other hand, if $\mathcal{T}_h$ is Delaunay, then $\mathcal{S}$, and hence also $\mathcal{D}+k\mathcal{S}$, is Stieltjes, which implies $(\mathcal{I} + k\mathcal{H})^{-1} = (\mathcal{D} + k\mathcal{S})^{-1}\mathcal{D} \ge 0$ for all $k \ge 0$, where we have used the fact that if $\mathcal{A}$ is a Stieltjes matrix, then $\mathcal{A}^{-1} \ge 0$. Hence

$$\mathcal{E}(t) = \lim_{n\to\infty} (\mathcal{I} + \frac{t}{n}\mathcal{H})^{-n} \ge 0, \text{ for all } t > 0.$$

We recall that if $\mathcal{A}$ is a Stieltjes matrix which is also irreducible, then $\mathcal{A}^{-1} > 0$. In particular, if $\mathcal{T}_h$ is Delaunay, we have $\mathcal{S}^{-1} > 0$. Returning to the general case,

we then also have $\mathcal{H}^{-1} = \mathcal{S}^{-1}\mathcal{M} > 0$. Since $\mathcal{G} = \mathcal{H}^{-1} = \mathcal{S}^{-1}\mathcal{M}$ is symmetric and positive definite with respect to the inner product $\mathcal{M}v \cdot w = \sum_{i=1}^{N} (\mathcal{M}v)_i w_i$, it has eigenvalues $\{\kappa_j\}_{j=1}^{N}$, with $0 < \kappa_{j+1} \le \kappa_j$, and orthonormal eigenvectors $\{\varphi_j\}_{j=1}^{N}$, with respect to this inner product. Recall that by the Perron-Frobenius theorem, if $\mathcal{G} > 0$, then $\varphi_1 > 0$ and $\kappa_j < \kappa_1$ for $j \ge 2$. Note that $\{\varphi_j\}_{j=1}^{N}$ are then also the eigenvectors of $\mathcal{H}$, with corresponding eigenvalues $\lambda_j = 1/\kappa_j,\ j = 1, \dots, N$, and thus $\lambda_j > \lambda_1$ for $j \ge 2$. We may write

$$\mathcal{E}(t)\widetilde{v} = \sum_{l=1}^{N} e^{-\lambda_l t} (\mathcal{M}\widetilde{v} \cdot \varphi_l)\ \varphi_l. \tag{17}$$

We now return to the general semidiscrete problem (4), or (5), and show that, if $\mathcal{G} = \mathcal{H}^{-1} > 0$, then there exists $t_0 \ge 0$ such that $\mathcal{E}(t) > 0$ for $t > t_0$. This result was incorrectly stated in [6], without the positivity assumption.

Theorem 3. *If $\mathcal{G} = \mathcal{H}^{-1} > 0$, then there is a $t_0 \ge 0$ such that the solution matrix $\mathcal{E}(t) = e^{-\mathcal{H}t}$ for (5) is positive for $t > t_0$.*

Proof. It suffices to show that $\mathcal{E}(t)e_j > 0$ for large t, for the finitely many unit vectors $\{e_j\}_{j=1}^{N}$. But, since $\varphi_1 > 0$ and $\mathcal{M}e_j \cdot \varphi_1 > 0$, we find by (17), for t large,

$$\mathcal{E}(t)e_j = \sum_{l=1}^{N} e^{-\lambda_l t} (\mathcal{M}e_j \cdot \varphi_l)\ \varphi_l = e^{-\lambda_1 t}\big((\mathcal{M}e_j \cdot \varphi_1)\ \varphi_1 + O(e^{-(\lambda_2 - \lambda_1)t})\big) > 0.$$

4 Fully Discrete Methods

In this section we study time discretizations of the semidiscrete problem (4), or (5). We thus consider approximations of the solution matrix $\mathcal{E}(t) = e^{-t\mathcal{H}}$ in (6) at $t_n = nk$, with k a time step, of the form $\mathcal{E}_k^n$, where $\mathcal{E}_k = r(k\mathcal{H})$, with $r(\xi)$ a rational function satisfying certain conditions.

We begin with the *Backward Euler* (BE) method, to find $U^n \in S_h$, $U^n \approx u_h(t_n)$, for $n \ge 0$, such that

$$\Big[\frac{U^n - U^{n-1}}{k}, \chi\Big] + A(U^n, \chi) = 0, \quad \forall \chi \in S_h, \text{ for } n \ge 1, \quad \text{with } U^0 = v_h. \tag{18}$$

In matrix formulation, with $U^n = \sum_{j=1}^{N} \alpha_j^n \Phi_j$, this takes the form

$$(\mathcal{M} + k\mathcal{S})\alpha^n = \mathcal{M}\alpha^{n-1} \quad \text{or} \quad \alpha^n = \mathcal{E}_k \alpha^{n-1}, \ \text{ for } n \ge 1, \quad \text{with } \alpha^0 = \widetilde{v},$$

where $\mathcal{E}_k$ the time stepping matrix

$$\mathcal{E}_k = (\mathcal{M} + k\mathcal{S})^{-1}\mathcal{M} = (\mathcal{I} + k\mathcal{H})^{-1} = r_{01}(k\mathcal{H}), \quad \mathcal{H} = \mathcal{M}^{-1}\mathcal{S}, \tag{19}$$

using $r(\xi) = r_{01}(\xi) = 1/(1+\xi)$. The fully discrete solution is thus $\alpha^n = \mathcal{E}_k^n \widetilde{v}$.

We first have the following time discrete analogue of Theorem 1, see [6].

Theorem 4. *Assume that $\mathcal{T}_h$ is normal and $\mathcal{M}$ nondiagonal. Then the BE time stepping matrix $\mathcal{E}_k = (\mathcal{I} + k\mathcal{H})^{-1}$ cannot be nonnegative for small $k > 0$.*

Proof. If we assume $\mathcal{E}_k \ge 0$ for $k > 0$ small, we would have, for any $t > 0$,

$$\mathcal{E}(t) = e^{-\mathcal{H}t} = \lim_{n\to\infty} (\mathcal{I} + \frac{t}{n}\mathcal{H})^{-n} = \lim_{n\to\infty} \mathcal{E}_{t/n}^n \ge 0, \tag{20}$$

in contradiction to Theorem 1.

For the Backward Euler Lumped Mass method the mass matrix is diagonal, and the following analogue of Theorem 2 was shown in [6].

Theorem 5. *For the BE LM method, $\mathcal{E}_k \ge 0$ for all $k > 0$ if and only if $\mathcal{T}_h$ Delaunay.*

For the nonnegativity of $\mathcal{E}_k$ for larger k, the following holds, where, as in the semi-discrete case, positivity properties of $\mathcal{H}^{-1}$ enter.

Theorem 6. *For $\mathcal{E}_k = (\mathcal{I} + k\mathcal{H})^{-1}$ to be nonnegative for k large, it is necessary that $\mathcal{H}^{-1} \ge 0$. If $\mathcal{H}^{-1} > 0$, then there exists $k_0 \ge 0$ such that $\mathcal{E}_k > 0$ for $k > k_0$.*

If $\mathcal{E}_{k_0} \ge 0$, then $\mathcal{E}_k \ge 0$ for $k \ge k_0$. Thus $\{k : \mathcal{E}_k \ge 0\}$ is an interval $[k_0, \infty)$.

Proof. We write $\mathcal{E}_k = \varepsilon(\varepsilon\mathcal{I} + \mathcal{H})^{-1}$, with $\varepsilon = 1/k$, and note that thus $\mathcal{E}_k \ge 0$ for k large implies $(\varepsilon\mathcal{I} + \mathcal{H})^{-1} \ge 0$ for $\varepsilon > 0$ small. But $(\varepsilon\mathcal{I} + \mathcal{H})^{-1} \to \mathcal{H}^{-1}$ as $\varepsilon \to 0$, and hence $\mathcal{H}^{-1} \ge 0$. On the other hand, if $\mathcal{H}^{-1} > 0$, then $(\varepsilon\mathcal{I} + \mathcal{H})^{-1} > 0$ for ε small, and hence $\mathcal{E}_k > 0$ for k large.

For the last statement in the theorem we show that if $(\varepsilon_0\mathcal{I} + \mathcal{H})^{-1} \ge 0$, with $\varepsilon_0 > 0$, then $(\varepsilon\mathcal{I} + \mathcal{H})^{-1} \ge 0$ for $\varepsilon \in [0, \varepsilon_0]$. With $\delta = \varepsilon_0 - \varepsilon > 0$, we may write

$$(\varepsilon\mathcal{I} + \mathcal{H})^{-1} = (\varepsilon_0\mathcal{I} + \mathcal{H} - \delta\mathcal{I})^{-1} = (\varepsilon_0\mathcal{I} + \mathcal{H})^{-1}(\mathcal{I} - \mathcal{K})^{-1}, \quad \text{where } \mathcal{K} = \delta(\varepsilon_0\mathcal{I} + \mathcal{H})^{-1}.$$

Here $\mathcal{K} \ge 0$, by assumption, and, if δ is so small that, for some matrix norm $|\cdot|$, $|\mathcal{K}| = \delta|(\varepsilon_0\mathcal{I} + \mathcal{H})^{-1}| < 1$, then $(\mathcal{I} - \mathcal{K})^{-1} = \sum_{j=0}^{\infty} \mathcal{K}^j \ge 0$, and therefore $(\varepsilon\mathcal{I} + \mathcal{H})^{-1} \ge 0$. But if $(\varepsilon\mathcal{I} + \mathcal{H})^{-1} \ge 0$ for $\varepsilon \in (\varepsilon_1, \varepsilon_0]$, with $\varepsilon_1 \ge 0$, then $(\varepsilon_1\mathcal{I} + \mathcal{H})^{-1} \ge 0$. Hence, by the above, $(\varepsilon\mathcal{I} + \mathcal{H})^{-1} \ge 0$ for some $\varepsilon < \varepsilon_1$, and thus the smallest such ε_1 has to be $\varepsilon_1 = 0$.

When $\mathcal{E}_k \ge 0$ for large k, we refer to the smallest k_0 such that $\mathcal{E}_k \ge 0$ for $k \ge k_0$ as the *threshold of positivity* for $\mathcal{E}_k$. Thus, by the last statement of Theorem 6, in the BE case the positivity threshold is the smallest k for which $\mathcal{E}_k \ge 0$.

The following result from [6] gives precise values of k for $\mathcal{E}_k$ to be guaranteed to be nonnegative, under a sharper conditions than $\mathcal{H}^{-1} > 0$, namely if $s_{ij} < 0$ for P_i, P_j neighbors, or $\alpha + \beta < \pi$ for each edge $\mathbf{e} = P_iP_j$ of $\mathcal{T}_h$ (see Fig. 1).

Theorem 7. *If $s_{ij} < 0$ for all neighbors P_iP_j, then $\mathcal{E}_k \ge 0$ if*

$$k|s_{ij}| \ge m_{ij}, \quad \forall j \ne i. \tag{21}$$

Proof. (21) implies that $m_{ij} + ks_{ij} \le 0$ for all $j \ne i$, so that $\mathcal{M} + k\mathcal{S}$ is a Stieltjes matrix. Hence $(\mathcal{M} + k\mathcal{S})^{-1} \ge 0$, and thus $\mathcal{E}_k \ge 0$ by (19).

Thus $\mathcal{E}_k \geq 0$ if $k \geq \max(m_{ij}/|s_{ij}|)$, with max taken over all neighbors P_i, P_j. If $\{\mathcal{T}_h\}$ is a quasiuniform family, and $\alpha + \beta \leq \gamma < \pi$ for all P_iP_j, then $\mathcal{E}_k \geq 0$ if $k \geq ch^2$ with $c = c(\{\mathcal{T}_h\})$. Note that since $m_{ij}^{SG} = \frac{7}{9} m_{ij}^{FVE}$ for P_i, P_j neighbors, by (15), the above lower bound is smaller for FVE than for SG, by a factor 7/9.

Now consider, more generally, a fully discrete solution $\alpha^n = \mathcal{E}_k^n \widetilde{v}_h$, $n \geq 0$, of (5) defined by a time stepping matrix $\mathcal{E}_k = r(k\mathcal{H})$, where $r(\xi)$ is a bounded rational function for $\xi \geq 0$ approximating $e^{-\xi}$ for small ξ, so that

$$r(\xi) = 1 - \xi + O(\xi^2), \quad \text{as } \xi \to 0. \tag{22}$$

We may write

$$\mathcal{E}_k \widetilde{v} = r(k\mathcal{H})\widetilde{v} = \sum_{l=1}^{N} r(k\lambda_l)(\mathcal{M}\widetilde{v} \cdot \varphi_l)\, \varphi_l.$$

As in Theorem 4, $\mathcal{E}_k$ cannot be nonnegative for small k and $\mathcal{M}$ nondiagonal [6].

Theorem 8. *Assume that $\mathcal{T}_h$ is normal and $\mathcal{M}$ nondiagonal. Let $\mathcal{E}_k = r(k\mathcal{H})$, with $r(\xi)$ satisfying (22). Then $\mathcal{E}_k$ cannot be nonnegative for small k.*

Proof. Using (22), the result follow as in Theorem 4 from

$$\lim_{n\to\infty} \mathcal{E}_{t/n}^n = \lim_{n\to\infty} \left(\mathcal{I} - \frac{t}{n}\mathcal{H} + O(\frac{t^2}{n^2})\right)^n = e^{-t\mathcal{H}} = \mathcal{E}(t), \quad \text{for any } t > 0.$$

For nonnegativity of $\mathcal{E}_k = r(k\mathcal{H})$ for larger k we first show that if $\mathcal{H}^{-1} > 0$, this requires that $r(\xi) \geq 0$ for large ξ.

Theorem 9. *Let $\mathcal{H}^{-1} > 0$ and let $\mathcal{E}_k = r(k\mathcal{H})$. Then a necessary condition for $\mathcal{E}_k$ to be nonnegative for large k is that $r(\xi) \geq 0$ for large ξ.*

Proof. With λ_1, φ_1 the first eigenvalue and the corresponding eigenvector of $\mathcal{H}$, we have $\mathcal{E}_k\varphi_1 = r(k\lambda_1)\varphi_1$, and thus, since $\lambda_1 > 0$, $\varphi_1 > 0$, for $\mathcal{E}_k\varphi_1$ to be nonnegative for large k it is necessary that $r(k\lambda_1)$ be nonnegative for large k, showing our claim.

A typical and interesting example is the $(0, 2)$ Padé approximation $r_{02}(\xi) = 1/(1 + \xi + \frac{1}{2}\xi^2)$. However, the Padé approximations $r_{11}(\xi) = (1 - \frac{1}{2}\xi)/(1 + \frac{1}{2}\xi)$ and $r_{12}(\xi) = (1 - \frac{1}{3}\xi)/(1 + \frac{2}{3}\xi + \frac{1}{6}\xi^2)$ are negative for large ξ, and hence the corresponding $\mathcal{E}_k$ cannot be nonnegative for large k when $\mathcal{H}^{-1} > 0$.

We now assume that $r(\infty) = 0$. If $r(\xi) \geq 0$ for large ξ, we may then write

$$r(\xi) = c\xi^{-q} + O(\xi^{-q-1}), \quad \text{as } \xi \to \infty, \quad \text{with } q \geq 1,\ c > 0. \tag{23}$$

We show the following result, generalizing the first part of Theorem 6.

Theorem 10. *Assume that (23) holds. Then $\mathcal{H}^{-q} \geq 0$ is a necessary condition for $\mathcal{E}_k = r(k\mathcal{H}) \geq 0$ for large k. If $\mathcal{H}^{-q} > 0$, then $\mathcal{E}_k = r(k\mathcal{H}) > 0$ for large k.*

Proof. Both statements of the theorem follow since, by (23),

$$\mathcal{E}_k = ck^{-q}(\mathcal{H}^{-q} + O(k^{-1})), \quad \text{as } k \to \infty. \tag{24}$$

The result shows, in particular, that $\mathcal{E}_k = r_{02}(k\mathcal{H}) > 0$ for large k if $\mathcal{H}^{-2} > 0$. We complete this section by showing that for this method, the negative conclusion of Theorem 8 holds also for the LM method, even though $\mathcal{M}$ is then diagonal, under the not very restrictive assumption that $\mathcal{T}_h$ is *4-connected* in the following sense: There exists a path $\mathcal{P}$ in Z_h consisting of four connected edges P_mP_n, with $s_{mn} \neq 0$, and such that the endpoints P_i, P_j of the path cannot be connected by a path with fewer than four edges.

Theorem 11. *Assume that $\mathcal{T}_h$ is Delaunay and 4-connected. Then, for the LM method, $\mathcal{E}_k = r_{02}(k\mathcal{H})$ cannot be nonnegative for small k.*

Proof. We have, by Taylor expansion of $r_{02}(\xi)$,

$$\mathcal{E}_k = r_{02}(k\mathcal{H}) = \mathcal{I} - k\mathcal{H} + \tfrac{1}{2}k^2\mathcal{H}^2 - \tfrac{1}{4}k^4\mathcal{H}^4 + O(k^5), \quad \text{as } k \to 0.$$

We shall show that if $P_iP_pP_qP_rP_j$ is a path $\mathcal{P}$ as above, then $(\mathcal{E}_k)_{ij} < 0$ for small k. For this we write $\mathcal{H} = \mathcal{D}^{-1}\mathcal{S} = \mathcal{V} - \mathcal{W}$, where $\mathcal{V}$ is a positive diagonal matrix and $\mathcal{W}$ has elements $w_{mn} = -s_{mn}/d_{mm} > 0$ when P_m, P_n are neighbors with $s_{mn} \neq 0$, with the remaining elements 0. (Recall that since $\mathcal{S}$ is Stieltjes, $\mathcal{W} \geq 0$.) It follows that $(\mathcal{H}^4)_{ij} = \sum_{l_1,l_2,l_3} h_{il_1}h_{l_1l_2}h_{l_2l_3}h_{l_3j}$ and, by our assumption on the path $\mathcal{P}$ connecting P_i and P_j, none of the nonzero terms have factors from $\mathcal{V}$. Hence $(\mathcal{H}^4)_{ij} \geq w_{ip}w_{pq}w_{qr}w_{rj} > 0$. In the same way, since P_j cannot be reached from P_i in less than four steps, $(\mathcal{H}^l)_{ij} = 0$ for $l = 0, 1, 2, 3$. Hence $(\mathcal{E}_k)_{ij} = -\frac{1}{4}k^4(\mathcal{H}^4)_{ij} + O(k^5) < 0$ for k small.

5 A Numerical Example

In this final section we present a numerical example to illustrate our theoretical results. For a family of uniform triangulations of the unit square $\Omega = (0,1) \times (0,1)$, we study the positivity properties of the spatially semidiscrete, the Backward Euler, and the $(0,2)$ Padé methods, using the SG, FVE and LM spatial discretizations. The triangulations $\mathcal{T}_h$ of Ω are defined as follows: Let M be a positive integer, $h = 1/(M+1)$, and set $x_j = y_j = jh$, for $j = 0, \ldots, M+1$. This partitions Ω into squares $(x_j, x_{j+1}) \times (y_m, y_{m+1})$, and we may define a triangulation $\mathcal{T}_h$, by connecting the nodes (x_j, y_m), (x_{j+1}, y_{m-1}). The number of interior vertices is $N = M^2$, and $\max_{\mathcal{T}_h} \operatorname{diam}(K) = \sqrt{2}h$. We note that $\mathcal{T}_h$ is normal, Delaunay, and 4-connected (if $M \geq 3$).

To determine the stiffness and mass matrices, let $\zeta_0 = (x_j, y_m)$ be an interior vertex of $\mathcal{T}_h$ and let $\{\zeta_j\}_{j=1}^6$ be the surrounding (including possibly boundary) vertices, numbered counterclockwise, with $\zeta_1 = (x_{j+1}, y_m)$, and $\{\Psi_j\}_{j=0}^6$ the corresponding basis functions, see Fig. 3. The contributions corresponding to ζ_0 to $\mathcal{S}$ are then given by (cf. (7))

$$(\nabla\Psi_0, \nabla\Psi_j) = \begin{cases} 4, & j = 0, \\ -1, & j = 1, 2, 4, 5, \\ 0, & j = 3, 6, \end{cases}$$

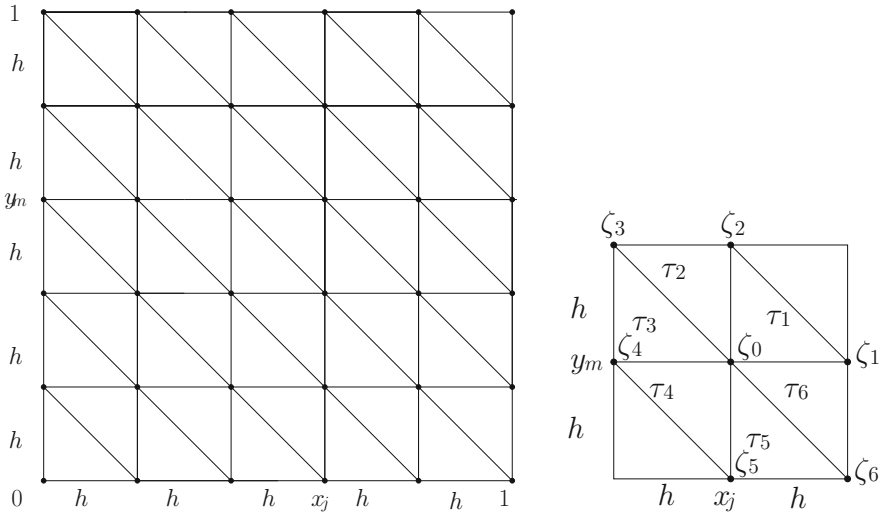

Fig. 3. *Left*: The unit square Ω with the symmetric triangulation $\mathcal{T}_h$. *Right*: The patch at ζ_0.

and to the mass matrices $\mathcal{M}$ for the SG and FVE methods by

$$(\Psi_0, \Psi_j) = \tfrac{1}{2}h^2 \begin{cases} 1, & j = 0, \\ \frac{1}{6}, & j = 1, \dots, 6, \end{cases} \quad \text{and} \quad \langle \Psi_0, \Psi_j \rangle = \tfrac{1}{2}h^2 \begin{cases} \frac{11}{9}, & j = 0, \\ \frac{7}{54}, & j = 1, \dots, 6. \end{cases}$$

Note that since the sum of the angles opposite a diagonal edge is π, the corresponding elements s_{ij} of the stiffness matrix vanish. We observe that $\mathcal{S}$ is an irreducible Stieltjes matrix, so that $\mathcal{S}^{-1} > 0$, and hence the matrices $\mathcal{H}^{-1} = \mathcal{S}^{-1}\mathcal{M}$ for the SG, FVE and LM methods are all positive. Thus the results of Theorems 3, 6, 9, and 10 concerning positivity for large t and k all apply. However, since some $s_{ij} = 0$ for P_i, P_j neighbors, this does not hold for Theorem 7.

Table 1. Positivity thresholds for the numerical example in Sect. 5.

	Semidiscrete		Backward Euler		(0,2) Padé		
h	SG	FVE	SG	FVE	SG	FVE	LM
0.10	0.046	0.043	0.0053	0.0045	0.025	0.024	0.020
0.05	0.035	0.031	0.0013	0.0011	0.023	0.023	0.021
0.025	0.021	0.019	0.0003	0.0003	0.022	0.022	0.022

In Table 1 we show some computed positivity thresholds t_0 for $\mathcal{E}(t)$, and k_0 for $\mathcal{E}_k = r_{01}(k\mathcal{H})$ and $\mathcal{E}_k = r_{02}(k\mathcal{H})$, for the SG, FVE, and in the case of $r_{02}(k\mathcal{H})$ also the LM method. The numbers indicate that for the spatially semidiscrete

problem, the positivity thresholds diminish with h, and are smaller for the FVE than for the SG method. For the BE method the thresholds are small, with the ratio k_0/h^2 approximately 0.54 for SG and 0.45 for FVE, even though Theorem 7 does not apply. For the $(0, 2)$ Padé method the thresholds do not appear to diminish with h, and also to be independent of the choice of the spatial discretization method.

References

1. Chatzipantelidis, P., Lazarov, R.D., Thomée, V.: Some error estimates for the lumped mass finite element method for a parabolic problem. Math. Comp. **81**, 1–20 (2012)
2. Chatzipantelidis, P., Lazarov, R.D., Thomée, V.: Some error estimates for the finite volume element method for a parabolic problem. Comput. Methods Appl. Math. **13**, 251–279 (2013)
3. Chou, S.H., Li, Q.: Error estimates in L^2, H^1 and L^∞ in covolume methods for elliptic and parabolic problems: a unified approach. Math. Comp. **69**, 103–120 (2000)
4. Drăgănescu, A., Dupont, T.F., Scott, L.R.: Failure of the discrete maximum principle for an elliptic finite element problem. Math. Comp. **74**, 1–23 (2004)
5. Fujii, H.: Some remarks on finite element analysis of time-dependent field problems. In: Theory and Practice in Finite Element Structural Analysis, pp. 91–106. University of Tokyo Press, Tokyo (1973)
6. Schatz, A.H., Thomée, V., Wahlbin, L.B.: On positivity and maximum-norm contractivity in time stepping methods for parabolic equations. Comput. Methods Appl. Math. **10**, 421–443 (2010)
7. Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems, 2nd edn. Springer, Heidelberg (2006)
8. Thomée, V., Wahlbin, L.B.: On the existence of maximum principles in parabolic finite element equations. Math. Comp. **77**, 11–19 (2008)

http://www.springer.com/978-3-319-15584-5

Numerical Methods and Applications
8th International Conference, NMA 2014, Borovets, Bulgaria,
August 20–24, 2014, Revised Selected Papers
Dimov, I.; Fidanova, S.; Lirkov, I. (Eds.)
2015, XII, 313 p. 109 illus., Softcover
ISBN: 978-3-319-15584-5