P. Chatzipantelidis, Z. Horváth, and V. Thomée

On preservation of positivity in some finite element methods for the heat equation

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 complement in a number of ways earlier studies of the possible extension of this fact to spatially semidiscrete and fully discrete piecewise linear finite element discretizations, based on the standard Galerkin method, the lumped mass method, and the finite volume element method. We also provide numerical examples that illustrate our findings.

Keywords: heat equation, finite element method, lumped mass, finite volume element method, spatially semidiscrete, fully discrete, positivity preserving

Classification: 65M12, 65M60

P. Chatzipantelidis: Department of Mathematics and Applied Mathematics, University of Crete, Heraklion, GR-70013, Greece, email:p.chatzipa@uoc.gr
Z. Horváth: Department of Mathematics and Computational Sciences, Széchenyi István University, 1 Egyetem square, Györ, H-9026, Hungary, email:horvathz@sze.hu
V. Thomée: Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg, SE-412 96 Göteborg, Sweden, and Institute of Applied and Computational Mathematics, FORTH, Heraklion GR-71110, Greece, email:thomee@chalmers.se

1 Introduction

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

$$u_t = \Delta u \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega, \quad \text{for } t \geqslant 0, \qquad \text{with } u(\cdot,0) = v \quad \text{in } \Omega, \tag{1.1}$$

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

$$v \geqslant 0 \quad \text{in } \Omega \quad \text{implies} \quad E(t)v \geqslant 0 \quad \text{in } \Omega, \text{ for } t \geqslant 0. \tag{1.2}$$

Our purpose here is to further investigate and extend known results from Thomée and Wahlbin [14], Schatz, Thomée and Wahlbin [11], and Thomée [13] concerning analogues of this property for some finite element methods, based on piecewise linear finite elements. We shall study, in particular, the Standard Galerkin (SG) method, the Lumped Mass (LM) method, and the Finite Volume Element (FVE) method. For general information about these methods, and especially error estimates, see, e.g., Thomée [12], Chou and Li [4] and Chatzipantelidis, Lazarov and Thomée [2] and [3]. We consider both spatially semidiscrete and fully discrete methods.

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 $t \geqslant 0$, such that

$$(u_t,\varphi) + A(u,\varphi) = 0, \quad \forall \varphi \in H_0^1, \text{ for } t \geqslant 0, \quad \text{with } u(0) = v, \tag{1.3}$$

where

$$(v,w) = (v,w)_{L_2(\Omega)}, \quad A(v,w) = (\nabla v, \nabla w).$$

The finite element methods studied 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

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

Following [13], the spatially semidiscrete methods considered here are based on using analogues of (1.3) restricted to S_h, in which the first term (u_t, φ) has been modified, or to find $u_h(t) \in S_h$ for $t \geqslant 0$, such that

$$[u_{h,t}, \chi] + A(u_h, \chi) = 0, \quad \forall \chi \in S_h, \text{ for } t \geqslant 0, \quad \text{with } u_h(0) = v_h, \tag{1.4}$$

where $[\cdot,\cdot]$ is an inner product in S_h, approximating $(\cdot,\cdot)$. The specific choices of $[\cdot,\cdot]$ for the SG, LM and FVE methods will be reviewed in Section 2.

We now formulate (1.4) in matrix form. Let $\{P_j\}_{j=1}^N$ be the interior nodes of $\mathcal{T}_h$, and $\{\Phi_j\}_{j=1}^N \subset S_h$ the corresponding nodal basis, thus with $\Phi_j(P_i) = \delta_{ij}$. We may then write

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

The semidiscrete problem (1.4) may then be expressed, with $\alpha = (\alpha_1, \dots, \alpha_N)^T$, as

$$\mathcal{M}\alpha' + \mathcal{S}\alpha = 0, \text{ for } t \geqslant 0, \quad \text{with } \alpha(0) = \tilde{v}, \tag{1.5}$$

where $\mathcal{M} = (m_{ij}),\ m_{ij} = [\Phi_j, \Phi_i],\ \mathcal{S} = (s_{ij}),\ s_{ij} = A(\Phi_j, \Phi_i)$, and $\tilde{v} = (\tilde{v}_1, \dots, \tilde{v}_N)^T$. Here $\mathcal{M}$ is the mass matrix and $\mathcal{S}$ the stiffness matrix; they are both symmetric, positive definite. The solution of (1.5) may be written, with $\mathcal{E}(t)$ the solution matrix,

$$\alpha(t) = \mathcal{E}(t)\,\tilde{v}, \quad \text{where } \mathcal{E}(t) = e^{-\mathcal{H}t}, \quad \mathcal{H} = \mathcal{M}^{-1}\mathcal{S}, \text{ for } t \geqslant 0. \tag{1.6}$$

We note that the semidiscrete solution $u_h(t)$ is $\geqslant 0$ (> 0) if and only if, elementwise, $\alpha(t) \geqslant 0$ (> 0), and that this holds for all $\tilde{v}$ if and only if $\mathcal{E}(t) \geqslant 0 (> 0)$ elementwise.

It was proved in [14] that, for the semidiscrete SG method, the discrete analogue of (1.2) is not valid for all $t \geqslant 0$, and this was generalized in [13] to methods of the form (1.4) with nondiagonal mass matrices, including the FVE method. However, in the case of the LM method, for which the mass matrix is diagonal, $\mathcal{E}(t) \geqslant 0$ for all $t \geqslant 0$ if and only if the triangulation is of Delaunay type - for triangulations with all angles $\leqslant \frac{1}{2}\pi$ this was shown already in Fujii [6]. When the solution matrix is not nonnegative for all positive times, the possible nonnegativity of $\mathcal{E}(t)$ for larger time was also discussed in [14] and [13], with t_0 such that $\mathcal{E}(t) \geqslant 0$ for $t \geqslant t_0 > 0$ referred to as a *threshold of positivity.*

In [11] some analogous results to those for the spatially semidiscrete SG and LM methods were obtained for one step fully discrete schemes, with time stepping matrices of the form $\mathcal{E}_k^n \approx \mathcal{E}(t_n),\ t_n = nk$, where $\mathcal{E}_k = r(k\mathcal{H})$, with $r(\xi)$ a rational function and k a time step. Some of these were extended in [13] to the present generality.

In this work we complement these investigations in a number of ways. After the introductory Sections 1 and 2, we will first discuss, in Section 3, the spatially semidiscrete methods, with a somewhat more precise study of the positivity threshold than in [14] and [13]. In Section 4, we treat fully discrete methods, starting with the Backward Euler method, and continuing with more general $\mathcal{E}_k = r(k\mathcal{H})$. We discuss the existence of a positivity threshold k_0 such that $\mathcal{E}_k \geqslant 0$ for $k \geqslant k_0$, and show that this requires $r(\xi) \geqslant 0$ for large ξ. This is satisfied, e.g., for the $(0,2)$-Padé method, which has a positivity threshold, but, for the $\theta-$method, with $r(\xi) = (1-(1-\theta)\xi)/(1+\theta\xi),\ 0 < \theta < 1$, we have $r(\xi) < 0$ for large ξ, and $\mathcal{E}_k \geqslant 0$ may then possibly hold in an interval $k_0 \leqslant k \leqslant k^0$.

In Section 5 we give concrete examples, using numerical computations in MATLAB to elucidate our theoretical results. The first example uses uniform triangulations $\mathcal{T}_h$ of the unit square, in which the stiffness matrices correspond to the 5-point finite difference Laplacian. In this case, the semidiscrete solution has a positivity threshold which decreases with h, and for the BE method, it is bounded below by ch^2, with $c > 0$. The second example illustrates the case of non-Delaunay triangulations, and the BE method then has no positivity threshold, but the semidiscrete and $(0,2)$-Padé methods behave reasonably. We finally give some examples using unstructured triangulations based on commercial software, namely a square, a disk and an L-shaped domain. In all cases, the positivity thresholds decrease for the semidiscrete method and as ch^2 for BE, but does not decrease for the $(0,2)$-Padé. In Section 6, for further insight, we consider the restriction of our above analysis to the case of a uniformly partition of the one dimensional interval $(0,1)$, and the analysis and computations confirm our 2D conclusions.

Our investigations indicate that positivity is preserved for the Backward Euler method in all cases considered with Delaunay triangulations, even for the SG and FVE spatial discretizations, with all reasonable choices of the time step. In fact, the positivity thresholds all decrease with h like ch^2 with $c > 0$. The behavior is less encouraging for the semidiscrete and other fully discrete methods.

In the final Section 7, we present a simple way to find a nonnegative approximate solution when a not necessarily nonnegative approximate solution is given, the *cutoff method.* If $u_h(t) \in S_h$ is a spatially semidiscrete approximate solution of (1.1), i.e., the solution of (1.4), then defining $u_h^+(t) \in S_h$ by using the pointwise positive parts of the already computed approximate solution $u_h(t)$ at the nodes of $\mathcal{T}_h$, or $u_h^+(P_j,t) = \max(u_h(P_j,t),0)$, we have at once

$$|u_h^+(P_j,t) - u(P_j,t)| \leqslant |u_h(P_j,t) - u(P_j,t)|, \quad \text{for } j = 1,\dots,N.$$

We show how L_2 error bounds for $u_h(t)$ imply such error bounds for $u_h^+(t)$. This procedure may also be applied to fully discrete approximate solutions. For another approach in this case, using cutoff at each time step of the computation, see [10].

2 The spatially semidiscrete methods

We begin our discussion of the semidiscrete problem (1.4), or (1.5), by observing that for the stiffness matrix $\mathcal{S} = (s_{ij})$, with $s_{ij} = (\nabla\Phi_i, \nabla\Phi_j)$, which is common to all our problems (1.4), simple calculations show, see, e.g., [5],

$$s_{ij} = \begin{cases} \sum_{K\subset\Pi_i} h_{K,i}^{-2}|K|, & \text{if } i = j, \\ -\frac{1}{2}\cot\alpha - \frac{1}{2}\cot\beta = -\frac{1}{2}\sin(\alpha+\beta)/(\sin\alpha\sin\beta), & \text{if } P_i, P_j \text{ neighbors,} \\ 0 & \text{if } P_i, P_j \text{ not neighbors,} \end{cases} \tag{2.1}$$

where $\Pi_i = \operatorname{supp}(\Phi_i)$, $h_{K,i}$ is the height of K with respect to the side opposite P_i and α and β are the angles opposite P_iP_j, see Fig. 1. We shall assume throughout that $\mathcal{S}$ is irreducible.

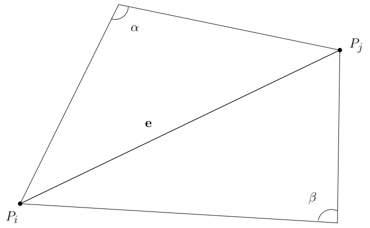

Fig. 1. An interior edge $\mathbf{e} = P_iP_j$ of $\mathcal{T}_h$.

We shall now present the three different versions of (1.4) mentioned above, defined by three different discrete inner products $[\cdot,\cdot]$ on S_h.

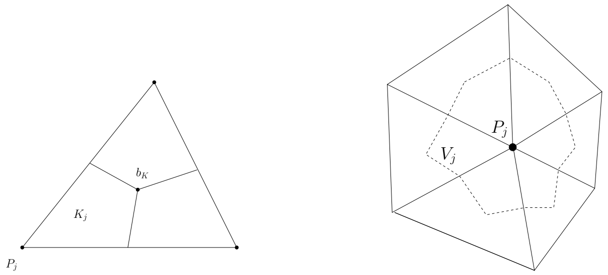

Fig. 2. A triangle $K \in \mathcal{T}_h$ and a patch $\Pi_j = \mathrm{supp}(\Phi_j)$ around a node P_j.

For the *Standard Galerkin* method we use $[\cdot,\cdot] = (\cdot,\cdot) = (\cdot,\cdot)_{L_2(\Omega)}$, and thus the mass matrix is $\mathcal{M} = \widehat{\mathcal{M}} = (\widehat{m}_{ij})$, where

$$\widehat{m}_{ij} = (\Phi_i, \Phi_j) = \begin{cases} \frac{1}{6}|\mathrm{supp}(\Phi_i)|, & \text{if } i = j, \\ \frac{1}{12}|\mathrm{supp}(\Phi_i\Phi_j)|, & \text{if } P_i, P_j \text{ neighbors}, \\ 0, & \text{if } P_i, P_j \text{ not neighbors}. \end{cases} \tag{2.2}$$

For the *Lumped Mass* method we employ $[\cdot,\cdot] = (\cdot,\cdot)_h$, where the latter is defined by quadrature,

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

with $\{P_{K,j}\}_{j=1}^3$ the vertices of the triangle K. In the matrix formulation (1.5) this means that $\mathcal{M} = \mathcal{D} = (d_{ij})$, with $d_{ii} = (\Phi_i,\Phi_i)_h = \frac{1}{3}|\,\mathrm{supp}(\Phi_i)|$, $d_{ij} = (\Phi_j,\Phi_i)_h = 0$ for $j \neq i$, so that $\mathcal{D}$ is a diagonal matrix.

To define the *Finite Volume Element* method, we note that a solution of the differential equation $u_t = \Delta u$ in (1.1) satisfies the local conservation law

$$\int_V u_t\,dx - \int_{\partial V} \frac{\partial u}{\partial n}\,ds = 0, \quad \text{for } t \geqslant 0, \tag{2.3}$$

for any $V \subset \Omega$ with piecewise smooth boundary ∂V, and n the unit exterior normal to ∂V. The spatially semidiscrete FVE method is then to find $\tilde{u}_h(t) \in S_h$, for $t \geqslant 0$, satisfying

$$\int_{V_j} \tilde{u}_{h,t}\,dx - \int_{\partial V_j} \frac{\partial \tilde{u}_h}{\partial n}\,ds = 0, \text{ for } j = 1,\dots,N, \quad t \geqslant 0, \quad \text{with } \tilde{u}_h(0) = v_h, \tag{2.4}$$

where the V_j are the so called control volumes, defined as follows, see Fig. 2. Let b_K be the barycenter of $K \in \mathcal{T}_h$, and connect b_K with the midpoints of the edges of K, thus partitioning K into three quadrilaterals K_l, $l = j, m, n$, if K has vertices P_j, P_m, P_n. The control volume V_j is then the union of the subregions K_j, sharing the vertex P_j. The equations (2.4) then preserve (2.3) for any union of control volumes.

To write (2.4) in weak form, we introduce the finite dimensional space

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

For $\eta \in Y_h$, we multiply (2.4) by $\eta(P_j)$, and sum over j, to obtain the Petrov–Galerkin formulation

$$(\tilde{u}_{h,t}, \eta) + a_h(\tilde{u}_h, \eta) = 0, \quad \forall \eta \in Y_h, \quad t \geqslant 0, \quad \text{with } \tilde{u}_h(0) = v_h, \tag{2.5}$$

where

$$a_h(\chi,\eta) = -\sum_{j=1}^{N} \eta(P_j) \int_{\partial V_j} \frac{\partial \chi}{\partial n}\,ds, \quad \forall \chi \in S_h, \ \eta \in Y_h. \tag{2.6}$$

In order to rephrase this as a 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 [4].

Lemma 2.1. *The bilinear form $\langle \chi, \psi \rangle = (\chi, J_h\psi)$ is symmetric, positive definite on S_h, and*

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

Setting $[\chi, \psi] = \langle \chi, \psi \rangle$, for $\chi, \psi \in S_h$, the Petrov-Galerkin equation (2.5), (2.6) may then be written as (1.4), and the mass matrix $\mathcal{M}$ in (1.5) is now $\widetilde{\mathcal{M}} = (\widetilde{m}_{ij})$ where

$$\widetilde{m}_{ij} = \langle \Phi_i, \Phi_j \rangle = \begin{cases} \frac{11}{54}|\mathrm{supp}(\Phi_i)|, & \text{if } i = j, \\ \frac{7}{108}|\mathrm{supp}(\Phi_i\Phi_j)|, & \text{if } P_i, P_j \text{ neighbors}, \\ 0, & \text{if } P_i, P_j \text{ not neighbors}. \end{cases} \tag{2.7}$$

We note that by (2.2) and (2.7), $\widetilde{m}_{ii} = \frac{11}{9}\widehat{m}_{ii}$, $\widetilde{m}_{ij} = \frac{7}{9}\widehat{m}_{ij}$, $i \neq j$, $i, j = 1, \dots, N$. Thus $\widetilde{\mathcal{M}}$ is more concentrated on the diagonal than $\widehat{\mathcal{M}}$.

3 Positivity preservation in the spatially semidiscrete methods

In this section we consider the general spatially semidiscrete problem (1.4) in the form (1.5). We shall first recall two results from [13] concerning the positivity of the solution matrices, and then discuss the positivity of the solution matrices for large t. We assume that $[\cdot, \cdot]$ is either such that $m_{ij} > 0$ 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 first have the following negative result, which was shown in [14] for the SG method, and generalized in [13] to the present framework. The proof depends on a technical assumption about the triangulation $\mathcal{T}_h$. First, a node of $\mathcal{T}_h$ is said to be *strictly interior* if all its neighbors are interior nodes, and then $\mathcal{T}_h$ is *normal* if it 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. This is satisfied, e.g., if all neighbors of P_j are strictly interior and the patch Π_j defined by P_j is convex.

Theorem 3.1. *Assume that $\mathcal{T}_h$ is normal, and $\mathcal{M}$ nondiagonal. Then the solution matrix for* (1.5), $\mathcal{E}(t) = e^{-\mathcal{H}t}$, *with $\mathcal{H} = \mathcal{M}^{-1}\mathcal{S}$, cannot be nonnegative for all $t > 0$.*

This result thus covers the SG method and the FVE method, but not the LM method. We recall that an edge $\mathbf{e}$ of $\mathcal{T}_h$ is a Delaunay edge if the sum of the angles opposite $\mathbf{e}$ is $\leqslant \pi$, see Fig. 1, and that $\mathcal{T}_h$ a Delaunay triangulation if all *interior* edges are Delaunay. Using (2.1) we see that an interior edge $\mathbf{e} = P_iP_j$ is a Delaunay edge if and only if $s_{ij} \leqslant 0$, and thus the triangulation $\mathcal{T}_h$ is of Delaunay type if and only if $s_{ij} \leqslant 0$ for all $i \neq j$, i.e., if and only if the stiffness matrix is a Stieltjes matrix, i.e., a symmetric positive definite matrix with nonpositive off diagonal elements. We may then cite the following theorem from [14].

Theorem 3.2. *The LM solution matrix $\bar{\mathcal{E}}(t) = e^{-\bar{\mathcal{H}}t}$, $\bar{\mathcal{H}} = \mathcal{D}^{-1}\mathcal{S}$, is nonnegative for all $t \geqslant 0$ if and only if $\mathcal{T}_h$ is Delaunay.*

Recall that, see e.g. [15, Corollary 3.24], if $\mathcal{A}$ is a Stieltjes matrix, then its inverse $\mathcal{A}^{-1} \geqslant 0$. Further, if $\mathcal{A}$ is also irreducible, then $\mathcal{A}^{-1} > 0$. In particular, if $\mathcal{T}_h$ is Delaunay, then, since $\mathcal{S}$ is irreducible, we have that $\mathcal{S}^{-1} > 0$, and hence also $\mathcal{G} = \mathcal{S}^{-1}\mathcal{M} > 0$. However, $\mathcal{T}_h$ Delaunay is not a necessary condition for $\mathcal{S}^{-1} > 0$.

Since $\mathcal{G} = \mathcal{S}^{-1}\mathcal{M} = \mathcal{H}^{-1}$ is symmetric 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 positive eigenvalues $\{\kappa_j\}_{j=1}^{N}$ and orthonormal eigenvectors $\{\varphi_j\}_{j=1}^{N}$ with respect to this inner product. We shall say that $\mathcal{G}$ is *eventually positive* if $\kappa_j < \kappa_1$ for $j \geqslant 2$ and $\varphi_1 > 0$. By the Perron-Frobenius theorem this holds if $\mathcal{G} > 0$, and more generally if $\mathcal{G}^q > 0$ for some $q \geqslant 1$.

We now return to the general semidiscrete problem (1.4) in matrix form (1.5), with solution matrix $\mathcal{E}(t) = e^{-t\mathcal{H}}$, where $\mathcal{H} = \mathcal{M}^{-1}\mathcal{S}$. We shall see that, if $\mathcal{G} = \mathcal{H}^{-1}$ is eventually positive, there exists a positivity threshold $t_0 \geqslant 0$ such that $\mathcal{E}(t) > 0$ for $t > t_0$. Following Horváth [7] we shall discuss this here in a somewhat more precise way than in [14] and [13].

Under the above assumptions, any $V \in \mathbb{R}^N$ has the eigenfunction expansion

$$V = \sum_{j=1}^{N} \eta_j \varphi_j, \quad \text{where } \eta_j = \mathcal{M}V \cdot \varphi_j, \tag{3.1}$$

and the solution of (1.5) with $\tilde{v} = V$ is

$$\mathcal{E}(t)V = \sum_{j=1}^{N} e^{-\lambda_j t} \eta_j \varphi_j, \quad \text{where } \lambda_j = 1/\kappa_j.$$

We now define the convex cone

$$\mathcal{P} = \{V \in \mathbb{R}^N;\ \sum_{j=2}^{N} |\eta_j|\sigma_j < \eta_1\}, \quad \text{where } \sigma_j = \sup_l(|\varphi_{j,l}|/\varphi_{1,l}), \quad \varphi_j = (\varphi_{j,1}, \dots, \varphi_{j,N})^T. \tag{3.2}$$

Note that since $1 = \mathcal{M}\varphi_j \cdot \varphi_j \leqslant \sigma_j^2 \mathcal{M}\varphi_1 \cdot \varphi_1 = \sigma_j^2$, we have $\sigma_j \geqslant 1$. For $V \in \mathcal{P}$ it is required that $\eta_1 > 0$, and $V \in \mathcal{P}$ implies $V > 0$, since

$$V_l \geqslant \eta_1\varphi_{1,l} - \sum_{j=2}^{N} |\eta_j|\,|\varphi_{j,l}| \geqslant \eta_1\varphi_{1,l} - \sum_{j=2}^{N} |\eta_j|\,\sigma_j\,\varphi_{1,l} = (\eta_1 - \sum_{j=2}^{N} |\eta_j|\sigma_j)\varphi_{1,l} > 0.$$

We now show that if the matrix $\mathcal{G} = \mathcal{H}^{-1} = \mathcal{S}^{-1}\mathcal{M}$ is eventually positive, then the solution matrix of (1.5) is positive for large t. A somewhat less precise result was shown in [13].

Theorem 3.3. *Let $\mathcal{E}(t) = e^{-\mathcal{H}t}$ be the solution matrix for (1.5), and let $\mathcal{H}^{-1}$ be eventually positive. Let κ_j, σ_j be as above, and $\lambda_j = 1/\kappa_j$. Then $\mathcal{E}(t) > 0$ if*

$$\sum_{j=2}^{N} e^{-\lambda_j t}\sigma_j^2 < e^{-\lambda_1 t}. \tag{3.3}$$

Proof. Let $V = \sum_j \eta_j\varphi_j$. If $V \geqslant 0,\ V \neq 0$, we have $\eta_1 = \mathcal{M}V \cdot \varphi_1 > 0$, since $\mathcal{M}V \geqslant 0$ and $\varphi_1 > 0$. Further,

$$|\eta_j| = |\mathcal{M}V \cdot \varphi_j| \leqslant (\mathcal{M}V \cdot \varphi_1)\ \sigma_j = \eta_1\ \sigma_j, \ \text{for } 2 \leqslant j \leqslant N. \tag{3.4}$$

Hence (3.3) implies $\sum_{j=2}^{N} e^{-\lambda_j t}|\eta_j|\sigma_j < e^{-\lambda_1 t}\eta_1$, and thus $\mathcal{E}(t)V \in \mathcal{P}$ and $\mathcal{E}(t)V > 0$. Hence $\mathcal{E}(t) > 0$. □

The decreasing function $\sum_{j=2}^{N} e^{-(\lambda_j-\lambda_1)t}\sigma_j^2 - 1$ has a unique zero t_1, which is > 0 since $\sigma_j \geqslant 1$ implies $\sum_{j=2}^{N} \sigma_j^2 \geqslant N-1$, and (3.3) then holds for $t > t_1$. Thus $\mathcal{E}(t) > 0$ for $t > t_1$. Clearly, $t_1 \geqslant t_0$, the positivity threshold.

4 Fully discrete methods

In this section we consider time discretization of the semidiscrete problem (1.4), or (1.5). We review results from [11] and [13] concerning nonnegativity for all positive k of time stepping matrices $\mathcal{E}_k = r(k\mathcal{H})$, where $r(\xi)$ is a bounded rational function for $\xi \geqslant 0$ and $\mathcal{H} = \mathcal{M}^{-1}\mathcal{S}$, and then discuss nonnegativity of such time stepping matrices for large time steps k.

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

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

In matrix formulation, with $U^n = \sum_{j=1}^N \alpha_j^n \Phi_j,\ \alpha^n = (\alpha_1^n, \dots, \alpha_N^n)^T$, this takes the form

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

This may also be written as $\alpha^n = \mathcal{E}_k^n \tilde{v}$, where $\mathcal{E}_k$ is the solution matrix defined by

$$\mathcal{E}_k = (\mathcal{M} + k\mathcal{S})^{-1}\mathcal{M} = (\mathcal{I} + k\mathcal{H})^{-1}, \quad \text{where } \mathcal{H} = \mathcal{M}^{-1}\mathcal{S}.$$

The following time discrete analogue of Theorem 3.1 was shown in [13], [14].

Theorem 4.1. *Assume that $\mathcal{T}_h$ is normal and $\mathcal{M}$ nondiagonal. Then $\mathcal{E}_k = (\mathcal{I} + k\mathcal{H})^{-1}$ cannot be nonnegative for small $k > 0$.*

The positivity of $\mathcal{E}_k$ for larger k is related to the positivity of the matrix $\mathcal{H}^{-1}$, and the following result was shown in [13].

Theorem 4.2. *If $\mathcal{E}_k = (\mathcal{I} + k\mathcal{H})^{-1} \geqslant 0$ for k large, then $\mathcal{H}^{-1} \geqslant 0$. If $\mathcal{H}^{-1} > 0$, then there exists $k_0 \geqslant 0$ such that $\mathcal{E}_k > 0$ for $k > k_0$. If $\mathcal{E}_{k_0} \geqslant 0$, then $\mathcal{E}_k \geqslant 0$ for $k \geqslant k_0$.*

We refer to the smallest k_0 such that $\mathcal{E}_k \geqslant 0$ for $k \geqslant k_0$ as the *threshold of positivity* for $\mathcal{E}_k$. In view of the last part of the theorem, in the BE case this is the smallest k for which $\mathcal{E}_k \geqslant 0$. In [13], and [11] in the case of SG, the following more precise result for values of k for which $\mathcal{E}_k \geqslant 0$ was derived, under a sharper condition than $\mathcal{H}^{-1} > 0$.

Theorem 4.3. *If $s_{ij} < 0$ for all neighbors P_i, P_j, then $\mathcal{E}_k \geqslant 0$ if*

$$k \geqslant k_1 = \max_{\mathcal{N}}(m_{ij}/|s_{ij}|), \quad \textit{where } \mathcal{N} = \{(i,j); P_i, P_j \textit{ neighbors}\}. \tag{4.1}$$

Since $\tilde{m}_{ij} = \frac{7}{9}\hat{m}_{ij}$ the bound k_1 in (4.1) is smaller for FVE than for SG. For instance, if the maximal angle of $\mathcal{T}_h$ is $\alpha < \frac{1}{2}\pi$, then, for all $\mathbf{e} = P_iP_j$, $|\operatorname{supp}(\Phi_i, \Phi_j)| \leqslant h^2 \sin\alpha$ and $|s_{ij}| \geqslant \cot\alpha$ by (2.1), and hence $\hat{k}_1 = \max_{\mathcal{N}}(\hat{m}_{ij}/|s_{ij}|) \leqslant \frac{1}{12}h^2 \sin^2\alpha/\cos\alpha$ and similarly, $\tilde{k}_1 \leqslant \frac{7}{108}h^2 \sin^2\alpha/\cos\alpha$. In both cases $\mathcal{E}_k \geqslant 0$ for $k \geqslant ch^2$, with the appropriate c. When all $K \in \mathcal{T}_h$ are equilateral, we find $\hat{k}_1 = \frac{1}{8}h^2$ and $\tilde{k}_1 = \frac{7}{72}h^2$.

For the Backward Euler LM method we have $\mathcal{M} = \mathcal{D}$ and

$$\bar{\mathcal{E}}_k = (\mathcal{D} + k\mathcal{S})^{-1}\mathcal{D} = (\mathcal{I} + k\mathcal{H})^{-1}, \quad \text{where } \bar{\mathcal{H}} = \mathcal{D}^{-1}\mathcal{S}.$$

In this case we have the following result, analogous to Theorem 3.2 in the semidiscrete case, cf. [11].

Theorem 4.4. *$\bar{\mathcal{E}}_k \geqslant 0$ for all $k > 0$ if and only if $\mathcal{T}_h$ Delaunay.*

Note that if $\mathcal{S}^{-1} > 0$ it follows from Theorem 4.2 that $\bar{\mathcal{E}}_k > 0$ for large k.

We now turn to more general time stepping methods and consider a time stepping matrix $\mathcal{E}_k = r(k\mathcal{H})$, where $r(\xi)$ is a bounded rational function for $\xi \geqslant 0$, approximating $e^{-\xi}$, so that $r(\xi) = 1 - \xi + O(\xi^2)$ as $\xi \to 0$. We define a single step time discretization $\mathcal{E}_k^n \tilde{v}_h$ of (1.5), by

$$\alpha^n = \mathcal{E}_k^n \tilde{v}_h, \text{ for } n \geqslant 0, \quad \text{where } \mathcal{E}_k = r(k\mathcal{H}),\ \mathcal{H} = \mathcal{M}^{-1}\mathcal{S}.$$

We recall from [13] that, as in Theorem 4.1, the time stepping matrix $\mathcal{E}_k$ cannot be nonnegative for small k when $\mathcal{M}$ is nondiagonal.

Theorem 4.5. *Assume that $\mathcal{T}_h$ is normal and $\mathcal{M}$ nondiagonal. Then $\mathcal{E}_k = r(k\mathcal{H})$ cannot be nonnegative for small k.*

For the possible nonnegativity of $\mathcal{E}_k = r(k\mathcal{H})$ for larger k, we quote the next theorem from [13].

Theorem 4.6. *Let $\mathcal{H}^{-1} > 0$. Then a necessary condition for $\mathcal{E}_k = r(k\mathcal{H})$ to be nonnegative for large k is that $r(\xi) \geqslant 0$ for large ξ.*

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

We assume now that $r(\infty) = 0$ and, more precisely,

$$r(\xi) = c\xi^{-q} + O(\xi^{-q-1}), \quad \text{as } \xi \to \infty, \quad \text{with } q \geqslant 1,\ c > 0. \tag{4.2}$$

The following result was shown in [13].

Theorem 4.7. *Assume that $r(\xi)$ satisfies* (4.2). *Then $\mathcal{H}^{-q} \geqslant 0$ is a necessary condition for $\mathcal{E}_k = r(k\mathcal{H}) \geqslant 0$ for large k. If $\mathcal{H}^{-q} > 0$, then $\mathcal{E}_k > 0$ for large k.*

In particular, $\mathcal{E}_k = r_{02}(k\mathcal{H}) > 0$ for large k, if $\mathcal{H}^{-2} > 0$. By Theorem 4.5, the positivity threshold k_0 has to be strictly positive if $\mathcal{M}$ is nondiagonal. However, even for the LM method, with a diagonal mass matrix, it was shown in [13] that $\bar{\mathcal{E}}_k$ cannot be nonnegative for small k if $\mathcal{T}_h$ is 4-connected in the sense of the following definition of p-connected: There exists a path $\mathcal{P}$ in the set of nodes of $\mathcal{T}_h$ consisting of p connected edges P_mP_n with $s_{mn} \neq 0$, and such that the endpoints of $\mathcal{P}$ cannot be connected by a path with fewer than p such edges. We now show the following more general result.

Theorem 4.8. *Assume that $\mathcal{T}_h$ is Delaunay and p-connected, and that $\bar{\mathcal{E}}_k = r(k\bar{\mathcal{H}}) \geqslant 0$ for small k. Then $(-1)^p r^{(p)}(0) \geqslant 0$.*

Proof. We have, by Taylor expansion of $r(\xi)$,

$$\bar{\mathcal{E}}_k = r(k\bar{\mathcal{H}}) = r(0)\mathcal{I} + r'(0)\,k\bar{\mathcal{H}} + \cdots + r^{(p)}(0)\,k^p\bar{\mathcal{H}}^p + O(k^{p+1}), \quad \text{as } k \to 0.$$

We shall show that if $P_iP_{l_1}P_{l_2}P_{l_{p-1}}P_j$ is a path $\mathcal{P}$ as above, then $(\bar{\mathcal{E}}_k)_{ij} < 0$ for small k. For this we write $\bar{\mathcal{H}} = \mathcal{D}^{-1}\mathcal{S} = \mathcal{V} - \mathcal{W}$, where $\mathcal{V}$ is a positive diagonal matrix and $\mathcal{W}$ has elements $w_{mn} = -s_{mn}/d_m > 0$ when P_m, P_n are neighbors with $s_{mn} \neq 0$, with the remaining elements 0. (Recall that since $\mathcal{S}$ is Stieltjes, $\mathcal{W} \geqslant 0$.) It follows that $((-\bar{\mathcal{H}})^p)_{ij} = \sum_{l_1,\dots,l_{p-1}}(-\bar{h}_{i,l_1})(-\bar{h}_{l_1,l_2})\dots(-\bar{h}_{l_{p-1},j})$ 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 $((-\bar{\mathcal{H}})^p)_{ij} \geqslant w_{i,l_1}\dots w_{l_{p-1},j} > 0$. In the same way, since P_j cannot be reached from P_i in less than p steps, $(\bar{\mathcal{H}}^l)_{ij} = 0$ for $l = 0, 1, \dots, p-1$. Hence, for k small,

$$(\bar{\mathcal{E}}_k)_{ij} = (-1)^p r^{(p)}(0)k^p((-\bar{\mathcal{H}})^p)_{ij} + O(k^{p+1}) \geqslant 0,$$

which implies our claim. □

Thus, as a particular case, if $\mathcal{T}_h$ is Delaunay and 4-connected, then $r_{02}(k\bar{\mathcal{H}})$ cannot be nonnegative for small $k > 0$ since $r_{02}(\xi) = 1-\xi+\frac{1}{2}\xi^2-\frac{1}{4}\xi^4+O(\xi^5)$ for ξ small, so $r_{02}^{(4)}(0) < 0$. Note that if the conclusion of the theorem holds for all $p \geqslant 1$, $r(\xi)$ is completely monotone for $\xi = 0$.

We recall that $\bar{\mathcal{E}}_k = r(k\bar{\mathcal{H}}) \geqslant 0$ for all $k > 0$ if $r(\xi)$ is of positive type, i.e., if $r(z) = \int_0^\infty g(t)\,e^{-zt}\,dt + r(\infty)$ for Re $z \geqslant 0$, with $g(t) \geqslant 0$, $r(\infty) \geqslant 0$, since $\bar{\mathcal{E}}(t) = e^{-t\mathcal{H}} \geqslant 0$ for $t \geqslant 0$, cf. Bolley and Crouzeix [1] and [11]. This holds for the Backward Euler method, but generally, since $r(\xi)$ is of positive type if and only if it is completely monotone, this cannot hold for approximations of higher order than first,

We now apply the technique of Theorem 3.3 to show a sufficient condition for the positivity of $\mathcal{E}_k = r(k\mathcal{H})$.

Theorem 4.9. *Let $\mathcal{E}_k = r(k\mathcal{H})$, where $r(\xi) \geqslant 0$ and $\mathcal{H}^{-1} > 0$. Let λ_j and σ_j be as in Section 3. Then $\mathcal{E}_k > 0$ if*

$$\sum_{j=2}^{N} r(k\lambda_j)\sigma_j^2 < r(k\lambda_1). \tag{4.3}$$

Proof. For V of the form (3.1) we have $\mathcal{E}_k V = \sum_{j=1}^N r(k\lambda_j)\eta_j\varphi_j$. If $V \geqslant 0$, with $V \neq 0$, then, since $|\eta_j| \leqslant \eta_1\sigma_j$ for $j \geqslant 2$ by (3.4), it follows from (4.3) that

$$\sum_{j=2}^N r(k\lambda_j)\,|\eta_j|\,\sigma_j \leqslant \eta_1 \sum_{j=2}^N r(k\lambda_j)\,\sigma_j^2 < r(k\lambda_1)\eta_1,$$

so that $\mathcal{E}_k V \in \mathcal{P}$. Hence $\mathcal{E}_k V > 0$ and thus $\mathcal{E}_k > 0$. □

We now use the same technique to demonstrate that, for any given k, the fully discrete solution is positive after a finite number of steps.

Theorem 4.10. *Let $\mathcal{E}_k = r(k\mathcal{H})$, where $r(\xi)$ is positive and decreasing for $\xi \geqslant 0$, and $\mathcal{H}^{-1} > 0$. Then, for any $k > 0$, there exists a $n_0(k)$ such that $\mathcal{E}_k^n > 0$ for $n \geqslant n_0(k)$.*

Proof. Setting $\rho_j = r(k\lambda_j)/r(k\lambda_1)$ we have $\rho_j \leqslant \rho_2 < 1$ for $j \geqslant 2$, and hence

$$\sum_{j=2}^N r(k\lambda_j)^n\,\sigma_j^2 \leqslant r(k\lambda_1)^n\rho_2^n \sum_{j=2}^N \sigma_j^2 < r(k\lambda_1)^n, \quad \text{for } n \geqslant n_0(k).$$

Hence $\mathcal{E}_k^n$ satisfies the analogue of (4.3) for large n, and thus $\mathcal{E}_k^n > 0$. □

We close this section with a short discussion of the $\theta-$method, thus with the time stepping matrix

$$\mathcal{E}_{\theta,k} = r_\theta(k\mathcal{H}) = (\mathcal{M} + k\theta\mathcal{S})^{-1}(\mathcal{M} - k(1-\theta)\mathcal{S}), \quad 0 < \theta < 1. \tag{4.4}$$

In this case $r_\theta(\infty) = -(1-\theta)/\theta < 0$, and $\mathcal{E}_{\theta,k}$ thus cannot be nonnegative for large k. We will show that the set of k for which $\mathcal{E}_{\theta,k} \geqslant 0$ is either and interval, possibly just a point, or empty.

Theorem 4.11. *Suppose that $\mathcal{H}^{-1} > 0$, and let k_0 be the positivity threshold for $\mathcal{E}_{1,k} = (\mathcal{I} + k\mathcal{H})^{-1}$. Then $\delta(k) := \min_{1\leqslant i\leqslant N}(\mathcal{E}_{1,k})_{ii}$ is a continuous, strictly decreasing function on $[0,\infty)$ with range $(0,1]$. With $0 < \theta < 1$ we have $\mathcal{E}_{\theta,k} \geqslant 0$ if and only if $\theta k \geqslant k_0$ and $\delta(\theta k) \geqslant 1-\theta$. If $\delta(k_0) \geqslant 1-\theta$ this is the interval $\theta^{-1}[k_0, \delta^{-1}(1-\theta)]$, and otherwise the empty set.*

Proof. We note that for $k_2 > k_1$, $\mathcal{E}_{1,k_1} - \mathcal{E}_{1,k_2} = (k_2 - k_1)\mathcal{H}(\mathcal{I} + k_1\mathcal{H})^{-1}(\mathcal{I} + k_2\mathcal{H})^{-1}$, which is a positive definite matrix and thus has positive diagonal elements. Hence the diagonal elements of the matrix on the left are positive and thus those of $\mathcal{E}_{1,k}$ strictly decreasing, so that $\delta(k)$ is strictly decreasing.

We now note that by the identity $\theta r_\theta(\xi) = r_1(\theta\xi) - (1-\theta)$ we have $\theta\,\mathcal{E}_{\theta,k} = \mathcal{E}_{1,\theta k} - (1-\theta)\mathcal{I}$. Hence $\mathcal{E}_{\theta,k} \geqslant 0$ if and only if $\mathcal{E}_{1,\theta k} \geqslant 0$ and, in addition, the diagonal elements of $\mathcal{E}_{1,\theta k}$ are $\geqslant 1-\theta$. By Theorem 4.2 the first condition is equivalent to $\theta k \geqslant k_0$, and the second holds if and only if $\delta(\theta k) \geqslant 1-\theta$, or $\theta k \leqslant \delta^{-1}(1-\theta)$. where δ^{-1} is the inverse function to δ. □

We now show the following result similar to Theorem 4.3.

Theorem 4.12. *If $s_{ij} < 0$ for all neighbors P_i, P_j, then $\mathcal{E}_{\theta,k} = r_\theta(k\mathcal{H}) \geqslant 0$ if $k \in [k_1, k^1]$, where*

$$k_1 = \theta^{-1}\max_{\mathcal{N}}(m_{ij}/|s_{ij}|), \quad k^1 = (1-\theta)^{-1}\min_i(m_{ii}/|s_{ii}|), \quad 0 < \theta < 1.$$

Proof. We easily find that if $k \geqslant k_1$ then $m_{ij} + k\theta s_{ij} \leqslant 0$ for P_i, P_j neighbors and $= 0$ if P_i, P_j not neighbors, so that $\mathcal{M} + k\theta\mathcal{S}$ is a Stieltjes matrix and hence $(\mathcal{M} + k\theta\mathcal{S})^{-1} \geqslant 0$. Also, if $k \leqslant k^1$, then $\mathcal{M} - k(1-\theta)\mathcal{S} \geqslant 0$. Thus taking the product we find $\mathcal{E}_k \geqslant 0$. □

Note that it could be the case that $k_1 > k^1$ in which case the interval is empty. This could happen if θ is small. For the LM method, $\mathcal{M} = \mathcal{D}$, and thus $k_1 = 0$, so that $\mathcal{E}_{\theta,k} \geqslant 0$ if $k \leqslant k^1$. When all $K \in \mathcal{T}_h$ are equilateral, we have, $s_{ii} = 2\sqrt{3}$ and $s_{ij} = -\sqrt{3}/3$ for $(i,j) \in \mathcal{N}$. For SG we find $m_{ij} = \frac{1}{24}\sqrt{3}\,h^2$, $(i,j) \in \mathcal{N}$, and $m_{ii} = \frac{1}{4}\sqrt{3}\,h^2$, and for FVE, $m_{ij} = \frac{7}{36}\sqrt{3}\,h^2$, $(i,j) \in \mathcal{N}$, and $m_{ii} = \frac{11}{36}\sqrt{3}\,h^2$. Hence for the Crank-Nicolson SG method, $\widehat{k}_1 = \widehat{k}^1 = \frac{1}{4}h^2$, so that the interval reduces to the point $k = \frac{1}{4}h^2$. For the CN LM method the interval becomes $[0, 2\widehat{k}^1] = [0, \frac{1}{2}h^2]$ and for the CN FVE method $[\frac{7}{9}\widehat{k}_1, \frac{11}{9}\widehat{k}^1] = [\frac{7}{36}h^2, \frac{11}{36}h^2]$.

5 Numerical examples

In this section we present some numerical examples, illustrating our theoretical results. We first consider a uniform Delaunay triangulation of standard type of the unit square, where the corresponding stiffness matrix $\mathcal{S}$ and mass matrix $\mathcal{M}$ are such that k_1 in (4.1) is not finite for the SG and FVE methods. We then consider a non-Delaunay triangulation of the unit square, thus with a corresponding stiffness matrix $\mathcal{S}$ which is not Stieltjes. Finally, we apply a software package to derive unstructured Delaunay triangulations of the unit square, and also of two other simple domains. We investigate the positivity of the spatially semidiscrete, the Backward Euler and $(0,2)-$Padé methods, for the SG, FVE and LM spatial discretization, and make some remarks about the $\theta-$method.

5.1 A standard triangulation of the unit square

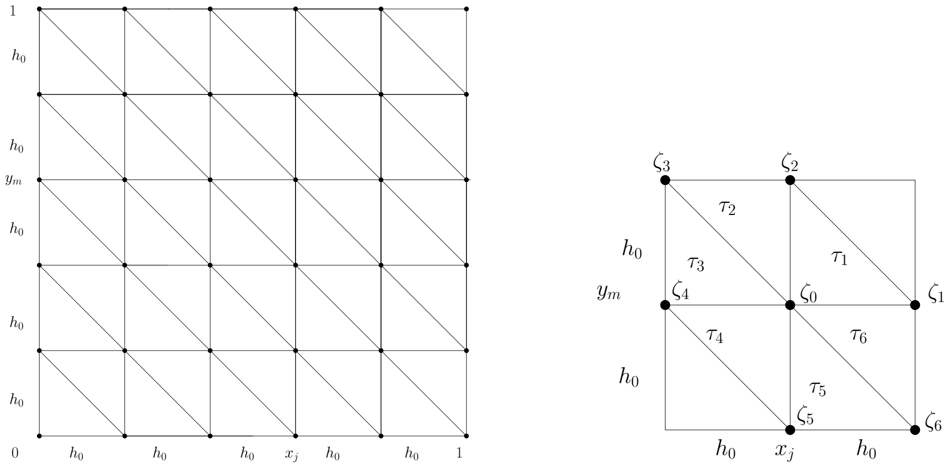

Fig. 3. *Left*: The unit square Ω with the symmetric triangulation $\mathcal{T}_h$. *Right*: The patch Π_0 around the vertex ζ_0.

In this first example we consider the unit square $\Omega = (0,1) \times (0,1)$ and introduce a uniform triangulation $\mathcal{T}_h$ of Ω as follows. Let M be a positive integer, $h_0 = 1/(M+1)$, and set for $j = 0, \dots, M+1$, $x_j = y_j = jh_0$. This partitions Ω into squares $(x_j, x_{j+1}) \times (y_m, y_{m+1})$, and we define $\mathcal{T}_h$ by connecting the nodes (x_j, y_m), (x_{j+1}, y_{m-1}), see Fig. 3. The number of interior nodes is $N = M^2$, and $h = \max_{\mathcal{T}_h} \operatorname{diam}(K) = \sqrt{2}h_0$. We note that $\mathcal{T}_h$ is a Delaunay triangulation, but since the sum of the angles opposite a diagonal edge is π, the corresponding elements s_{ij} of the stiffness matrix vanish.

Let now $\zeta_0 = (x_j, y_m)$ be an interior node of $\mathcal{T}_h$ and let $\{\zeta_j\}_{j=1}^{6}$ be the surrounding nodes, see Fig. 3. Let K_j be the triangle with vertices ζ_0, ζ_j, ζ_{j+1}, where $\zeta_7 = \zeta_1$. We then have $|K_j| = \frac{1}{2}h_0^2$, for $j = 1, \dots, 6$. It is easy to form the stiffness matrix $\mathcal{S}$ and the mass matrices $\widehat{\mathcal{M}}$ and $\widetilde{\mathcal{M}}$, for the SG and FVE methods. Indeed, since only the surrounding nodes to ζ_0 contribute to the corresponding row of the matrices, we get for $\mathcal{S}$,

$$(\nabla\Phi_0, \nabla\Phi_j) = \begin{cases} 4, & j = 0, \\ -1, & j = 1,2,4,5, \\ 0, & j = 3,6, \end{cases} \tag{5.1}$$

and for $\widehat{\mathcal{M}}$ and $\widetilde{\mathcal{M}}$, using (2.2) and (2.7), and $h = \sqrt{2}\,h_0$,

$$(\Phi_0, \Phi_j) = \tfrac{1}{4}h^2 \begin{cases} 1, & j = 0, \\ \frac{1}{6}, & j = 1, \dots, 6, \end{cases} \quad \text{and} \quad \langle \Phi_0, \Phi_j \rangle = \tfrac{1}{4}h^2 \begin{cases} \frac{11}{9}, & j = 0, \\ \frac{7}{54}, & j = 1, \dots, 6. \end{cases} \tag{5.2}$$

Thus, with $\mathcal{I}$ the identity matrix and $(\mathcal{J})_{ij} = 1$ if P_i, P_j neighbors and 0 for other i, j,

$$\widehat{\mathcal{M}} = \tfrac{1}{4}h^2(\mathcal{I} + \tfrac{1}{6}\mathcal{J}) \quad \text{and} \quad \widetilde{\mathcal{M}} = \tfrac{11}{36}h^2(\mathcal{I} + \tfrac{7}{66}\mathcal{J}).$$

It follows that $\widetilde{\mathcal{M}} = \frac{1}{9}h^2\mathcal{I} + \frac{7}{9}\widehat{\mathcal{M}}$. For the LM method the mass matrix is the diagonal matrix $\mathcal{D} = \frac{1}{2}h^2\mathcal{I}$.

We note that $\mathcal{S}$ is a Stieltjes matrix, so that $\mathcal{S}^{-1} > 0$, and hence the matrices $\mathcal{H}^{-1} = \mathcal{S}^{-1}\mathcal{M} > 0$ for the SG, FVE and LM methods. Thus the results in Section 4 concerning positivity for large t and k apply.

In addition to our computational results, we want to study, somewhat more precisely, the Backward Euler method, and recall that since some $s_{ij} = 0$ for P_i, P_j neighbors, Theorem 4.3 does not apply. We want to show that nevertheless $\mathcal{E}_k = (\mathcal{M} + k\mathcal{S})^{-1}\mathcal{M} \geqslant 0$, for $\lambda = k/h_0^2$ bounded below appropriately. This follows from the following lemma.

Lemma 5.1. *We have* $(\widehat{\mathcal{M}} + k\mathcal{S})^{-1} \geqslant 0$ *for* $k \geqslant \widehat{k}_1 \approx 0.46\,h^2$ *and* $(\widetilde{\mathcal{M}} + k\mathcal{S})^{-1} \geqslant 0$ *for* $k \geqslant \widetilde{k}_1 \approx 0.38\,h^2$.

Proof. We write $\mathcal{J} = \mathcal{J}_0 + \mathcal{J}_1$ where $(\mathcal{J}_0)_{ij} = 1$ if P_i, P_j vertical or horizontal neighbors and 0 for other i, j and $(\mathcal{J}_1)_{ij} = 1$ if P_i, P_j diagonal neighbors and 0 for other i, j. Then $\mathcal{S} = 4\mathcal{I} - \mathcal{J}_0$. We note that $\mathcal{J}_0^2 \geqslant 2\,\mathcal{J}_1$, since $\mathcal{J}_0^2$ has nonzero value m in the line corresponding to P_i at nodes P_j that can be reached in m ways from P_i in two steps, vertical or horizontal, and a diagonal neighbor can be reached either by first going vertically and then horizontally, or first horizontally and then vertically.

We want to determine k such that $(\mathcal{M} + k\mathcal{S})^{-1} \geqslant 0$, and begin with the SG method. By (2.2) we now have $\widehat{\mathcal{M}} = \frac{1}{2}h^2(\frac{1}{2}\mathcal{I} + \frac{1}{12}\mathcal{J}_0 + \frac{1}{12}\mathcal{J}_1)$. Thus, with $\lambda = k/h^2$, $\widehat{\Lambda} = 4\lambda + \frac{1}{4}$,

$$\begin{aligned} \widehat{\mathcal{M}} + k\mathcal{S} &= (4k + \tfrac{1}{4}h^2)\mathcal{I} - (k - \tfrac{1}{24}h^2)\mathcal{J}_0 + \tfrac{1}{24}h^2\mathcal{J}_1 = \widehat{\Lambda}h^2\Big(\mathcal{I} - \frac{\lambda - \frac{1}{24}}{\widehat{\Lambda}}\mathcal{J}_0 + \frac{1}{24\widehat{\Lambda}}\mathcal{J}_1\Big) \\ &= \widehat{\Lambda}h^2\Big(\mathcal{I} - \frac{1}{4}(1 - \widehat{\delta})\mathcal{J}_0 + \frac{\widehat{\delta}}{10}\mathcal{J}_1\Big), \quad \text{where } \widehat{\delta} = \frac{5}{12\widehat{\Lambda}}. \end{aligned} \tag{5.3}$$

With $\mu = \frac{1}{4}(1 - \delta)$ and $\nu = \delta/10$, where $\delta = \widehat{\delta}$, we would thus like to determine δ such that

$$(\mathcal{I} - \mu\,\mathcal{J}_0 + \nu\,\mathcal{J}_1)^{-1} \geqslant 0. \tag{5.4}$$

We shall first find δ so that $(\mathcal{I} - \mu\,\mathcal{J}_0 + \frac{1}{2}\nu\,\mathcal{J}_0^2)^{-1} \geqslant 0$, using the following lemma, to be shown below.

Lemma 5.2. *Let* $0 < 2\nu \leqslant \mu^2$. *Then the zeros* $x_{1,2}$ *of* $P(x) = 1 - \mu x + \frac{1}{2}\nu x^2$ *are positive, and*

$$\frac{1}{P(x)} = \sum_{n=0}^{\infty} \omega_n x^n, \ \textit{for } 0 \leqslant x \leqslant \frac{1}{\mu}, \quad \textit{with } \omega_n > 0.$$

Note that $2\nu \leqslant \mu^2$ is equivalent to $\delta/5 \leqslant (1-\delta)^2/16$, or $\delta^2 - 5.2\delta + 1 \geqslant 0$, which is true for $0 < \delta \leqslant 0.2$. Since $\|\mathcal{J}_0\| = 4 < 4/(1 - \delta) = 1/\mu$, with $\|\cdot\|$ the matrix maximum–norm, it follows from Lemma 5.2 that

$$\mathcal{L} = (\mathcal{I} - \mu\mathcal{J}_0 + \tfrac{1}{2}\,\nu\mathcal{J}_0^2)^{-1} = \sum_{n=0}^{\infty} \omega_n \mathcal{J}_0^n \geqslant 0. \tag{5.5}$$

Further,

$$\|\mathcal{L}\| \leqslant \sum_{n=0}^{\infty} \omega_n \|\mathcal{J}_0\|^n \leqslant \sum_{n=0}^{\infty} \omega_n 4^n = \frac{1}{P(4)} = \frac{1}{1 - 4\mu + 8\nu} = \frac{1}{1.8\,\delta}. \tag{5.6}$$

We may write

$$(\mathcal{I} - \mu\mathcal{J}_0 + \nu\mathcal{J}_1)^{-1} = (\mathcal{I} - \mu\mathcal{J}_0 + \tfrac{1}{2}\nu\mathcal{J}_0^2 - \tfrac{1}{2}\nu(\mathcal{J}_0^2 - 2\mathcal{J}_1))^{-1} = (\mathcal{I} - \mathcal{N})^{-1}\,\mathcal{L}, \quad \mathcal{N} = \tfrac{1}{2}\nu\mathcal{L}(\mathcal{J}_0^2 - 2\mathcal{J}_1). \tag{5.7}$$

Here $\mathcal{N} \geqslant 0$ since $\mathcal{J}_0^2 - 2\mathcal{J}_1 \geqslant 0$ and $\mathcal{L} \geqslant 0$ by (5.5), and using (5.6), $\nu = \delta/10$, and $\|\mathcal{J}_0^2 - 2\mathcal{J}_1\| \leqslant 16$,

$$\|\mathcal{N}\| \leqslant \tfrac{1}{2}\,(\tfrac{1}{10}\,\delta)\,(1/1.8\delta)\,16 = \tfrac{4}{9} < 1,$$

and hence $(\mathcal{I} - \mathcal{N})^{-1} = \sum_{j=0}^{\infty} \mathcal{N}^j \geqslant 0$, which shows (5.4). Since $\delta = 5/(12\widehat{\Lambda}) \leqslant 0.2$ if $\widehat{\Lambda} \geqslant 5/2.4$ and hence if $\lambda = (\widehat{\Lambda} - 0.25)/4 \geqslant 11/24 \approx 0.458$, we have thus shown $\widehat{\mathcal{M}} + k\mathcal{S} \geqslant 0$ for $k \geqslant \widehat{k}_1 \approx 0.46\,h^2$.

For the FVE method we have $\widetilde{\mathcal{M}} = \frac{1}{2}h^2(\frac{11}{18}\mathcal{I} + \frac{7}{108}\mathcal{J})$, and thus, similarly to (5.3), with $\lambda = k/h^2$,

$$\begin{aligned}\widetilde{\mathcal{M}} + k\mathcal{S} &= (4k + \tfrac{11}{36}h^2)\mathcal{I} - (k - \tfrac{7}{216}h^2)\mathcal{J}_0 + \tfrac{7}{216}h^2\mathcal{J}_1 \\ &= \widetilde{\Lambda}h^2\Big(\mathcal{I} - \tfrac{1}{4}(1-\widetilde{\delta})\mathcal{J}_0 + \tfrac{7\widetilde{\delta}}{94}\mathcal{J}_1\Big), \quad \text{where } \widetilde{\Lambda} = 4\lambda + \tfrac{11}{36}, \quad \widetilde{\delta} = \tfrac{47}{108}\widetilde{\Lambda}.\end{aligned}$$

This time we would like to determine δ so that (5.4) holds when $\mu = \frac{1}{4}(1-\delta)$ and $\nu = 7\delta/94$, where $\delta = \widetilde{\delta}$. Note that $2\nu \leqslant \mu^2$ now means $7\delta/47 \leqslant (1-\delta)^2/16$, or $\delta^2 - (206/47)\,\delta + 1 \geqslant 0$, which is true for $0 \leqslant \delta \leqslant 0.2414$. The positivity results (5.5) and (5.7) remain valid. Thus, since $\delta = (47/108)/(4\lambda{+}33/108)$, we find that $\mathcal{E}_k$ is now positive for $\lambda \geqslant \widetilde{k}_1 \approx 0.38\,h^2$. □

Proof of Lemma 5.2. The zeros of $P(x)$ are $x_{1,2} = \frac{\mu}{\nu} \pm \sqrt{\frac{\mu^2}{\nu^2} - \frac{2}{\nu}}$, so for the smallest zero, $x_2 = \frac{\mu}{\nu}(1 - \sqrt{1 - \frac{2\nu}{\mu^2}}) > \frac{\mu}{\nu}(1 - 1 + \frac{\nu}{\mu^2}) = \frac{1}{\mu} > 0$, where we have used the inequality $\sqrt{1-x} < 1 - \frac{1}{2}x$, for $0 < x \leqslant 1$. Hence, for $x < x_2$,

$$\begin{aligned}\frac{1}{P(x)} &= \frac{2}{\nu(x-x_1)(x-x_2)} = \frac{2}{\nu x_2(x_1-x_2)}\cdot\frac{1}{1-x/x_2} - \frac{2}{\nu x_1(x_1-x_2)}\cdot\frac{1}{1-x/x_1} \\ &= \frac{2}{\nu(x_1-x_2)}\sum_{n=0}^{\infty}\left(x_2^{-n-1} - x_1^{-n-1}\right)x^n = \sum_{n=0}^{\infty}\omega_n\,x^n, \quad \text{with } \omega_n > 0.\end{aligned}$$

But, by the above, $x \leqslant 1/\mu$ implies $x < x_2$, which completes the proof. □

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 for the LM methods, when M=10, 20, and 40. The numbers indicate that for 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 smaller, and the ratio k_0/h^2 is approximately 0.27 for SG and 0.23 for FVE, which is better than the above theoretical results. For the $(0,2)-$Padé method the thresholds do not appear to diminish with h, and are also independent of the choice of the finite element discretization. In Table 2 we exhibit similar results for $\mathcal{E}_k^m, m = 4$, for $\mathcal{E}_k = r_{01}(k\mathcal{H})$ and $\mathcal{E}_k = r_{02}(k\mathcal{H})$, respectively. Comparing with the results for $m = 1$ in Table 1 we see that, for the BE method, the improvement in using $m = 4$ is moderate, but a little better for the $(0,2)-$Padé method. The thresholds for BE become smaller with h as in Table 1 whereas this is not the case for the $(0,2)-$Padé.

			$e^{-t\mathcal{H}}$		$r_{01}(k\mathcal{H})$		$r_{02}(k\mathcal{H})$		
h_0	h	N	$\widehat{t}_0$	$\widetilde{t}_0$	$\widehat{k}_0$	$\widetilde{k}_0$	$\widehat{k}_0$	$\widetilde{k}_0$	$\overline{k}_0$
0.100	0.141	81	0.046	0.043	0.0053	0.0045	0.025	0.024	0.020
0.050	0.070	361	0.035	0.031	0.0013	0.0011	0.023	0.023	0.021
0.025	0.035	1521	0.021	0.019	0.0003	0.0003	0.022	0.022	0.022

Table 1. Positivity thresholds in Ex. 5.1, for $\mathcal{E}(t) = e^{-t\mathcal{H}}$ and $\mathcal{E}_k = r_{0i}(k\mathcal{H})$, with $i = 1, 2$, for SG, FVE and LM.

We end with a remark about the $\theta-$method (4.4), and consider first the SG spatial discretization. By Lemma 5.1 we have $(\widehat{\mathcal{M}} + k\theta\mathcal{S})^{-1} \geqslant 0$ for $k\theta \geqslant \widehat{k}_1 \approx 0.46\,h^2$ and clearly $\widehat{\mathcal{M}} - (1-\theta)\,k\mathcal{S} \geqslant 0$ if $(1-\theta)\,k\,s_{ii} \leqslant \widehat{m}_{ii}$ for all i, or, since $s_{ii} = 4$ by (5.1) and $\widehat{m}_{ii} = \frac{1}{24}\,h^2$ by (5.2), if $(1-\theta)\,k \leqslant 0.06\,h^2$. Thus $\widehat{\mathcal{E}}_{\theta,k} \geqslant 0$ for $0.46\,\theta^{-1}\,h^2 \leqslant k \leqslant 0.06\,(1-\theta)^{-1}\,h^2$. Note that this interval is nonempty if $0.46\,(1-\theta) \leqslant 0.06\,\theta$

		$r_{01}(k\mathcal{H})$		$r_{02}(k\mathcal{H})$		
h_0	N	$\widehat{k}_0$	$\widetilde{k}_0$	$\widehat{k}_0$	$\widetilde{k}_0$	$\bar{k}_0$
0.100	81	0.0048	0.0040	0.010	0.009	0.020
0.050	361	0.0013	0.0011	0.009	0.008	0.021
0.025	1521	0.0003	0.0003	0.008	0.008	0.022

Table 2. Positivity thresholds in Ex. 5.1, for $\mathcal{E}_k^m = r_{0i}(k\mathcal{H})^m$, with $i = 1, 2$, m=4, for SG, FVE and LM.

or if $\theta \geqslant 0.89$. For the FVE method, $(\widetilde{\mathcal{M}} + k\mathcal{S})^{-1} \geqslant 0$ for $k\theta \geqslant \widetilde{k}_1 \approx 0.38\,h^2$ and $\widetilde{\mathcal{M}} - (1-\theta)\mathcal{S} \geqslant 0$ for $(1-\theta)k \leqslant 0.07\,h^2$ so that $\widetilde{\mathcal{E}}_{\theta,k} \geqslant 0$ for $0.38\,\theta^{-1}\,h^2 \leqslant k \leqslant 0.07\,(1-\theta)^{-1}\,h^2$, which interval is nonempty if $\theta \geqslant 0.85$. The Crank-Nicolson method ($\theta = \frac{1}{2}$) is not covered in any of these cases, and numerical calculations show that $\mathcal{E}_{1/2,k} \geqslant 0$ does not hold for any of our three triangulations.

5.2 A non-Delaunay triangulation of the unit square

In this example we consider again the unit square $\Omega = (0,1) \times (0,1)$ and introduce a triangulation $\mathcal{T}_h$ of Ω as follows. Let M be a positive integer, $h_0 = 1/(2M)$, and set $x_j = jh_0$ for $j = 0, \dots, 2M$, and $y_m = 2mh_0$ for $m = 0, \dots, M$. This partitions Ω into rectangles $(x_j, x_{j+1}) \times (y_m, y_{m+1})$, and we now connect the nodes (x_j, y_m), (x_{j+1}, y_{m+1}) and (x_{j+1}, y_m), (x_j, y_{m+1}), see Fig. 4. This introduces into $\mathcal{T}_h$ also the nodes $(x_{j+1/2}, y_{m+1/2})$, with $x_{j+1/2} = (x_j + x_{j+1})/2$ and $y_{m+1/2} = (y_m + y_{m+1})/2$. The number of interior nodes is then $N = 2M^2 + (2M-1)(M-1)$, and $h = \max_{\mathcal{T}_h} \operatorname{diam}(K) = 2h_0$. We note that $\mathcal{T}_h$ is not a Delaunay triangulation.

To construct the stiffness matrix $\mathcal{S}$ and the mass matrix $\mathcal{M}$ we distinguish between two kinds of patches, Π_0 and $\check{\Pi}_0$, centered at $\zeta_0 = (x_k, y_m)$ and $\check{\zeta}_0 = (x_{\ell+1/2}, y_{m+1/2})$, respectively. For the patch Π_0 we denote by $\{\zeta_j\}_{j=1}^8$ the surrounding nodes, numbered counterclockwise as in the patch of the previous subsection, starting with $\zeta_1 = (x_k + h_0, y_m)$. Letting K_j be the triangle with vertices ζ_0, ζ_j, ζ_{j+1}, where $\zeta_9 = \zeta_1$, we then have $|K_j| = \frac{1}{2}h_0^2$, $j = 1, \dots, 8$. Similarly, for the patch $\check{\Pi}_0$, let $\{\check{\zeta}_j\}_{j=1}^4$ be the surrounding nodes, numbered counterclockwise, starting with $\check{\zeta}_1 = (x_\ell + h_0, y_m)$. With $\check{K}_j$ the triangle with vertices $\check{\zeta}_0$, $\check{\zeta}_j$, $\check{\zeta}_{j+1}$, where $\check{\zeta}_5 = \check{\zeta}_1$ we have $|\check{K}_j| = \frac{1}{2}h_0^2$, $j = 1, \dots, 4$. To form $\mathcal{S}$ it suffices to calculate $(\nabla\Phi_0, \nabla\Phi_j)$, $j = 0, \dots, 8$, and $(\nabla\check{\Phi}_0, \nabla\check{\Phi}_j)$, $j = 0, \dots, 4$, and we obtain easily, using (2.1),

$$(\nabla\Phi_0, \nabla\Phi_j) = \begin{cases} 5, & j = 0, \\ -\frac{3}{4}, & j = 1, 5, \\ -\frac{5}{4}, & j = 2, 4, 6, 8, \\ \frac{3}{4}, & j = 3, 7, \end{cases} \quad \text{and} \quad (\nabla\check{\Phi}_0, \nabla\check{\Phi}_j) = \begin{cases} 5, & j = 0, \\ -\frac{5}{4}, & j = 1, 2, 3, 4, \end{cases}$$

Correspondingly, for the mass matrix $\widehat{\mathcal{M}}$, for SG,

$$(\Phi_0, \Phi_j) = \tfrac{1}{2}h_0^2 \begin{cases} \frac{4}{3}, & j = 0, \\ \frac{1}{6}, & j = 1, \dots, 8, \end{cases} \quad \text{and} \quad (\check{\Phi}_0, \check{\Phi}_j) = \tfrac{1}{2}h_0^2 \begin{cases} \frac{2}{3}, & j = 0, \\ \frac{1}{6}, & j = 1, \dots, 4. \end{cases}$$

Similarly, for $\widetilde{\mathcal{M}}$, for FVE, we have

$$\langle\Phi_0, \Phi_j\rangle = \tfrac{1}{2}h_0^2 \begin{cases} \frac{44}{27}, & j = 0, \\ \frac{7}{54}, & j = 1, \dots, 8, \end{cases} \quad \text{and} \quad \langle\check{\Phi}_0, \check{\Phi}_j\rangle = \tfrac{1}{2}h_0^2 \begin{cases} \frac{22}{27}, & j = 0, \\ \frac{7}{54}, & j = 1, \dots, 4, \end{cases}$$

and for the LM diagonal matrix $\bar{\mathcal{M}} = \mathcal{D}$, $(\Phi_0, \Phi_0)_h = \frac{4}{3}h_0^2$, $(\check{\Phi}_0, \check{\Phi}_0)_h = \frac{2}{3}h_0^2$.

Since $\mathcal{T}_h$ is not Delaunay, the LM solution matrices $\bar{\mathcal{E}}(t)$ and $\bar{\mathcal{E}}_k$ cannot be nonnegative for all $t > 0$ and $k > 0$, respectively, by Theorems 3.2 and 4.4. Further $\mathcal{S}$ is not a Stieltjes matrix, and therefore $\mathcal{H}^{-1}$

may not be nonnegative. Even if this is so, $\mathcal{H}^{-2}$ could be positive, and therefore $\mathcal{H}^{-1}$ eventually positive, so that $\mathcal{E}_k = r_{01}(k\mathcal{H})$ has no threshold of positivity, but $\mathcal{E}(t)$ and $\mathcal{E}_k = r_{02}(k\mathcal{H})$ do, by Theorems 3.3 and 4.7. In Table 3 we consider the cases M=5, 10, and 20, for the SG, FVE and LM methods. In these cases the matrices $\mathcal{H}^{-1}$ are not nonnegative but $\mathcal{H}^{-2} > 0$, and hence the BE matrices have no positivity threshold, but those for $\mathcal{E}(t)$ and $r_{02}(k\mathcal{H})$ do, and these then diminish slowly with h.

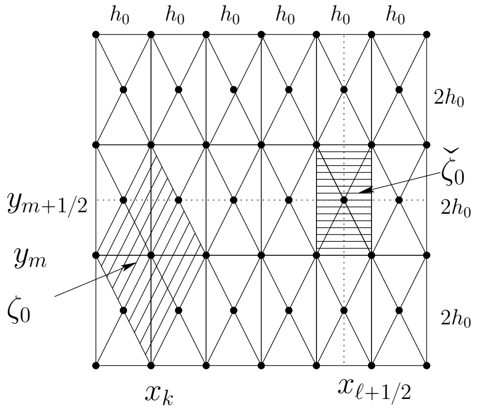

Fig. 4. The non–Delaunay triangulation $\mathcal{T}_h$ and patches Π_0 and $\check{\Pi}_0$ around ζ_0 and $\check{\zeta}_0$, respectively ($M = 3$).

			$e^{-t\mathcal{H}}$			$r_{02}(k\mathcal{H})$		
h_0	h	N	$\widehat{t}_0$	$\widetilde{t}_0$	$\overline{t}_0$	$\widehat{k}_0$	$\widetilde{k}_0$	$\overline{k}_0$
0.100	0.200	86	0.050	0.046	0.028	0.037	0.029	0.026
0.050	0.100	371	0.043	0.040	0.022	0.020	0.019	0.015
0.025	0.050	1541	0.028	0.026	0.014	0.012	0.011	0.010

Table 3. Positivity thresholds for $\mathcal{E}(t) = e^{-t\mathcal{H}}$ and $\mathcal{E}_k = r_{02}(k\mathcal{H})$, for SG, FVE and LM.

5.3 Unstructured Delaunay triangulations

We consider again the unit square $\Omega = (0,1) \times (0,1)$, which is now partitioned by unstructured Delaunay triangulations. This was done by means of the commercial software Hypermesh [8], a finite element preprocessor used in various platforms in industry and research projects. In order to be able to compare with our earlier computations, we applied its automatic 2D triangular mesh generator, with parameters chosen to produce three triangulations, with maximal side lengths close to those in the computations in Section 5.1, i.e. $h \approx 0.14$, 0.07, and 0.035. The parameters chosen included the maximum and minimum side lengths and angles, and the resulting triangulations were then modified manually to improve their quality. In all cases the maximal angle was less than 80.4°. The mass and stiffness matrices corresponding to the triangulations were then assembled using the MATLAB geometry preprocessing tool [9].

As earlier we then computed positivity thresholds for $\mathcal{E}(t)$ and 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 methods. The results are displayed in Table 4. In addition to the computed values k_0 for the BE, we also give k_1, computed according to Theorem 4.3.

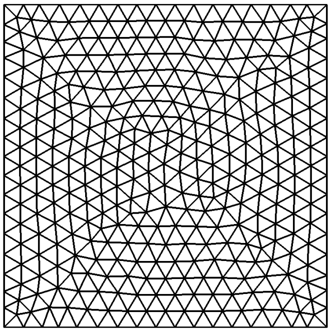

Fig. 5. An unstructured Delaunay triangulation for the unit square ($N = 295$)

		$e^{-t\mathcal{H}}$		$r_{01}(k\mathcal{H})$				$r_{02}(k\mathcal{H})$		
h	N	$\widehat{t}_0$	$\widetilde{t}_0$	$\widehat{k}_0$	$\widehat{k}_1$	$\widetilde{k}_0$	$\widetilde{k}_1$	$\widehat{k}_0$	$\widetilde{k}_0$	$\overline{k}_0$
0.140	65	0.038	0.033	0.0030	0.0039	0.0024	0.0030	0.022	0.021	0.017
0.068	295	0.026	0.022	0.0007	0.0008	0.0005	0.0007	0.022	0.022	0.021
0.035	1170	0.015	0.013	0.0002	0.0002	0.0002	0.0002	0.022	0.022	0.022

Table 4. Positivity thresholds for $\mathcal{E}_h(t) = e^{-t\mathcal{H}}$ and $\mathcal{E}_k = r_{0i}(k\mathcal{H})$, $i = 1, 2$, for SG, FVE and LM on the unit square.

We also study in the same manner two other simple domains, namely a disk with diameter 1, and an L-shaped domain, the unit square with the bottom right quarter deleted, see Fig. 6. The results are exhibited in Tables 5 and 6. We see that the numerical experiments show the same behavior of the positivity thresholds as for structured triangulations.

As in Section 5.1, the Crank-Nicolson method did not have any interval of positivity for the SG and FVE method, for any of the domains studied.

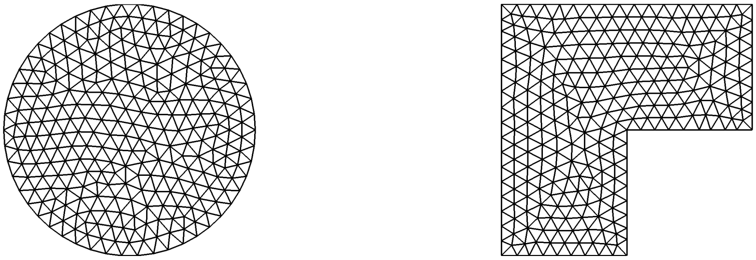

Fig. 6. Unstructured triangulations for the disk with radius $1/2$ ($N = 231$) and the L-chaped domain ($N = 212$).

		$e^{-t\mathcal{H}}$		$r_{01}(k\mathcal{H})$				$r_{02}(k\mathcal{H})$		
h	N	$\widehat{t}_0$	$\widetilde{t}_0$	$\widehat{k}_0$	$\widehat{k}_1$	$\widetilde{k}_0$	$\widetilde{k}_1$	$\widehat{k}_0$	$\widetilde{k}_0$	$\overline{k}_0$
0.138	50	0.025	0.022	0.0028	0.0029	0.0022	0.0022	0.015	0.014	0.010
0.070	231	0.018	0.015	0.0007	0.0009	0.0006	0.0007	0.015	0.014	0.013
0.034	1186	0.010	0.009	0.0002	0.0004	0.0002	0.0003	0.014	0.014	0.014

Table 5. Positivity thresholds for $\mathcal{E}_h(t) = e^{-t\mathcal{H}}$ and $\mathcal{E}_k = r_{0i}(k\mathcal{H})$, $i = 1, 2$, for SG, FVE and LM on disk with radius $1/2$.

		$e^{-t\mathcal{H}}$		$r_{01}(k\mathcal{H})$				$r_{02}(k\mathcal{H})$		
h	N	$\widehat{t}_0$	$\widetilde{t}_0$	$\widehat{k}_0$	$\widehat{k}_1$	$\widetilde{k}_0$	$\widetilde{k}_1$	$\widehat{k}_0$	$\widetilde{k}_0$	$\overline{k}_0$
0.138	56	0.043	0.036	0.0029	0.0038	0.0023	0.0029	0.022	0.022	0.017
0.069	212	0.026	0.022	0.0008	0.0009	0.0006	0.0007	0.021	0.020	0.019
0.035	992	0.014	0.012	0.0002	0.0003	0.0002	0.0002	0.020	0.020	0.020

Table 6. Positivity thresholds for $\mathcal{E}_h(t) = e^{-t\mathcal{H}}$ and $\mathcal{E}_k = r_{0i}(k\mathcal{H})$, $i = 1, 2$, for SG, FVE and LM on an L-shaped domain.

6 A special case in one space dimension

In this section we consider the discretization of the initial-boundary value problem in one space dimension,

$$u_t = u_{xx}, \quad \text{in } \Omega = (0,1), \quad u(0,t) = u(1,t) = 0, \text{ for } t > 0, \qquad \text{with } u(x,0) = v(x).$$

We partition $\Omega = (0,1)$ uniformly into subintervals $I_j = (x_{j-1}, x_j)$ by $x_j = jh$, $j = 0, \dots, N+1$, $h = 1/(N+1)$, and let S_h be the continuous piecewise linear functions χ on this partition, with $\chi_0 = \chi_{N+1} = 0$, where $\chi_j = \chi(x_j)$. The basis functions $\{\Phi_i\}_{i=1}^N \subset S_h$ are defined by $\Phi_i(x_j) = \delta_{ij}$.

We consider first the spatially semidiscrete case, and then give some results for fully discrete methods.

6.1 The spatially semidiscrete problem

With $[\cdot,\cdot]$ an appropriate inner product on S_h, and $(\cdot,\cdot) = (\cdot,\cdot)_{L_2(0,1)}$, the spatially semidiscrete problem is

$$[u_{h,t}, \chi] + (u_h', \chi') = 0, \quad \forall \chi \in S_h, \quad \text{for } t \geqslant 0, \quad \text{with } u_h(0) = v_h. \tag{6.1}$$

For the Standard Galerkin method we use $[\cdot,\cdot] = (\cdot,\cdot)$, and for the Lumped Mass method, we approximate $\int_{I_j} f(x)\,dx$ by $\frac{1}{2}h(f(x_{j-1}) + f(x_j))$ and thus employ

$$[\psi, \chi] = (\psi, \chi)_h = \tfrac{1}{2}h \sum_{j=1}^{N} \psi(x_j)\,\chi(x_j).$$

For the Finite Volume Element method the control volumes are now the intervals $V_j = (x_{j-1/2}, x_{j+1/2})$, $j = 1, \dots, N$, where $x_{j\pm1/2} = x_j \pm \frac{1}{2}h$, and the analogue of the FVE equation (2.4) is

$$\int_{V_j} \widetilde{u}_{h,t}dx - \big(\widetilde{u}_h'(x_{j+1/2}) - \widetilde{u}_h'(x_{j-1/2})\big) = 0, \quad j = 1, \dots, N,$$

or

$$\int_{V_j} \widetilde{u}_{h,t}dx - h\Delta_h\widetilde{u}_h(x_j) = 0, \quad \text{where } \Delta_h\chi_j = h^{-2}\big(\chi_{j+1} - 2\chi_j + \chi_{j-1}\big), \quad j = 1, \dots, N. \tag{6.2}$$

For $\chi \in S_h$, letting $J_h\chi$ be the piecewise constant function on the V_j with $(J_h\chi)(x_j) = \chi(x_j)$, we may multiply (6.2) by $J_h\chi(x_j)$ and sum to obtain, with $\partial\chi_j = (\chi_{j+1} - \chi_j)/h$,

$$(\tilde{u}_{h,t}, J_h\chi) = h\sum_{j=1}^{N} \Delta_h\tilde{u}_h(x_j)\chi(x_j) = -h\sum_{j=0}^{N} \partial\tilde{u}_{h,j}\,\partial\chi_j = -(\tilde{u}_h', \chi'), \quad \forall\chi \in S_h,$$

which is (6.1), with $[\psi, \chi] = \langle\psi, \chi\rangle = (\psi, J_h\chi)$.

In matrix form, (6.1) may be written as

$$\mathcal{M}\alpha' + \mathcal{S}\alpha = 0, \text{ for } t \geqslant 0, \quad \text{with } \alpha(0) = \tilde{v}, \tag{6.3}$$

with the mass matrix $\mathcal{M} = ([\Phi_i, \Phi_j])$ and the stiffness matrix $\mathcal{S} = ((\Phi_i', \Phi_j'))$. As in (1.6) the solution matrix is $\mathcal{E}(t) = e^{-\mathcal{H}t}$, with $\mathcal{H} = \mathcal{M}^{-1}\mathcal{S}$. Here $\mathcal{M}$ and $\mathcal{S}$ take the form, with $2m_1+m_0 = 1$, $0 \leqslant 2m_1 < m_0$,

$$\mathcal{M} = h\begin{pmatrix} m_0 & m_1 & 0 & \dots & 0 \\ m_1 & m_0 & m_1 & \dots & 0 \\ \vdots & \vdots & \ddots & & \\ 0 & 0 & \dots & & m_0 \end{pmatrix} \quad \text{and} \quad \mathcal{S} = h^{-1}\begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & & \\ 0 & 0 & \dots & & 2 \end{pmatrix},$$

where $m_1 = \hat{m}_1 = 1/6$ for SG, $m_1 = \tilde{m}_1 = 1/8$ for FVE, and $m_1 = \bar{m}_1 = 0$ for LM. Note that $\widetilde{\mathcal{M}}$ is more concentrated on the diagonal than $\widehat{\mathcal{M}}$. The eigenvectors and –values of $\mathcal{H}$ are, for $j = 1, \dots, N$,

$$\{\varphi_j(x_l)\}_{l=1}^N = \{\sqrt{2h}\sin(j\pi x_l)\}_{l=1}^N \quad \text{and} \quad \lambda_j = \frac{2}{h^2}\frac{1 - \cos(\pi jh)}{m_0 + 2\,m_1\cos(\pi jh)}. \tag{6.4}$$

Analogously to Theorem 3.1 we have the following.

Theorem 6.1. *If $\mathcal{M}$ is nondiagonal (or $m_1 > 0$), then the solution matrix $\mathcal{E}(t) = e^{-\mathcal{H}t}$ cannot be nonnegative for small $t > 0$.*

Proof. Assume $\mathcal{E}(t) \geqslant 0$ for all $t > 0$. Then, as in two space dimensions,

$$\mathcal{E}(t) = e^{-\mathcal{H}t} = \mathcal{I} - \mathcal{H}\,t + O(t^2) \geqslant 0, \quad \text{as } t \to 0, \quad \text{with } \mathcal{H} = \mathcal{M}^{-1}\mathcal{S},$$

so that $h_{ij} \leqslant 0,\ j \neq i$. Here $\mathcal{M} = h\,m_0(\mathcal{I} + \mu\mathcal{J})$, and $\mu = m_1/m_0 < 1/2$, where the elements of $\mathcal{J}$ are 1 in the two main bidiagonals, with all other elements 0. Thus

$$\mathcal{M}^{-1} = h^{-1}m_0^{-1}(\mathcal{I} + \mu\mathcal{J})^{-1} = h^{-1}m_0^{-1}\sum_{j=0}^{\infty}(-1)^j\mu^j\mathcal{J}^j,$$

where the series converges since, in maximum norm, $\|\mathcal{J}\| = 2$. Further, $\mathcal{J}^j$ only has nonzero elements in bidiagonals of even order when j is even, and of odd order when j is odd. It follows that the elements of $\mathcal{M}^{-1}$ are positive in bidiagonals of even order and negative in those of odd order. Since $\mathcal{S} = 2h^{-1}(\mathcal{I} - \frac{1}{2}\mathcal{J})$, the same holds for $\mathcal{H} = \mathcal{M}^{-1}\mathcal{S}$, in contradiction to $h_{ij} \leqslant 0$ for $j \neq i$. □

For the LM method the situation is more positive.

Theorem 6.2. *The LM solution matrix $\bar{\mathcal{E}}(t) = e^{-\bar{\mathcal{H}}t}$ is nonnegative for $t \geqslant 0$.*

Proof. Since the mass matrix for LM is $\bar{\mathcal{M}} = h\mathcal{I}$, and thus $\bar{\mathcal{H}} = h^{-1}\mathcal{S} = 2h^{-2}(\mathcal{I} - \frac{1}{2}\mathcal{J})$, we have

$$\mathcal{E}(t) = e^{-t\bar{\mathcal{H}}} = e^{-2th^{-2}}e^{th^{-2}\mathcal{J}} = e^{-2th^{-2}}\sum_{j=0}^{\infty}\frac{1}{j!}(th^{-2})^j\mathcal{J}^j \geqslant 0. \qquad \square$$

Note that, since $\mathcal{S}^{-1} > 0$ we have $\mathcal{H}^{-1} = \mathcal{S}^{-1}\mathcal{M} > 0$. It therefore follows as in Theorem 3.3, since in 1D the analogues of the σ_j in (3.2) are bounded by $\sigma_j = \max_l|\sin(j\pi x_l)|/\sin(\pi x_l) \leqslant j$, that $\mathcal{E}(t) > 0$ if

$$\sum_{j=2}^{N} e^{-\lambda_j t}j^2 < e^{-\lambda_1 t}. \tag{6.5}$$

Some values of the thresholds t_0 for positivity and the infimum t_1 of t such that (6.5) holds are given in Table 7. As in Section 5 the thresholds t_0 diminish with h, here almost linearly, for both the SG and FVE methods. However, this is not true for t_1, which is seen to be a very pessimistic estimate for t_0.

h	$\widehat{t}_0$	$\widehat{t}_1$	$\widetilde{t}_0$	$\widetilde{t}_1$
0.020	0.0082	0.0522	0.0067	0.0523
0.010	0.0044	0.0523	0.0036	0.0524
0.005	0.0023	0.0523	0.0019	0.0524

Table 7. Positivity thresholds and smallest t_1 for (6.5) to hold for $\mathcal{E}(t) = e^{-t\mathcal{H}}$, for SG and FVE.

We end this section with a comment on the relation between nonnegativity and the validity of (6.5). For the semidiscrete solution matrix $\mathcal{E}(t) = (e_{ij}(t))$ of (6.3) we have by eigenvector expansion

$$e_{ij}(t) = \sum_{l=1}^{N} e^{-t\lambda_l}\varphi_i(x_l)\varphi_j(x_l) = 2h\sum_{l=1}^{N} e^{-t\lambda_l}\sin(i\pi x_l)\sin(j\pi x_l).$$

In particular, since $N\pi x_l = l\pi - \pi x_l$, we have $\sin(\pi x_l)\sin(N\pi x_l) = (-1)^{l+1}\sin^2(\pi x_l)$, and hence

$$e_{1N}(t) = 2h\sum_{l=1}^{N}(-1)^{l+1}e^{-t\lambda_l}\sin^2(l\pi h). \tag{6.6}$$

There is numerical evidence that the positivity threshold t_0 is the maximal zero of $e_{1N}(t)$. At any rate, positivity of $e_{1N}(t)$ is necessary for positivity of $\mathcal{E}(t)$. In view of (6.5) and (6.6) we have

$$\begin{aligned}\sin^2(\pi h)\Big(e^{-\lambda_1 t} - \sum_{j=2}^{N} e^{-\lambda_j t}j^2\Big) &\leqslant \sin^2(\pi h)e^{-\lambda_1 t} - \sum_{j=2}^{N} e^{-\lambda_j t}\sin^2(j\pi h) \\ &= (2h)^{-1}e_{1N}(t) - 2\sum_{j=1}^{[(N-1)/2]} e^{-\lambda_{2j+1}t}\sin^2((2j+1)\pi h).\end{aligned} \tag{6.7}$$

This inequality shows that if (6.5) holds, then $e_{1N}(t)$ is positive. It also shows, through the last positive sum in (6.7), that (6.5) is a much stronger property than nonnegativity.

6.2 Fully discrete methods.

We consider time stepping matrices $\mathcal{E}_k = r(k\mathcal{H})$ where, as in Section 4, $r(\xi)$ is a bounded rational function for $\xi \geqslant 0$ such that $r(\xi) = 1 - \xi + O(\xi^2)$ as $\xi \to 0$. In the same way as in Theorem 4.5 we have at once the following.

Theorem 6.3. *With $r(\xi)$ as above and $\mathcal{M}$ nondiagonal, $\mathcal{E}_k = r(k\mathcal{H})$ cannot be nonnegative for small k.*

This applies, in particular, to the Backward Euler method,

$$\mathcal{E}_k = r_{01}(k\mathcal{H}) = (\mathcal{M} + k\mathcal{S})^{-1}\mathcal{M}, \quad \mathcal{H} = \mathcal{M}^{-1}\mathcal{S}, \tag{6.8}$$

with $\mathcal{M}, \mathcal{S}$ as in (6.3). It also applies to the $(0, 2)-$Padé method, with $r(\xi) = r_{02}(\xi)$. For the LM method $\mathcal{M}$ is diagonal, but in spite of Theorem 6.2 it was shown in [11] that the matrix $\bar{\mathcal{E}}_k = r_{02}(k\bar{\mathcal{H}})$ cannot be nonnegative for small $\lambda = k/h^2$ if $N \geqslant 4$.

In order to discuss positivity for larger k, we recall that $\mathcal{H}^{-1} > 0$. Hence, as in Theorem 4.7, we may show the following result.

Theorem 6.4. *With $r(\xi)$ as above, if $\mathcal{E}_k = r(k\mathcal{H}) \geqslant 0$ for large k, then $r(\xi) \geqslant 0$ for large ξ. If $r(\infty) = 0$ and $r(\xi) \geqslant 0$ for large ξ, then $\mathcal{E}_k = r(k\mathcal{H}) > 0$ for large k.*

Proof. If $\mathcal{E}_k \geqslant 0$ for large k, then $\mathcal{E}_k\varphi_1 = r(k\lambda_1)\varphi_1 \geqslant 0$ for large k, and since $\varphi_1 > 0$, we have $r(k\lambda_1) \geqslant 0$ for large k, which shows the first part of the theorem.

If $r(\infty) = 0$ and $r(\xi) \geqslant 0$ for large ξ, then $r(\xi) = c\xi^{-q} + O(\xi^{-q-1})$ for large ξ, with $c > 0,\ q \geqslant 1$. We then have $\mathcal{E}_k = k^{-q}(c\mathcal{H}^{-q} + O(k^{-1}))$, and hence, since $\mathcal{H}^{-q} > 0$ we conclude $\mathcal{E}_k > 0$ for large k. □

In particular, this shows the nonnegativity for large k of $\mathcal{E}_k = r_{01}(k\mathcal{H})$ and $\mathcal{E}_k = r_{02}(k\mathcal{H})$. In the same way as in Theorem 4.9 we find, more precisely, that $\mathcal{E}_k = r(k\mathcal{H}) > 0$ if

$$\sum_{j=2}^{N} r(k\lambda_j)j^2 < r(k\lambda_1). \tag{6.9}$$

This was used in [11] to show that for the SG method, $\mathcal{E}_k = r_{02}(k\mathcal{H}) > 0$ for $k > 0.5$, independently of h. We remark that the lower bound k_1 for k for (6.9) to hold is very pessimistic, as can be seen in Table 8. We also note that this condition will not be useful for the BE method. In fact, since λ_1, λ_2 are close to the first two eigenvalues in the continuous case, π^2 and $4\pi^2$, even for the first term in (6.9) we have approximately $4r(k4\pi^2)/r(k\pi^2) = 4(1 + k\pi^2)/(1 + k4\pi^2) > 1$.

We now show a precise result concerning the nonnegativity for large k for the BE method as in (6.8). We note that, with $\lambda = k/h^2$,

$$\mathcal{M} + k\mathcal{S} = (m_0h + 2k/h)\mathcal{I} + (m_1h - k/h)\mathcal{J} = h(m_0 + 2\lambda)(\mathcal{I} + \varepsilon\mathcal{J}), \text{ with } \varepsilon = \frac{m_1 - \lambda}{m_0 + 2\lambda}.$$

Theorem 6.5. *For $\mathcal{E}_k$ defined in (6.8), we have $\mathcal{E}_k \geqslant 0$ if and only if $\lambda \geqslant m_1$.*

Proof. We first show the analogue of Theorem 4.2 in this 1D case, i.e., that the set of k with $\mathcal{E}_k \geqslant 0$ is an interval $[k_0, \infty)$. If this is not so, there is a largest $k_1 \geqslant k_0$ such that $\mathcal{E}_{k_1} \geqslant 0$, or, equivalently, a smallest $\kappa_1 > 0$ such that $(\kappa_1\mathcal{I} + \mathcal{H})^{-1} \geqslant 0$. With $\kappa = \kappa_1 - \delta < \kappa_1$, we may write

$$(\kappa\mathcal{I} + \mathcal{H})^{-1} = (\kappa_1\mathcal{I} + \mathcal{H} - \delta\mathcal{I})^{-1} = (\kappa_1\mathcal{I} + \mathcal{H})^{-1}(\mathcal{I} - \mathcal{K})^{-1}, \quad \text{where } \mathcal{K} = \delta(\kappa_1\mathcal{I} + \mathcal{H})^{-1}.$$

Here $\mathcal{K} \geqslant 0$, by assumption, and, if δ is so small that $\|\mathcal{K}\| = \delta\|(\kappa_0\mathcal{I} + \mathcal{H})^{-1}\| < 1$, then $(\mathcal{I} - \mathcal{K})^{-1} = \sum_{j=0}^{\infty} \mathcal{K}^j \geqslant 0$, and therefore $(\kappa\mathcal{I} + \mathcal{H})^{-1} \geqslant 0$. Since $\kappa < \kappa_1$ this is in contradiction to the definition of κ_1.

For $\lambda \geqslant m_1$ we have $\varepsilon \leqslant 0$ and hence, since $|\varepsilon| < 1/2$, $\mathcal{I} + \varepsilon\mathcal{J}$ is a Stieltjes matrix, so that $(\mathcal{M} + k\mathcal{S})^{-1} \geqslant 0$. Since $\mathcal{M} \geqslant 0$ it follows from (6.8) that $\mathcal{E}_k \geqslant 0$.

On the other hand, if $\lambda < m_1$, then $\varepsilon > 0$ and the elements in the first bidiagonals of $(\mathcal{I} + \varepsilon\mathcal{J})^{-1}$ are $-\varepsilon + O(\varepsilon^2)$ and those in the j^{th} bidiagonals are of order $O(\varepsilon^2)$ for $j > 1$. Hence, the elements of the second bidiagonal of $(\mathcal{I} + \varepsilon\mathcal{J})^{-1}\mathcal{M}$ are $-m_1\varepsilon + O(\varepsilon^2)$, and thus, by (6.8), the corresponding elements of $\mathcal{E}_k$ negative for $\varepsilon > 0$ small. The above therefore shows that $k_0 = m_1h^2$ and completes the proof. □

The positivity threshold is thus $k_0 = m_1h^2$. We note that since $\widehat{m}_1 > \tilde{m}_1 > 0$, the condition in Theorem 6.5 is weaker for FVE than for SG, and that since $\bar{m}_1 = 0$, the LM Backward Euler solution matrix $\bar{\mathcal{E}}_k$ is nonnegative for $k \geqslant 0$. This is illustrated in Table 8.

We observe the following 1D analogue of Theorem 4.10, showing the positivity of the discrete solution operator for any k, after a finite number of steps.

Theorem 6.6. *Let $\mathcal{E}_k = r(k\mathcal{H})$, with $r(\xi)$ as above, and positive and decreasing for $\xi \geqslant 0$. Then, for any $k > 0$, there exists a $n_0(k)$ such that $\mathcal{E}_k^n \geqslant 0$ for $n \geqslant n_0(k)$.*

We close this section by noting that as in Theorem 4.12 one may show the following result for the $\theta-$method, thus with

$$\mathcal{E}_{\theta,k} = r_\theta(k\mathcal{H}) = (\mathcal{M} + k\theta\mathcal{S})^{-1}(\mathcal{M} - k(1-\theta)\mathcal{S}), \quad 0 < \theta < 1.$$

	$r_{01}(k\mathcal{H})$		$r_{02}(k\mathcal{H})$					
h	$\widehat{k}_0$	$\widetilde{k}_0$	$\widehat{k}_0$	$\widehat{k}_1$	$\widetilde{k}_0$	$\widetilde{k}_1$	$\overline{k}_0$	$\overline{k}_1$
0.020	0.000067	0.000050	0.017	0.30	0.017	0.35	0.016	0.44
0.010	0.000017	0.000013	0.017	0.34	0.017	0.37	0.017	0.41
0.005	0.000005	0.000004	0.017	0.36	0.017	0.38	0.017	0.40

Table 8. Positivity thresholds for $r_{01}(k\mathcal{H})$ for SG, FVE and lower bound k_1 for (6.9) for $r_{02}(k\mathcal{H})$, for SG, FVE, LM.

Theorem 6.7. *We have $\mathcal{E}_{\theta,k} \geqslant 0$ if $k \in \big[\theta^{-1} m_1\, h^2, (1-\theta)^{-1}\frac{1}{2} m_0\, h^2\big]$.*

For the Crank-Nicolson method ($\theta = \frac{1}{2}$) this gives the nonnegativity interval $[\frac{1}{3}\,h^2, \frac{2}{3}\,h^2]$ for SG and $[\frac{1}{4}\,h^2, \frac{3}{4}\,h^2]$ for FVE.

7 The cutoff method

Consider first the spatially semidiscrete problem (1.4), with $\tilde{v} \geqslant 0$. The cutoff method then defines the approximate nonnegative solution $u_h^+(t) \in S_h$ for $t \geqslant 0$ by taking for its nodal values the nonnegative parts of those of $u_h(t)$, or $u_h^+(P_j, t) = \max(u_h(P_j, t), 0)$, for $j = 1, \dots, N$. Since the exact solution $u(t)$ of (1.1) is nonnegative, we have

$$|u_h^+(P_j, t) - u(P_j, t)| \leqslant |u_h(P_j, t) - u(P_j, t)|, \ \text{for } j = 1, \dots, N. \qquad (7.1)$$

We shall see that, as a result of (7.1), an error bound in $\|\cdot\| = \|\cdot\|_{L_2(\Omega)}$ for $u_h(t)$ implies an error bound in $\|\cdot\|$ for $u_h^+(t)$.

Similarly, for a fully discrete solution $U^n \in S_h$, for $n \geqslant 0$, we define the nonnegative cutoff solution $(U^n)^+ \in S_h$ by $(U^n)^+(P_j) = \max(U^n(P_j), 0)$, for $j = 1, \dots, N$, and now find

$$|(U^n)^+(P_j, t) - u(P_j, t)| \leqslant |U^n(P_j) - u(P_j)|, \ \text{for } j = 1, \dots, N.$$

We first show the following lemma.

Lemma 7.1. *Let $\chi, \psi \in S_h$, and $|\chi(P_j)| \leqslant |\psi(P_j)|$ for $j = 1, \dots, N$. Then $\|\chi\| \leqslant 2\|\psi\|$.*

Proof. We have for any $K \in \mathcal{T}_h$, with vertices $P_{K,l}$, and setting $\chi_l = \chi(P_{K,l}),\ l = 1, 2, 3$,

$$\|\chi\|^2_{L_2(K)} = \frac{1}{3}|K|\big[\big(\frac{\chi_1 + \chi_2}{2}\big)^2 + \big(\frac{\chi_2 + \chi_3}{2}\big)^2 + \big(\frac{\chi_3 + \chi_1}{2}\big)^2\big].$$

One easily shows

$$\chi_1^2 + \chi_2^2 + \chi_3^2 \leqslant (\chi_1 + \chi_2)^2 + (\chi_2 + \chi_3)^2 + (\chi_3 + \chi_1)^2 \leqslant 4(\chi_1^2 + \chi_2^2 + \chi_3^2).$$

Hence

$$\|\chi\|^2_{L_2(K)} \leqslant \frac{1}{3}|K|(\chi_1^2 + \chi_2^2 + \chi_3^2) \leqslant \frac{1}{3}|K|(\psi_1^2 + \psi_2^2 + \psi_3^2) \leqslant 4\|\psi\|^2_{L_2(K)}.$$

Summation over $K \in \mathcal{T}_h$ implies our claim. □

Let $I_h : \mathcal{C}(\Omega) \to S_h$ be defined by $I_h v(P_j) = v(P_j)$ for $j = 1, \dots, N$. Application of the lemma with $\chi = u_h^+(t) - I_h u(t),\ \psi = u_h(t) - I_h u(t)$, together with the triangle inequality, and correspondingly for the fully discrete solution immediately shows the following result.

Theorem 7.1. *We have, for the semidiscrete and fully discrete cutoff solutions of* (1.4)*,*

$$\|u_h^+(t) - u(t)\| \leqslant 2\|u_h(t) - u(t)\| + 3\|I_h u(t) - u(t)\|, \ \textit{for } t \geqslant 0,$$
$$\|(U^n)^+ - u(t_n)\| \leqslant 2\|U^n - u(t_n)\| + 3\|I_h u(t_n) - u(t_n)\|, \ \textit{for } t_n \geqslant 0.$$

As an application, let U^n be the fully discrete $(0,2)-$Padé approximation of the solution of (1.1), with discrete initial data $P_h v$ where $P_h : L_2(\Omega) \to S_h$ is the L_2-projection. Then

$$\|U^n - u(t_n)\| \leqslant C\big(h^2 t_n^{-1} + k^2 t_n^{-2}\big)\|v\|, \text{ for } t_n = nk > 0, \tag{7.2}$$

see, e.g., [12], Theorem 7.7. Since

$$\|I_h u(t_n) - u(t_n)\| \leqslant Ch^2 \|u(t_n)\|_{H^2(\Omega)} \leqslant Ch^2 t_n^{-1}\|v\|, \text{ for } t_n > 0,$$

Theorem 7.1 shows that the error bound (7.2) holds also with U^n replaced by the nonnegative cutoff solution $(U^n)^+$.

Acknowledgement: The research of P. Chatzipantelidis was partly supported by the award KA 4179 funded by the University of Crete Research Committee. The research of Z. Horváth was supported by the European Social Fund and Hungary under the grant of TÁMOP-4.2.2.A-11/1/KONV-2012-0012: Basic research for the development of hybrid and electric vehicles.

References

[1] C.Bolley and M. Crouzeix, *Conservation de la positivité lors de la discrétisation des problems d'évolution paraboliques*, RAIRO Anal. numér. 12 (1978), 237–245.

[2] P. Chatzipantelidis, R.D. Lazarov, and V. Thomée, *Some error estimates for the lumped mass finite element method for a parabolic problem*, Math. Comp. 81 (2012), 1–20.

[3] P. Chatzipantelidis, R. D. Lazarov, and V. Thomée, *Some error estimates for the finite volume element method for a parabolic problem*, Comput. Methods Appl. Math. 13 (2013), 251–279.

[4] S. H. Chou and Q. Li, *Error estimates in L^2, H^1 and L^∞ in covolume methods for elliptic and parabolic problems: A unified approach*, Math. Comp. 69, (2000), 103–120.

[5] A. Drăgănescu, T.F. Dupont, and L.R. Scott, *Failure of the discrete maximum principle for an elliptic finite element problem*, Math. Comp. 74 (2004), 1–23.

[6] H. Fujii, *Some remarks on finite element analysis of time-dependent field problems*, in Theory and Practice in Finite Element Structural Analysis, University of Tokyo Press, Tokyo, 1973, pp. 91–106.

[7] Z. Horváth, *On the positivity of matrix-vector products*, Linear Algebra Appl. 393 (2004), 253-258.

[8] Altair Ltd., HyperMesh. http://www.altairhyperworks.com/Product,7,HyperMesh.aspx

[9] T.A. Kocsis, *Matlab framework for simulation of 2D flow problems with FE and FV methods.* Research report and software, Széchenyi István University, 2014.

[10] C. Lu, W. Huang, and E. S. Van Vleck. *The cutoff method for the numerical computation of nonnegative solutions of parabolic PDEs with application to anisotropic diffusion and Lubrication-type equations*, J. Comput. Phys., 242 (2013), 24–36.

[11] A.H.Schatz, V. Thomée, and L.B.Wahlbin, *On Positivity and maximum-norm Contractivity in time stepping methods for parabolic equations*, Comput. Methods Appl. Math. 10 (2010), 421-443.

[12] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, Second Edition, Berlin, 2006.

[13] V. Thomée, *On positivity preservation in some finite element methods for the heat equation*, Numerical Methods and Applications, Lecture Notes in Computer Science, 8962, Springer, Berlin, 2015 (to appear).

[14] V. Thomée and L.B. Wahlbin, *On the existence of maximum principles in parabolic finite element equations*, Math. Comp. 77 (2008), 11–19.

[15] R.S. Varga, *Matrix Iterative Analysis*, Springer-Verlag, Second Edition, Berlin, 2000