DISCRETE MAXIMUM PRINCIPLE AND ADEQUATE DISCRETIZATIONS OF LINEAR PARABOLIC PROBLEMS

ISTVÁN FARAGÓ * AND RÓBERT HORVÁTH †

Abstract. In this paper, we analyze the connections between the different qualitative properties of numerical solutions of linear parabolic problems with Dirichlet-type boundary condition. First we formulate the qualitative properties for the differential equations and shad light on their relations. Then we show how the well-known discretization schemes can be written in the form of a one-step iterative process. We give necessary and sufficient conditions of the main qualitative properties of these iterations. We apply the results to the finite difference and Galerkin finite element solutions of linear parabolic problems. In our main result we show that the nonnegativity preservation property is equivalent to the maximum-minimum principle and they imply the maximum norm contractivity. In one, two and three dimensions, we list sufficient a priori conditions that ensure the required qualitative properties. Finally, we demonstrate the above results on numerical examples.

Keywords: Parabolic problems, heat conduction, numerical solution, qualitative properties, discrete maximum principle, nonnegativity, contractivity in maximum norm.

AMS subject classifications. 65M06, 65M50, 65M60, 65F10

1. Introduction.

Let us consider the partial differential equation $\mathcal{L}v = f$, where $\mathcal{L}$ denotes some linear partial differential operator, f is a given function, and v is the unknown function to be determined. A function g is also given, which prescribes the values of v on certain parts of the boundary of the solution domain.

Several problems for partial differential equations possess some characteristic qualitative properties, which are typical of the phenomenon the partial differential equation describes. The most important three of them are the *maximum-minimum principle* (MP), the *nonnegativity preservation* (NP) and the *maximum norm contractivity* (MNC). The MP states that the solution of a differential equation can be estimated from above and below by means of the given functions f and g. For example, when $f = 0$, the solution takes its extremal values on the boundary of the solution domain as well. The NP property means that the nonnegativity of f and g implies the nonnegativity of the solution v. The MNC property, which makes sense for time-dependent problems only, says that for arbitrary two initial functions the maximum norm of the difference of the solutions at every time level is not greater than the maximum norm of the difference of the initial functions.

In most cases, a partial differential equation cannot be solved in a useful form. In some particular cases, the solution can be written in the form of an infinite Fourier series but this is also useless from the practical point of view. Thus, the use of numerical methods is a necessary step to obtain an approximate solution. It is a natural requirement of an adequate numerical method that it has to possess the discrete equivalents of the qualitative properties the continuous problem satisfies.

The discrete version of the maximum-minimum principle is the so-called *discrete maximum-minimum principle* (DMP). The topic of validity of various DMPs arose already 30 years ago and was first investigated for elliptic problems (see e.g. [4, 5,

*Department of Applied Analysis, Eötvös Loránd University, Pf. 120, H-1518, Budapest, Hungary (faragois@cs.elte.hu). The author was supported by the National Scientific Research Found (OTKA) N. T043765.

†Corresponding author, Institute of Mathematics and Statistics, University of West-Hungary, Erzsébet u. 9, H-9400, Sopron, Hungary (rhorvath@ktk.nyme.hu). The author was supported by the National Scientific Research Found (OTKA) N. T043765 and K61800.

18, 23]). The DMP was generally guaranteed by some geometrical conditions for the meshes. The DMP for parabolic problems was discussed in [8, 10, 11, 20]. In [11], based on the acuteness of the tetrahedral meshes, a sufficient condition of the DMP was obtained for the Galerkin finite element solution of certain parabolic problems. In paper [8], a necessary and sufficient condition of the DMP was derived for Galerkin finite element methods and sufficient conditions were given for hybrid meshes. About the DMP, a comprehensive survey can be found in papers [2] and [3]; [17] investigates nonlinear problems.

The conditions of the *discrete nonnegativity preservation* (DNP) was discussed in [7, 12] for linear finite elements in one, two and three dimensions, and in [6] in one dimension with the combination of the finite difference and finite element methods. The DNP is investigated for nonlinear problems in [24].

The *discrete maximum norm contractivity* (DMNC) was analyzed for one-dimensional parabolic problems in [13, 19]. In both references the necessary and sufficient conditions were given. In the first one, the dependence on the spatial discretization was also discussed.

For one-dimensional problems, we can deduce some other remarkable qualitative properties such as the preservation of the shape and the monotonicity of the initial function, and the sign-stability (see e.g. [9, 14, 15, 16, 22]).

Albeit some sufficient conditions (and in certain cases the necessary ones as well) are known for the above listed three main qualitative properties, the relations and the implications between them have not been revealed yet. The main goal of this paper is to establish the connections between the maximum-minimum principle, the nonnegativity preservation and the maximum norm contractivity both for linear parabolic problems and their numerical solutions. In this paper, we will show the implications

$$(D)NP \Longleftrightarrow (D)MP \Longrightarrow (D)MNC,$$

that is the maximum-minimum principle is equivalent to the nonnegativity preservation property and both imply the maximum norm contractivity. We give necessary and sufficient conditions for the adequate (qualitatively correct) discrete models. Furthermore, we also give some useful sufficient conditions.

The paper is organized as follows. In §2, we discuss the qualitative properties of continuous problems. In §3, we give the linear algebraic form of the finite difference and finite element discretization of the investigated problem. We prove the existence and the uniqueness of the numerical solution. In §4, we construct a one-step iterative process, we define its qualitative properties and we formulate necessary and sufficient conditions of their preservation. In §5, we apply the results of §4 to the discretizations given in §3. In §6, some sufficient conditions are listed in order to guarantee the numerical qualitative properties a priori. In the last section, we demonstrate our results on numerical examples.

For simplicity, we denote zero matrices and zero vectors by the symbol $\mathbf{0}$, whose size is always chosen according to the context. The ordering relation for vectors and matrices is always meant elementwise.

2. Qualitative Properties of Linear Parabolic Problems. Let Ω and $\partial\Omega$ denote, respectively, a bounded domain in $\mathbb{R}^d$ ($d \in \mathbb{N}^+$) and its boundary and we introduce the sets

$$Q_\tau = \Omega \times (0,\tau), \qquad Q_{\bar{\tau}} = \Omega \times (0,\tau], \qquad \Gamma_\tau = (\partial\Omega \times [0,\tau]) \cup (\Omega \times \{0\})$$

for any arbitrary positive number τ. For some fixed number $T > 0$, we consider the linear partial differential operator

$$\mathcal{L} \equiv \frac{\partial}{\partial t} - \sum_{0<|\varsigma|\leq\delta} a_\varsigma \frac{\partial^{|\varsigma|}}{\partial^{\varsigma_1} x_1 \dots \partial^{\varsigma_d} x_d} \equiv \frac{\partial}{\partial t} - \sum_{0<|\varsigma|\leq\delta} a_\varsigma D^\varsigma, \tag{2.1}$$

where $\delta \geq 1, \varsigma_1, \dots, \varsigma_d$ denote nonnegative integers, $|\varsigma|$ is defined as $|\varsigma| = \varsigma_1 + \dots + \varsigma_d$ for the multi-index $\varsigma = (\varsigma_1, \dots, \varsigma_d)$, and the coefficient functions $a_\varsigma : Q_T \to \mathbb{R}$ are bounded on the set Q_T. We define the domain of the operator $\mathcal{L}$, denoted by $\operatorname{dom}\mathcal{L}$, as the space of functions $v \in C(Q_T \cup \Gamma_T)$, for which the derivatives $D^\varsigma v$ $(0 < |\varsigma| \leq \delta)$ and $\partial v/\partial t$ exist in Q_T and they are bounded. It can be seen easily that $\mathcal{L}v$ is bounded on $Q_{\bar{t}_1}$ for each $v \in \operatorname{dom}\mathcal{L}$ and $0 < t_1 < T$. Thus $\inf_{Q_{\bar{t}_1}} \mathcal{L}v$ and $\sup_{Q_{\bar{t}_1}} \mathcal{L}v$ are finite.

Definition 2.1. *We say that the operator $\mathcal{L}$ satisfies the maximum-minimum principle if for any function $v \in \operatorname{dom}\mathcal{L}$ the inequality*

$$\min_{\Gamma_{t_1}} v + t_1 \cdot \min\{0, \inf_{Q_{\bar{t}_1}} \mathcal{L}v\} \leq v(\mathbf{x}, t_1) \leq \max_{\Gamma_{t_1}} v + t_1 \cdot \max\{0, \sup_{Q_{\bar{t}_1}} \mathcal{L}v\}$$

is satisfied for all $\mathbf{x} \in \Omega$, $0 < t_1 < T$.

The following statement shows that, in case of the maximum-minimum principle, the functions from $\operatorname{dom}\mathcal{L}$ are uniquely determined by their values on the boundary Γ_T.

Theorem 2.2. *Let $t_1 \in (0, T)$ be any arbitrary fixed number. If $\mathcal{L}$ satisfies the maximum-minimum principle, then for any two functions $v^\star, v^{\star\star} \in \operatorname{dom}\mathcal{L}$ the relations $\mathcal{L}v^\star|_{Q_{t_1}} = \mathcal{L}v^{\star\star}|_{Q_{t_1}}$ and $v^\star|_{\Gamma_{t_1}} = v^{\star\star}|_{\Gamma_{t_1}}$ imply $v^\star|_{Q_{t_1}} = v^{\star\star}|_{Q_{t_1}}$.*

Proof. Let us choose an arbitrary number satisfying the condition $0 < t_0 < t_1$, and we consider the function $\bar{v} = v^\star - v^{\star\star}$. For this function the relations $\mathcal{L}\bar{v}|_{Q_{\bar{t}_0}} = 0$ and $\bar{v}|_{\Gamma_{t_0}} = 0$ are valid. Thus, based on the maximum-minimum principle we have $\bar{v}(\mathbf{x}, t_0) = 0$ for all $\mathbf{x} \in \Omega$ and $0 < t_0 < t_1$. This completes the proof. □

Definition 2.3. *The operator $\mathcal{L}$ is called nonnegativity preserving when the relations* $\min_{\Gamma_{t_1}} v \geq 0$ *and* $\mathcal{L}v|_{Q_{\bar{t}_1}} \geq 0$ *imply* $v|_{Q_{\bar{t}_1}} \geq 0$ *for all* $0 < t_1 < T$.

Definition 2.4. *The operator $\mathcal{L}$ is called contractive in maximum norm when for all arbitrary two functions $v^\star$, $v^{\star\star} \in \operatorname{dom}\mathcal{L}$ with $\mathcal{L}v^\star|_{Q_{\bar{t}_1}} = \mathcal{L}v^{\star\star}|_{Q_{\bar{t}_1}}$ and $v^\star|_{\partial\Omega\times[0,t_1]} = v^{\star\star}|_{\partial\Omega\times[0,t_1]}$ the relation*

$$\max_{\mathbf{x}\in\bar{\Omega}} |v^\star(\mathbf{x}, t_1) - v^{\star\star}(\mathbf{x}, t_1)| \leq \max_{\mathbf{x}\in\bar{\Omega}} |v^\star(\mathbf{x}, 0) - v^{\star\star}(\mathbf{x}, 0)|$$

is valid for all $0 < t_1 < T$.

The main connections between the above properties of the operator $\mathcal{L}$ are listed in the next two theorems.

Theorem 2.5. *The operator $\mathcal{L}$ defined in (2.1) satisfies the maximum-minimum principle if and only if it preserves the nonnegativity.*

Proof. The necessity of the condition is trivial. In order to show the sufficiency, we choose an arbitrary function $v \in \operatorname{dom}\mathcal{L}$ and apply the operator $\mathcal{L}$ to the function $\bar{v} = v - \min_{\Gamma_{t_1}} v - t \cdot \min\{0, \inf_{Q_{\bar{t}_1}} \mathcal{L}v\}$. Clearly, $\bar{v}|_{\Gamma_{t_1}} \geq 0$. Moreover, we obtain that

$$\mathcal{L}\bar{v}|_{Q_{\bar{t}_1}} = (\mathcal{L}v - \min\{0, \inf_{Q_{\bar{t}_1}} \mathcal{L}v\})|_{Q_{\bar{t}_1}} \geq 0,$$

which implies that $\bar{v}$ is nonnegative on $Q_{\bar{t}_1}$ by virtue of the nonnegativity preservation assumption. Thus the lower estimation $\min_{\Gamma_{t_1}} v + t_1 \cdot \min\{0, \inf_{Q_{\bar{t}_1}} \mathcal{L}v\} \leq v(\mathbf{x}, t_1)$ is satisfied.

By choosing

$$\bar{v} = \max_{\Gamma_{t_1}} v - v + t \cdot \max\{0, \sup_{Q_{\bar{t}_1}} \mathcal{L}v\},$$

the upper bound is proved similarly. This completes the proof. □

THEOREM 2.6. *If the operator $\mathcal{L}$ defined in (2.1) is nonnegativity preserving, then it is contractive in maximum norm.*

Proof. Let $v^\star$ and $v^{\star\star} \in \operatorname{dom}\mathcal{L}$ be two arbitrary functions with $\mathcal{L}v^\star|_{Q_{\bar{t}_1}} = \mathcal{L}v^{\star\star}|_{Q_{\bar{t}_1}}$ and $v^\star|_{\partial\Omega\times[0,t_1]} = v^{\star\star}|_{\partial\Omega\times[0,t_1]}$. We consider the functions $\bar{v}_\pm = \zeta \pm (v^\star - v^{\star\star})$ with $\zeta = \max_{\mathbf{x}\in\bar{\Omega}} |v^\star(\mathbf{x},0) - v^{\star\star}(\mathbf{x},0)|$. For these functions, the estimations $\mathcal{L}\bar{v}_\pm|_{Q_{\bar{t}_1}} = 0 \geq 0$ and $\min_{\Gamma_{t_1}} \bar{v}_\pm \geq 0$ are true, which implies the nonnegativity of $\bar{v}_\pm$ on $Q_{\bar{t}_1}$. Thus, we have

$$\max_{\mathbf{x}\in\bar{\Omega}} |v^\star(\mathbf{x},t_1) - v^{\star\star}(\mathbf{x},t_1)| \leq \max_{\mathbf{x}\in\bar{\Omega}} |v^\star(\mathbf{x},0) - v^{\star\star}(\mathbf{x},0)|.$$

This completes the proof. □

In the sequel, we consider the initial-boundary value problem in the form

$$\mathcal{L}v = f \quad \text{in } Q_T, \tag{2.2}$$

$$v|_{\Gamma_T} = g, \tag{2.3}$$

where $g : \Gamma_T \to \mathrm{I\!R}$, $f : Q_T \to \mathrm{I\!R}$ are arbitrary given functions. The operator $\mathcal{L}$ is defined in (2.1). (In the usual context, $g|_{\Omega\times\{0\}}$ is the initial and $g|_{\partial\Omega\times[0,T]}$ is the boundary condition, respectively.) We are interested in finding a function $v \in \operatorname{dom}\mathcal{L}$ that satisfies equalities (2.2)-(2.3). We say that the problem (2.2)-(2.3) is nonnegativity preserving / contractive in maximum norm / satisfies the maximum-minimum principle if the operator $\mathcal{L}$ in the equation (2.2) possesses the same properties. According to Theorem 2.5 and Theorem 2.6, the maximum-minimum principle is valid for the problem (2.2)-(2.3) if and only if it is nonnegativity preserving; in this case the problem is contractive in maximum norm as well. If the maximum-minimum principle is valid for the problem, then its solution (when it exists) is unique (cf. Theorem 2.2).

In this paper we investigate the parabolic problem

$$C\frac{\partial v}{\partial t} - \nabla(\kappa \cdot \nabla v) = f, \quad (\mathbf{x},t) \in Q_T \tag{2.4}$$

$$v|_{\Gamma_T} = g, \tag{2.5}$$

where $C : \Omega \to \mathrm{I\!R}$ is a known bounded function with the property $0 < C_{\min} \leq C \equiv C(\mathbf{x}) \leq C_{\max}$, the known bounded function $\kappa : \Omega \to \mathrm{I\!R}$ has continuous first derivatives and fulfills the property $0 < \kappa_{\min} \leq \kappa \equiv \kappa(\mathbf{x}) \leq \kappa_{\max}$, the function $g : \Gamma_T \to \mathrm{I\!R}$ is continuous on Γ_T, the function $f : Q_T \to \mathrm{I\!R}$ is bounded in Q_T, ∇ denotes the usual nabla operator, and the solution v is sought in $C^{2,1}(Q_T \cup \Gamma_T)$.

Remark 2.7. The problem (2.4)-(2.5) is generally applied for the description of heat conduction processes, where v denotes the temperature, C is the product of the specific heat and the density, κ is the thermal conductivity and f gives the density of heat sources. The variables $\mathbf{x}$ and t play the role of the space and time variables, respectively. Problem (2.4)-(2.5) is suitable to describe diffusion and transport phenomena as well.

We show that the problem (2.4)-(2.5) is nonnegativity preserving, contractive in maximum norm and satisfies the maximum-minimum principle. Because (2.4)-(2.5) can be written in the form of the problem (2.2)-(2.3) dividing both sides of the equation by the positive function C (that is $\mathcal{L} \equiv \partial/\partial t - (1/C)\nabla(\kappa\nabla)$), we have to show only the validity of the nonnegativity preservation (see Theorem 2.5 and Theorem 2.6). The proof will be based on the next theorem.

THEOREM 2.8. *For any $t_1 \in (0,T)$ and $\mathbf{x} \in \Omega$, the solution v of the problem (2.4)-(2.5) satisfies the inequality*

$$\sup_{\lambda>0}\left(e^{\lambda t_1}\min\left\{\min_{\Gamma_{t_1}} ve^{-\lambda t}, \frac{1}{\lambda}\inf_{Q_{\bar{t}_1}}\frac{fe^{-\lambda t}}{C}\right\}\right) \leq \tag{2.6}$$

$$\leq v(\mathbf{x},t_1) \leq \inf_{\lambda>0}\left(e^{\lambda t_1}\max\left\{\max_{\Gamma_{t_1}} ve^{-\lambda t}, \frac{1}{\lambda}\sup_{Q_{\bar{t}_1}}\frac{fe^{-\lambda t}}{C}\right\}\right),$$

where $Q_{\bar{t}_1} = Q_{t_1} \cup (\Omega \times \{t_1\})$.

Proof. For any arbitrary number $\lambda > 0$ we define the function $\hat{v}(\mathbf{x},t) = v(\mathbf{x},t)e^{-\lambda t}$, where v stands for the solution of (2.4). It can be seen easily that $\hat{v} \in C(Q_T \cup \Gamma_T)$, $\hat{v}|_{\Gamma_T} = (ve^{-\lambda t})|_{\Gamma_T}$, and $\hat{v}$ satisfies the equation

$$C\frac{\partial \hat{v}}{\partial t} - \nabla(\kappa\nabla\hat{v}) + C\lambda\hat{v} = fe^{-\lambda t} \tag{2.7}$$

in Q_T. As $\hat{v}$ is continuous on $\bar{Q}_{t_1}$, it takes its maximum either on the boundary Γ_{t_1} or in $\Omega \times (0,t_1]$ at some point $(\mathbf{x}^0,t^0)$. In the first case we trivially have

$$\sup_{Q_{t_1}} \hat{v} \leq \max_{\Gamma_{t_1}} \hat{v}. \tag{2.8}$$

In the second case, the relations

$$\sup_{Q_{t_1}} \hat{v} \leq \hat{v}(\mathbf{x}^0,t^0),\ \frac{\partial \hat{v}}{\partial t}(\mathbf{x}^0,t^0) \geq 0,\ \frac{\partial \hat{v}}{\partial x_i}(\mathbf{x}^0,t^0) = 0,\ \frac{\partial^2 \hat{v}}{\partial x_i^2}(\mathbf{x}^0,t^0) \leq 0 \tag{2.9}$$

hold for $i = 1,\ldots,d$, where the last two relations imply that $\nabla(\kappa\nabla\hat{v})(\mathbf{x}^0,t^0) \leq 0$. This relation, combined with the second one in (2.9) and with (2.7), results in

$$0 \leq f(\mathbf{x}^0,t^0)e^{-\lambda t^0} - C(\mathbf{x}^0)\hat{v}(\mathbf{x}^0,t^0)\lambda,$$

which can be rewritten in the form

$$\hat{v}(\mathbf{x}^0,t^0) \leq \frac{1}{\lambda C(\mathbf{x}^0)}f(\mathbf{x}^0,t^0)e^{-\lambda t^0} \leq \sup_{Q_{\bar{t}_1}}\frac{fe^{-\lambda t}}{C\lambda}. \tag{2.10}$$

Thus, in general case, using the upper bounds (2.8) and (2.10) we obtain the estimation

$$\hat{v}(\mathbf{x},t_1) \leq \sup_{Q_{t_1}} \hat{v} \leq \max\left\{\max_{\Gamma_{t_1}} \hat{v}, \sup_{Q_{\bar{t}_1}}\frac{fe^{-\lambda t}}{C\lambda}\right\}.$$

Multiplying both sides by $e^{\lambda t_1}$ and taking into account that the relation is true for all positive numbers $\lambda > 0$, we obtain the inequality on the right-hand side of (2.6). The lower bound can be proved similarly. □

Remark 2.9. The proof of the above theorem is based on the proof of Theorem 2.1 in [20]. Let us notice, however, that we did not confine ourselves to second order operators, and leaving the zeroth order derivative out of the expression of $\mathcal{L}$, we arrived at a stronger estimation.

THEOREM 2.10. *The problem (2.4)-(2.5) preserves the nonnegativity, and consequently it is contractive in maximum norm and satisfies the maximum-minimum principle.*

Proof. The proof is based on the previous theorem. Let $t_1 \in (0,T)$ be an arbitrary number, and let us suppose that $\mathcal{L}v|_{Q_{t_1}} \equiv (f/C)|_{Q_{t_1}} \geq 0$ and $v|_{\Gamma_{t_1}} \equiv g|_{\Gamma_{t_1}} \geq 0$. Then, for any $t_0 \in (0,t_1)$, we have $(f/C)|_{Q_{\bar{t}_0}} \geq 0$ and $v|_{\Gamma_{t_0}} \geq 0$, which result in $0 \leq v(\mathbf{x},t_0)$ in view of (2.6). That is v is nonnegative in Q_{t_1}. □

3. Numerical Models of Linear Parabolic Problems. The two most widely used numerical methods for solving the problem (2.4)-(2.5) are the finite difference and the Galerkin finite element methods. Finite difference methods are typically applied when the solution domain Ω is relatively simple (an interval, a rectangle or a block), while finite element methods can be used also with more complicated geometrical structures (e.g. polyhedrons). Both methods start with the discretization of the computational space $\Omega \subset \mathbb{R}^d$ in order to get a system of ordinary differential equations. This is the so-called semi-discretization. Then the system of ordinary differential equations is solved by some time-integration method, such as the well-known θ-method. In this paper we consider only the 3D case; the 1D and 2D cases can be derived in the same manner by omitting the corresponding terms.

3.1. Spatial Semi-Discretization with the Finite Difference Method. In the case of the finite difference method, the function v is approximated at the points of a rectangular mesh defined on the rectangular domain Ω. Let us denote the interior mesh points by $P_1,\ldots,P_N \in \Omega$, and the points on the boundary by $P_{N+1},\ldots,P_{N+N_\partial} \in \partial\Omega$, respectively. For the sake of simplicity, we also use the notation P_{i+x} (P_{i-x}) for the grid point adjoint to P_i in positive (negative) x-direction. We also define $\bar{N} = N + N_\partial$. After denoting the semi-discrete approximation of $v(\mathbf{x},t)$ at a grid point $\mathbf{x} = P_i$ by $v_i(t)$, we can approximate (2.4) at each inner point P_i of Ω as

$$C_i \frac{\mathrm{d}v_i(t)}{\mathrm{d}t} - (v_{ixx}(t) + v_{iyy}(t) + v_{izz}(t)) = f_i(t), \quad i = 1,\ldots,N. \tag{3.1}$$

Here $f_i(t)$ denotes some approximation to $f(\mathbf{x},t)$ at the point P_i (the most typical choice is $f_i(t) = f(P_i,t)$), furthermore $v_{ixx}(t)$ approximates the derivative

$$\frac{\partial}{\partial x}\left(\kappa \frac{\partial v}{\partial x}\right)$$

of the function v at the point P_i and at the time instant t, and has the form

$$v_{ixx}(t) = \frac{2}{h_{i+x} + h_{i-x}} \left(\kappa_{i+x/2} \frac{v_{i+x}(t) - v_i(t)}{h_{i+x}} - \kappa_{i-x/2} \frac{v_i(t) - v_{i-x}(t)}{h_{i-x}} \right). \tag{3.2}$$

The distances between the points P_{i+x}, P_i and P_{i-x}, P_i are denoted by h_{i+x} and h_{i-x}, respectively. The values $\kappa_{i-x/2}$ and $\kappa_{i+x/2}$ denote the approximate values of the material parameter κ on the segments $[P_{i-x}, P_i]$ and $[P_i, P_{i+x}]$ (typically the midpoint values), C_i denotes the approximate value of C at the point P_i (typically $C_i = C(P_i)$ for continuous functions). The terms $v_{iyy}(t)$ and $v_{izz}(t)$ are defined similarly.

Hence the semi-discretization (3.1) and the boundary condition (2.5) result in a Cauchy problem for the system of ordinary differential equations

$$\mathbf{M}\frac{\mathrm{d}\mathbf{v}(t)}{\mathrm{d}t} + \mathbf{K}\mathbf{v}(t) = \mathbf{f}(t), \tag{3.3}$$

$$\mathbf{v}(0) = [g(P_1,0),\dots,g(P_N,0),g(P_{N+1},0),\dots,g(P_{\bar{N}},0)]^\top,$$

where

$$\begin{aligned} \mathbf{v}(t) &\equiv [v_1(t),\dots,v_N(t),v_{N+1}(t),\dots,v_{\bar{N}}(t)]^\top \\ &= [v_1(t),\dots,v_N(t),g(P_{N+1},t),\dots,g(P_{\bar{N}},t)]^\top, \\ \mathbf{f}(t) &\equiv [f_1(t),\dots,f_N(t)]^\top \end{aligned} \tag{3.4}$$

and $\mathbf{M}$ and $\mathbf{K}$ are sparse $N \times \bar{N}$ matrices. For the entries of $\mathbf{M}$ we have

$$M_{ij} = \begin{cases} C_i, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}, \quad i = 1,\dots,N;\ j = 1,\dots,\bar{N} \tag{3.5}$$

and the entries of $\mathbf{K}$ can be computed using equations (3.1) and (3.2). The nonzero elements of the i-th row of $\mathbf{K}$ are $K_{i,i-x}, K_{i,i+x}, K_{i,i-y}, K_{i,i+y}, K_{i,i-z}, K_{i,i+z}, K_{i,i}$, where

$$K_{i,i-x} = \frac{-2\kappa_{i-x/2}}{h_{i-x}(h_{i-x}+h_{i+x})}, \quad K_{i,i+x} = \frac{-2\kappa_{i+x/2}}{h_{i+x}(h_{i-x}+h_{i+x})}, \tag{3.6}$$

$$K^x_{i,i} = \frac{2\kappa_{i+x/2}}{h_{i+x}(h_{i-x}+h_{i+x})} + \frac{2\kappa_{i-x/2}}{h_{i-x}(h_{i-x}+h_{i+x})},$$

$K_{i,i-y}, K_{i,i+y}, K_{i,i-z}, K_{i,i+z}, K^y_{i,i}, K^z_{i,i}$ are defined similarly and

$$K_{i,i} = K^x_{i,i} + K^y_{i,i} + K^z_{i,i}. \tag{3.7}$$

3.2. Semi-Discretization with the Galerkin Finite Element Method. In the case of the Galerkin finite element method we cover the domain Ω with the following meshes $\mathcal{T}_h$ (h is a discretization parameter): the 1D mesh consists of intervals, the 2D mesh is a so-called hybrid mesh, which is a combination of triangles and rectangles, and the 3D mesh is a tetrahedron or a block mesh. Figure 3.1 shows a 2D hybrid mesh. As before, $P_1,\dots,P_N$ denote the interior nodes of $\mathcal{T}_h$, and $P_{N+1},\dots,P_{\bar{N}}$ the boundary ones. Let $\phi_1,\dots,\phi_{\bar{N}}$ be basis functions defined as follows: each ϕ_i is required to be continuous, piecewise linear (over intervals, triangular or tetrahedral elements) or multi-linear (over rectangular or block elements), such that $\phi_i(P_j) = \delta_{ij}$, $i,j = 1,\dots,\bar{N}$, where δ_{ij} is Kronecker's symbol. (Multi-linearity means that it is bilinear in case of rectangles, and trilinear in case of blocks.) It is obvious that such basis functions have the properties

$$\phi_i \geq 0,\ i = 1,...,\bar{N},$$

$$\sum_{i=1}^{\bar{N}} \phi_i \equiv 1 \text{ in } \bar{\Omega}.$$

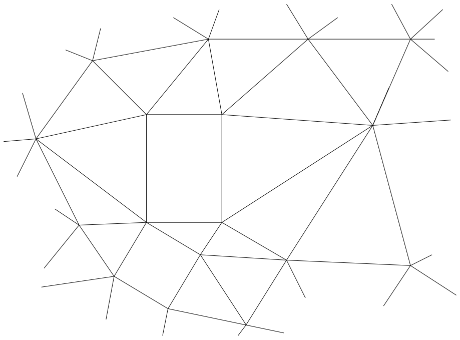

FIG. 3.1. Hybrid mesh in 2D.

We search for the semi-discrete solution of (2.4)-(2.5) in the form

$$\sum_{i=1}^{N} v_i(t)\phi_i(\mathbf{x}) + \sum_{i=1}^{N_\partial} g(P_{N+i},t)\phi_{N+i}(\mathbf{x}).$$

Using the weak formulation of the problem, we arrive at a Cauchy problem again that has the form (3.1) with

$$\mathbf{M} = [M_{ij}]_{N\times\bar{N}}, \quad M_{ij} = \int_\Omega C(\mathbf{x})\phi_j(\mathbf{x})\phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x},$$

$$\mathbf{K} = [K_{ij}]_{N\times\bar{N}}, \quad K_{ij} = \int_\Omega \kappa(\mathbf{x})\ \mathrm{grad}\phi_j(\mathbf{x})\ \mathrm{grad}\phi_i(\mathbf{x})\ \mathrm{d}\mathbf{x},$$

$$\mathbf{f}(t) = [f_i(t)]_{N\times 1}, \quad f_i(t) = \int_\Omega f(\mathbf{x},t)\phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

(see [8]). The above defined matrices $\mathbf{M}$ and $\mathbf{K}$ are called *mass and stiffness matrices*, respectively.

3.3. Fully Discretized Model. To get a fully discrete numerical scheme, we choose a time-step $\Delta t > 0$. We denote the approximation to $\mathbf{v}(n\Delta t)$ by $\mathbf{v}^n$, and we set $\mathbf{f}^n = \mathbf{f}(n\Delta t)$ $(n = 0, 1, ...)$. The time-discretization of (3.1) with the θ-method ($\theta \in [0,\ 1]$ is a given parameter) can be written in the form of the systems of linear algebraic equations

$$\mathbf{M}\frac{\mathbf{v}^{n+1}-\mathbf{v}^n}{\Delta t} + \theta\mathbf{K}\mathbf{v}^{n+1} + (1-\theta)\mathbf{K}\mathbf{v}^n = \mathbf{f}^{(n,\theta)} := \theta\mathbf{f}^{n+1} + (1-\theta)\mathbf{f}^n, \tag{3.8}$$

where $n = 0, 1, \ldots$ and $\theta \in [0,\ 1]$ is a given parameter. Clearly, (3.8) can be rewritten as

$$(\mathbf{M} + \theta\Delta t\mathbf{K})\mathbf{v}^{n+1} = (\mathbf{M} - (1-\theta)\Delta t\mathbf{K})\mathbf{v}^n + \Delta t\ \mathbf{f}^{(n,\theta)}. \tag{3.9}$$

In the sequel, the matrices $\mathbf{M} + \theta\Delta t\mathbf{K}$ and $\mathbf{M} - (1-\theta)\Delta t\mathbf{K}$ will be denoted by $\mathbf{A}$ and $\mathbf{B}$, respectively. We shall use the following partitions of the matrices and vectors:

$$\mathbf{A} = [\mathbf{A}_0|\mathbf{A}_\partial], \quad \mathbf{B} = [\mathbf{B}_0|\mathbf{B}_\partial], \quad \mathbf{v}^n = \begin{bmatrix} \mathbf{u}^n \\ \mathbf{g}^n \end{bmatrix},$$

where $\mathbf{A}_0$ and $\mathbf{B}_0$ are square matrices from $\mathrm{IR}^{N\times N}$; $\mathbf{A}_\partial, \mathbf{B}_\partial \in \mathrm{IR}^{N\times N_\partial}$, $\mathbf{u}^n \equiv [u_1^n, ..., u_N^n]^\top \in \mathrm{IR}^N$ and $\mathbf{g}^n \equiv [g_1^n, ..., g_{N_\partial}^n]^\top \in \mathrm{IR}^{N_\partial}$. Similar partition is used for the matrices $\mathbf{M}$ and $\mathbf{K}$. Then, iteration (3.9) can be written as

$$[\mathbf{A}_0|\mathbf{A}_\partial]\left[\begin{array}{c}\mathbf{u}^{n+1}\\ \mathbf{g}^{n+1}\end{array}\right] = [\mathbf{B}_0|\mathbf{B}_\partial]\left[\begin{array}{c}\mathbf{u}^{n}\\ \mathbf{g}^{n}\end{array}\right] + \Delta t\, \mathbf{f}^{(n,\theta)}. \tag{3.10}$$

The relation (3.10) defines a one-step iterative process with respect to the unknown vector $\mathbf{u}^{n+1}$.

Remark 3.1. For the finite difference method, the choice $\theta = 0$ results in an explicit scheme, due to the diagonality of $\mathbf{M}$. This is the so-called explicit Euler method. The choice $\theta = 1$ gives the so-called implicit Euler method and $\theta = 1/2$ is the Crank-Nicolson method. However, the schemes obtained by the Galerkin finite element method are always implicit for any choice of θ.

In order to guarantee the existence and the uniqueness of $\mathbf{u}^{n+1}$ in (3.10), we have to show that $\mathbf{A}_0$ is a nonsingular matrix.

THEOREM 3.2. *The matrix $\mathbf{A}_0 \in \mathrm{IR}^{N\times N}$ is nonsingular for both the finite difference and the Galerkin finite element methods.*

Proof. The case of the finite difference method: Because of the relation $\mathbf{A}_0 = \mathbf{M}_0 + \theta\Delta t\mathbf{K}_0$ and the equalities in (3.6), the off-diagonal elements of $\mathbf{A}_0$ are nonpositive and the diagonal ones are positive. Moreover, considering (3.6) and (3.7) we have

$$K_{i1} + \ldots + K_{i\bar{N}} = 0 \quad (i = 1, \ldots, N).$$

Thus, the estimate

$$\begin{aligned}(\mathbf{A}_0)_{i1} + \ldots + (\mathbf{A}_0)_{iN} &= C_i + \theta\Delta t(K_{i1} + \ldots + K_{iN})\\ &\geq C_i + \theta\Delta t(K_{i1} + \ldots + K_{i\bar{N}}) = C_i \geq C_{\min} > 0\end{aligned}$$

is true $(i = 1, \ldots, N)$, which shows that $\mathbf{A}_0$ is a strictly diagonally dominant matrix, hence it is nonsingular.

The case of the Galerkin finite element method: The elements of $\mathbf{A}_0$ are

$$(\mathbf{A}_0)_{ij} = \int_\Omega (C(\mathbf{x})\phi_j(\mathbf{x})\phi_i(\mathbf{x}) + \theta\Delta t\ \kappa(\mathbf{x})\ \mathrm{grad}\phi_j(\mathbf{x})\ \mathrm{grad}\phi_i(\mathbf{x}))\,\mathrm{d}\mathbf{x}, \quad (i, j = 1, \ldots, N).$$

The right-hand side defines a scalar product on the vector space $\mathrm{span}\{\phi_1, \ldots, \phi_N\}$. Thus $\mathbf{A}_0$ is a Gram-matrix of the linearly independent functions $\phi_1, \ldots, \phi_N$, hence it is nonsingular and also positive definite. □

Remark 3.3. It is important to notice that in the case of the finite difference method, $\mathbf{A}_0$ is a so-called M-matrix (see [1, p. 137]). This can be seen from the sign-structure and the strict diagonal dominance of $\mathbf{A}_0$. That is, in addition to the nonsingularity, $\mathbf{A}_0$ has a nonnegative inverse too. In the finite element method the nonnegativity of $\mathbf{A}_0^{-1}$ is not satisfied automatically.

4. Qualitative Properties of One-Step Iterative Models. As we mentioned in the Introduction, it is a natural requirement for the numerical solution that it has to possess some basic qualitative properties. The numerical solution can be obtained by the iteration (3.10). Hence, the qualitative properties of the numerical solution will be defined as the qualitative properties of such iteration processes.

We use the denotations

$$\mathbf{e} = \mathbf{e}^{(\bar{N})},\ \mathbf{e}_0 = \mathbf{e}^{(N)},\ \mathbf{e}_\partial = \mathbf{e}^{(N_\partial)} \text{ with } \mathbf{e}^{(k)} = [1, \ldots, 1]^\top \in \mathrm{IR}^k.$$

4.1. PO-iterations. Based on the iteration (3.10), we will investigate the general one-step iteration

$$[\mathbf{A}_0|\mathbf{A}_\partial]\begin{bmatrix}\mathbf{u}^{n+1}\\ \mathbf{g}^{n+1}\end{bmatrix}=[\mathbf{B}_0|\mathbf{B}_\partial]\begin{bmatrix}\mathbf{u}^{n}\\ \mathbf{g}^{n}\end{bmatrix}+\mathbf{h}^n,\ n=0,1,\ldots \tag{4.1}$$

or equivalently

$$\mathbf{A}\mathbf{v}^{n+1}=\mathbf{B}\mathbf{v}^n+\mathbf{h}^n,\ n=0,1,\ldots. \tag{4.2}$$

Here $\mathbf{h}^n\equiv[h_1^n,\ldots,h_N^n]^\top\in\mathrm{I\!R}^N$ and all matrices and vectors have the same dimension as in the previous section. We emphasize that now they are assumed to be *arbitrary* with the only restriction that the matrix $\mathbf{A}_0$ is nonsingular. Iteration (4.2) is called *partitioned one-step iteration* or shortly PO-iteration.

4.2. Qualitative properties of PO-iterations. Comparing the form

$$\mathbf{u}^{n+1}-\mathbf{u}^n=(\mathbf{A}_0^{-1}\mathbf{B}_0-\mathbf{I})\mathbf{u}^n-\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{g}^{n+1}+\mathbf{A}_0^{-1}\mathbf{B}_\partial\mathbf{g}^n+\mathbf{A}_0^{-1}\mathbf{h}^n$$

of (4.1) with the equation (2.4), we define the equivalents of the qualitative properties for the PO-iterations.

Definition 4.1. *A PO-iteration is said to satisfy the discrete nonnegativity preservation property (DNP) if the inequality* $\mathbf{u}^{n+1}\geq\mathbf{0}$ *is valid for all nonnegative vectors* $\mathbf{u}^n,\mathbf{g}^n,\mathbf{g}^{n+1}$ *and* $\mathbf{A}_0^{-1}\mathbf{h}^n$.

Theorem 4.2. *A PO-iteration satisfies the DNP if and only if the inequalities* $-\mathbf{A}_0^{-1}\mathbf{A}_\partial\geq\mathbf{0}$ *and* $\mathbf{A}_0^{-1}\mathbf{B}\geq\mathbf{0}$ *hold.*

Proof. The statement follows directly from the following equivalent form of (4.1)

$$\mathbf{u}^{n+1}=-\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{g}^{n+1}+\mathbf{A}_0^{-1}\mathbf{B}\begin{bmatrix}\mathbf{u}^{n}\\ \mathbf{g}^{n}\end{bmatrix}+\mathbf{A}_0^{-1}\mathbf{h}^n,\quad n=0,1,\ldots. \tag{4.3}$$

□

Definition 4.3. *A PO-iteration is said to satisfy the discrete maximum norm contractivity property (DMNC) if for all vectors* $\hat{\mathbf{u}}^n,\tilde{\mathbf{u}}^n,\mathbf{g}^n,\mathbf{g}^{n+1},\mathbf{h}^n$ *the relation*

$$\|\hat{\mathbf{u}}^{n+1}-\tilde{\mathbf{u}}^{n+1}\|_\infty\leq\|\hat{\mathbf{u}}^n-\tilde{\mathbf{u}}^n\|_\infty \tag{4.4}$$

is valid, where $\hat{\mathbf{u}}^{n+1}$ *and* $\tilde{\mathbf{u}}^{n+1}$ *are computed from (4.1) by setting* $\mathbf{u}^n=\hat{\mathbf{u}}^n$ *and* $\mathbf{u}^n=\tilde{\mathbf{u}}^n$.

Theorem 4.4. *A PO-iteration satisfies the DMNC if and only if* $\|\mathbf{A}_0^{-1}\mathbf{B}_0\|_\infty\leq 1$.

Proof. We apply (4.3) with the two different vectors $\mathbf{u}^n=\hat{\mathbf{u}}^n$ and $\mathbf{u}^n=\tilde{\mathbf{u}}^n$. Hence we obtain

$$\hat{\mathbf{u}}^{n+1}-\tilde{\mathbf{u}}^{n+1}=\mathbf{A}_0^{-1}\mathbf{B}_0(\hat{\mathbf{u}}^n-\tilde{\mathbf{u}}^n).$$

The relation (4.4) is valid for all $\hat{\mathbf{u}}^n$ and $\tilde{\mathbf{u}}^n$ vectors if and only if $\|\mathbf{A}_0^{-1}\mathbf{B}_0\|_\infty\leq 1$. □

Definition 4.5. *A PO-iteration is said to satisfy the discrete maximum-minimum principle (DMP) if the relation*

$$\min\{\mathbf{u}^n,\mathbf{g}^n,\mathbf{g}^{n+1}\}+\min\{0,\mathbf{A}_0^{-1}\mathbf{h}^n\}\leq u_i^{n+1}\leq\max\{\mathbf{u}^n,\mathbf{g}^n,\mathbf{g}^{n+1}\}+\max\{0,\mathbf{A}_0^{-1}\mathbf{h}^n\} \tag{4.5}$$

is valid for each choice $\mathbf{u}^n,\mathbf{g}^n,\mathbf{g}^{n+1},\mathbf{h}^n$, $i=1,\ldots,N$ *and* $n\geq 0$. *(The max and min are understood elementwise.)*

Now some necessary conditions are listed for the DMP.

LEMMA 4.6. *For PO-iterations DMP implies DNP.*

Proof. Let us suppose that the DMP is valid for the PO-iteration and $\mathbf{u}^n$, $\mathbf{g}^n$, $\mathbf{g}^{n+1}$, $\mathbf{A}_0^{-1}\mathbf{h}^n$ are nonnegative vectors. Then the left inequality in (4.5) shows the nonnegativity of $\mathbf{u}^{n+1}$. □

LEMMA 4.7. *If a PO-iteration satisfies the DMP, then* $\mathbf{Ae} = \mathbf{Be}$.

Proof. Choosing the vectors $\mathbf{h}^n = \mathbf{0}$, $\mathbf{u}^n = \mathbf{e}_0$, $\mathbf{g}^n = \mathbf{g}^{n+1} = \mathbf{e}_\partial$, (4.5) results in $\mathbf{u}^{n+1} = \mathbf{e}_0$. That is $\mathbf{Ae} = \mathbf{Be}$ in view of (4.1), which completes the proof. □

LEMMA 4.8. *For PO-iterations the DNP and condition* $\mathbf{Ae} = \mathbf{Be}$ *imply the DMNC.*

Proof. The DNP ensures the relations $-\mathbf{A}_0^{-1}\mathbf{A}_\partial \geq \mathbf{0}$ and $\mathbf{A}_0^{-1}\mathbf{B} \geq \mathbf{0}$. Hence

$$\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{e}_0 \leq \mathbf{A}_0^{-1}\mathbf{Be} = \mathbf{A}_0^{-1}\mathbf{Ae} = \mathbf{A}_0^{-1}\mathbf{A}_0\mathbf{e}_0 + \mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e}_\partial \leq \mathbf{e}_0.$$

This shows that $\|\mathbf{A}_0^{-1}\mathbf{B}_0\|_\infty \leq 1$, that is the DMNC is valid. □

Remark 4.9. Because the DMP implies both the DNP and condition $\mathbf{Ae} = \mathbf{Be}$, the DMP implies the DMNC.

LEMMA 4.10. *If a PO-iteration satisfies the DNP and condition* $\mathbf{Ae} = \mathbf{Be}$*, then* $\mathbf{0} \leq -\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e}_\partial \leq \mathbf{e}_0$.

Proof. The DNP property ensures the relations $-\mathbf{A}_0^{-1}\mathbf{A}_\partial \geq \mathbf{0}$ and $\mathbf{A}_0^{-1}\mathbf{B} \geq \mathbf{0}$. Thus, the relation $\mathbf{0} \leq -\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e}_\partial$ is trivial. Let us multiply the equality $\mathbf{Ae} = \mathbf{A}_0\mathbf{e}_0 + \mathbf{A}_\partial\mathbf{e}_\partial = \mathbf{Be}$ by the matrix $\mathbf{A}_0^{-1}$, and let us express the term $-\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e}_\partial$. We obtain the inequality

$$\mathbf{0} \leq -\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e}_\partial = \mathbf{e}_0 - \mathbf{A}_0^{-1}\mathbf{Be} \leq \mathbf{e}_0$$

by virtue of the nonnegativity of the vector $\mathbf{A}_0^{-1}\mathbf{Be}$. This completes the proof. □

4.3. Qualitatively adequate one-step iterations. We have seen in the previous section that the properties

$$\begin{array}{ll} (P1) & \mathbf{A}_0^{-1}\mathbf{B} \geq \mathbf{0}, \\ (P2) & -\mathbf{A}_0^{-1}\mathbf{A}_\partial \geq \mathbf{0}, \\ (P3) & \mathbf{Ae} = \mathbf{Be} \end{array}$$

are necessary conditions of the DMP for PO-iterations. In the next theorem, we prove that they are sufficient conditions as well. In order to make the expressions much simpler, we introduce some notations. We define the values

$$v^n_{\min} = \min\{\mathbf{v}^n\}, \quad v^n_{\max} = \max\{\mathbf{v}^n\},$$

and vectors

$$\mathbf{v}^n_{\max} = v^n_{\max}\mathbf{e} \in \mathbb{R}^{\bar{N}}, \quad \mathbf{v}^n_0 = v^n_{\max}\mathbf{e}_0 \in \mathbb{R}^N, \quad \mathbf{v}^n_\partial = v^n_{\max}\mathbf{e}_\partial \in \mathbb{R}^{N_\partial}.$$

THEOREM 4.11. *A PO-iteration satisfies the DMP if and only if it satisfies the conditions* $(P1) - (P3)$.

Proof. The necessity of the condition follows directly from Theorem 4.2 and Lemmas 4.6 and 4.7. To show the sufficiency, we first prove the inequality on the right-hand side in (4.5). It follows from $(P3)$ that $\mathbf{Bv}^n_{\max} = \mathbf{Av}^n_{\max}$. Using $(P1)$, we obtain

$$\begin{aligned}
\mathbf{u}^{n+1} &= -\mathbf{A}_0^{-1}\mathbf{A}_\partial\, \mathbf{g}^{n+1} + \mathbf{A}_0^{-1}\mathbf{B}\mathbf{v}^n + \mathbf{A}_0^{-1}\mathbf{h}^n \\
&\le -\mathbf{A}_0^{-1}\mathbf{A}_\partial\, \mathbf{g}^{n+1} + \mathbf{A}_0^{-1}\mathbf{B}\mathbf{v}^n_{\max} + \mathbf{A}_0^{-1}\mathbf{h}^n \\
&= -\mathbf{A}_0^{-1}\mathbf{A}_\partial\, \mathbf{g}^{n+1} + \mathbf{A}_0^{-1}\mathbf{A}\mathbf{v}^n_{\max} + \mathbf{A}_0^{-1}\mathbf{h}^n \\
&\le -\mathbf{A}_0^{-1}\mathbf{A}_\partial\, \mathbf{g}^{n+1} + \mathbf{A}_0^{-1}[\mathbf{A}_0|\ \mathbf{A}_\partial]\mathbf{v}^n_{\max} + \max\{0, \mathbf{A}_0^{-1}\mathbf{h}^n\}\mathbf{e}_0 \\
&= -\mathbf{A}_0^{-1}\mathbf{A}_\partial\, \mathbf{g}^{n+1} + \mathbf{v}_0^n + \mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{v}_\partial^n + \max\{0, \mathbf{A}_0^{-1}\mathbf{h}^n\}\mathbf{e}_0.
\end{aligned}$$

Regrouping the above inequality, we get

$$\mathbf{u}^{n+1} - \mathbf{v}_0^n - \max\{0, \mathbf{A}_0^{-1}\mathbf{h}^n\}\mathbf{e}_0 \le -\mathbf{A}_0^{-1}\mathbf{A}_\partial \left(\mathbf{g}^{n+1} - \mathbf{v}_\partial^n\right).$$

Hence, for the i-th coordinate of both sides we obtain

$$\begin{aligned}
u_i^{n+1} - v^n_{\max} - \max\{0, \mathbf{A}_0^{-1}\mathbf{h}^n\} &\le \sum_{j=1}^{N_\partial} \left(\left(-\mathbf{A}_0^{-1}\mathbf{A}_\partial\right)_{ij} \cdot \left(g_j^{n+1} - v^n_{\max}\right)\right) \\
&\le \left(\sum_{j=1}^{N_\partial} \left(-\mathbf{A}_0^{-1}\mathbf{A}_\partial\right)_{ij} \right) \cdot \max\{0, \max_j\{g_j^{n+1} - v^n_{\max}\}\} \\
&\le \max\{0, \max_j\{g_j^{n+1} - v^n_{\max}\}\},
\end{aligned}$$

where we applied property $(P2)$ and Lemma 4.10. Finally, expressing u_i^{n+1} we obtain the required inequality. The inequality on the left-hand side of (4.5) can be proved similarly. This completes the proof. □

Theorem 4.11, Lemma 4.6 and Remark 4.9 yield that a PO-iteration possesses all the three qualitative properties (DMP, DNP and DMNC) if and only if it satisfies the properties $(P1)-(P3)$. Hence the iterative process (4.1) with the properties $(P1)-(P3)$ is called *qualitatively adequate one-step iteration*.

5. Qualitative Properties of the Numerical Solutions of Parabolic Problems. We apply the results, which were formulated for arbitrary PO-iterations, of the previous section for the fully discretized numerical model (3.9). Based on Theorem 3.2, the finite difference and Galerkin finite element methods can be written in the form of a PO-iteration with the choice $\mathbf{h}^n = \Delta t\mathbf{f}^{(n,\theta)}$. This is why the qualitative properties (DMP, DNP and DMNC) of the numerical methods can be defined in the same manner like for the PO-iterations.

THEOREM 5.1. *Relation $(P3)$ holds both for the finite difference and finite element methods.*

Proof. Owing to the relation $\mathbf{A} - \mathbf{B} = (\mathbf{M} + \theta\Delta t\mathbf{K}) - (\mathbf{M} - (1-\theta)\Delta t\mathbf{K}) = \Delta t\mathbf{K}$, property $(P3)$ is equivalent to the equality $\mathbf{K}\mathbf{e} = \mathbf{0}$. For the finite difference method the statement follows directly from the equalities in (3.6). For the finite element method, we have

$$\begin{aligned}
(\mathbf{K}\mathbf{e})_i = \sum_{j=1}^{\bar{N}} K_{ij} &= \sum_{j=1}^{\bar{N}} \int_\Omega \kappa(\mathbf{x}) \cdot \mathrm{grad}\phi_j(\mathbf{x}) \cdot \mathrm{grad}\phi_i(\mathbf{x})\ \mathrm{d}\mathbf{x} \\
&= \int_\Omega \kappa(\mathbf{x}) \cdot \mathrm{grad}\left(\sum_{j=1}^{\bar{N}} \phi_j(\mathbf{x})\right) \cdot \mathrm{grad}\phi_i(\mathbf{x})\ \mathrm{d}\mathbf{x} \\
&= \int_\Omega \kappa(\mathbf{x}) \cdot \mathrm{grad}1 \cdot \mathrm{grad}\phi_i(\mathbf{x})\ \mathrm{d}\mathbf{x} = 0.
\end{aligned}$$

□

THEOREM 5.2. *For the finite difference and Galerkin finite element models of the problem (2.4)-(2.5), the implications*

$$DNP \Longleftrightarrow DMP \Longrightarrow DMNC$$

are valid. That is the DMP and DNP are equivalent and both imply the DMNC.

Proof. The first part of the statement follows from Theorem 4.2, Lemma 4.6 and Theorem 5.1. The second one is a consequence of Lemma 4.8. □

COROLLARY 5.3. *A finite difference or a Galerkin finite element method for the problem (2.4)-(2.5) is qualitatively adequate if and only if it satisfies properties $(P1)-(P2)$, that is if the method preserves the nonnegativity.*

Remark 5.4. In view of the results of [13], there exist parameter choices when the finite difference or the finite element method is contractive in maximum norm, but it does not preserves the nonnegativity.

Remark 5.5. Let us notice that our results can be applied not only for the finite difference and the Galerkin finite element method but also for any numerical method that can be written in the form of a PO-iteration.

6. Sufficient a Priori Conditions of the Qualitative Properties. The necessary and sufficient conditions in the previous sections can not be checked easily, without computing the elements of the matrices. More precisely, the theorems do not state anything explicitly about the choice of the mesh and the choice of the parameters θ and Δt (which define the elements of the matrices $\mathbf{M}$ and $\mathbf{K}$) in order to guarantee the qualitative properties. Our aim is to find conditions that can be checked a priori and imply the DNP / DMP. Then, these conditions, according to Theorem 5.2, imply the qualitative adequateness of the discrete models. In this section, we will give some a priori conditions. In order to present a complete work for researchers involved in scientific computing, we also added two corresponding results from the literature.

THEOREM 6.1. *The Galerkin finite element solution of problem (2.4)–(2.5), combined with the θ-method in the time discretization, satisfies the DMP if the conditions*

$$\begin{array}{ll}(S1) & K_{ij} \leq 0, \quad i \neq j,\ i=1,...,N,\ j=1,...,\bar{N},\\ (S2) & A_{ij} = M_{ij} + \theta\Delta t K_{ij} \leq 0, \quad i \neq j,\ i=1,...,N,\ j=1,...,\bar{N},\\ (S3) & B_{ii} = M_{ii} - (1-\theta)\Delta t K_{ii} \geq 0, \quad i=1,...,N,\end{array}$$

are fulfilled. The DMP is valid for the finite difference solution of the problem if $(S3)$ holds.

Proof. We have to show that under the assumptions of the theorem the properties $(P1)-(P2)$ hold. For the finite difference method $\mathbf{A}_0^{-1} \geq \mathbf{0}$ and conditions $(S1)$ and $(S2)$ are fulfilled automatically. This implies property $(P2)$. Condition $(S3)$ implies that the main diagonal of $\mathbf{B}$ is nonnegative. The off-diagonal of $\mathbf{B}$ is also nonnegative because the off-diagonal of $\mathbf{K}$ is non-positive. Hence, by virtue of $\mathbf{A}_0^{-1} \geq \mathbf{0}$, we obtain property $(P1)$.

In the case of the finite element method, relations $(S1)$ and $(S3)$ yield $\mathbf{B} \geq \mathbf{0}$. Condition $\mathbf{A}_\partial \leq \mathbf{0}$ follows from $(S2)$. $\mathbf{A}_0$ is a Gram-matrix, and because of $(S2)$, the off-diagonal elements of $\mathbf{A}_0$ are non-positive. This implies that $\mathbf{A}_0$ is an M-matrix (see [1, p. 134]), thus $\mathbf{A}_0^{-1} \geq \mathbf{0}$. These facts yield $(P1)$ and $(P2)$, thus the DMP. □

In the expressions of this section, for the sake of simplicity, fractions with zero denominators are understood as infinity.

6.1. Conditions for the finite difference method. Let us consider the finite difference method, where, because of the relations (3.5)-(3.7), condition ($S3$) is valid for one-dimensional problems if

$$C_i - (1-\theta)\Delta t\left(\frac{2\kappa_{i+x/2}}{(h_{i+x}+h_{i-x})h_{i+x}} + \frac{2\kappa_{i-x/2}}{(h_{i+x}+h_{i-x})h_{i-x}}\right) \geq 0, \quad i = 1,\ldots,N.$$

The condition

$$\Delta t \leq \frac{C_{\min} h_{\min}^2}{2\kappa_{\max}(1-\theta)}$$

guarantees the validity of the above relations, where $h_{\min}$ is the minimal spatial step size in the mesh and $\kappa_{\max} = \sup_{\mathbf{x}\in\Omega}\{\kappa(\mathbf{x})\}$. In 2D, similarly, we obtain the sufficient condition

$$\Delta t \leq \frac{C_{\min} h_{\min}^2}{4\kappa_{\max}(1-\theta)},$$

and in 3D

$$\Delta t \leq \frac{C_{\min} h_{\min}^2}{6\kappa_{\max}(1-\theta)}.$$

The above results can be summarized as follows.

THEOREM 6.2. *The DMP is always satisfied for the implicit Euler ($\theta = 1$) finite difference method. For other finite difference methods, the DMP can be guaranteed by the condition*

$$\Delta t \leq \frac{C_{\min} h_{\min}^2}{2d\kappa_{\max}(1-\theta)} \quad (\theta \neq 1)$$

(d is the dimension of Ω).

Remark 6.3. For the explicit Euler method, Theorem 6.2 yields the condition

$$\Delta t \leq \frac{C_{\min} h_{\min}^2}{2d\kappa_{\max}},$$

and for the Crank-Nicolson method we obtain

$$\Delta t \leq \frac{C_{\min} h_{\min}^2}{d\kappa_{\max}}.$$

We remark that for 1D problems on uniform meshes of step-size h, with constant material parameters (C_0, κ_0), the necessary and sufficient condition of the DMP is ([6])

$$\Delta t \leq \frac{C_0 h^2}{2\kappa_0(1-\theta)}.$$

6.2. Conditions for the finite element method. For finite element methods, condition ($S1$) is generally guaranteed by some geometrical requirements for the mesh. Moreover, conditions ($S2$) and ($S3$) can be achieved by some lower and upper bounds for the time-step.

6.2.1. Finite element method in 1D. Computing the elements of the stiffness matrix, we can check that the off-diagonal elements are non-positive, thus condition $(S1)$ is satisfied. Computing additionally the elements of the mass matrix we obtain that $(S2)$ and $(S3)$ are valid if

$$\frac{C_{\max} h_{\max}^2}{6\theta\kappa_{\min}} \le \Delta t \le \frac{C_{\min} h_{\min}^2}{3(1-\theta)\kappa_{\max}}.$$

Thus we obtain the next theorem.

THEOREM 6.4. *The Galerkin finite element solution of the one-dimensional problem (2.4)–(2.5) with piecewise linear elements satisfies the DMP if*

$$\frac{C_{\max} h_{\max}^2}{6\theta\kappa_{\min}} \le \Delta t \le \frac{C_{\min} h_{\min}^2}{3(1-\theta)\kappa_{\max}}, \quad \theta \in [\theta_l, 1)\,,$$

$$\frac{C_{\max} h_{\max}^2}{6\kappa_{\min}} \le \Delta t, \quad \theta = 1,$$

where $\theta_l = C_{\max}\kappa_{\max}h_{\max}^2/(C_{\max}\kappa_{\max}h_{\max}^2 + 2C_{\min}\kappa_{\min}h_{\min}^2)$.

Let us notice that unlike in the case of finite difference methods, the above theorem yields a lower bound for the time-step. This is a general phenomenon for finite element methods. This lower bound exists not only in sufficient conditions but in necessary and sufficient conditions, too. This is shown by the next theorem obtained in [6] for the DNP.

THEOREM 6.5. *The Galerkin finite element solution of the one-dimensional problem (2.4)–(2.5) with piecewise linear elements on uniform grid and with constant material parameters satisfies the DMP if and only if the relations*

$$\frac{C_0 h^2}{6\theta\kappa_0} \le \Delta t \le \frac{C_0 h^2}{3(1-\theta)\kappa_0}, \quad \theta \in [1/3, 1)\,,$$

$$\frac{C_0 h^2}{6\kappa_0} \le \Delta t, \quad \theta = 1$$

hold.

6.2.2. Finite element method in 2D. In 2D we apply the hybrid mesh in the spatial discretization. On triangular elements linear and on rectangular elements bilinear basis functions are used.

Let us consider a hybrid mesh $\mathcal{T}_h$. Let us define the number

$$\sigma_T = \min\{\cot\alpha_1, \cot\alpha_2, \cot\alpha_3\}$$

for each triangle T of the mesh, where α_1, α_2 and α_3 denote the angles of the triangle. Let us define $\sigma = \min_{T\in\mathcal{T}_h}\sigma_T$. Furthermore, let us introduce the value

$$\mu_R = \frac{2\min^2\{a,b\} - \max^2\{a,b\}}{ab}$$

for each rectangle R, where a and b denote the edges of the rectangle. Let us define $\mu = \min_{R\in\mathcal{T}_h}\mu_R$. The hybrid mesh $\mathcal{T}_h$ is called to be of strictly compact type if $\sigma > 0$

and $\mu > 0$. The condition $\sigma > 0$ means that each angle of the triangles in the mesh is of acute type, and $\mu > 0$ means that the longer edges of the rectangles are not greater than $\sqrt{2}$ times the shorter ones. The main result of [8] is formulated as follows.

THEOREM 6.6. *The Galerkin finite element solution of the two-dimensional problem (2.4)–(2.5) using linear/bilinear basis functions on a hybrid mesh of strictly compact type satisfies the DMP if*

$$\frac{C_{\max}}{6\,\theta\,\kappa_{\min}\,\min\{\frac{\mu}{3\lambda^{\#}_{\max}},\frac{\sigma}{\lambda^{\triangle}_{\max}}\}} \le \Delta t \le \frac{C_{\min}}{9\,(1-\theta)\,\kappa_{\max}\,\max\{\frac{\gamma^{\#}_{\max}}{3\lambda^{\#}_{\min}},\frac{\gamma^{\triangle}_{\max}}{4\,\lambda^{\triangle}_{\min}}\}} \tag{6.1}$$

is fulfilled, where

$$\gamma^{\triangle}_{\max} = \max_{T\in\mathcal{T}_h}\left\{\frac{l^2_{\max}}{\text{area}(T)}\right\},\quad \gamma^{\#}_{max} = \max_{R\in\mathcal{T}_h}\left\{\frac{a^2+b^2}{\text{area}(R)}\right\},$$

$$\lambda^{\triangle}_{\min} = \min_{T\in\mathcal{T}_h}\text{area}(T),\ \ \lambda^{\triangle}_{\max} = \max_{T\in\mathcal{T}_h}\text{area}(T),$$

$$\lambda^{\#}_{\min} = \min_{R\in\mathcal{T}_h}\text{area}(R),\ \ \lambda^{\#}_{\max} = \max_{R\in\mathcal{T}_h}\text{area}(R).$$

The symbols area(T) *and* area(R) *denote the area of the triangle T and the rectangle R, respectively. l_{max} is the length of the longest edge in T.*

Thus, for strictly compact meshes, the off-diagonal elements of the stiffness matrix are non-positive, hence condition $(S1)$ is fulfilled. Relation (6.1) implies the conditions $(S2)$ and $(S3)$.

Remark 6.7. For purely rectangular meshes, we have the weaker lower bound for Δt in the form

$$\Delta t \ge \frac{C_{\max}\lambda^{\#}_{max}}{3\,\theta\,\kappa_{\min}\,\mu}.$$

For a square mesh with step size h and with constant material parameters a sufficient condition of the DMP is

$$\frac{C_0 h^2}{3\theta\kappa_0} \le \Delta t \le \frac{C_0 h^2}{6(1-\theta)\kappa_0},\quad \theta \in [2/3, 1).$$

(In case of $\theta = 1$ the upper bound disappears.) This shows that in this case the DMP can be guaranteed only (with our sufficient condition) for methods with $\theta \ge 2/3$.

6.2.3. Finite element method in 3D. In 3D we investigate two different meshes. The first one is the rectangular and the second one is the tetrahedral mesh. Let us start with the first one, where trilinear basis functions are applied. With simple but tedious calculations we can compute the elements of the mass and stiffness matrices, and we can notice that the condition $(S1)$ can be valid only for uniform meshes. Moreover, we can observe that condition $(S2)$ cannot be guaranteed, because there are also positive off-diagonal elements in the matrix $\mathbf{A}$. Thus $\mathbf{A}_0$ is not an M-matrix in this case. We have to guarantee the nonnegative matrix inverse by means of other tools. In work [12], the so-called Lorenz-criterion ([21]) was applied.

Theorem 6.8. *The Galerkin finite element solution of the three-dimensional problem (2.4)–(2.5) using trilinear basis functions on a uniform rectangular mesh with constant material parameters satisfies the DMP if*

$$\frac{(259+13\sqrt{409})C_0h^2}{36\theta\kappa_0} \le \Delta t \le \frac{C_0h^2}{9(1-\theta)\kappa_0}, \quad \theta \in [0.9924, 1),$$

$$14.4975\frac{C_0h^2}{\kappa_0} \approx \frac{(259+13\sqrt{409})C_0h^2}{36\kappa_0} \le \Delta t, \quad \theta = 1.$$

For tetrahedral meshes the DMP was discussed in paper [11]. We will recall this result. Let us consider a tetrahedral mesh $\mathcal{T}_h$ of the solution domain Ω, and define the number σ_T for each tetrahedron T of $\mathcal{T}_h$ as $\sigma_T = \min\{\cos\beta_1, \dots, \cos\beta_6\}$, where $\beta_1, \dots, \beta_6$ denote the angles made by any two faces of the tetrahedron T. If each angle is less than $\pi/2$, then $\sigma_T > 0$. Let us set $\sigma = \min_{T\in\mathcal{T}_h} \sigma_T$. If $\sigma > 0$, then the mesh is called strictly acute type.

Theorem 6.9. *The Galerkin finite element solution of the three-dimensional problem (2.4)–(2.5) with constant material parameters using linear basis functions on a strictly acute type tetrahedral mesh satisfies the DMP if the condition*

$$\frac{C_0p_{\max}^2}{20\sigma\theta\kappa_0} \le \Delta t \le \frac{C_0p_{\min}^2}{10(1-\theta)\kappa_0}$$

holds, where $p_{\min}$ and $p_{\max}$ denote the minimal and maximal perpendicular length in the mesh.

7. Numerical Examples.

In this section, we demonstrate the validity of our results on some numerical examples in one, two and three dimensions. For the sake of simplicity, we suppose that the problem (2.4)-(2.5) describes a heat conduction process (see Remark 2.7). The material parameters are set to be constant one. Furthermore, we suppose that there are no heat sources or sinks present in the computational domain.

7.1. Examples in 1D.

In the first example, the 1D heat equation is solved with the Crank-Nicolson ($\theta = 1/2$) finite difference method on the interval [0,1]. We are interested in the approximation of the temperature at the time-level $t = 1$. We apply an equidistant mesh with the spatial step-size $h = 1/30$. The temperatures at the left and right boundaries are given by the functions $\mu_1(t) = 1$ and $\mu_2(t) = 0$, respectively. A nonnegative approximation of a continuous and nonnegative initial function can be seen in Figure 7.1. Using this initial grid-function, the approximations of the temperature at the time-level $t = 1$ can be seen in Figure 7.2. The figures on the left-hand and right-hand sides were obtained using the time-steps $\Delta t = 1/11$ and $\Delta t = 1/31$, respectively. These approximations show that neither the nonnegativity preservation nor the maximum-minimum principle is valid for the above chosen time-steps. In order to get a qualitatively correct approximation, Δt has to satisfy the condition

$$\Delta t \le \frac{h^2}{2(1-\theta)} = \frac{1}{900}$$

(c.f. equation (6.3)). This bound shows that Δt was chosen to be too large in the previous numerical example. Naturally, the above upper bound is the necessary and

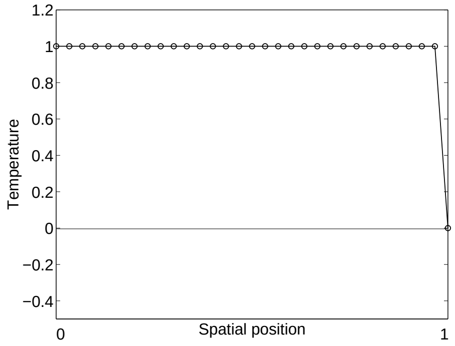

Fig. 7.1. Approximation of the initial function.

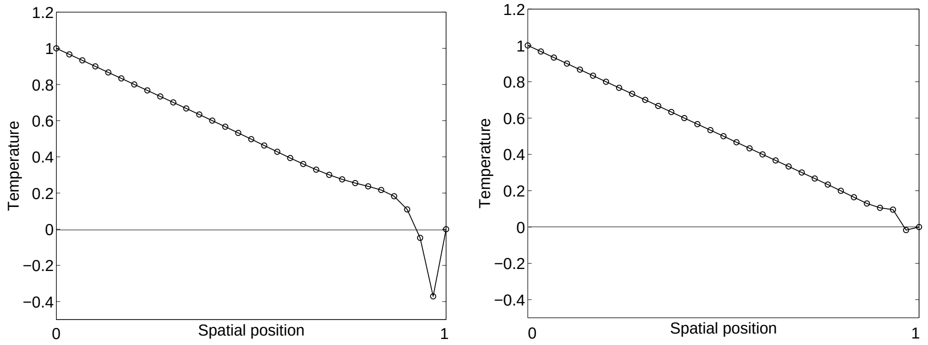

Fig. 7.2. Approximations at the time-level $t = 1$ with the time-steps $\Delta t = 1/11$ and $\Delta t = 1/31$.

sufficient condition of the DMP (DNP) for all initial functions. This means that for certain initial functions, larger time steps can also result in a qualitatively correct numerical solution. For example, Figure 7.3 shows a nonnegative approximation with the time-step $\Delta t = 1/51\,1/900$ at the time-level $t = 1$.

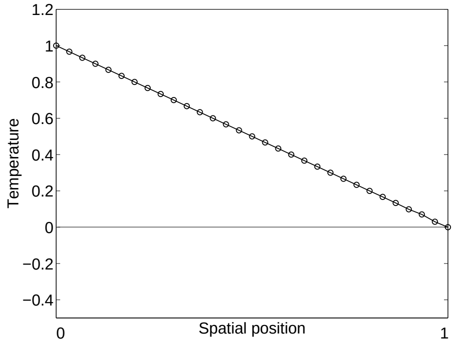

Fig. 7.3. Approximation at the time-level $t = 1$ with the time-step $\Delta t = 1/51$.

One can think that choosing the time-step to be sufficiently small we can obtain a qualitatively adequate solution. The next example shows that this is generally not the case. Namely, as we have shown, for finite element methods not only upper but lower bounds do exist.

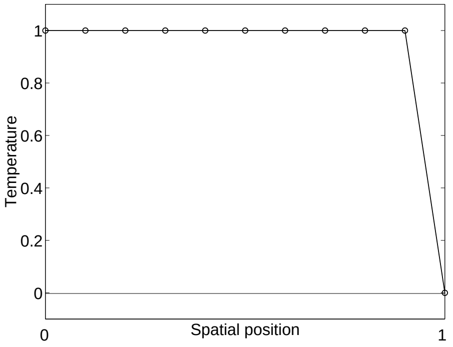

FIG. 7.4. Approximation of the initial function.

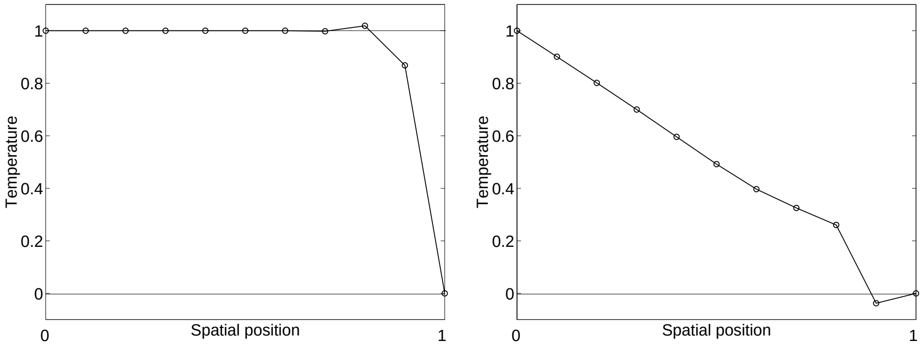

FIG. 7.5. Approximations at the time-levels $t = 1/1000$ and $t = 1$ applying the time-steps $\Delta t = 1/100000$ and $\Delta t = 1/9$, respectively.

Let us solve the 1D heat equation with the Galerkin finite element method on the interval [0,1] using the Crank-Nicolson ($\theta = 1/2$) scheme in the time integration. Let us suppose that the initial function is a nonnegative continuous function and it takes on values only from the interval $[0, 1]$. The discretization of the initial function is carried out on an equidistant mesh with $h = 1/10$ (Figure 7.4). The temperatures at the left and right boundaries are given by the functions $\mu_1(t) = 1$ and $\mu_2(t) = 0$, respectively. The numerical solutions calculated at the time-levels $t = 1/1000$ and $t = 1$ with the time steps $\Delta t = 1/100000$ and $\Delta t = 1/9$ can be seen, respectively, on the left-hand and right-hand sides of Figure 7.5. Both numerical solutions are qualitatively incorrect. Namely, the maximum-minimum principle is broken because the solutions have nonnegative values and values that are greater than one. In order to obtain qualitatively adequate solutions we can apply Theorem 6.4, which states that the maximum-minimum principle is valid providing that Δt is chosen according to the relation $1/300 = 0.0033 \leq \Delta t \leq 0.0067 = 2/300$. In Figure 7.6, the numerical solution is calculated at the time-level $t = 1$ with the time-step $\Delta t = 0.005$, which falls into the above interval. As it was expected, the solution has values only from the interval $[0, 1]$.

In the third example we solve again the 1D heat equation on the interval [0,1]. The temperature at the right-hand side is equal to one ($\mu_2(t) = 1$), while the temperature

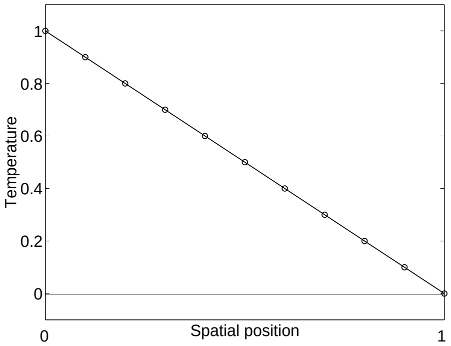

FIG. 7.6. Approximation at the time-level $t = 1$ with the time-step $\Delta t = 0.005$.

at the left-hand side changes periodically according to the function

$$\mu_1(t) = 2|-\mathrm{frac}(100t/2) + 0.5|$$

($\mathrm{frac}(100t/2)$ means the fractional part of $100t/2$). Naturally, $0 \le \mu_1(t) \le 1$. Let the initial function be the constant one function in $(0, 1)$.

Our aim is to approximate the solution at the time-level $t = 10$. We apply the finite difference method on the equidistant mesh with $h = 1/30$. We choose the Crank-Nicolson method for the time-integration with $\Delta t = 1/30$. The numerical solution is depicted on the left-hand side of Figure 7.7. Two time-levels further, that is at $t = 302/30$, we obtain the solution shown on the right-hand side of the same figure. In the case of the numerical solutions shown in Figure 7.7, the maximum-minimum principle is broken. The time-step was chosen to be too large (c.f. (6.3)). We can observe that the numerical solution will be periodic after the transient period ($t > 3$) (the boundary condition is periodic and the effect of the initial function disappears). The approximations shown in Figure 7.7 appear regularly in every third time-steps. Thus the numerical example demonstrates that it is possible that the qualitatively bad behavior of the numerical solution does not disappear while increasing the time-level at which the numerical solution is computed.

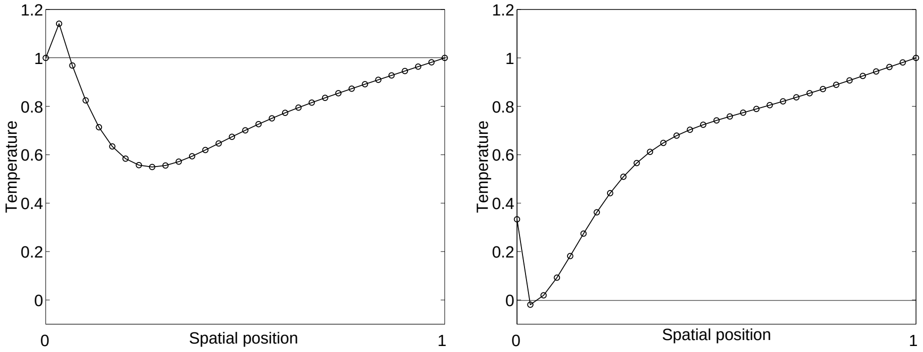

FIG. 7.7. Approximations at the time-levels $t = 300/30$ and $t = 302/30$ applying the time-step $\Delta t = 1/30$.

7.2. Examples in 2D. We solve the two-dimensional heat equation on the unit square $[0,1]\times[0,1]$ with homogeneous boundary condition. We apply the finite element method with bilinear elements on a rectangular mesh with $\Delta x = 1/10$ and $\Delta y = 1/12$. In the time-discretization the θ-method is used with $\theta = 0.9$ and with a fixed time-step Δt. A nonnegative discretization of a nonnegative and continuous initial function can be seen on the left-hand side of Figure 7.8. With this discretization, however, the approximation of the temperature at the first time-level has negative value both for $\Delta t = 0.0005$ and $\Delta t = 0.5$ (see the left-hand and right-hand sides in Figure 7.9). We remark that, with $\Delta t = 0.0005$, the numerical solution has negative components at the first 11 time-levels, while with $\Delta t = 0.5$, the solution has negative components at every time-level. Choosing the time-step to be $\Delta t = 0.01$, we obtain an adequate approximation (right-hand side of Figure 7.8).

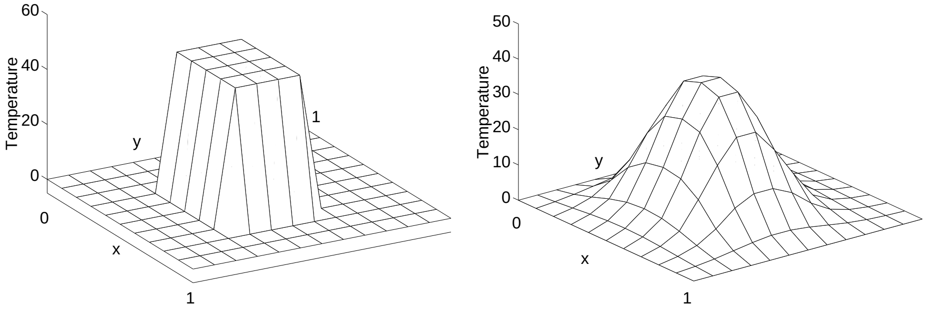

FIG. 7.8. Approximation of the initial function. Approximation at the first time-level with $\Delta t = 0.01$.

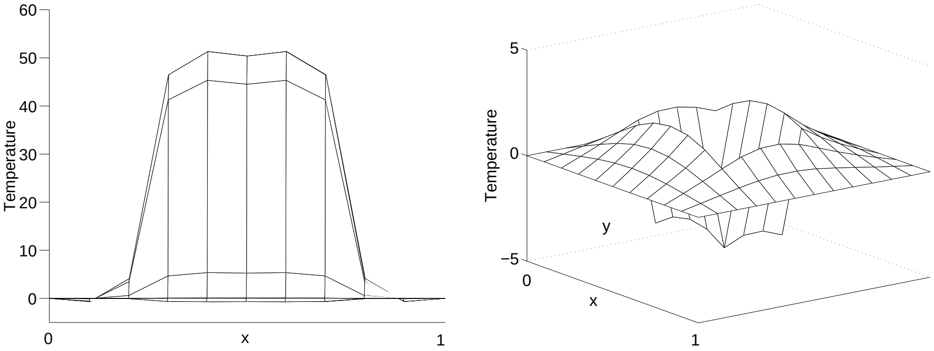

FIG. 7.9. Approximations at the first time-level with $\Delta t = 0.0005$ and $\Delta t = 0.5$.

We can calculate the necessary and sufficient condition of the DNP using Theorem 4.2, which results in the condition $0.0050 \le \Delta t \le 0.0155$. This shows that the chosen time-steps were too small or too large. A sufficient condition obtained applying the results of Theorem 6.6 is $0.0066 \le \Delta t \le 0.0136$.

In the second example, we apply the finite difference method with $\theta = 0.9$ for the initial approximation seen on the left-hand side of Figure 7.8. The results at the first time-levels with $\Delta t = 0.5$ and $\Delta t = 0.01$ can be seen, respectively, on the left-hand and right-hand sides of Figure 7.10. In order to obtain a qualitatively adequate numerical approximation, it is enough to choose Δt to be not greater than 0.017 (Theorem 6.2).

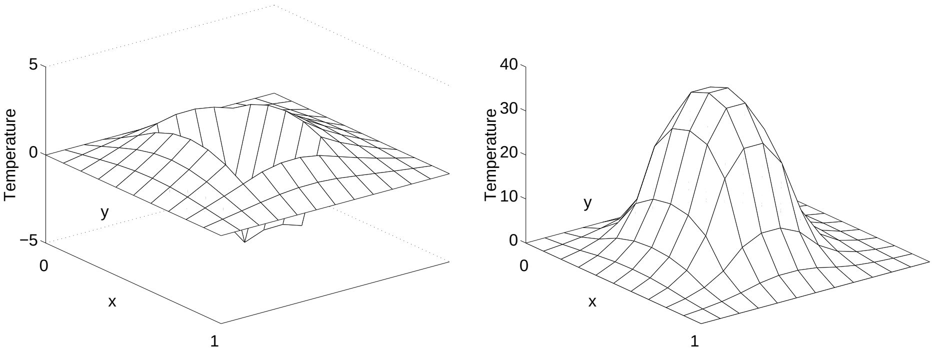

FIG. 7.10. Approximations at the first time-level with $\Delta t = 0.5$ and $\Delta t = 0.01$.

7.3. Example in 3D. In order to give a 3D example, we solve the homogeneous heat equation on the unit cube. We apply the finite difference method with the Crank-Nicolson time-integration. The spatial discretization is performed with an equidistant mesh, where $\Delta x = \Delta y = \Delta z = 1/10$. The approximation of a continuous nonnegative initial function is zero in every grid point excepting 27 grid points in the middle of the region, where the temperature is approximated by 50. The approximations of the temperature at the point $(5/10, 4/10, 4/10)$ using the time-steps $\Delta t = 0.05$ and $\Delta t = 0.01$ can be seen in Figure 7.11. The time-step $\Delta t = 0.05$ results in negative temperatures, which contradicts to the nonnegativity preservation property. Theorem 6.2 gives a sufficient condition of the nonnegativity, namely Δt must be not greater than 1/300.

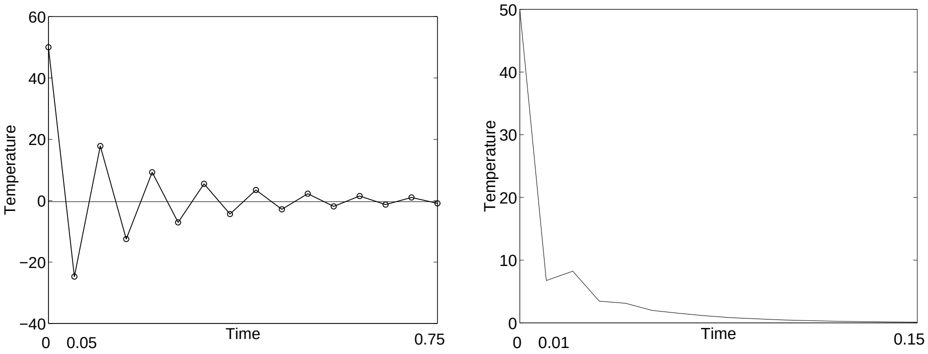

FIG. 7.11. Approximations of the temperature at the point $(5/10, 4/10, 4/10)$ using the time-steps $\Delta t = 0.05$ and $\Delta t = 0.01$.

8. Conclusions. In this paper, we have shad light on the connections between the main qualitative properties of parabolic partial differential equations and their numerical solution methods. We have found that the maximum-minimum principle, the nonnegativity preservation and the maximum norm contractivity are equivalent properties for the continuous problem, but this is not the case when we consider the finite difference or Galerkin finite element solutions of the problem. In this case, the maximum-minimum principle is equivalent to the nonnegativity preservation, and the maximum norm contractivity yields a weaker condition as this property is implied by the properties mentioned earlier. Thus, to achieve a qualitatively adequate method, we have to apply a nonnegativity preserving method, which can be guaranteed by the sufficient conditions of Theorem 6.1 and by its special cases listed in §6.

Acknowledgments. The authors thank the referees for reading the manuscript thoroughly. All of their suggestions, especially the ones regarding the inclusion of demonstrative numerical tests, contributed to the improvement of the paper.

REFERENCES

[1] A. Berman, R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York 1979.

[2] V. S. Borisov, *On Discrete Maximum Principles for Linear Equation Systems and Monotonicity of Difference Schemes*, SIAM J. Matrix Anal. Appl. 24 (2003), 1110–1135.

[3] V. S. Borisov, S. Sorek, *On the Monotonicity of Difference Schemes for Computational Physics*, SIAM J. Sci. Comput. 25 (2004), 1557–1584.

[4] P. G. Ciarlet, *Discrete Maximum Principle for Finite Difference Operators*, Aequationes Math. 4 (1970) 338–352.

[5] P. G. Ciarlet, P. A. Raviart, *Maximum Principle and Uniform Convergence for the Finite Element Method*, Comput. Methods Appl. Mech. Engrg. 2 (1973) 17–31.

[6] I. Faragó, *Nonnegativity of the Difference Schemes*, Pure Math. Appl. 6 (1996), 38–56.

[7] I. Faragó, R. Horváth, *On the Nonnegativity Conservation of Finite Element Solutions of Parabolic Problems. Proc. Conf. Finite Element Methods: Three-Dimensional Problems* (eds. P. Neittaanmäki, M. Krizek), GAKUTO Internat. Series Math. Sci. Appl., 15, Gakkotosho, Tokyo, 2001, 76-84.

[8] I. Faragó, R. Horváth, S. Korotov, *Discrete Maximum Principle for Linear Parabolic Problems Solved on Hybrid Meshes*, Appl. Num. Math. 53 (2005), 249-264.

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

[10] A. Friedmann, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Inc. Englewood Cliffs, N.J. 1964.

[11] H. Fujii, *Some Remarks on Finite Element Analysis of Time-Dependent Field Problems*, Theory and Practice in Finite element Structural Analysis, Univ. Tokyo Press, Tokyo (1973) 91–106.

[12] Hariton A. H., *Some Qualitative Properties of the Numerical Solution to the Heat Conduction Equation*, Thesis for Cand. of Math. Science, Eötvös Loránd University Budapest, 1995.

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

[14] R. Horváth, *On the Sign-Stability of the Numerical Solution of the Heat Equation*, Pure Math. Appl. 11 (2000), 281-291.

[15] R. Horváth, *Some Integral Properties of the Heat Equation*, J. Comp. Math. Appl. 42 (2001), 1135-1141.

[16] R. Horváth, *On the Monotonicity Conservation in Numerical Solutions of the Heat Equation*, Appl. Num. Math. 42 (2002), 189-199.

[17] J. Karátson, S. Korotov, *Discrete Maximum Principles in Finite Element Solutions of Nonlinear Problems with Mixed Boundary Conditions*, Numer. Math. 99 (2005), 669-698.

[18] S. Korotov, M. Krizek, P. Neittaanmäki, *Weakened Acute Type Condition for Tetrahedral Triangulations and the Discrete Maximum Principle*, Math. Comp. 70 (2001) 107–119.

[19] J.F.B.M. Kraaijevanger, *Maximum Norm Contractivity of Discretization Schemes for the Heat Equation*, Appl. Num. Math. 9 (1992), 475-492.

[20] O.A. Ladyženskaja, V.A. Solonnikov, N.N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Translations of Mathematical Monographs, Vol. 23.

[21] J. Lorenz, *Zur Inverzmonotonie discreter Probleme*, Numer. Math. 27 (1976/77), 227-238.

[22] T. Pfeil, *On the Monotonicity in Time of the Solutions of Linear Second Order Homogeneous Parabolic Equations*, Ann. Univ. Sci. Budapest. Eötvös Sect. Math. 36 (1993), 139-146.

[23] V. Ruas Santos, *On the Strong Maximum Principle for Some Piecewise Linear Finite Element Approximate Problems of Non-Positive Type*, J. Fac. Sci. Univ. Tokyo Sect. IA Math. 29 (1982), 473–491.

[24] T. Vejchodsky, *On the Nonnegativity Conservation in Semidiscrete Parabolic Problems*, in: M. Krizek at all eds; Conjugate Gradient Algorithms and Finite Element Methods, Springer Verlag, Berlin, 2004, 282-295.