Maximum Principles in the Theory of Numerical Methods

Ph.D. Dissertation

Miklós Emil Mincsovics

2014

Maximum Principles in the Theory of Numerical Methods

Miklós Emil Mincsovics

Ph.D. Dissertation

Supervisor: Prof. István Faragó, DHAS

Eötvös Loránd University, Faculty of Science
Ph.D. School for Mathematics, Applied Mathematics Program

School Leader: Prof. Miklós Laczkovich, MHAS
Program Leader: Prof. György Michaletzky, DHAS

Department of Applied Analysis
and Computational Mathematics

2014

Contents

Introduction

This dissertation consists of two parts. The topic of the first part is the Lax theory of the numerical solution of linear and nonlinear equations, see Chapter 1. The second part deals with discrete elliptic and parabolic maximum principles, see Chapters 2–4. Chapter 5 is the Appendix, which contains the necessary basics we build upon.

Now, we introduce the two topics of the thesis consecutively.

Lax-type theorems were already used when the application of some numerical method was necessary in order to approximate the solution of linear or nonlinear equations. The first paper was [36, Lax and Richtmyer, 1956], which generalized the preceding theorems and brought them to an abstract level. It contained the Lax equivalence theorem, which was later reformulated for semigroups by the same author [35, Lax, 2002]. This famous theorem was formulated for linear initial value problems. The paper [44, Palencia and Sanz-Serna, 1985] gave a framework applicable both for initial value and boundary value problems.

The theory was generalized for the nonlinear case in many papers. The primary difference between these papers is in their stability definitions. See [33, Keller, 1975], [48, Stetter, 1973], [37, López-Marcos and Sanz-Serna, 1988] and [54, Trenogin, 1980]. [47, Samarskii, Matus, Vabishchevich, 2002] collected many different types of stability notions.

In some of the works the error (i.e., the distance between the solution and the numerical approximation) is measured in the space of the solution using interpolation, see, e.g., the results of Aubin in [53, Temam, 1977], while most of the earlier mentioned works made the comparison in the space of the approximate solution using restriction.

When we want to approximate the solution $\bar{u}$ of the equation $F(u) = 0$, where $\mathcal{X}$ and $\mathcal{Y}$ are normed spaces, $\mathcal{D} \subset \mathcal{X}$ and $F : \mathcal{D} \to \mathcal{Y}$ is a (nonlinear) operator, usually a numerical method is used. Section 1.1 addresses the general (nonlinear) case. This section is based on the paper [23, Faragó, Mincsovics, Fekete, 2012]. In Subsection 1.1.1 we gave the definitions of the notions "problem", "numerical method" and "discretization". The success of a numerical method can be measured by the notion

of convergence. Even this notion can be defined in different ways, using interpolation or restriction as we already mentioned. Our choice is using restriction, but we shortly investigate the pros and cons of the other possible choice, too.

The definition of convergence is theoretical since it contains the unknown solution $\bar{u}$. Lax's idea was to substitute this unverifiable notion with the notions of consistency and stability. In our framework we use the stability notion of Keller. Firstly, the idea works since stability and consistency together implies convergence, which was firstly proven by Stetter for the nonlinear case. Secondly, for the applications the following recipe works: it is sufficient to check consistency for a set of elements, which can be done in parallel, and it is enough to check stability "near to the solution". In Subsection 1.1.3 we formulate these results at an abstract level. In Subsection 1.1.4 we investigate the relation of the basic notions (consistency, stability and convergence) providing numerous examples.

Section 1.2 contains the linear part of the framework. Note that we use the name affine instead of linear since we formulated the problem otherwise. In this case stability and convergence are equivalent under the consistency assumption. This is the Lax equivalence theorem, which we present in the form given by Palencia and Sanz-Serna. We compare the basic notions of the linear (affine) case with the basic notions of the general case as well. Finally, in Subsection 1.2.2 we present examples showing how the framework can be applied for approximating the solutions of elliptic and parabolic PDE's.

The second part of my dissertation deals with discrete elliptic and parabolic maximum principles.

What is the relevance of the discrete maximum principles?

When choosing a numerical method to approximate the solution of a continuous mathematical problem, the first thing to consider is which method results in an good approximation from a quantitative point of view. This is investigated in the first part of the thesis. However, in most of the cases it is not enough. The original problem (which is usually some model of a phenomenon) possesses important qualitative properties and a natural requirement from the numerical solution is to preserve these qualitative properties. E.g., when we seek an approximation of the Laplace's equation where the boundary condition is defined to be nonnegative, then the solution is nonnegative, too, and a good approximation should be nonnegative as well. For linear elliptic and parabolic problems the main qualitative properties are the various maximum principles.

The first paper in which a discrete elliptic maximum principle was formulated is probably [56, Varga, 1966]. The definition of the discrete weak maximum principle

which is used today appeared first in [5, Ciarlet, 1970] (but it was named differently). While the discrete weak maximum principle was extensively investigated in the last decades, see, e.g., the works [25, Hannukainen, Korotov, Vejchodský, 2009], [57, Vejchodsky, 2011], the discrete strong maximum principles have not been thoroughly analysed. In [30, Ishihara, 1987] and in [34, Knabner-Angermann, 2003] a sufficient algebraic condition was given, while in [8, Draganescu, Dupont, Scott, 2005] the positivity of the discrete Green function was investigated (this is in a close relation with the discrete strong maximum principles) in a special case. However, a sufficient and necessary algebraic condition was missing.

The first paper on a discrete parabolic maximum principle was [32, Keller, 1960], and from the early years the paper [24] should also be mentioned. From the recent years the works [11, Faragó, 2008], [17, Faragó and Horváth, 2009] contain a detailed investigation of a whole family of discrete (and continuous) parabolic maximum principles.

Discrete maximum principles can be investigated at two levels. One is purely algebraic (and theoretical), the other is more related to application. Namely, for a certain continuous problem (which possesses some continuous maximum principle) some discretization is applied. Then the question is how we should choose the mesh and the parameters of the discretization to get a discrete problem which possesses the corresponding discrete maximum principle. This latter case is naturally dependent both on the problem and on the discretization. As a consequence, there are countless papers of this sort. In our work both types of investigation can be found, the purely algebraic, and the other when for a problem a certain discretization is applied.

We present a short introduction on elliptic and parabolic maximum principles in Chapter 2. We note that we define maximum principles for an operator and not for an equation. Chapters 3 and 4 contain our work on discrete elliptic and parabolic maximum principles, respectively.

In Section 3.1 and 4.1 we give an algebraic framework on discrete elliptic and discrete parabolic maximum principles, respectively. At the elliptic case we focused on the differences between the weak and strong discrete maximum principles, see Section 3.2.

In Section 3.3 we investigate some elliptic problem where an interior penalty discontinuous Galerkin method is applied as discretization. We give sufficient conditions on the discretization parameters and on the mesh fulfilling the most important discrete elliptic maximum principles.

In Section 4.2 we investigate a parabolic problem when some FEM + θ-method

discretization is used. We derive practical conditions under which the most important discrete parabolic maximum principles can be preserved.

In Section 4.3 we introduce a new notion, the discrete stabilization property (DSP), and we present our results on the relation of the DSP and the discrete elliptic and discrete parabolic maximum principles. These results explain the property that a non-adequate mesh can already hinder the fulfilment of discrete parabolic maximum principles.

Throughout the thesis we use the following convention. We give references next to every result, lemma or theorem, except if it is our result. In this latter case, we supply the references at the beginning of the chapter/section/subsection, which contains the result. In that chapter/section/subsection all of the results without reference are from the same work unless the result has not been published yet.

Chapter 1

Basic notions of numerical analysis

This chapter contains an introduction on the basic notions of numerical analysis, defining in an exact way the mostly intuitively used notions including the discretization and the numerical method. Their important properties (convergence, consistency and stability) are introduced and the relation of these properties is investigated in the nonlinear and in the linear case, respectively.

1.1 Nonlinear theory

We consider a general nonlinear equation in an abstract (Banach space) setting. We seek an approximate solution of this equation. The usual way to proceed is to discretize the problem obtaining a simpler equation which can be solved already. This is how we can get one approximate solution which is usually enough in practice.

However, from a theoretical point of view it is better to define the notion of discretization as it results in a sequence of simpler problems which will be called numerical method. The main aim is to guarantee the convergence of the approximate solutions to the exact solution of the original problem. However, the convergence is difficult to treat directly.

It will be shown that this notion can be guaranteed by two other notions: the consistency and the stability together ensure the convergence, see Theorem 1.1.24 and Theorem 1.1.36, and these two notions can be checked directly. In the linear case this result is well known as the Lax (or sometimes Lax-Richtmyer-Kantorovich) theorem, which states more, actually, see Section 1.2.

The necessity of these conditions is investigated by giving suitable examples that show that neither consistency, nor stability is necessary for the convergence, in general. (The linear theory is different from this viewpoint.) All the notions and the results on

these are illustrated by showing their meaning for the numerical solution of a Cauchy problem of ordinary differential equation by means of the explicit Euler method.

The section is based on the paper [23].

1.1.1 Introduction

When we describe some real-life phenomenon with a mathematical model, it results in a – usually nonlinear – problem of the form

$$F(u) = 0\,, \tag{1.1}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are assumed to be normed spaces, $\mathcal{D} \subset \mathcal{X}$ and $F : \mathcal{D} \to \mathcal{Y}$ is assumed to be a (nonlinear) operator. Moreover, it is assumed that there exists a unique solution, which will be denoted by $\bar{u}$.

However, we note that, for any concrete applied problems we must prove the existence of a unique $\bar{u} \in \mathcal{D}$. In most cases the proof is not constructive, c.f. [33].

Even if it is possible to solve directly, the realization of the solving process is very difficult or even impossible. However, in practice, we need only a good approximation for the solution of problem (1.1), since our model is usually already a simplification of a real-life phenomenon. Therefore we use some discretization, which results in a sequence of simpler problems, i.e., a numerical method, see Definition 1.1.3 and Definition 1.1.5 for the exact definition of these notions.

With this approach we need to face the following difficulties:

- we need to compare the solutions of the simpler problems with the solution of the original problem (1.1), which might be found in different spaces;
- naturally, this comparison seems to be impossible, since the solution of the original problem (1.1) is unknown.

To get rid of the latter difficulty, the usual trick is to introduce the notions of consistency and stability, which do not require the knowledge of the solution of the original problem (1.1) and can be verified. Thus, the convergence can be replaced with these two notions. Sometimes this popular "recipe" is summarized in the "formula"

$$\text{Consistency} + \text{Stability} \Rightarrow \text{Convergence}\,. \tag{1.2}$$

In the following we introduce and investigate these notions in an abstract framework, and we try to shed some light on the formula (1.2). Namely:

- how to define consistency and stability to ensure the formula (1.2);
- is it consistency or/and stability that is necessary for the convergence (in the linear case the Lax equivalence theorem deals with this question, too, see Section 1.2).

The following is mainly devoted to answer these questions. First, we start with some definitions and notations, by giving an example.

Definition 1.1.1. Problem (1.1) can be given as a triplet $\mathscr{P} = (\mathcal{X}, \mathcal{Y}, F)$. We will refer to it as *problem* $\mathscr{P}$.

Example 1.1.2. Consider the following initial value problem:

$$u'(t) = f(u(t)) \tag{1.3}$$

$$u(0) = u_0\,, \tag{1.4}$$

where $t \in [0,1]$, $u_0 \in \mathbb{R}$ and $f \in C(\mathbb{R}, \mathbb{R})$ is a Lipschitz continuous function.

The operator F and the spaces $\mathcal{X}, \mathcal{Y}$ are defined as follows.

- $\mathcal{X} = C^1[0,1], \quad \|u\|_{\mathcal{X}} = \max\limits_{t\in[0,1]} |u(t)|;$
- $\mathcal{Y} = C[0,1] \times \mathbb{R}, \quad \left\| \begin{pmatrix} u \\ u_0 \end{pmatrix} \right\|_{\mathcal{Y}} = \max\limits_{t\in[0,1]} (|u(t)|) + |u_0|;$
- $F(u) = \begin{pmatrix} u'(t) - f(u(t)) \\ u(0) - u_0 \end{pmatrix}.$

Definition 1.1.3. We say that the sequence $\mathscr{N} = (\mathcal{X}_n, \mathcal{Y}_n, F_n)_{n\in\mathbb{N}}$ is a *numerical method* if it generates a sequence of problems

$$F_n(u_n) = 0\,, \quad n = 1, 2, \dots\,, \tag{1.5}$$

where

- $\mathcal{X}_n, \mathcal{Y}_n$ are normed spaces;
- $\mathcal{D}_n \subset \mathcal{X}_n$ and $F_n : \mathcal{D}_n \to \mathcal{Y}_n\,.$

If there exists a unique solution of the (approximating) problems (1.5), it will be denoted by $\bar{u}_n\,.$

Example 1.1.4. For $n \in \mathbb{N}$ we define the following sequence of triplets:

- $\mathcal{X}_n = \mathbb{R}^{n+1}$, $\mathbf{v}_n = (v_0, v_1, \ldots, v_n) \in \mathcal{X}_n : \|\mathbf{v}_n\|_{\mathcal{X}_n} = \max\limits_{i=0,\ldots,n} |v_i|$;
- $\mathcal{Y}_n = \mathbb{R}^{n+1}$, $\mathbf{y}_n = (y_0, y_1, \ldots, y_n) \in \mathcal{Y}_n : \|\mathbf{y}_n\|_{\mathcal{Y}_n} = |y_0| + \max\limits_{i=1,\ldots,n} |y_i|$;
- $F_n : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$, and for any $\mathbf{v}_n = (v_0, v_1, \ldots, v_n) \in \mathbb{R}^{n+1}$ it acts as

$$(F_n(\mathbf{v}_n))_i = \begin{cases} n\,(v_i - v_{i-1}) - g\,(v_{i-1})\,, & i = 1, \ldots, n\,, \\ \\ v_0 - c\,, & i = 0\,. \end{cases} \tag{1.6}$$

Here $g : \mathbb{R} \to \mathbb{R}$ and $c \in \mathbb{R}$ are arbitrary given data and one can see that the defined numerical method is the explicit Euler method.

Definition 1.1.5. We say that the sequence $\mathscr{D} = (\varphi_n, \psi_n, \Phi_n)_{n\in\mathbb{N}}$ is a *discretization* if

- the φ_n-s (respectively ψ_n-s) are operators from $\mathcal{X}$ into $\mathcal{X}_n$ (respectively from $\mathcal{Y}$ into $\mathcal{Y}_n$), where $\mathcal{X}, \mathcal{X}_n, \mathcal{Y}, \mathcal{Y}_n$ are normed spaces;
- $\Phi_n : \{F : \mathcal{D} \to \mathcal{Y} \,|\, \mathcal{D} \subset \mathcal{X}\} \to \{F_n : \mathcal{D}_n \to \mathcal{Y}_n \,|\, \mathcal{D}_n \subset \mathcal{X}_n\}$.

Example 1.1.6. Based on Examples 1.1.2 and 1.1.4, in Definition 1.1.5 we define $\mathcal{X} = C^1[0,1]$, $\mathcal{Y} = C[0,1] \times \mathbb{R}$, and $\mathcal{X}_n = \mathcal{Y}_n = \mathbb{R}^{n+1}$. $\mathbb{G}_n := \{t_i = \frac{i}{n},\ i = 0, \ldots, n\}$. Then, we define the triplet of the operators as follows.

- For any $v \in \mathcal{X}$ we put $(\varphi_n v)_i = v(t_i)$, $i = 0, 1, \ldots, n$.
- For any $y \in \mathcal{Y}$ we put

$$(\psi_n y)_i = \begin{cases} y\,(t_{i-1})\,, & 1, \ldots, n\,, \\ \\ y(t_0), & i = 0\,. \end{cases}$$

- In order to give Φ_n, we define the mapping $\Phi_n : C^1[0,1] \to \mathbb{R}^{n+1}$ in the following way:

$$[(\Phi_n(F))\,v]_i = \begin{cases} n\,(v(t_i) - v(t_{i-1})) - g(v(t_{i-1}))\,, & i = 1, \ldots, n\,, \\ \\ v(t_0) - c\,, & i = 0\,. \end{cases} \tag{1.7}$$

We note that the introduced notions of problem $\mathscr{P}$ and numerical method $\mathscr{N}$ are independent of each other. However, for our purposes only those numerical methods $\mathscr{N}$ are interesting which are obtained when some discretization method $\mathscr{D}$ is applied to a certain problem $\mathscr{P}$. We introduce the notation $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ for the sentence "the discretization $\mathscr{D}$ is applied to problem $\mathscr{P}$ resulting in the numerical method $\mathscr{N}$". Thus, this notation denotes the whole process.

Remark 1.1.7. Theoretically, the normed spaces $\mathcal{X}$ and $\mathcal{Y}$ in the definitions of the problem and of the discretization might be different. However the application of the discretization to the problem is possible only when these normed spaces are the same. In the sequel this will be always assumed.

Example 1.1.8. Let us define the numerical method $\mathscr{N}$ for problem $\mathscr{P}$ from Example 1.1.2, and for the discretization $\mathscr{D}$ from Example 1.1.6. Then we solve the sequence of problems in the form (1.5), where in the discretization for g and c we put f and u_0 from problem (1.3)-(1.4), respectively. This yields that the mapping $F_n : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ is defined as follows: for the vector $\mathbf{v}_n = (v_0, v_1, \ldots, v_n) \in \mathbb{R}^{n+1}$ we have

$$(F_n(\mathbf{v}_n))_i = \begin{cases} n\,(v_i - v_{i-1}) - f(v_{i-1})\,, & i = 1, \ldots, n\,, \\ \\ v_0 - u_0\,, & i = 0\,. \end{cases} \tag{1.8}$$

Hence, using the notation $h = 1/n$, the equation (1.5) for (1.8) results in the task: we seek the vector $\mathbf{v}_n = (v_0, v_1, \ldots, v_n) \in \mathbb{R}^{n+1}$ such that

$$\begin{cases} \dfrac{v_i - v_{i-1}}{h} = f(v_{i-1})\,, & i = 1, \ldots, n\,, \\ \\ v_0 = u_0\,, & i = 0\,. \end{cases} \tag{1.9}$$

Hence, the obtained numerical method is the well-known explicit Euler method on the mesh $\mathbb{G}_n$ with uniform step-size h.

In the sequel the following assumption will be used.

Assumption 1.1.9. $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ possesses the following properties.

(a1) $\mathscr{N}$ possesses the property $\dim \mathcal{X}_n = \dim \mathcal{Y}_n < \infty$.

(a2) F_n is continuous on the ball $B_R(\varphi_n(\bar{u}))$ from some index.

(a3) $\psi_n(0) = 0$ holds from some index.

Obviously, when ψ_n are linear operators, then (a3) is automatically satisfied.

1.1.2 Basic notions and theoretical results

In this part we introduce the important properties (convergence, consistency and stability) related to the process $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$.

Convergence. Our aim is to guarantee the existence of the solutions $\bar{u}_n$ and its closeness to $\bar{u}$. We define the distance between these elements, which will be called global discretization error. Since these elements belong to different spaces, this is not straightforward.

There are two possible options for where to compare the solutions: in $\mathcal{X}$, which might appear the more natural at first sight, or in the spaces where the solutions of the simpler problems can be found, i.e. in the spaces $\mathcal{X}_n$. We choose this latter possibility, however, both possibilities will be investigated shortly, giving their pros and cons.

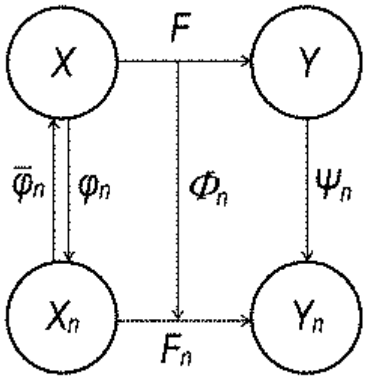

Figure 1.1: The general scheme of numerical methods with interpolation operators.

- It is possible to define the distance between the elements $\bar{u}$ and $\bar{u}_n$ in the space $\mathcal{X}$, with the help of (interpolation) operators $\bar{\varphi}_n : \mathcal{X}_n \to \mathcal{X}$, by the quantity $\|\bar{u} - \bar{\varphi}_n \bar{u}_n\|_{\mathcal{X}}$. For such an approach see Figure 1.1. In this approach the convergence means that the numerical sequence $\|\bar{u} - \bar{\varphi}_n \bar{u}_n\|_{\mathcal{X}}$ tends to zero.

  At first sight this approach seems to be more natural, however to deal with it on an abstract level is more difficult. The difficulty is that the convergence depends on two processes, on the numerical method and on the interpolation.

Example 1.1.10. Let us choose the numerical method so that we choose an arbitrary $\bar{u}_1$ from an arbitrary space $\mathcal{X}_1$ and $\bar{u}_n := \bar{u}_1$, $\mathcal{X}_n := \mathcal{X}_1$. We use the interpolation $\bar{\varphi}_n$ defined as $\bar{\varphi}_n(v_n) = \bar{u}$ for all $v_n \in \mathcal{X}_n$, for all n.

Then clearly, $\|\bar{u} - \bar{\varphi}_n \bar{u}_n\|_{\mathcal{X}}$ tends to zero.

This degenerate example shows that the whole process is convergent in spite of the fact that the numerical method is simply unacceptable. To avoid such cases, usually it is assumed that $\lim(\bar{\varphi}_n \circ \varphi_n)v = v$ for any $v \in \mathcal{X}$ (or some similar property). We note that this relation does not mean that $\bar{\varphi}_n$ is the inverse of φ_n, because φ_n is not invertible, typically it represents some interpolation.

However, on the basis of all this it seems to be more appropriate to handle the numerical method and the interpolation separately. This leads to the other approach.

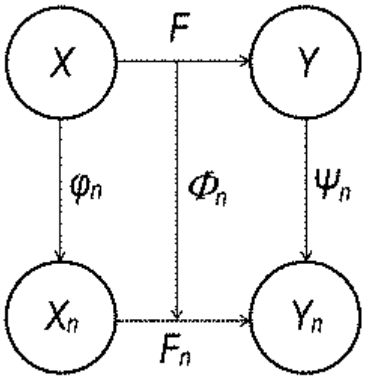

Figure 1.2: Second approach (which is our choice): The general scheme of numerical methods without interpolation.

- The general scheme of this approach is illustrated in Figure 1.2.

Definition 1.1.11. The element $e_n = \varphi_n(\bar{u}) - \bar{u}_n \in \mathcal{X}_n$ is called *global discretization error*.

Clearly, our aim is to guarantee that the global discretization error is arbitrarily small, by increasing n. That is, we require the following property.

Definition 1.1.12. $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is called *convergent* if

$$\lim \|e_n\|_{\mathcal{X}_n} = 0 \tag{1.10}$$

holds. When

$$\|e_n\|_{\mathcal{X}_n} = \mathcal{O}(n^{-p})$$

we say that the *order of the convergence* is p.

Thus, the whole process is split into two tasks, into the numerical method and into the interpolation. Naturally, for a convergent $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ it is much easier to find an appropriate interpolation.

Thus, this approach is more appropriate if the numerical method is in our focus (without the interpolation process) and this is the reason why we choose this one. However, it does not mean that the interpolation process (or the possibility of this process, which depends on the approximation capabilities of the space-sequence $(\mathcal{X}_n)_{n\in\mathbb{N}}$) is less important. To underline this statement the next example is shown.

Example 1.1.13. Let us choose the numerical method so that we choose an arbitrary $\bar{u}_1$ from an arbitrary space $\mathcal{X}_1$ and $\bar{u}_n := \bar{u}_1$, $\mathcal{X}_n := \mathcal{X}_1$ with the norm $\|\cdot\|_{\mathcal{X}_n} := \frac{1}{n}\|\cdot\|_{\mathcal{X}_1}$. Moreover, we choose an arbitrary φ_1 and $\varphi_n := \varphi_1$.

Then clearly, e_n tends to 0 thanks to the factor $1/n$.

On the other hand, nobody would call it a convergent numerical method. To avoid such an example, some kind of norm-consistency could be assumed, e.g., $\lim \|\varphi_n(v)\|_{\mathcal{X}_n} = \|v\|_{\mathcal{X}}$ for all $v \in \mathcal{X}$.

Independently of the form of the definition of the global error, it is hardly applicable in practice, because the knowledge of the exact solution $\bar{u}$ is assumed. Therefore, we introduce some further notions (consistency, stability), which help us in getting information about the behavior of the global discretization error.

Consistency. Consistency is the connecting link between the problem $\mathscr{P}$ and the numerical method $\mathscr{N}$.

Definition 1.1.14. $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is called *consistent at the element* $v \in \mathcal{D}$ if

- $\varphi_n(v) \in \mathcal{D}_n$ holds from some index,
- the relation

$$\lim \|F_n(\varphi_n(v)) - \psi_n(F(v))\|_{\mathcal{Y}_n} = 0 \tag{1.11}$$

holds.

$\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is called *consistent* if it is consistent at the element $\bar{u}$.

The element $l_n(v) = F_n(\varphi_n(v)) - \psi_n(F(v)) \in \mathcal{Y}_n$ in (1.11) plays an important role in the numerical analysis. When we fix some element $v \in \mathcal{D}$, we can transform it into the space $\mathcal{Y}_n$ in two different ways (with the help of the operators F, ψ_n and φ_n, F_n): $\mathcal{X} \to \mathcal{Y} \to \mathcal{Y}_n$ and $\mathcal{X} \to \mathcal{X}_n \to \mathcal{Y}_n$ (c.f. Figure 1.2). The magnitude $l_n(v)$ characterizes the difference of these two directions for the element v. Hence, the consistency at the element v yields that in limit the diagram of Figure 1.2 is commutative. A special role is played by the behaviour of $l_n(v)$ on the solution of the problem (1.1), that are the elements $l_n(\bar{u})$. Later on we will use the following notions.

Definition 1.1.15. The element $l_n(v) = F_n(\varphi_n(v)) - \psi_n(F(v)) \in \mathcal{Y}_n$ is called *local discretization error at the element* v. Assuming (a3) of Assumption 1.1.9, the element $l_n =: l_n(\bar{u}) = F_n(\varphi_n(\bar{u})) - \psi_n(F(\bar{u})) = F_n(\varphi_n(\bar{u}))$ is called *local discretization error*. When

$$\|l_n(v)\|_{\mathcal{X}_n} = \mathcal{O}(n^{-p}),$$

we say that the *order of the consistency at* v is p (analogously simply *order of the consistency* for $v = \bar{u}$).

One might ask whether consistency implies convergence. The following simple example shows that this is not true in general.

Example 1.1.16. Let us consider the case $\mathcal{X} = \mathcal{X}_n = \mathcal{Y} = \mathcal{Y}_n = \mathbb{R}$, $\varphi_n = \psi_n =$ *identity*. Our aim is to solve the scalar equation $F(x) = 0$, where we assume that it has a unique solution $\bar{x} = 0$. We define the numerical method as $F_n(x) = (1 - x)/n$. Clearly, due to the linearity of φ_n and ψ_n, we have $l_n = F_n(0) - 0 = F_n(0)$. Since $F_n(0) \to 0$, therefore this discretization is consistent. However, it is not convergent, since the solution of each problem $F_n(x) = 0$ is $\bar{x}_n = 1$.

Thus, convergence cannot be replaced by consistency in general.

Stability. As we have already seen, consistency in itself is not enough for convergence. Assuming the existence of the inverse operator F_n^{-1}, we can easily get to the relation

$$e_n = \varphi_n(\bar{u}) - \bar{u}_n = F_n^{-1}(F_n(\varphi_n(\bar{u}))) - F_n^{-1}(0) = F_n^{-1}(l_n) - F_n^{-1}(0)\,,$$

which shows the connection between the global and local discretization errors. This relation suggests that the consistency (i.e., the convergence of the local discretization error l_n to zero) can provide the convergence (i.e., the approach of e_n to zero) when $(F_n^{-1})_{n\in\mathbb{N}}$ has good behavior. Such a property is the Lipschitz continuity: it would be useful to assume that the functions F_n^{-1} uniformly satisfy the Lipschitz condition at

the points $0 \in \mathcal{Y}_n$. However, generally at this point we have no guarantee even to the existence of F_n^{-1}, thus we provide this with some property of the functions F_n, without assuming their invertibility. The first step in this direction is done by introducing a simplified form of the notion of semistability in [37].

Definition 1.1.17. $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is called *semistable* if there exist $S \in \mathbb{R}$, $R \in (0, \infty]$ such that

- $B_R(\varphi_n(\bar{u})) \subset \mathcal{D}_n$ holds from some index;
- $\forall (v_n)_{n\in\mathbb{N}}$ which satisfy $v_n \in B_R(\varphi_n(\bar{u}))$ from some index, the relation

$$\|\varphi_n(\bar{u}) - v_n\|_{\mathcal{X}_n} \leq S \|F_n(\varphi_n(\bar{u})) - F_n(v_n)\|_{\mathcal{Y}_n} \tag{1.12}$$

holds.

Semistability is a purely theoretical notion, which, similarly to the consistency, cannot be checked directly, due to the fact that $\bar{u}$ is unknown. However, the following statement clearly shows the relation of the three important notions.

Lemma 1.1.18. *We assume that $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is such that*

- *(a3) of Assumption 1.1.9 is satisfied;*
- *it is consistent and semistable with stability threshold R;*
- *equation* (1.5) *has a solution in $B_R(\varphi_n(\bar{u}))$ from some index.*

Then the sequence of these solutions of equation (1.5) *converges to the solution of problem $\mathscr{P}$, and the order of convergence is not less than the order of consistency.*

Proof. First, using the semistability gives

$$\|e_n\|_{\mathcal{X}_n} = \|\varphi_n(\bar{u}) - \bar{u}_n\|_{\mathcal{X}_n} \leq S \|F_n(\varphi_n(\bar{u})) - F_n(\bar{u}_n)\|_{\mathcal{Y}_n} = S \|F_n(\varphi_n(\bar{u}))\|_{\mathcal{Y}_n} = S\|l_n\|_{\mathcal{Y}_n}$$

from some index. Finally, using the consistency proves the statement. □

This lemma has some drawbacks. Firstly, we cannot verify its conditions because this requires the knowledge of the solution. Secondly, we have no guarantee that equation (1.5) has a (possibly unique) solution in $B_R(\varphi_n(\bar{u}))$ from some index. By using the following modified stability notion, see [33], we can get rid of the second problem.

Definition 1.1.19. $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is called *stable* at the element $v \in \mathcal{X}$ if there exist $S \in \mathbb{R}$, $R \in (0, \infty]$ such that

- for the stability neighbourhood $B_R(\varphi_n(v)) \subset \mathcal{D}_n$ holds from some index;
- $\forall (v_n^1)_{n\in\mathbb{N}}, (v_n^2)_{n\in\mathbb{N}}$ which satisfy $v_n^i \in B_R(\varphi_n(v))$, the estimate

$$\left\|v_n^1 - v_n^2\right\|_{\mathcal{X}_n} \leq S \left\|F_n(v_n^1) - F_n(v_n^2)\right\|_{\mathcal{Y}_n} \tag{1.13}$$

  holds.

$\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is called *stable* if it is stable at the element $\bar{u} \in \mathcal{X}$.

Remark 1.1.20. Obviously, stability implies semistability.

The immediate profit of this definition is injectivity as it is formulated in the next statement.

Corollary 1.1.21. *If $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is stable at the element $v \in \mathcal{X}$ with stability threshold R, then F_n is injective on $B_R(\varphi_n(v))$ from some index.*

The following statements demonstrate the usefulness of the stability notion, given in Definition 1.1.19. These results first appeared in [48], however, based on a different notion of stability, see the Paragraph "Notes on the notion of stability – other possibilities." of this dissertation. Thus, these results are converted in order to fit in our framework and are presented here in this converted form.

Lemma 1.1.22. *[48, Version of Lemma 1.2.1.]*
We assume that

- *$\mathcal{V}$, $\mathcal{W}$ are normed spaces with the property $\dim \mathcal{V} = \dim \mathcal{W} < \infty$;*
- *$G : B_R(v) \to \mathcal{W}$ is continuous for some $v \in \mathcal{V}$ and $R \in (0, \infty]$;*
- *for all v^1, v^2 which satisfy $v^i \in B_R(v)$, the stability estimate*

$$\left\|v^1 - v^2\right\|_{\mathcal{V}} \leq S \left\|G(v^1) - G(v^2)\right\|_{\mathcal{W}} \tag{1.14}$$

  holds.

Then

- *G is invertible, and $G^{-1} : B_{R/S}(G(v)) \to B_R(v)$;*
- *G^{-1} is Lipschitz continuous with the constant S.*

The proof of this lemma is rather technical, thus it is placed into the Appendix.

Lemma 1.1.23. *[48, Version of Theorem 1.2.3.]*
For $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ we assume that

- *(a1) and (a2) of Assumption 1.1.9 are satisfied;*
- *it is consistent and stable with stability threshold R and constant S.*

Then $\mathscr{D}(\mathscr{P})$ generates a numerical method $\mathscr{N}$ such that equation (1.5) *has a unique solution in $B_R(\varphi_n(\bar{u}))$ from some index.*

Proof. Due to Lemma 1.1.22, F_n is invertible, and $F_n^{-1} : B_{R/S}(F_n(\varphi_n(\bar{u}))) \to B_R(\varphi_n(\bar{u}))$. Note that $F_n(\varphi_n(\bar{u})) = l_n \to 0$, due to the consistency. This means that $0 \in B_{R/S}(F_n(\varphi_n(\bar{u})))$ holds from some index. This proves the statement. □

Hence, we can formulate the following theorem.

Theorem 1.1.24. *[48, Version of Theorem 1.2.4.]*
For $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ we assume that

- *(a1)–(a3) of Assumption 1.1.9 are true;*
- *it is consistent and stable with stability threshold R and constant S.*

Then $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is convergent, and the order of the convergence is not less than the order of consistency.

Proof. The statement is a consequence of Lemmas 1.1.23 and 1.1.18. □

Notes on the notion of stability – other possibilities. We complete this subsection with some remarks w.r.t. the stability notion by Definition 1.1.19.

There are other definitions for stability in the literature, these are mostly generalizations of the stability notion of Keller. We list some of them.

- The first one of them is the following one, which is given in [48].

  Definition 1.1.25. $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is called *stable in the sense of Stetter* if there exist $S \in \mathbb{R}$, $R \in (0, \infty]$ and $r \in (0, \infty]$ such that

  - $B_R(\varphi_n(\bar{u})) \subset \mathcal{D}_n$ holds from some index;
  - for all $(v_n^1)_{n\in\mathbb{N}}, (v_n^2)_{n\in\mathbb{N}}$ such that $v_n^i \in B_R(\varphi_n(\bar{u}))$, and the inclusion $F_n(v_n^i) \in B_r(F_n(\varphi_n(\bar{u})))$ is true, the estimate (1.13) holds.

Note that the stability notion by Stetter is less restrictive than the one given in Definition 1.1.19: if we put $r = \infty$ in Definition 1.1.25, then we re-obtain the stability definition by Keller, given in Definition 1.1.19.

- The second one was given in the paper [37].

  Definition 1.1.26. $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is called *stable in the sense of López-Marcos and Sanz-Serna* if there exist $S \in \mathbb{R}$ and $(R_n)_{n\in\mathbb{N}}$, $R_n \in (0, \infty]$ such that

  - $B_{R_n}(\varphi_n(\bar{u})) \subset \mathcal{D}_n$ holds from some index;
  - $\forall (v_n^1)_{n\in\mathbb{N}}, (v_n^2)_{n\in\mathbb{N}}$ which satisfy $v_n^i \in B_{R_n}(\varphi_n(\bar{u}))$ from that index, the estimate (1.13) holds.

  This stability notion allows us to vary the radius of the balls which could be necessary as it has been shown in [37], where an example is presented for which this is the appropriate notion, while the others fail.

- Finally we mention another generalization which was introduced in [54] (actually, here we present a version of it).

  Definition 1.1.27. $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is called *stable in Trenogin's sense* if there exist $S \in \mathbb{R}$ and $R \in (0, \infty]$ such that

  - $B_R(\varphi_n(\bar{u})) \subset \mathcal{D}_n$ holds from some index;
  - there exists a continuous at a neighbourhood of zero, strictly monotonically increasing function $\omega(t)$ defined on $t \geq 0$ such that $\omega(0) = 0$ and

    $$\omega\left(\left\|v_n^1 - v_n^2\right\|_{\mathcal{X}_n}\right) \leq \left\|F_n(v_n^1) - F_n(v_n^2)\right\|_{\mathcal{Y}_n} \tag{1.15}$$

    holds for all $v_n^1, v_n^2 \in B_R(\varphi_n(\bar{u}))$.

If we choose ω as *identity*$/L$, we re-obtain the Definition 1.1.19.

We mention that similarly to that definition of stability and the corresponding built-up we choose the whole construction can be carried through choosing the above mentioned stability definitions, too.

Naturally it is possible to construct further types of stability notions, e.g., mixing the above mentioned ones. But this would be fruitful only from a theoretical point of view, the real question is always that of how these could work in practice. Even the stability notion of Stetter and that of Trenogin's seem to be too theoretical until now.

1.1.3 Basic notions – revisited from the application point of view

Theorem 1.1.24 is not yet suitable for our purposes: the condition requires to check the stability and the consistency at the unknown element $v = \bar{u}$. Therefore, this statement is not applicable for real problems. Since we are able to verify the above properties on some set of points (sometimes on the entire $\mathcal{D}$), we convert the previously given framework into another one which fits more for the application and is based on global properties instead of the local (pointwise) ones.

Definition 1.1.28. $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is called *densely consistent* if there exists a set $\mathcal{D}^0 \subset \mathcal{D}$ whose image $F(\mathcal{D}^0)$ is dense in some neighbourhood of the point $0 \in \mathcal{Y}$, and it is consistent at each element $v \in \mathcal{D}^0$.

The order of the dense consistency on $\mathcal{D}^0$ is defined as $\inf\{p_v : v \in \mathcal{D}^0\}$, where p_v denotes the order of consistency at the point v.

Example 1.1.29. Let us consider the explicit Euler method, given in Examples 1.1.4, 1.1.6 and 1.1.8. We apply it to the Cauchy problem of Example 1.1.2, i.e., to the problem (1.3)-(1.4). We verify the consistency and its order on the set $\mathcal{D}^0 \subset \mathcal{D}$, where $\mathcal{D} := C^1[0,1]$ and $\mathcal{D}^0 := C^2[0,1]$. Then for the local discretization error we obtain

$$[F_n(\varphi_n(v)) - \psi_n(F(v))](t_i) = \begin{cases} \frac{1}{2n} v''(\theta_i)\,, & i = 1, \dots, n\,, \\ 0\,, & i = 0\,, \end{cases} \tag{1.16}$$

where $\theta_i \in (t_{i-1}, t_i)$ are given numbers and $v \in \mathcal{D}^0$ is an arbitrary element. Then $\|l_n(v)\|_{\mathcal{X}_n} = \mathcal{O}(n^{-1})$ on $\mathcal{D}^0$.

Hence, for the class of problems (1.3)-(1.4) with Lipschitz continuous right-hand side f, the explicit Euler method is densely consistent, and the order of the dense consistency on $\mathcal{D}^0 := C^2[0,1]$ equals one.

In the paragraph "Consistency." in Subsection 1.1.2 (c.f. Example 1.1.16) we showed that (pointwise) consistency in itself is not enough for the convergence. One may think that the notion of dense consistency, given by Definition 1.1.28, ensures convergence. The following example shows that this is not true.

Example 1.1.30. Let us choose the normed spaces as $\mathcal{X} = \mathcal{X}_n = \mathcal{Y} = \mathcal{Y}_n = \mathbb{R}$, $\varphi_n, = \psi_n = \textit{identity}$. Our aim is to solve the scalar equation $F(x) = 0$, where the

function $F \in C(\mathbb{R}, \mathbb{R})$ is given as

$$F(x) = \begin{cases} |x| \,, \text{ if } x \in (-1, 1) \,, \\ 1 \,, \text{ if } x \in (-\infty, -1] \cup [1, \infty) \,. \end{cases}$$

Clearly this problem has a unique solution $\bar{x} = 0$. We define the numerical method as

$$F_n(x) = \begin{cases} \frac{1}{n} \,, \text{ if } x \in \left[-\frac{1}{n}, \frac{1}{n}\right] \,, \\ x \,, \text{ if } x \in \left(\frac{1}{n}, 1\right) \,, \\ 1 \,, \text{ if } x \in (-\infty, -1] \cup [1, n) \cup [n+2, \infty) \,, \\ -x \,, \text{ if } x \in \left(-1, -\frac{1}{n}\right) \,, \\ |x - (n+1)| \,, \text{ if } x \in [n, n+2) \,. \end{cases}$$

For the given problem this $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is consistent on the entire $\mathbb{R}$, however, it is not convergent, since the solutions of the discrete problems $F_n(x) = 0$ are $\bar{x}_n = n + 1$, and therefore $\bar{x}_n \nrightarrow \bar{x}$.

What is more interesting is that dense consistency does not imply consistency either as the following examples show.

Example 1.1.31. Let us choose the normed spaces as $\mathcal{X} = \mathcal{X}_n = \mathcal{Y} = \mathcal{Y}_n = \mathbb{R}$, $\varphi_n = \psi_n = \textit{identity}$. Our aim is to solve the scalar equation $F(x) = 0$, where the function $F \in C(\mathbb{R}, \mathbb{R})$ is the identity. Clearly this problem has a unique solution $\bar{x} = 0$. We define the numerical method as

$$F_n(x) = \begin{cases} 1 - n|x| \,, \text{ if } x \in \left(-\frac{1}{n-1}, \frac{1}{n+1}\right) \,, \\ x \,, \text{ if } x \in \left(-\infty, -\frac{1}{n-1}\right] \cup \left[\frac{1}{n+1}, \infty\right) \,. \end{cases}$$

It can be seen that in this case $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is densely consistent, since it is consistent at all $x \in \mathbb{R} \setminus \{0\}$, however, it is not consistent.

Example 1.1.32. We modify Example 1.1.31 only at some points. We choose the function $F \in C(\mathbb{R}, \mathbb{R})$ as $F(x) = |x|$. We define the numerical method as

$$F_n(x) = \begin{cases} 1 - n|x| \,, \text{ if } x \in \left(-\frac{1}{n+1}, \frac{1}{n+1}\right) \,, \\ |x| \,, \text{ if } x \in \left(-\infty, -\frac{1}{n+1}\right] \cup \left[\frac{1}{n+1}, \infty\right) \,. \end{cases}$$

Here we can conclude the same as in the last example.

The alarming difference is that in Example 1.1.31 we have a unique solution of the equation $F_n(x) = 0$ for all n, moreover $\bar{x}_n \to \bar{x} = 0$, while here $F_n(x) > 0$ for all n.

We note that both examples fail in the stability test due to the lack of injectivity.

In spite of all this, the notion of dense consistency is extremely useful as the Reader will see below.

In the sequel, besides Assumption 1.1.9, we will use the following new assumptions.

Assumption 1.1.33. $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ possesses the following properties.

(a4) The problem $\mathscr{P}$ is such that F^{-1} exists in some neighbourhood of $0 \in \mathcal{Y}$ and is continuous at the point $0 \in \mathcal{Y}$.

(a5) There exists $K_1 > 0$ such that for all $v \in \mathcal{D}$ the relation

$$\|\varphi_n(\bar{u}) - \varphi_n(v)\|_{\mathcal{X}_n} \leq K_1 \|\bar{u} - v\|_{\mathcal{X}}$$

holds for all $n \in \mathbb{N}$.

(a6) There exists $K_2 > 0$ such that for all $y \in \mathcal{Y}$ the relation

$$\|\psi_n(y) - \psi_n(0)\|_{\mathcal{Y}_n} \leq K_2 \|y - 0\|_{\mathcal{Y}}$$

holds for all $n \in \mathbb{N}$.

Lemma 1.1.34. *We assume that $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ possesses the following properties.*

- *(a1)–(a3) of Assumption 1.1.9 hold.*
- *(a4) and (a6) of Assumption 1.1.33 hold.*
- *It is densely consistent and stable with stability threshold R and constant S.*

Then F_n is invertible at the point $\psi_n(0) = 0$, i.e., there exists $F_n^{-1}(\psi_n(0))$ for sufficiently large indices n.

Proof. We can choose a sequence $(y^k)_{k\in\mathbb{N}}$ such that $y^k \to 0 \in \mathcal{Y}$ and $F^{-1}\left(y^k\right) =: u^k \to \bar{u}$, due to the continuity of F^{-1}. Then the discretization $\mathscr{D}$ on problem $\mathscr{P}$ at the element u^k is stable with stability threshold $R/2$ and constant S, for some sufficiently large indices k. Moreover, F_n is continuous on $B_{R/2}(\varphi_n(u^k))$. Thus, for these indices k and also for sufficiently large n there exists $F_n^{-1} : B_{R/2S}(F_n(\varphi_n(u^k))) \to B_{R/2}(\varphi_n(u^k))$ moreover, it is Lipschitz continuous with constant S, according to Lemma 1.1.22. Let us write a trivial upper estimate:

$$\left\|F_n(\varphi_n(u^k))\right\|_{\mathcal{Y}_n} \leq \left\|F_n(\varphi_n(u^k)) - \psi_n(F(u^k))\right\|_{\mathcal{Y}_n} + \left\|\psi_n(F(u^k))\right\|_{\mathcal{Y}_n}.$$

Here the first term tends to 0 as $n \to \infty$, due to the consistency. For the second term, based on (a3) and (a6) we have the estimate $\left\|\psi_n(y^k)\right\|_{\mathcal{Y}_n} \leq K_2 \left\|y^k\right\|_{\mathcal{X}_n}$. Since the right-hand side tends to zero as $k \to \infty$, this means that the centre of the ball $B_{R/2}(F_n(\varphi_n(u^k)))$ tends to $0 \in \mathcal{Y}_n$, which proves the statement. □

Corollary 1.1.35. *Under the conditions of Lemma 1.1.34, for sufficiently large indices k and n, the following results are true.*

- *There exists $F_n^{-1}(\psi_n(y^k))$, since $\psi_n(y^k) \in B_{R/2S}(F_n(\varphi_n(u^k)))$.*
- *$F_n^{-1}(\psi_n(y^k)) \in B_{R/2}(\varphi_n(\bar{u}))$,*

moreover, under (a5) of Assumption 1.1.33

- *$\varphi_n(F^{-1}(y^k)) \in B_{R/2}(\varphi_n(\bar{u}))$ holds, too.*

Now we are in the position to formulate our basic result.

Theorem 1.1.36. *We assume that $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ possesses the following properties.*

- *(a1)–(a3) of Assumption 1.1.9 and (a4)–(a6) of Assumption 1.1.33 hold.*
- *It is densely consistent and stable with stability threshold R and constant S.*

Then it is convergent, and the order of the convergence can be estimated from below by the order of consistency on the corresponding set $\mathcal{D}^0$.

Proof. By use of the triangle inequality and Corollary 1.1.35, we can write

$$\begin{aligned}
&\|\varphi_n(\bar{u}) - \bar{u}_n\|_{\mathcal{X}_n} = \left\|\varphi_n(F^{-1}(0)) - F_n^{-1}(\psi_n(0))\right\|_{\mathcal{X}_n} \leq \\
&\underbrace{\left\|\varphi_n(F^{-1}(0)) - \varphi_n(F^{-1}(y^k))\right\|_{\mathcal{X}_n}}_{I.} + \\
&\underbrace{\left\|\varphi_n(F^{-1}(y^k)) - F_n^{-1}(\psi_n(y^k))\right\|_{\mathcal{X}_n}}_{II.} + \\
&\underbrace{\left\|F_n^{-1}(\psi_n(y^k)) - F_n^{-1}(\psi_n(0))\right\|_{\mathcal{X}_n}}_{III.},
\end{aligned} \tag{1.17}$$

where the elements $y^k \in \mathcal{Y}$ are defined in the proof of Lemma 1.1.34.

In the next step we estimate the different terms on the right-hand side of (1.17).

I. For the first term, on the basis of (a5) of Assumption 1.1.33, we have the estimate

$$\left\|\varphi_n(F^{-1}(0)) - \varphi_n(F^{-1}(y^k))\right\|_{\mathcal{X}_n} \leq K_1 \left\|F^{-1}(0) - F^{-1}(y^k)\right\|_{\mathcal{X}}.$$

Since $y^k \to 0$ as $k \to \infty$, and F^{-1} is continuous at the point $0 \in \mathcal{Y}$, therefore this term tends to zero, independently of n.

II. Due to Corollary 1.1.35, we can use the stability estimate, therefore for this term we have the estimate

$$\left\|\varphi_n(F^{-1}(y^k)) - F_n^{-1}(\psi_n(y^k))\right\|_{\mathcal{X}_n} \leq$$
$$S\left\|F_n(\varphi_n(F^{-1}(y^k))) - \psi_n(y^k)\right\|_{\mathcal{Y}_n} = S\left\|F_n(\varphi_n(u^k)) - \psi_n(F(u^k))\right\|_{\mathcal{Y}_n}.$$

In this estimate the term on the right-hand side tends to zero because of the consistency at u^k.

III. For the estimation of the third term we can use the Lipschitz continuity of F_n^{-1}, due to Lemma 1.1.34 and Corollary 1.1.35. Hence, by using (a3) and (a6) of Assumption 1.1.9 and Assumption 1.1.33, respectively, we have

$$\left\|F_n^{-1}(\psi_n(y^k)) - F_n^{-1}(\psi_n(0))\right\|_{\mathcal{X}_n} \leq S\left\|\psi_n(y^k) - \psi_n(0)\right\|_{\mathcal{Y}_n} \leq SK_2\left\|y^k\right\|_{\mathcal{Y}}.$$

The right-hand side of the above estimate tends to zero, independently of the index n.

These estimations complete the proof. □

There is only one job left, to ensure the stability. Analogously to the consistency, in the stability the lack of knowledge of the solution $\bar{u}$ makes the direct application of the Definition 1.1.19 impossible. Thus, we need a condition which can be easily checked and implies stability. The following trivial lemma gives a helping hand.

Lemma 1.1.37. *We assume that $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ possesses (a5) of Assumption 1.1.33 and it is stable with stability threshold R and constant S. Then it is stable at all $v \in \mathcal{D} \cap B_R(\bar{u})$ with stability constant S.*

As a consequence, we need to check stability on a set of elements that the union of their stability neighbourhoods contains $\varphi_n(\bar{u})$ and the infimum of their stability constants is positive.

Example 1.1.38. [54, Version of Paragraph 38.2]
Let us analyse the stability property of the explicit Euler method, given in Example 1.1.8.

Let $\mathbf{v}^{(1)}, \mathbf{v}^{(2)} \in \mathcal{X}_n = \mathbb{R}^{n+1}$ be two arbitrary vectors, and we use the notation $\epsilon = \mathbf{v}^{(1)} - \mathbf{v}^{(2)} \in \mathbb{R}^{n+1}$. We define the vector $\delta = F_n\left(\mathbf{v}^{(1)}\right) - F_n\left(\mathbf{v}^{(2)}\right) \in \mathbb{R}^{n+1}$, where F_n is defined in (1.6). (In the notation, for simplicity, we omit the use of the subscript n for the vectors. We recall that the coordinates of the vectors are numbered from $i = 0$ to $i = n$.)

For the coordinates of the vector δ we have the following relations.

- For the first coordinate ($i = 0$) we obtain: $\delta_0 = \left(F_n\left(\mathbf{v}^{(1)}\right)\right)_0 - \left(F_n\left(\mathbf{v}^{(2)}\right)\right)_0 = \left(v_0^{(1)} - u_0\right) - \left(v_0^{(2)} - u_0\right) = \epsilon_0\,.$
- For the other coordinates $i = 1, \ldots, n$ we have

$$\begin{aligned}\delta_i = v_i^{(1)} - v_i^{(2)} = \\ n(v_i^{(1)} - v_{i-1}^{(1)}) - f(v_{i-1}^{(1)}) - n(v_i^{(2)} - v_{i-1}^{(2)}) + f(v_{i-1}^{(2)}) = \\ n(v_i^{(1)} - v_i^{(2)}) - n(v_{i-1}^{(1)} - v_{i-1}^{(2)}) - (f(v_{i-1}^{(1)}) - f(v_{i-1}^{(2)})) = \\ n\epsilon_i - n\epsilon_{i-1} - (f(v_{i-1}^{(1)}) - f(v_{i-1}^{(2)}))\,.\end{aligned}$$

We can express ϵ_i from this relation as follows:

$$\epsilon_i = \epsilon_{i-1} + \frac{1}{n}\left(f(v_{i-1}^{(1)}) - f(v_{i-1}^{(2)})\right) + \frac{1}{n}\delta_i\,.$$

Under our assumption, $f \in C(\mathbb{R}, \mathbb{R})$ is a Lipschitz continuous function, therefore we have the estimation $|f(v_{i-1}^{(1)}) - f(v_{i-1}^{(2)})| \leq L|v_{i-1}^{(1)} - v_{i-1}^{(2)}|$. Hence, we get

$$|\epsilon_i| \leq |\epsilon_{i-1}| + \frac{1}{n}L|v_{i-1}^{(1)} - v_{i-1}^{(2)}| + \frac{1}{n}|\delta_i| = |\epsilon_{i-1}|\left(1 + \frac{L}{n}\right) + \frac{1}{n}|\delta_i|\,.$$

If we apply this estimate consecutively to $|\epsilon_{i-1}|$, $|\epsilon_{i-2}|$, etc., we obtain:

$$\begin{aligned}|\epsilon_i| \leq |\epsilon_{i-2}|\left(1 + \frac{L}{n}\right)^2 + \frac{1}{n}|\delta_i| + \left(1 + \frac{L}{n}\right)\frac{1}{n}|\delta_{i-1}| \leq \ldots \\ |\epsilon_0|\left(1 + \frac{L}{n}\right)^n + \frac{1}{n}\sum_{i=1}^{n}|\delta_i|\left(1 + \frac{L}{n}\right)^{n-i}.\end{aligned} \tag{1.18}$$

Since $\delta_0 = \epsilon_0$ and $\left\|\mathbf{v}^{(1)} - \mathbf{v}^{(2)}\right\|_{\mathcal{X}_n} = \max\limits_{i=0,\ldots,n}|\epsilon_i|$, hence we can write our estimation in the form

$$\left\|\mathbf{v}^{(1)} - \mathbf{v}^{(2)}\right\|_{\mathcal{X}_n} \leq |\delta_0|\left(1 + \frac{L}{n}\right)^n + \frac{1}{n}\sum_{i=1}^{n}|\delta_i|\left(1 + \frac{L}{n}\right)^{n-i} \tag{1.19}$$

$$< e^L(\delta_0 + \max_{i=1,\ldots,n}|\delta_i|) = e^L\left\|\delta\right\|_{\mathcal{Y}_n} = e^L\left\|F_n\left(\mathbf{v}^{(1)}\right) - F_n\left(\mathbf{v}^{(2)}\right)\right\|_{\mathcal{Y}_n}\,. \tag{1.20}$$

This shows us that the discretization (1.8) applied to the problem given in Example 1.1.2 resulting in the explicit Euler method given in Example 1.1.8 is stable on the whole set $\mathcal{X} = C^1[0,1]$ with $S = e^L$ and $R = \infty$ for this problem.

Hence, on the basis of Theorem 1.1.36, the results of this example and Example 1.1.29, we can conclude that the explicit Euler method is convergent, and the order of its convergence is one.

We note that the whole process can be done (with small modifications) when f is only locally Lipschitz continuous.

1.1.4 Relation between the basic notions

Theorems 1.1.24 and 1.1.36 show that, under the assumptions (a1)–(a3) and (a1)–(a6), the consistency or dense consistency and stability result in the convergence, i.e., consistency and stability together are a sufficient condition for convergence. (Roughly speaking, this implication is shown in (1.2).) However, from this observation we cannot get an answer to the question of the necessity of these conditions.

In the sequel, we raise a more general question: What is the general relation between the above listed three basic notions? Since each of them can be true (T) or false (F), we have to consider eight different cases, listed in Table 1.1.

	consistency/ dense consistency	stability	convergence
1	T	T	T
2	T	T	F
3	T	F	T
4	T	F	F
5	F	T	T
6	F	T	F
7	F	F	T
8	F	F	F

Table 1.1: The list of the different cases (T: true, F: false).

Before giving the answer, we consider some examples. In each examples $\mathcal{X} = \mathcal{X}_n = \mathcal{Y} = \mathcal{Y}_n = \mathbb{R}$, $\mathcal{D} = \mathcal{D}_n = [0, \infty)$, $\varphi_n = \psi_n = \mathit{identity}$. Our aim is to solve the scalar equation

$$F(x) \equiv x^2 = 0\,, \tag{1.21}$$

which has the unique solution $\bar{x} = 0$.

Example 1.1.39. For solving equation (1.21) we choose the numerical method defined by the n-th Lagrangian interpolation, i.e., $F_n(x)$ is the Lagrangian interpolation polynomial of order n. Since the Lagrangian interpolation is exact for $n \geq 2$, therefore $F_n(x) = x^2$ holds for all $n \geq 2$. Hence, clearly the numerical method is consistent and convergent. The operator F_n^{-1} can be defined easily, and it is $F_n^{-1}(x) = \sqrt{x}$. One can see that if F_n^{-1} exits and it is differentiable, then for the stability $(F_n^{-1})'$ needs to be bounded around the solution $\bar{x_n}$ from some index. Since in this case it is not fulfilled, the numerical method is not stable.

Example 1.1.40. For solving equation (1.21) we choose now the numerical method $F_n(x) = 1 - nx$. The roots of the discrete equations $F_n(x) = 0$ are $\bar{x}_n = 1/n$, therefore $\bar{x}_n \to \bar{x} = 0$ as $n \to \infty$. This means that the numerical method is convergent. We observe that $\varphi_n(F_n(0)) = \varphi_n(1) = 1$, and $\psi_n(F(0)) = \psi_n(0) = 0$. Hence, for the local discretization error we have $|l_n| = 1$, for any index n. This means that the numerical method is not consistent. One can easily check that F_n is invertible, and $F_n^{-1}(x) = -x/n + 1/n$. Hence the derivative of the inverse operators are uniformly bounded on $[0, \infty)$ by 1 for any n. Therefore, the numerical method is stable.

Example 1.1.41. For solving equation (1.21) we choose the following numerical method: $F_n(x) = 1 - nx^2$. Then $\bar{x}_n = 1/\sqrt{n}$, and hence $\bar{x}_n \to \bar{x} = 0$ as $n \to \infty$. This means that the numerical method is convergent. Due to the relations $\varphi_n(F_n(0)) = \varphi_n(1) = 1$ and $\psi_n(F(0)) = \psi_n(0) = 0$, this method is not consistent. Since for this numerical method $F_n^{-1}(x) = \sqrt{(1-x)/n}$, therefore the derivatives are not bounded. Therefore, the numerical method is not stable.

number of the case	answer	reason
1	always true	Theorem 1.1.24 and 1.1.36
2	always false	Theorem 1.1.24 and 1.1.36
3	possible	Example 1.1.39/ Example 1.1.31
4	possible	Examples 1.1.16 and 1.1.30/ Example 1.1.32
5	possible	Example 1.1.40
6	uninteresting	uninteresting
7	possible	Example 1.1.41
8	uninteresting	uninteresting

Table 1.2: The possibility of the different cases.

Now, we are in the position to answer the question, raised at beginning of this section. Using the numeration of the different cases in Table 1.1, the answers are included in Table 1.2. (We note that two cases (cases 6 and 8 in Table 1.1) are uninteresting from a practical point of view, therefore we have neglected their investigation.) The results particularly show that neither consistency/ dense consistency, nor stability is a necessary condition for the convergence.

1.2 Linear theory

It should be made clear in the first place that the name linear in the title could be misleading, this section contains the case where F is an affine operator $Fu = Lu - f$, where L is a linear operator and $f \in \mathcal{Y}$. The name comes from the linear inhomogeneous equation of the form $Lu = f$.

Comparing the nonlinear and the linear theory some introductory remarks are mentioned. First, the linear theory is a special case contained by the nonlinear theory. As we have seen in the nonlinear case, stability with consistency (under some assumptions) implies convergence, but nothing more can be stated, see Subsection 1.1.4. However, in the special case where the operator F is affine, something more can be stated. Finally, we mention that in the nonlinear theory there is a large variety of the definitions (c.f. the Paragraph "Notes on the notion of stability – other possibilities."), while the linear theory is more fixed, we will see that every stability notion of the nonlinear case is simplified to one stability notion in the linear case. On the other hand, this (and other) simplifications provide a possibility to handle parallel a family of affine operators (differing only in the constant part) by defining consistency, stability and convergence.

The linear theory is more elaborated, the foundations of the theory are already laid in the famous paper [36] and later developed, e.g., in the papers [42, 43, 44]. We also rely on the results of these papers.

1.2.1 Problem setting, basic notions and theoretical results

Problem, discretization and numerical method. In this paragraph we follow the paper [44]. When F is an affine operator, the equation (1.1) to be solved reads as

$$Lu = f\,, \tag{1.22}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are assumed to be normed spaces, $f \in \mathcal{Y}$, $\mathcal{D} \subset \mathcal{X}$, $\mathcal{R} \subset \mathcal{Y}$ and $L : \mathcal{D} \to \mathcal{R}$ is assumed to be an (unbounded) linear operator.

It is supposed that the problem (1.22) is well-posed in the following sense. The range $\mathcal{R}$ of L is dense in $\mathcal{Y}$ and there exists an operator $E \in B(\mathcal{Y}, \mathcal{X})$ such that EA is the identity in $\mathcal{D}$.

This yields that for $f \in \mathcal{R}$ the unique solution is Ef. If $f \notin \mathcal{R}$, then Ef can be regarded as a generalized solution, since E is the unique bounded extension to $\mathcal{Y}$ of $L^{-1} : \mathcal{R} \to \mathcal{D}$. In each cases the unique solution (corresponding to f) will be denoted by u_f.

We assume that $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ (defined above) generates a sequence of problems in the special form

$$L_n u_n = f_n\,, \quad n = 1, 2, \dots\,, \tag{1.23}$$

where $\mathcal{X}_n$ and $\mathcal{Y}_n$ are assumed to be normed spaces, $f_n \in \mathcal{Y}_n$, and $L_n : \mathcal{X}_n \to \mathcal{Y}_n$ is a linear operator.

We assume that the problems (1.23) are well-posed in the same sense as problem (1.22) with solution operators $E_n = L_n^{-1}$.

Note that (a1) of Assumption 1.1.9 with stability implies well-posedness, but here we do not want to restrict ourself to the case where the spaces $\mathcal{X}_n$, $\mathcal{Y}_n$ are finite dimensional.

On $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ we make some further assumptions.

Assumption 1.2.1. $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ possesses the following properties.

(A1) φ_n, ψ_n are bounded linear operators for all n.

(A2) For the operators φ_n, ψ_n the estimates

$$\|\varphi_n\|_{B(\mathcal{X},\mathcal{X}_n)} \le C_1\,, \quad \|\psi_n\|_{B(\mathcal{Y},\mathcal{Y}_n)} \le C_2$$

hold with the constants C_1, C_2 independently of n.

(A3) The relation $\psi_n f = f_n$ holds.

Note that (a5) and (a6) of Assumption 1.1.33 with (A1) of Assumption 1.2.1 implies (A2).

We recall that from now on (in this section) we assume that the problem $\mathscr{P}$ is linear and it has the form (1.22) with the properties given above, moreover, that $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is such that the numerical method $\mathscr{N}$ generates a sequence of problems in the special form (1.23) with the properties given there as well.

Basic notions in the linear case. Basic notions as stability, consistency and convergence are already defined in the nonlinear case. Here it is shown how those definitions can be transformed to the definitions of the linear case. We begin with some observations, and finally we give the adequate definitions.

- Stability: due to the special form of F, the relation (1.13) can be rewritten as

$$F_n(v_n^1) - F_n(v_n^2) = L_n v_n^1 - f_n - (L_n v_n^2 - f_n) = L_n(v_n^1 - v_n^2)\,.$$

Using the notation $v_n^1 - v_n^2 := w_n$ it can be written as

$$\|w_n\|_{\mathcal{X}_n} \leq S \, \|L_n w_n\|_{\mathcal{Y}_n} \, ,$$

thus, the relation (1.13) reads as $\|E_n\|_{B(\mathcal{Y}_n,\mathcal{X}_n)} \leq S$ in the linear case.

Note that it means stability is entirely independent of f_n, which enables us to handle a complete family of problems (differing only in the term f). Furthermore, the notion of stability is the property of the numerical method $\mathscr{N}$ only.

- Consistency: for a given f, the local discretization error can be transformed as follows.

$$l_{n,f}(v) := F_n(\varphi_n(v)) - \psi_n(F(v)) = L_n\varphi_n(v) - f_n - (\psi_n(Lv - f)) \, .$$

Using (A1) and (A3) of Assumption 1.2.1 implies that

$$l_{n,f}(v) = L_n\varphi_n v - \psi_n L v \, .$$

As we can see, consistency can be defined for a family of problems, too.

- Convergence: using (A1) of Assumption 1.2.1 the global discretization error reads as

$$e_{n,f} := \varphi_n u_f - u_{n,f} = \varphi_n E f - E_n \psi_n f \, .$$

In the light of the previous items we reformulate the basic notions. First, we introduce the notations $\{\mathscr{P}\} = \{(\mathcal{X}, \mathcal{Y}, F) \, : \, Fu = Lu - f, f \in \mathcal{Y}\}$ and $\mathscr{D}(\{\mathscr{P}\}) \rightsquigarrow \{\mathscr{N}\}$ for the notions that we apply the discretization on the family of problems $\{\mathscr{P}\}$ resulting in the family of numerical methods $\{\mathscr{N}\}$.

Definition 1.2.2. $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is called *convergent at the element* $f \in \mathcal{Y}$ if

$$\lim \|\varphi_n E f - E_n \psi_n f\|_{\mathcal{X}_n} = 0 \tag{1.24}$$

holds. When it is convergent for all $f \in \mathcal{Y}$, we say that $\mathscr{D}(\{\mathscr{P}\}) \rightsquigarrow \{\mathscr{N}\}$ is *convergent*.

Definition 1.2.3. $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is called *consistent at the element* $v \in \mathcal{D}$ if the relation

$$\lim \|L_n\varphi_n v - \psi_n L v\|_{\mathcal{Y}_n} = 0 \tag{1.25}$$

holds. We call $\mathscr{D}(\{\mathscr{P}\}) \rightsquigarrow \{\mathscr{N}\}$ *consistent* if there exists a set $\mathcal{D}^0 \subset \mathcal{D}$ for which $L\mathcal{D}^0$ is dense in $\mathcal{Y}$ and $\mathscr{D}(\mathscr{P}) \rightsquigarrow \mathscr{N}$ is consistent at each element in $\mathcal{D}^0$.

Definition 1.2.4. $\mathscr{D}(\{\mathscr{P}\}) \rightsquigarrow \{\mathscr{N}\}$ is called *stable* if the inequality

$$\|E_n\|_{B(\mathcal{Y}_n,\mathcal{X}_n)} \leq S \tag{1.26}$$

holds with a constant S (independently of n).

Results. A generalization of the Lax equivalence theorem is presented.

Theorem 1.2.5. *[44, Equivalence theorem, part 1]*
We assume that $\mathscr{D}(\{\mathscr{P}\}) \rightsquigarrow \{\mathscr{N}\}$ possesses the following properties.

- *(A1)–(A3) of Assumption 1.2.1 is valid.*
- *It is consistent and stable.*

Then it is convergent.

This theorem is analogous to Theorem 1.1.24, however the main task by that theorem was to ensure the existence of the discrete solutions. Thus, Theorem 1.2.5 is rather similar to Lemma 1.1.18. The proof is almost the same (since Theorem 1.2.5 can be viewed as a special case of Lemma 1.1.18), but we need to handle generalized solutions, too. This makes the proof similar to the proof of Theorem 1.1.36 (but simpler). Another difference is that, due to the linearity, Theorem 1.2.5 deals with a family of problems, while in the nonlinear case this was impossible.

Proof. If $f \in \mathcal{R}$, then

$$\|\varphi_n E f - E_n \psi_n f\|_{\mathcal{X}_n} = \|E_n(L_n \varphi_n u_f - \psi_n L u_f)\|_{\mathcal{X}_n} \leq S \|L_n \varphi_n u_f - \psi_n L u_f\|_{\mathcal{Y}_n} \to 0\,.$$

If $f \notin \mathcal{R}$, we can choose a sequence $\left(f^k\right)_{k\in\mathbb{N}}$, with $f^k \in \mathcal{R}$ and $\lim f^k = f$. Then

$$\|\varphi_n E f - E_n \psi_n f\|_{\mathcal{X}_n} \leq$$
$$\underbrace{\|\varphi_n E f - \varphi_n E f^k\|_{\mathcal{X}_n}}_{I.} + \underbrace{\|\varphi_n E f^k - E_n \psi_n f^k\|_{\mathcal{X}_n}}_{II.} + \underbrace{\|E_n \psi_n f^k - E_n \psi_n f\|_{\mathcal{X}_n}}_{III.}\,.$$

I. and III. tend to 0 independently of n. II. tends to 0 independently of k because of the first part of the proof. □

Before moving on to the second part of the equivalence theorem we take preparation.

Assumption 1.2.6. We assume that $\mathscr{D}(\{\mathscr{P}\}) \rightsquigarrow \{\mathscr{N}\}$ possesses the following properties.

(A4) $\mathcal{Y}$ is a Banach space.

(A5) There exists a constant L such that, for all n and for all $g_n \in \mathcal{Y}_n$ with $\|g_n\|_{\mathcal{Y}_n} \leq 1$, there exists an element $g \in \mathcal{Y}$ such that $\|g\|_{\mathcal{Y}} \leq L$ and $\psi_n g = g_n$.

(A5) establishes a connection between the norms of the spaces $\mathcal{Y}$ and $\mathcal{Y}_n$, see [44, Rem.2.2.] and c.f. with the Paragraph "Convergence." in Subsection 1.1.2.) The second part of the equivalence theorem is based mainly on the following lemma.

Lemma 1.2.7. *[42] Let $\mathcal{Z}$ be a Banach space, $(\mathcal{W}_n)_{n\in\mathbb{N}}$ a sequence of normed spaces and $T_n : \mathcal{Z} \to \mathcal{W}_n$ linear operators. If for each $z \in \mathcal{Z}$, $\sup \|T_n z\|_{\mathcal{W}_n} \leq \infty$, then $\sup \|T_n\|_{B(\mathcal{Z},\mathcal{W}_n)} \leq \infty$.*

This is a generalization of the Banach-Steinhaus theorem. (The proof can be done in the same way as by the original theorem.)

Ready with the preparation, the second part of the equivalence theorem is presented.

Theorem 1.2.8. *[44, Equivalence theorem, part 2]*
Assume that $\mathscr{D}(\{\mathscr{P}\}) \rightsquigarrow \{\mathscr{N}\}$ possesses the following properties.

- *(A1)–(A3) of Assumption 1.2.1 and (A4)–(A5) of Assumption 1.2.6 are valid.*
- *It is convergent.*

Then it is stable.

This part contains the novelty compared to the nonlinear case, i.e. convergence is necessary for stability.

Proof. For each $f \in \mathcal{Y}$ the sequences $(\|\varphi_n E f - E_n \psi_n f\|_{\mathcal{X}_n})_{n\in\mathbb{N}}$, $(\|\varphi_n E f\|_{\mathcal{X}_n})_{n\in\mathbb{N}}$ are bounded due to the convergence and Assumption 1.2.1, respectively. This implies that the sequence $(\|E_n \psi_n f\|_{\mathcal{X}_n})_{n\in\mathbb{N}}$ is bounded as well.

The generalized Banach-Steinhaus lemma 1.2.7 implies that there exists a constant K_1 such that $\|E_n \psi_n\|_{B(\mathcal{Y},\mathcal{X}_n)} \leq K_1$. Choosing a sequence $(g_n)_{n\in\mathbb{N}}$, $g_n \in \mathcal{Y}_n$ with $\|g_n\|_{\mathcal{Y}_n} \leq 1$ and $\psi_n g = g_n$, then $\|E_n g_n\|_{\mathcal{X}_n} = \|E_n \psi_n g\|_{\mathcal{X}_n} \leq \|E_n \psi_n\|_{B(\mathcal{Y},\mathcal{X}_n)} \|g\|_{\mathcal{Y}} = K_1 L$ by (A5) of Assumption 1.2.6. Thus, $\|E_n\|_{B(\mathcal{X},\mathcal{X}_n)} \leq K_1 L$. □

Remark 1.2.9. Here we note the following.

- Theorems1.1.24 and 1.1.36 contained the essence of the nonlinear theory, that result can be illustrated with the formula (1.2). Meanwhile, the heart of the linear theory is summarized in Theorems 1.2.5 and 1.2.8. This result can be illustrated by the formula

$$\begin{aligned} &\text{Consistency} + \text{Stability} \Rightarrow \text{Convergence} \\ &\text{moreover,} \\ &\text{Convergence} \Rightarrow \text{Stability}\,. \end{aligned} \tag{1.27}$$

This explains the name "equivalence theorem" (i.e., stability is equivalent to convergence under the assumption of consistency).

- (A4) and (A5) of Assumption 1.2.6 are necessary, see [43] and [44], respectively for the details.

1.2.2 Examples

Until now we have shown the linear framework on the abstract level. In the following we illustrate these abstract results with various examples.

Problem 1. Let $\Omega \subset \mathbb{R}^d$ be an open and bounded domain with a smooth boundary $\partial\Omega$. We investigate the elliptic equation

$$\begin{cases} Ku = f\,, & \text{in } \Omega\,, \\ u = g\,, & \text{at } \partial\Omega\,, \end{cases} \tag{1.28}$$

where K is an elliptic operator given in a divergence form as

$$Ku = -\sum_{i,j=1}^{d} \frac{\partial u}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + \sum_{i=1}^{d} b_i \frac{\partial u}{\partial x_i} + cu\,, \tag{1.29}$$

where a_{ij}, b_i, c are smooth coefficient functions.

Remark 1.2.10. [10, Ch.6.1.1] We can model physical processes with PDE's. The solution of the previously defined problem can be interpreted as a chemical concentration (or the density of some quantity) at equilibrium within a region Ω.

Then the second-order term represents the *diffusion*, the first-order term represents the *transport* within Ω and the zeroth-order term describes the local *creation* or *depletion* of the chemical (simply saying the *reaction* term). (The coefficients a_{ij} describe the anisotropic heterogeneous nature of the medium.)

Example 1.2.11. This example is based on [44, Paragraph 3.4.]. We set a homogeneous Dirichlet boundary condition (i.e., $g \equiv 0$), moreover, we assume that $K : L^2(\Omega) \hookrightarrow L^2(\Omega)$ is a symmetric, uniformly positive operator (this can be ensured by some restrictions on the coefficients) whose domain is $\operatorname{dom} K = H_0^1(\Omega) \cap H^2(\Omega)$. $f \in L^2(\Omega)$ is a given function.

In this case there exists a unique weak (generalized) solution $u_f = Ef \in H_0^1(\Omega)$, and $E : L^2(\Omega) \to H_0^1(\Omega)$ is characterized by the variational formula

$$a(Ef, v) = (f, v)\,, \quad \forall v \in H_0^1(\Omega)\,, \tag{1.30}$$

where $a(\cdot,\cdot)$ is the bilinear form corresponding to K which is defined as

$$a(u,v) = \int_\Omega \sum_{i,j=1}^{d} a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j} + \sum_{i=1}^{d} b_i \frac{\partial u}{\partial x_i} v + cuv \, \mathrm{d}\mathbf{x} \tag{1.31}$$

and $(\cdot,\cdot)$ denotes the L^2 inner product.

This means the setting $\mathcal{X} = H_0^1(\Omega)$ with the energy norm, $\mathcal{D} = H_0^1(\Omega) \cap H^2(\Omega)$, $\mathcal{Y} = L^2(\Omega)$ with the L^2 norm, and consequently the problem is well-posed in the sense explained at the beginning of this subsection.

To get an approximation $u_{n,f}$ of the solution u_f, a finite dimensional subspace S_n in $H_0^1(\Omega)$ is chosen and $u_{n,f}$ is defined by the equality

$$a(u_{n,f}, v) = (f, v)\,, \quad \forall v \in S_n\,. \tag{1.32}$$

It is known that in this case $u_{n,f}$ exits uniquely.

We set $\mathcal{X}_n$ as S_n with the energy norm and $\mathcal{Y}_n$ as S_n with the L^2 norm, $\varphi_n : \mathcal{X} \to \mathcal{X}_n$ and $\psi_n : \mathcal{Y} \to \mathcal{Y}_n$ as the $a(\cdot,\cdot)$- and $(\cdot,\cdot)$-orthogonal projections, respectively. With this choice Assumptions 1.2.1 and 1.2.6 are fulfilled. The discrete problems are well-posed with solution operators $E_n : \mathcal{Y}_n \to \mathcal{X}_n$ defined as

$$a(E_n h, v) = (h, v)\,, \quad \forall v \in S_n\,. \tag{1.33}$$

This means that

$$a(u_f - u_{n,f}, v) = 0\,, \quad \forall v \in S_n\,, \tag{Galerkin-orthogonality}$$

consequently, $\varphi_n E f = \varphi_n u_f = u_{n,f} = E_n f_n = E_n \psi_n f$, thus the global discretization error is 0, which means that this method is convergent independently of the choice of the subspaces S_n.

This may sound odd, but reflects well on the argumentation of the Paragraph "Convergence" in Subsection 1.1.2 i.e. the success of the whole procedure depends on two tasks, on the numerical method (in our terminology convergence is a notion related only to the numerical method) and on the approximation capabilities of the subspaces $\mathcal{X}_n$. The second task depends on the choice of the subspaces S_n. This can be explained by the relation

$$\|u_f - u_{n,f}\|_{\mathcal{X}} \leq \underbrace{\|u_f - \varphi_n u_f\|_{\mathcal{X}}}_{\text{approximation capabilities}} + \underbrace{\|\varphi_n u_f - u_{n,f}\|_{\mathcal{X}_n}}_{=0 \Rightarrow \text{convergence}}\,. \tag{1.34}$$

Thus, in this case

$$\|u_f - u_{n,f}\|_{\mathcal{X}} = \underbrace{\|u_f - \varphi_n u_f\|_{\mathcal{X}}}_{\text{approximation capabilities}}\,. \tag{1.35}$$

FEM. A question is: how to implement the numerical method described above? One possible way is the finite element method (FEM). The FEM is well-known in the numerical analysis community, detailed descriptions can be found in many textbooks, here we provide a short introduction in order to introduce some notations which will be used later.

To realize (1.32) we need to define S_n, which can be done by giving a basis of this subspace. There are many ways to do this, here is only one approach presented.

The first step is to define a mesh on Ω. A 1D mesh consists of intervals. The 2D mesh is a regular triangle mesh and the 3D one is a tetrahedron mesh. A given mesh determines the sets $\mathcal{P} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ and $\mathcal{P}_\partial = \{\mathbf{x}_{N+1}, \mathbf{x}_{N+2}, \dots, \mathbf{x}_{N+N_\partial}\}$ containing the vertices in Ω and on $\partial\Omega$, respectively. Let us introduce two more notations: $\overline{N} = N + N_\partial$ and $\overline{\mathcal{P}} = \mathcal{P} \cup \mathcal{P}_\partial$.

The basis functions are denoted by $\phi_i(\mathbf{x})$, $i = 1, \dots N$. One possibility is the use of the so-called hat functions which are defined with the following properties:

1. the basis functions are continuous functions ;

2. the basis functions are piecewise linear functions over intervals/triangles/tetrahedrons ;

3. $\phi_i(\mathbf{x}_i) = 1$ for $i = 1, \dots N$;

4. $\phi_i(\mathbf{x}_j) = 0$ for $i = 1, \dots N$, $j = 1, \dots \overline{N}$, $i \neq j$.

We set $n = N$ and seek $u_{n,f}$ in the form $\sum_{i=1}^{N} u_i \phi_i$. Then the coefficients u_i can be determined by solving the linear algebraic equation

$$\mathbf{K}_0 \mathbf{u}_0 = \mathbf{f} \,, \tag{1.36}$$

where $\mathbf{K}_0 \in \mathbb{R}^{N \times N}$ is the so-called stiffness matrix with entries

$$\mathbf{K}_{0ij} = a\left(\phi_j, \phi_i\right) \,, \tag{1.37}$$

$\mathbf{u}_0 \in \mathbb{R}^N$ contains the unknowns, and f_i is determined by the formula (f, ϕ_i).

Returning to Example 1.2.11, where a convergent numerical method was described, Theorem 1.2.8 (Equivalence theorem, part 2) implies that the inverses of the matrices $\mathbf{K}_0$ are uniformly bounded (stability), i.e. there exists a constant S independently of n such that

$$\|\mathbf{K}_0^{-1}\| \leq S \tag{1.38}$$

holds in some suitable norm. Note that (1.36) and (1.32) can be viewed as two different forms of the same equation, however, in this case this means that the norms corresponding to the second form (1.36) are determined, i.e. $\|\mathbf{f}\|_{\square}$ is defined as $\|\psi_n f\|_{L^2}$ and $\|\mathbf{u}\|_{\triangleleft}$ is defined as $\|\varphi_n u\|_{H_0^1}$, and the suitable norm in (1.38) is the $\|\cdot\|_{\square,\triangleleft}$ norm.

If we want to depart from this choice e.g. we choose $\mathcal{X}_n = \mathcal{Y}_n = (\mathbb{R}^N, \|\cdot\|_2)$, then we need to check (A2) of Assumption 1.2.1 and (A5) of Assumption 1.2.6. When $f = \sum_{i=1}^{N} r_i \phi_i$, then $\|f\|_{L^2} = \|\mathbf{r}\|_{\mathbf{M}}$ and $\mathbf{f} = \mathbf{M}\mathbf{r}$, thus $\|\psi_n\| = \|\mathbf{M}\|_2^{\frac{1}{2}}$ and similarly $\|\varphi_n\| = \|\mathbf{H}^{-1}\|_2^{\frac{1}{2}}$, where $H_{ij} = (\mathrm{grad}\phi_j, \mathrm{grad}\phi_i)$. This means that for (A2) it is needed to show that $\|\mathbf{H}^{-1}\|_2^{\frac{1}{2}} \leq C_1$ and $\|\mathbf{M}\|_2^{\frac{1}{2}} \leq C_2$ hold. This can be done, however, (A5) does not generally hold since

$$\frac{1}{\|\mathbf{M}\|_1} \leq \frac{1}{\|\mathbf{M}\|_2} \leq \|\mathbf{M}^{-1}\|_2 \,,$$

and e.g. for the one-dimensional uniform mesh $\|\mathbf{M}\|_1 = h$, where the meshsize is h, shows us that $\|\mathbf{M}^{-1}\|_2 \to \infty$ when $h \to 0$. This means that in this case the framework is not applicable.

We note that the FEM can be easily extended to the case where a nonhomogeneous Dirichlet boundary condition is prescribed. In this case the set of the basis functions need to be supplemented by the functions ϕ_i, $i = N+1, \ldots \overline{N}$ with the properties listed earlier. The equation to be solved reads as

$$\mathbf{K}\mathbf{u} = \mathbf{f} \,, \tag{1.39}$$

where $\mathbf{K} = (\mathbf{K}_0|\mathbf{K}_\partial) \in \mathbb{R}^{N\times\overline{N}}$, $\mathbf{u} = (\mathbf{u}_0|\mathbf{u}_\partial)^T \in \mathbb{R}^{\overline{N}}$ and $\mathbf{u}_\partial$ can be determined by using the boundary condition.

FDM. There are other ways to approximate the solution of the equation (1.28). In the following we overview the finite difference method (FDM) in a same short way as earlier for the FEM. To make easier the presentation of the FDM we simplify the problem (1.28) into the simple problem

$$\begin{cases} -u''(x) = f(x) \,, \quad x \in (0,1) \,, \\ u(0) = u(1) = 0 \,. \end{cases} \tag{1.40}$$

In the first step a mesh is defined (similarly as for the FEM), here we choose a uniform mesh which determines the sets $\mathcal{P} = \{\mathbf{x}_1 = h, \mathbf{x}_2 = 2h, \ldots, \mathbf{x}_N = Nh\}$ and $\mathcal{P}_\partial = \{\mathbf{x}_0 = 0, \mathbf{x}_{N+1} = 1\}$ containing the vertices in Ω and on $\partial\Omega$, respectively, with $h = \dfrac{1}{N+1}$.

Then we use the approximations

$$\begin{cases} -u''(x) \approx \dfrac{-u(x-h)+2u(x)-u(x+h)}{h^2}\,, & x = x_i\,, i = 1,\dots,N\,, \\ u(x) = 0\,, \quad x = x_0, x_{N+1}\,, \end{cases} \tag{1.41}$$

resulting in the linear algebraic equation

$$\mathbf{K}_0\mathbf{u}_0 = \mathbf{f}\,,$$

where $\mathbf{K}_0 = (N+1)^2\text{tridiag}\,(-1,2,-1) \in \mathbb{R}^{N\times N}$ and $\mathbf{u}_0 \in \mathbb{R}^N$ contains the unknowns whose coordinates u_i approximate the values of the function u at the points x_i and similarly $f_i = f(x_i)$.

Example 1.2.12. To be precise we set

$$\mathcal{X} = C^4(0,1) \cap \{u \in C\,[0,1] : u(0) = u(1) = 0\}$$

with the maximum norm and $\mathcal{Y} = C^2(0,1)$ with the maximum norm, too. We mention that this choice is needed to gain the usual second order consistency (and with that the possibility of the second order convergence) but for the consistency (and for the convergence) instead of C^4 C^3 would be sufficient.

$$\mathcal{X}_n = \left\{\mathbf{u} \in \mathbb{R}^{N+2} : u_0 = u_{N+1} = 0\right\}$$

and $\mathcal{Y}_n = \mathbb{R}^N$, both with the maximum norm. $\varphi_n : \mathcal{X} \to \mathcal{X}_n$ and $\psi_n : \mathcal{Y} \to \mathcal{Y}_n$ are defined as $u \mapsto \mathbf{u} : u(x_i) = u_i$ and $f \mapsto \mathbf{f} : f(x_i) = f_i$, respectively.

Note that here $\mathcal{Y}$ is not a Banach space, but this is not very interesting from a practical point of view, since we want to define a convergent numerical method. To ensure convergence we need to show that the procedure described above is consistent and stable, c.f. Theorem 1.2.5 (Equivalence theorem, part 1).

Consistency can be obtained easily using the Taylor series theorem. The main task is to prove stability.

To prove stability the notions of Z- and M-matrix and related basic results are used, which can be found in the Appendix.

The matrix $\mathbf{K}_0 = (N+1)^2\text{tridiag}\,(-1,2,-1)$ is a Z-matrix, moreover it is a non-singular M-matrix. To show that it can be used the 2nd point of the Theorem 5.0.14 which is usually called "dominant vector condition". We choose $\mathbf{d}$ as $d_i = x_i(1-x_i)$, $i = 1,\dots,N$. Then $\mathbf{d} > \mathbf{0}$ and $\min(\mathbf{K}_0\mathbf{d})_i = 2$ hold independently of N.

Using Lemma 5.0.15, the choice $\mathbf{d} : d_i = x_i(1-x_i)$, $i = 1, \ldots, N$ means that $\|\mathbf{d}\|_\infty \leq \frac{1}{4}$ independently of N, hence

$$\|\mathbf{K}_0^{-1}\|_\infty \leq \frac{\|\mathbf{d}\|_\infty}{\min(\mathbf{K}_0\mathbf{d})_i} \leq \frac{\frac{1}{4}}{2} \leq \frac{1}{8} \tag{1.42}$$

holds independently of N, and this yields the stability.

Finally, returning to Example 1.2.12 we obtained that the FDM applied to the problem 1.40 is consistent and stable, and so it is convergent as well.

Remark 1.2.13. We obtained above that $\mathbf{K}_0^{-1}$ is a nonnegative matrix (i.e., each entries are nonnegative) and this has an important consequence. Namely,

$$\mathbf{f} \leq \mathbf{0} \Rightarrow \mathbf{u}_0 = \mathbf{K}_0^{-1}\mathbf{f} \leq \mathbf{0}\,. \tag{1.43}$$

This property is called discrete nonpositivity preservation property. If the data is nonpositive, then the solution is nonpositive as well.

It is important to note that the original equation (1.40) possesses this property (continuous nonpositivity preservation property), too. Naturally, a numerical method which can reflect this property is a better choice, than another one which lacks this property.

We note that the most important difference between the notions convergence and qualitative properties (such as the discrete nonpositivity preservation property) is as follows. Convergence is a property of a sequence and a qualitative property is related to one member of the sequence. However, it can have the same importance.

We gave a quick look at a qualitative property in order to prepare the Reader for the subject of the forthcoming chapters, which deal with maximum principles, a generalization of the nonpositivity preservation property.

We note that the FDM can be extended to nonhomogeneous Dirichlet boundary conditions. The problem

$$\begin{cases} -u''(x) = f(x)\,, \quad x \in (0,1)\,, \\ u(0) = a \\ u(1) = b \end{cases} \tag{1.44}$$

transforms under the FDM into the system of linear equations $\mathbf{Ku} = \mathbf{f}$, where $\mathbf{K} =$

$(\mathbf{K}_0|\mathbf{K}_\partial)$ with

$$\mathbf{K}_\partial = \begin{pmatrix} 0 & -1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ -1 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{u} = (\mathbf{u}_0|b,a)^T \ . \tag{1.45}$$

Naturally, consistency, stability and convergence can be verified similarly to the homogeneous case.

Problem 2. In this paragraph we will study the linear parabolic problem

$$\frac{\partial v}{\partial t} + Kv = f \quad \text{in } \Omega_T \tag{1.46}$$

with the *Dirichlet boundary condition*

$$v = g \quad \text{on } \partial\Omega \times [0,T] \tag{1.47}$$

and with the *initial condition*

$$v = v_0 \quad \text{on } \Omega \times \{t = 0\} \ , \tag{1.48}$$

where $\Omega_T = \Omega \times (0,T]$ for some fixed $T > 0$. As in the previous paragraph, $\Omega \subset \mathbb{R}^d$ is open and bounded with boundary $\partial\Omega$, $\overline{\Omega} = \Omega \cup \partial\Omega$. $u : \overline{\Omega}_T \to \mathbb{R}$, $v \equiv v(\mathbf{x},t)$ is the unknown, $f : \Omega_T \to \mathbb{R}$, $f \equiv f(\mathbf{x},t)$, $g : \partial\Omega \times [0,T] \to \mathbb{R}$, $g \equiv g(\mathbf{x},t)$ and $u_0 : \Omega \to \mathbb{R}$, $v_0 \equiv u_0(\mathbf{x})$ are given. The differential operator K is given in *divergence form* as

$$Kv = -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_j}\left(a_{ij}\frac{\partial v}{\partial x_i}\right) + \sum_{i=1}^{d} b_i \frac{\partial v}{\partial x_i} + cv \ , \tag{1.49}$$

with sufficiently smooth coefficient functions $a_{ij}(\mathbf{x},t), b_i(\mathbf{x},t), c(\mathbf{x},t) : \Omega_T \to \mathbb{R}$, $i,j = 1,\ldots,d$.

Remark 1.2.14. [10, Ch.7.1.1] In Rem.1.2.10 we mentioned that we can model physical processes with PDE's. The solution of the above defined problem can be interpreted as the time evolution of a chemical concentration (or the density of some quantity) within a region Ω.

Switching to semigroup viewpoint, we assume that the coefficients of the operator K are time-independent, and K generates a strongly continuous semigroup in the Banach space B. We choose $\mathcal{X} = (C([0,T],B), \|\cdot\|_\infty)$, $L : v(\cdot) \mapsto (v(0), \frac{dv}{dt} - Kv)$, $\mathcal{D} = \left\{v \in \mathcal{X} : \exists \frac{dv}{dt}, \frac{dv}{dt} - Kv \in F([0,T],B)\right\}$, $\mathcal{Y} = B \times F([0,T],B)$ and we assume that $f \in F([0,T],B)$, where $F = L^p$, $1 \le p \le \infty$ or $F = C$.

The Reader can find information about the well-posedness of the above defined parabolic problem in [44, Paragraph 3.2.] and in [10, Ch.7.1.1 and Thm.3 in Ch.5.9.2, Thm.3 and 4 in Ch.7.1.2c, Thm.5 in Ch.7.1.3].

Discretization with FEM + θ-method. For the sake of simplicity we assume a homogeneous Dirichlet boundary condition i.e. $g \equiv 0$ on $\partial\Omega \times [0,T]$. We choose $B = L^2(\Omega)$. By using the weak formulation

$$\left(\frac{\partial v}{\partial t}, w\right) - a(v,w) = (f,w)\,,$$

where $(\cdot,\cdot)$ denotes the L^2 inner product, and $a(\cdot,\cdot)$ is the bilinear form corresponding to K (defined similarly as in the elliptic case). Choosing a subspace (defined with the basis functions $\phi_i(\mathbf{x})$, $i = 1,\dots N$) we arrive at the equations

$$\sum_{i=1}^{N} \dot{v}_i(t)(\phi_i,\phi_j) - \sum_{i=1}^{N} v_i(t) a(\phi_i,\phi_j) = f_j(t)\,, \quad j = 1,\dots,N\,,$$

where $f_j(t) = (f(t),\phi_j)$, which can be written in the matrix form

$$\mathbf{M}_0 \dot{\mathbf{v}}_0(t) - \mathbf{K}_0 \mathbf{v}_0(t) = \mathbf{f}(t)\,,$$

where $\mathbf{v}_0(t) = (v_1(t),\dots,v_N(t))^T$, $\mathbf{f}(t) = (f_1(t),\dots,f_N(t))^T$, $M_{0ij} = (\phi_j,\phi_i)$ is the mass matrix and $K_{0ij} = a(\phi_j,\phi_i)$ is the stiffness matrix.

To obtain the fully discretized form from the semidiscrete form one possible option is to apply the θ-method.

$$\mathbf{M}_0 \frac{\mathbf{v}_0^{n+1} - \mathbf{v}_0^n}{\Delta t} = \theta \mathbf{K}_0 \mathbf{v}_0^{n+1} + (1-\theta)\mathbf{K}_0 \mathbf{v}_0^n + \mathbf{f}^{\theta,n+1}\,, \quad n = 0,\dots,M\,,$$

where a uniform mesh is used with $T = M\Delta t$, $\theta \in [0,1]$. $\mathbf{v}_0^n$ approximates $\mathbf{v}(n\Delta t)$, $\mathbf{f}^{\theta,n+1} = \theta\mathbf{f}((n+1)\Delta t) + (1-\theta)\mathbf{f}(n\Delta t)$ in case of $F = C$, and $\mathbf{f}^{\theta,n+1} = \dfrac{1}{\Delta t}\int\limits_{n\Delta t}^{(n+1)\Delta t} \mathbf{f}(t)\,\mathrm{d}t$ in case of $F = L^p$.

Using the notation $\mathbf{X}_{10} = \frac{1}{\Delta t}\mathbf{M}_0 - \theta\mathbf{K}_0$ and $\mathbf{X}_{20} = \frac{1}{\Delta t}\mathbf{M}_0 + (1-\theta)\mathbf{K}_0$ it can be rewritten as

$$\mathbf{X}_{10}\mathbf{v}_0^{n+1} - \mathbf{X}_{20}\mathbf{v}_0^n = \mathbf{f}^{\theta,n+1}\,, \quad n = 0,\dots,M\,.$$

In the following, for the sake of simplicity, we drop the superscript θ from the expression $\mathbf{f}^{\theta,n+1}$. Using the notations

$$\mathcal{L}_0 = \begin{pmatrix} \mathbf{I} & \mathbf{0} & \dots & \dots & \mathbf{0} \\ -\mathbf{X}_{20} & \mathbf{X}_{10} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & -\mathbf{X}_{20} & \mathbf{X}_{10} & \mathbf{0} & \dots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & -\mathbf{X}_{20} & \mathbf{X}_{10} \end{pmatrix}, \qquad \nu_0 = \begin{pmatrix} \mathbf{v}_0^0 \\ \mathbf{v}_0^1 \\ \vdots \\ \vdots \\ \mathbf{v}_0^M \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mathbf{v}_0^0 \\ \mathbf{f}^1 \\ \vdots \\ \vdots \\ \mathbf{f}^M \end{pmatrix},$$

it can be written in the compact form

$$\mathcal{L}_0 \nu_0 = \mu \,,$$

or

$$\nu_0 = \mathcal{L}_0^{-1} \mu \,,$$

where

$$\mathcal{L}_0^{-1} = \begin{pmatrix} \mathbf{I} & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{T} & \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{T}^2 & \mathbf{T} & \mathbf{I} & \mathbf{0} & \dots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{T}^M & \dots & \mathbf{T}^2 & \mathbf{T} & \mathbf{I} \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mathbf{v}_0^0 \\ \mathbf{0} \\ \vdots \\ \vdots \\ \mathbf{0} \end{pmatrix} \tag{1.50}$$

in case of $f \equiv 0$, otherwise

$$\mathcal{L}_0^{-1} = \begin{pmatrix} \mathbf{I} & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{T} & \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{T}^2 & \mathbf{T} & \mathbf{I} & \mathbf{0} & \dots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{T}^M & \dots & \mathbf{T}^2 & \mathbf{T} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{10}^{-1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X}_{10}^{-1} & \mathbf{0} & \dots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{X}_{10}^{-1} \end{pmatrix}. \tag{1.51}$$

We choose $\mathcal{X}_n = ((\mathbb{R}^N, \|\cdot\|_\triangleleft)^{M+1}, \|\cdot\|_\infty)$ and $\mathcal{Y}_n = ((\mathbb{R}^N, \|\cdot\|_\square)^{M+1}, \|\cdot\|_1)$. Stability means that

$$\sup\left\{ \|\mathbf{T}^i\|_{\square,\triangleleft} : 1 \le i \le M, \right\} < \infty$$

holds for all n (note that the notation n was omitted earlier, here the procedure was presented for a fixed n) in case of $f \equiv 0$. This is similar to the stability condition obtained in the famous paper [36] and in the paper [44] for the semidiscrete form.

This type of stability is usually called stability with respect to the initial data c.f. [47, Paragraph 2.3.]. In the general case stability means that

$$\sup\left\{\|\mathbf{T}^i\mathbf{X}_{10}^{-1}\|_{\square,\triangleleft} : 1 \leq i \leq M, \right\} < \infty\,.$$

holds for all n. This type of stability is usually called stability with respect to the initial data and to the right hand side, c.f. [47, Paragraph 2.4.].

Here the framework was presented for a homogeneous boundary condition, but it is extendible to the nonhomogeneous case. Moreover, here the FEM + θ method is used, but FEM can be substituted with e.g. FDM, too.

Summary of the chapter. In this chapter we gave a framework on the numerical treatment of approximating the solution of the equation $F(u) = 0$, where $\mathcal{X}$ and $\mathcal{Y}$ are normed spaces, $\mathcal{D} \subset \mathcal{X}$ and $F : \mathcal{D} \to \mathcal{Y}$ is a (nonlinear) operator. The framework was split into two parts, the first contained the general (nonlinear) case while the second contained the affine case. Both parts were based on Lax's idea, namely on the statement that consistency and stability implies convergence. Moreover, in the affine case stability and convergence are equivalent under the consistency assumption (Lax equivalence theorem).

Section 1.1 contained the nonlinear theory and this was based on the paper [23, Faragó, Mincsovics, Fekete, 2012]. Our framework contained a theoretical part, where we rephrased Stetter's results in order to fit it into our framework, and we illustrated the basic notions and results for the explicit Euler method, see Subsection 1.1.2. We extended the framework for applications, see Subsection 1.1.3, including our results, namely Lemma 1.1.34, Theorem 1.1.36 etc. Finally, in the general case we investigated the relation of the basic notions with numerous examples.

Section 1.2 contained the affine part of the framework. Here we compared the basic notions of this special case to the basic notions of the general case, and we gave an overview by using the results of Palencia and Sanz-Serna. Finally, we presented examples for the case where the framework was applied to elliptic and parabolic PDE's.

†

Chapter 2

Maximum principles

In this chapter we overview the most important pieces of informations on maximum principles based mainly on the book [10].

2.1 Elliptic maximum principles

In this section we list the definitions of continuous maximum principles for linear elliptic operators and the important theorems about them, based mainly on [10, Ch.6.4.1–Ch.6.4.3]. We study elliptic operators, and not elliptic PDE's, since this approach is more comfortable, and clearly the qualitative properties of some PDE's depend on the qualitative properties of the corresponding operators.

Let $\Omega \subset \mathbb{R}^d$ be an open and bounded domain with boundary $\partial\Omega$, and $\overline{\Omega} = \Omega \cup \partial\Omega$. We investigate the elliptic operator K, $\operatorname{dom} K = C^2(\Omega) \cap C(\overline{\Omega})$, defined in non-divergence form as

$$Ku = -\sum_{i,j=1}^{d} a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i \frac{\partial u}{\partial x_i} + cu\,, \tag{2.1}$$

where $a_{ij}(\mathbf{x}), b_i(\mathbf{x}), c(\mathbf{x}) \in C(\overline{\Omega})$, moreover, for the sake of simplicity we assume that $a_{ij}(\mathbf{x}) \in C^1(\Omega)$, which enables us to rewrite the non-divergence form to divergence form and vice versa, c.f. (1.29).

The family of maximum principles consist of many members, the most known are the non-positivity preservation property (which was mentioned already earlier) and the weak and strong maximum principle. Here, besides these maximum principles we investigate less frequently used ones including newly introduced, too. This is done with the purpose to make the discussion clearer. Their similarity will provide the possibility

we can exploit mostly in Section 3 in the proofs. Thus, in other words, we can consider the family of maximum principles as variations on a theme.

First we define the weak and strong non-positivity preservation properties.

Definition 2.1.1. We say that the operator K, defined in (2.1), possesses

- the *weak non-positivity preservation property* (nP) if the following implication holds:

$$Ku \leq 0 \text{ in } \Omega\,, \quad \max_{\partial\Omega} u \leq 0 \quad \Rightarrow \quad \max_{\overline{\Omega}} u \leq 0\,. \tag{2.2}$$

- the *strong non-positivity preservation property* (NP) if it possesses the nP, moreover, the following implication holds:

$$Ku \leq 0 \text{ in } \Omega \quad \text{and} \quad \max_{\Omega} u = \max_{\overline{\Omega}} u = 0 \quad \Rightarrow \quad u \equiv 0 \text{ in } \overline{\Omega}\,. \tag{2.3}$$

We could call these two the "parents" in the family of maximum principles. These are clearly maximum principles and with a (relatively) mild expectation. For those operators possessing the nP we can give an upper bound (which is 0) for the function u under some conditions, namely the K-image of u is non-positive and u is non-positive at the boundary. For those operators possessing the NP we can state that if the K-image of u is non-positive, and u attains its maximum at an interior point, and this maximum is 0, then $u \equiv 0$.

To define further and less mild maximum principles we proceed in the following way. We push some condition from the left side of the implication (2.2) to the right side resulting in something like this: $Ku \leq 0$ in $\Omega \Rightarrow \max_{\overline{\Omega}} u \leq \max\{0, \max_{\partial\Omega} u\}$. It means that for those operators fulfilling this principle, if the K-image of u is non-positive, then u is bounded from above, and this bound is defined by the boundary values of u and the zero, more precisely, u attains its non-negative maximum at the boundary. We could make this notion more restrictive omitting the 0 from the upper bound $\max\{0, \max_{\partial\Omega} u\}$. (This means that for those operators fulfilling this principle, if the K-image of u is non-positive then u attains its maximum at the boundary.) Naturally, we want to proceed similarly with the implication (2.3), but in this case we modify the right side of it.

We summarize these "descendants" in the following definition.

Definition 2.1.2. We say that the operator K, defined in (2.1), possesses

- the *weak maximum principle* (wMP) if the following implication holds:

$$Ku \leq 0 \text{ in } \Omega \quad \Rightarrow \quad \max_{\overline{\Omega}} u \leq \max\{0, \max_{\partial\Omega} u\}\,;$$

- the *strictly weak maximum principle* (WMP) if the following implication holds:

$$Ku \leq 0 \text{ in } \Omega \quad \Rightarrow \quad \max_{\overline{\Omega}} u = \max_{\partial\Omega} u\,;$$

- the *strong maximum principle* (sMP) if it possesses the wMP, moreover, the following implication holds:

$$Ku \leq 0 \text{ in } \Omega \quad \text{and} \quad \max_{\Omega} u = \max_{\overline{\Omega}} u = m \geq 0 \quad \Rightarrow \quad u \equiv m \text{ in } \overline{\Omega}\,;$$

- the *strictly strong maximum principle* (SMP) if it possesses the WMP, moreover, the following implication holds:

$$Ku \leq 0 \text{ in } \Omega \quad \text{and} \quad \max_{\Omega} u = \max_{\overline{\Omega}} u = m \quad \Rightarrow \quad u \equiv m \text{ in } \overline{\Omega}\,.$$

We note that in the definition of the sMP and SMP m is a constant. The meaning of SMP (sMP) is the following. For those operators fulfilling this principle, if the K-image of u is non-positive and u attains its (non-negative) maximum at an interior point, then u is a constant function. We can see that the operator $-\Delta$ possesses all the above defined maximum principles when Ω is connected.

Remark 2.1.3. We make some comments on the above defined maximum principles.

- It is clear that the relation of the above defined notions are the following.

$$\begin{array}{ccccc} \text{WMP} & \Rightarrow & \text{wMP} & \Rightarrow & \text{nP} \\ \Uparrow & & \Uparrow & & \Uparrow \\ \text{SMP} & \Rightarrow & \text{sMP} & \Rightarrow & \text{NP} \end{array}$$

- Sometimes the case $c = 0$ is called strong elliptic maximum principle, see e.g. [10], but we wanted to reserve this name to another property.

- We mention that it is possible to define minimum principles similarly. E.g. the weak minimum principle (the twin of the wMP) reads as $Ku \leq 0$ in $\Omega \quad \Rightarrow \min_{\overline{\Omega}} u \geq \min\{0, \min_{\partial\Omega} u\}$. However, due to the linearity of the operator K, it requires the same restriction for an operator to fulfil it.

- To define maximum principles we followed a recipe. The Reader could ask whether we could make it further, pushing Ku, too, somehow to the right side of the implication of the wMP. This can be done and it can be found in the series of papers [20, 21, 22], and in collected form in [52] (however, we note that these papers discuss only a case of a special operator).

- We defined maximum principles for those operators whose domain is $\operatorname{dom} K = C^2(\Omega) \cap C(\overline{\Omega})$. It is possible to proceed similarly for a wider class of operators, namely, for those defined on $H^1(\Omega)$ (containing less smooth functions). This can be found in [57].

We collected the results on maximum principles in the following theorem.

Theorem 2.1.4. *[10, Thm.2. and Thm.1. in Ch.6.4.1, Thm.4. and Thm.3. in Ch.6.4.2] If operator K, defined in (2.1), is uniformly elliptic and*

- $c \geq 0$, *then it possesses the wMP;*
- $c = 0$, *then it possesses the WMP;*
- $c \geq 0$, *moreover Ω is connected, then it possesses the sMP;*
- $c = 0$, *moreover Ω is connected, then it possesses the SMP.*

Remark 2.1.5. We make some comments on this result.

- The definition of uniform ellipticity can be found in the Appendix.
- $c \geq 0$ is not necessary for the wMP and for the sMP.
- The requirements under which the operator possesses a weak maximum principle can be weakened, see, e.g. [5].
- One can see that the connectedness of Ω is necessary, too, for the sMP and SMP as well.

The Reader can find more information about maximum and minimum principles in [10, Ch.6.4.1–Ch.6.4.3].

2.2 Parabolic maximum principles

In this section we could proceed similarly to the elliptic case, namely, we could introduce a whole family of maximum principles, which is more plentiful in members. However, here we restrict ourselves to the most important ones, only. Besides this we skip the details (which are similar to the elliptic case), thus we switch to the brief style. This section is based mainly on [10, Ch.7.1.4]. To a more concise style introduction containing various types of parabolic maximum principles we recommend the works [11, 17, 19] besides [10].

We assume that the domain $\Omega \subset \mathbb{R}^d$ is open and bounded with boundary $\partial\Omega$, as before. Let T be a positive real number. For $t \in (0,T]$ we introduce the notations $Q_t = \Omega\times(0,t)$, $\overline{Q_t} = \overline{\Omega}\times[0,t]$ and $\Gamma_t = (\partial\Omega\times[0,t])\cup(\Omega\times\{0\})$ for a piece of the parabolic boundary. We investigate the parabolic operator L, $\operatorname{dom} L = C^{2,1}(Q_T)\cap C(\overline{Q_t})$ – where the symbol $C^{2,1}$ means: twice continuously differentiable with respect to the space variable and continuously differentiable with respect to the time variable – defined in non-divergence form as

$$Lv = \frac{\partial v}{\partial t} - \sum_{i,j=1}^{d} a_{ij}\frac{\partial^2 v}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i \frac{\partial v}{\partial x_i} + cv, \tag{2.4}$$

where $a_{ij}(\mathbf{x},t), b_i(\mathbf{x},t), c(\mathbf{x},t) \in C(\overline{\Omega}\times[0,T])$ and $a_{ij}(\mathbf{x},t)$ is continuously differentiable with respect to the space variable, this enables us to rewrite the non-divergence form to divergence form and vice versa, c.f., the paragraph "Problem 2" in the last chapter.

Definition 2.2.1. We say that the operator L, defined in (2.4), possesses

- the *non-positivity preservation property* (nP) if the following implication holds for all $t \in (0,T]$:

$$Lv \le 0 \text{ in } Q_t\,, \quad \max_{\Gamma_t} v \le 0 \quad \Rightarrow \quad \max_{\overline{Q_t}} v \le 0\,. \tag{2.5}$$

- the *maximum principle* (mP) if the following implication holds for all $t \in (0,T]$:

$$Lv \le 0 \text{ in } Q_t \quad \Rightarrow \quad \max_{\overline{Q_t}} v \le \max\{0, \max_{\Gamma_t} v\}\,. \tag{2.6}$$

- the *strict maximum principle* (MP) if the following implication holds for all $t \in (0,T]$:

$$Lv \le 0 \text{ in } Q_t \quad \Rightarrow \quad \max_{\overline{Q_t}} v = \max_{\Gamma_t} v\,. \tag{2.7}$$

We can see that these maximum principles are of the week type (c.f. the elliptic maximum principles) but we omitted this attribute in order to simplify the naming procedure and the notations. It is clear that their relation can be explained with the same recipe as the construction of the family of elliptic maximum principles.

Theorem 2.2.2. *[10, Ch.7.1.4, Thm.9. and Thm.8.] If operator L, defined in (2.4), is uniformly parabolic and*

- *$c \ge 0$, then it possesses the mP (and the nP);*
- *$c = 0$, then it possesses the MP (and both the mP and the nP).*

The definition of uniform parabolicity can be found in the Appendix.

Summary of the chapter. In this chapter an overview on elliptic and parabolic maximum principles was presented based mostly on the book [10]. From didactical considerations we introduced a new notion: the weak non-positivity preservation property (nP).

†

Chapter 3

Discrete elliptic maximum principles

In this chapter we present an algebraic framework for discrete maximum principles for matrices where we define these in accordance with the continuous case and we investigate their applicability. We give algebraic results on discrete maximum principles and we present numerical examples demonstrating the differences between them. These results are (mostly) from the paper [41]. Finally we end this chapter with a thorough investigation of how we can handle a discrete maximum principle when a discontinuous Galerkin method is applied as discretization on a special operator. This final part is based on the paper [28].

3.1 Algebraic framework

3.1.1 Discrete elliptic maximum principles

First we introduce some notations. We use the following typesetting: $\mathbf{A}$ for matrices, $\mathbf{a}$ for vectors. $\mathbf{0}$ denotes the zero matrix (or vector), $\mathbf{e}$ is the vector all coordinates of which are equal to 1. The dimensions of these vectors and matrices should be clear from the context. $\mathbf{A} \geq \mathbf{0}$ ($\mathbf{A} > \mathbf{0}$) or $\mathbf{a} \geq \mathbf{0}$ ($\mathbf{a} > \mathbf{0}$) means that all the elements of $\mathbf{A}$ or $\mathbf{a}$ are non-negative (positive). The symbol $\max \mathbf{a}$ stands for the maximal element of the vector $\mathbf{a}$ and $\max\{0, \mathbf{a}\}$ denotes $\max\{0, \max \mathbf{a}\}$.

We will use the notions of different types of matrices, such as Z-, M-, irreducible, diagonally dominant (DD), irreducibly diagonally dominant (IDD) and Stieltjes matrix. All of these notions and related basic results can be found in the Appendix.

In the following we define discrete maximum principles for a discrete operator, i.e.,

for a matrix in the partitioned form

$$\mathbf{K} = (\mathbf{K}_0|\mathbf{K}_\partial) \in \mathbb{R}^{N\times\overline{N}}\,, \tag{3.1}$$

where $\mathbf{K}_0 \in \mathbb{R}^{N\times N}$, $\mathbf{K}_\partial \in \mathbb{R}^{N\times N_\partial}$, $\overline{N} = N + N_\partial$, acting on the vector

$$\mathbf{u} = (\mathbf{u}_0|\mathbf{u}_\partial)^T \in \mathbb{R}^{\overline{N}}\,, \tag{3.2}$$

where $\mathbf{u}_0 \in \mathbb{R}^N$, $\mathbf{u}_\partial \in \mathbb{R}^{N_\partial}$. We assume that $N, N_\partial \geq 1$.

We choose the natural (which is at the same time the simplest) way to define discrete maximum principles for this matrix. Later, in Subsection 3.1.3 we investigate the applicability of the definitions in the light of different discretization methods.

The natural way means the following.

Definition 3.1.1. We say that the matrix $\mathbf{K}$, given in the form (3.1), possesses

- the *discrete weak non-positivity preservation property* (DnP) if the following implication holds:

$$\mathbf{Ku} \leq \mathbf{0}\,, \quad \max \mathbf{u}_\partial \leq 0 \quad \Rightarrow \quad \max \mathbf{u} \leq 0\,.$$

- the *discrete strong non-positivity preservation property* (DNP) if it possesses the DnP, moreover, the following implication holds:

$$\mathbf{Ku} \leq \mathbf{0} \quad \text{and} \quad \max \mathbf{u} = \max \mathbf{u}_0 = 0 \quad \Rightarrow \quad \mathbf{u} = \mathbf{0}\,.$$

Definition 3.1.2. We say that the matrix $\mathbf{K}$, given in the form (3.1), possesses

- the *discrete weak maximum principle* (DwMP) if the following implication holds:

$$\mathbf{Ku} \leq \mathbf{0} \quad \Rightarrow \quad \max \mathbf{u} \leq \max\{0, \mathbf{u}_\partial\}\,; \tag{3.3}$$

- the *discrete strictly weak maximum principle* (DWMP) if the following implication holds:

$$\mathbf{Ku} \leq \mathbf{0} \quad \Rightarrow \quad \max \mathbf{u} = \max \mathbf{u}_\partial\,;$$

- the *discrete strong maximum principle* (DsMP) if it possesses the DwMP, moreover, the following implication holds:

$$\mathbf{Ku} \leq \mathbf{0} \quad \text{and} \quad \max \mathbf{u} = \max \mathbf{u}_0 = m \geq 0 \quad \Rightarrow \quad \mathbf{u} = m\mathbf{e}\,;$$

- the *discrete strictly strong maximum principle* (DSMP) if it possesses the DWMP, moreover, the following implication holds:

$$\mathbf{Ku} \leq \mathbf{0} \quad \text{and} \quad \max \mathbf{u} = \max \mathbf{u}_0 = m \quad \Rightarrow \quad \mathbf{u} = m\mathbf{e}\,.$$

Here we also note (as in the continuous case) that m is a real number, representing the value of the maximal entry of the vector $\mathbf{u}$. These definitions correspond clearly to the Definitions 2.1.1 and 2.1.2. The relation between these discrete maximum principles is the same as that between the corresponding continuous ones.

$$\begin{array}{ccccc} \text{DWMP} & \Rightarrow & \text{DwMP} & \Rightarrow & \text{DnP} \\ \Uparrow & & \Uparrow & & \Uparrow \\ \text{DSMP} & \Rightarrow & \text{DsMP} & \Rightarrow & \text{DNP} \end{array}$$

Remark 3.1.3. However, there are other ways to define discrete maximum principles. About this we collected some information.

- The first paper in which a discrete maximum principle was formulated is probably [56], but that definition given there contains $\mathbf{Ku} = \mathbf{0}$ at the left side of the implication (3.3) instead of $\mathbf{Ku} \leq \mathbf{0}$. On the other hand, $\mathbf{K}$ was allowed to have complex entries.

- The definition of the discrete weak maximum principle which is used today (in the same form as we defined it) appeared first in [5] (but it was named differently).

- In Remark 2.1.3 we mentioned that it is possible to define more restrictive continuous maximum principles following the previously given recipe further. In the works [20, 21, 22], collected in [52], the Reader can find information about a discrete case, too.

- There are other types of discrete maximum principles based on other continuous models. We mention the papers [49, 50], which contain the definition of a discrete maximum principle suitable for input-output models. In [50] the connection of the two different discrete maximum principles is investigated, too.

3.1.2 Algebraic results on discrete elliptic maximum principles

Our aim is to give necessary and sufficient conditions for the above defined discrete maximum principles, moreover, by means of which we would also like to shed some

light on the relations and differences between them. Naturally, we also touch upon useful practical conditions which can be useful from an application point of view.

We begin with the DnP and the DNP.

Lemma 3.1.4. *The matrix* $\mathbf{K}$ *(given in the form* (3.1)*) possesses the DnP if and only if the following two conditions hold:*

(n1) $\mathbf{K}_0^{-1} \geq \mathbf{0}$; (n2) $-\mathbf{K}_0^{-1}\mathbf{K}_\partial \geq \mathbf{0}$.

Proof. – First, we assume (n1)–(n2) and $\mathbf{Ku} \leq \mathbf{0}$, $\mathbf{u}_\partial \leq \mathbf{0}$. Then $\mathbf{K}_0^{-1}$ exists by (n1) and we can use the identity

$$\mathbf{u}_0 = \mathbf{K}_0^{-1}\mathbf{Ku} - \mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{u}_\partial\,, \tag{3.4}$$

which gives immediately $\mathbf{u}_0 \leq \mathbf{0}$, the required relation of the DnP.

– Second, we assume the validity of the DnP. We use the setting $\mathbf{Ku} = \mathbf{0}$, $\mathbf{u}_\partial = \mathbf{0}$, which results in $\max \mathbf{u}_0 \leq 0$. We use the same setting in $-\mathbf{Ku} = \mathbf{K}(-\mathbf{u})$ in order to get $\max -\mathbf{u}_0 \leq 0$, thus $\ker \mathbf{K}_0 = \{\mathbf{0}\}$, and this means that the identity (3.4) can be applied. Then (n1) follows from setting $\mathbf{u}_\partial = \mathbf{0}$, while (n2) follows from setting $\mathbf{Ku} = \mathbf{0}$.

□

Lemma 3.1.5. *We assume that* $N \geq 2$. *The matrix* $\mathbf{K}$ *(given in the form* (3.1)*) possesses the DNP if and only if the following two conditions hold:*

(N1) $\mathbf{K}_0^{-1} > \mathbf{0}$; (N2) $-\mathbf{K}_0^{-1}\mathbf{K}_\partial > \mathbf{0}$.

Proof. – First, we assume (N1)–(N2). We have to show that the relations $\mathbf{Ku} \leq \mathbf{0}$ and $\max \mathbf{u} = \max \mathbf{u}_0 = 0$ together imply $\mathbf{u} = \mathbf{0}$. Then $\mathbf{u}_0 \leq \mathbf{0}$ have a 0 coordinate. Using the identity (3.4), (N1)–(N2) and the fact that $\mathbf{u}_0$ has a 0 coordinate yields that $\mathbf{Ku} = \mathbf{0}$ and $\mathbf{u}_\partial = \mathbf{0}$. These imply $\mathbf{u}_0 = \mathbf{0}$.

– Second, we assume the DNP. Then the DnP holds, thus (n1)–(n2) hold. We can choose freely $\mathbf{Ku} \leq \mathbf{0}$, $\mathbf{u}_\partial \leq \mathbf{0}$ in (3.4).

First, we set $\mathbf{u}_\partial = \mathbf{0}$ and we assume that $\mathbf{K}_0^{-1}$ has a 0 element, let it be the ij-th entry of the matrix. We choose the j-th coordinate of $\mathbf{Ku}$ as -1, the others as 0, then the i-th coordinate of $\mathbf{u}_0$ is 0. If in the j-th column of $\mathbf{K}_0^{-1}$ there is a positive entry, then $\mathbf{u}_0 \neq \mathbf{0}$, which is a contradiction. Otherwise, the matrix $\mathbf{K}_0^{-1}$ has a zero column, which is a contradiction, too, since it is invertible. Thus, we have proven that (N1) holds.

Second, we set $\mathbf{Ku} = \mathbf{0}$, and assume that $-\mathbf{K}_0^{-1}\mathbf{K}_\partial$ has a 0 element, let it be the ij-th entry of the matrix. We choose the j-th coordinate of $\mathbf{u}_\partial$ as -1, the others as 0, then the i-th coordinate of $\mathbf{u}_0$ is 0, but $\mathbf{u}_\partial \neq \mathbf{0}$, which is a contradiction. Thus, we have proven that (N2) holds, too.

□

Note that the following proofs in this section will be similar to the proofs of Lemma 3.1.4 and 3.1.5. Next we investigate the DwMP. The next lemma was first proven by Ciarlet, but we give here a slightly different proof exploiting Lemma 3.1.4.

Lemma 3.1.6. *[5] The matrix* $\mathbf{K}$ *possesses the DwMP if and only if the following three conditions hold:*

(w1) $\mathbf{K}_0^{-1} \geq \mathbf{0}$; (w2) $-\mathbf{K}_0^{-1}\mathbf{K}_\partial \geq \mathbf{0}$; (w3) $-\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e} \leq \mathbf{e}$.

Proof. We can observe that (w1) and (w2) are identical with (n1) and (n2).

– First we assume (w1)–(w3), then

$$\mathbf{Ku} \leq \mathbf{0} \quad \Rightarrow \quad \mathbf{u}_0 \leq -\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{u}_\partial \leq -\mathbf{K}_0^{-1}\mathbf{K}_\partial \max\{0, \mathbf{u}_\partial\}\mathbf{e} \leq \max\{0, \mathbf{u}_\partial\}\mathbf{e}.$$

– Second, to prove the reverse direction we assume the DwMP. DwMP implies DnP and that gives (w1) and (w2) (≡ (n1) and (n2)). (w3) follows from putting $\mathbf{Ku} = \mathbf{0}$, $\mathbf{u}_\partial = \mathbf{e}$ in (3.4).

□

Earlier in Chapter 2 we created the definition of wMP from the definition of nP following a recipe. Now we present a useful result in order to explain this recipe (and relation) from a deeper point of view.

Lemma 3.1.7. *[11, L.2.3.26] The matrix* $\mathbf{K}$ *possesses the DwMP if and only if the following two implication hold.*

$$\mathbf{Ku} \leq \mathbf{0}, \quad \mathbf{u}_\partial \leq \mathbf{0} \quad \Rightarrow \quad \max \mathbf{u}_0 \leq 0$$

and

$$\mathbf{Ku} \leq \mathbf{0}, \quad \mathbf{u}_\partial \geq \mathbf{0} \quad \Rightarrow \quad \max \mathbf{u}_0 \leq \max \mathbf{u}_\partial .$$

Proof. It is trivial that the DwMP implies the two implications given above. The converse is almost trivial because the first implication is equivalent to (w1) and (w2), since it is the DnP. To prove (w3) we set $\mathbf{u}_0 = -\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e}$ and $\mathbf{u}_\partial = \mathbf{e}$. With this setting we can apply the second implication, since $\mathbf{Ku} = \mathbf{0}$, thus $\max -\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e} \leq 1$, and this is exactly (w3). □

Practical algebraic conditions for the DwMP. Lemma 3.1.6 is a theoretical result which cannot usually be applied directly. There are several ways to get practical conditions under which the DwMP holds.

- The condition (w2) is usually replaced by the assumption

  (w2') $\mathbf{K}_\partial \leq \mathbf{0}$,

  as suggested in [5]. Then (w2') with (w1) clearly implies (w2), but the converse is not true.

- The condition (w3) is usually replaced by the assumption

  (w3') $\mathbf{Ke} \geq \mathbf{0}$,

  as suggested, also in [5]. (w3') with (w1) implies (w3) but the converse is not true again. (w3') corresponds to $0 \leq c$ c.f. with the continuous case.

- But the major task is to guarantee (w1). Inverse non-negativity is a more difficult notion. In most cases (w1) is relaxed by

  (w1') $\mathbf{K}_0$ is a non-singular M-matrix.

  But (w1') in this form is a theoretic condition, too. [5] (see also [34, Thm.1.9]) gives the condition

  (w1'a) $\mathbf{K}_0$ is an IDD Z-matrix with positive diagonal entries.

  (w1'a) implies $\mathbf{K}_0^{-1} > \mathbf{0}$, see [55, Cor. 3.20.]. Both the assumption and the result seem to be too much. Actually,

  (w1'b) $\mathbf{K}_0$ is an irreducible DD non-singular Z-matrix

  is enough to guarantee $\mathbf{K}_0^{-1} > \mathbf{0}$. These can be proven using [3, Thm. 2.7. in Ch. 6.2.] (see the Appendix). We can generalize this result with

  (w1'bb) $\mathbf{K}_0$ consists of diagonal blocks with the property (w1'b) (or (w1'a)) (elsewhere 0)

c.f. [34, argumentation below the Thm.1.9, and Ex.1.13]. Thus we can see that irreducibility is far not necessary and assuming this we get the "only" required condition $\mathbf{K}_0^{-1} \geq \mathbf{0}$.

In [25] (see [8], too) the condition

(w1'c) $\mathbf{K}_0$ is a Stieltjes matrix

is proposed since (w1'a) seems to be too restrictive in some cases in the practice. Naturally, it is recommended only if K is symmetric.

We can use the "dominant-vector" condition

(w1'd) $\mathbf{K}_0$ is a Z-matrix for which $\exists \mathbf{v} > \mathbf{0}$ with $\mathbf{K}_0\mathbf{v} > \mathbf{0}$,

too, as it is demonstrated in Subsection 3.3.2. We note that this condition is equivalent to (w1') in fact, see [3, Thm.2.3 in Ch.6.2] (see the Appendix).

$\mathbf{K}_0$ does not need to be an M-matrix. For other possibilities see [25] and the references therein.

After this, the Reader might think that it is needed to choose a triplet of practical conditions in order to guarantee the DwMP. This is right, but we note that the listed practical conditions are not entirely independent from each other. E.g., if we choose the following triplet

(w1'b–) $\mathbf{K}_0$ is an irreducible Z-matrix

(w2'+) $\mathbf{K}_\partial \lneq \mathbf{0}$

(w3') $\mathbf{K}\mathbf{e} \geq \mathbf{0}$,

then (w2'+) and (w3') "can help" the condition (w1'b–), since then $\mathbf{K}_0$ is IDD, too.

Note that the condition (w2'+) is wholly natural in practical situations and we will see later that a similar condition (s2') plays an important role in order to guarantee the discrete strong maximum principles.

We are going further with the DWMP.

Theorem 3.1.8. *[11, L.2.3.29 and L.2.3.30] or [12] The matrix* $\mathbf{K}$ *possesses the DWMP if and only if the following three conditions hold:*

(W1) $\mathbf{K}_0^{-1} \geq \mathbf{0}$; (W2) $-\mathbf{K}_0^{-1}\mathbf{K}_\partial \geq \mathbf{0}$; (W3) $-\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e} = \mathbf{e}$.

Proof. We can observe that (W1) and (W2) are identical with (w1) and (w2) (with (n1) and (n2), too).

– We assume (W1)–(W3) and $\mathbf{Ku} \leq \mathbf{0}$. Then

$$\mathbf{u}_0 = \mathbf{K}_0^{-1}\mathbf{Ku} - \mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{u}_\partial \leq -\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{u}_\partial \leq -\mathbf{K}_0^{-1}\mathbf{K}_\partial(\max \mathbf{u}_\partial)\mathbf{e} = (\max \mathbf{u}_\partial)\mathbf{e}\,,$$

which implies $\max \mathbf{u}_0 \leq \max \mathbf{u}_\partial$.

– In the reverse direction it is clear that (W1) and (W2) ($\equiv$ (w1) and (w2)) holds since DWMP implies DwMP.

To prove (W3) we poceed similarly as in the proof of Lemma (3.1.7). First, we set $\mathbf{u}_0 = -\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e}$ and $\mathbf{u}_\partial = \mathbf{e}$. With this setting we can apply the implication of the definition of DWMP, since $\mathbf{Ku} = \mathbf{0}$, thus $\max -\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e} \leq 1$. Second, we set $\mathbf{u}_0 = \mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e}$ and $\mathbf{u}_\partial = -\mathbf{e}$. With this setting we get $\max \mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e} \leq -1$, which is equivalent to $\min -\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e} \geq 1$. Finally, the relation

$$1 \leq \min -\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e} \leq \max -\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e} \leq 1$$

implies (W3).

□

- Note that (W3) is equivalent to $\mathbf{Ke} = \mathbf{0}$, and this corresponds to $c = 0$ c.f. the continuous case.

- In [34, Thm.1.10] it was proven that (w1'bb), (w2'), (W3) imply the DWMP, thus they form a practical triplet of conditions to guarantee it. (It is trivial, since we saw earlier that (w1'bb) implies (W1), and (w2') with (W1) implies (W2).)

We complete this part with the two strong maximum principles.

Theorem 3.1.9. *We assume that $N \geq 2$. The matrix $\mathbf{K}$ possesses the DSMP if and only if the following three conditions hold:*

(S1) $\mathbf{K}_0^{-1} > \mathbf{0}$; (S2) $-\mathbf{K}_0^{-1}\mathbf{K}_\partial > \mathbf{0}$; (S3) $-\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e} = \mathbf{e}$.

Proof. Note that (S1) and (S2) are identical with (N1) and (N2), moreover, (S3) are identical with (W3).

– First, we assume (S1)–(S3), $\mathbf{Ku} \leq \mathbf{0}$ and $\max \mathbf{u} = \max \mathbf{u}_0 = m$. We write $\mathbf{u}_0 = m\mathbf{e} - \mathbf{h}_0$, $\mathbf{u}_\partial = m\mathbf{e} - \mathbf{h}_\partial$, where both $\mathbf{h}_0, \mathbf{h}_\partial \geq \mathbf{0}$ have a 0 coordinate. We put these into the identity (3.4) resulting in

$$m\mathbf{e} - \mathbf{h}_0 = \mathbf{K}_0^{-1}\mathbf{Ku} - \mathbf{K}_0^{-1}\mathbf{K}_\partial m\mathbf{e} + \mathbf{K}_0^{-1}\mathbf{K}_\partial \mathbf{h}_\partial. \tag{3.5}$$

Using (S3) we get

$$-\mathbf{h}_0 = \mathbf{K}_0^{-1}(\mathbf{Ku}) - \mathbf{K}_0^{-1}\mathbf{K}_\partial(-\mathbf{h}_\partial)\,. \tag{3.6}$$

Using (S1), (S2) and the fact that $\mathbf{h}_0$ has a 0 coordinate yields that $\mathbf{Ku} = \mathbf{0}$ and $\mathbf{h}_\partial = \mathbf{0}$. These imply $\mathbf{h}_0 = \mathbf{0}$.

– Second, we assume the DSMP. DSMP implies both of DNP and DWMP, thus (S1)–(S2) (≡ (N1)–(N2)) and (S3) (≡ (W3)) hold, too.

□

Theorem 3.1.10. *We assume that $N \geq 2$. The matrix $\mathbf{K}$ possesses the DsMP if and only if the following three conditions hold:*

(s1) $\mathbf{K}_0^{-1} > \mathbf{0}$; (s2) $-\mathbf{K}_0^{-1}\mathbf{K}_\partial > \mathbf{0}$;
(s3) $-\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e} < \mathbf{e}$ *or* $-\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e} = \mathbf{e}$.

Proof. Note that (s1) and (s2) are identical with (N1) and (N2).

– First, we assume (s1)–(s3).

If $-\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e} = \mathbf{e}$ holds, then we can adopt the proof of the DSMP case.

If $-\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e} < \mathbf{e}$ holds and $m = 0$, then we can adopt the proof of the DSMP case again.

If $-\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e} < \mathbf{e}$ holds and $m > 0$, then (3.6) is modified as

$$-\mathbf{h}_0 < \mathbf{K}_0^{-1}(\mathbf{Ku}) - \mathbf{K}_0^{-1}\mathbf{K}_\partial(-\mathbf{h}_\partial)\,,$$

which excludes the possibility that $\mathbf{h}_0$ has a 0 coordinate. (This means that the left side of the implication in the definition of the DsMP is never fulfilled and consequently it is always true.)

– Second, we assume the DsMP. DsMP implies DNP, thus (s1)–(s2) (≡ (N1)–(N2)) hold. DsMP implies DwMP, too, thus (w3) holds, which can be rewritten as $\mathbf{e} + \mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e} \geq \mathbf{0}$.

To get (s3), we assume that $-\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e} \not\leq \mathbf{e}$ and $-\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e} \neq \mathbf{e}$, i.e., $\mathbf{e}+\mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e}$ has a 0 and a positive coordinate, too. Choosing $m = 1$, $\mathbf{Ku} = \mathbf{0}$, $\mathbf{h}_\partial = \mathbf{0}$ in (3.5) yields

$$\mathbf{h}_0 = \mathbf{e} + \mathbf{K}_0^{-1}\mathbf{K}_\partial\mathbf{e}\,,$$

thus $\mathbf{h}_0$ has a 0 and a positive coordinate, too, which is a contradiction.

□

Practical algebraic conditions for the discrete strong maximum principles. While the discrete weak maximum principle was extensively investigated in the last decades, the discrete strong maximum principle has not been thoroughly analysed.

- As we saw earlier, (w1'a) or (w1'b) is a sufficient condition for (s1) (≡(S1)).
- We can see that (s1) implies the irreducibility property of $\mathbf{K}_0$. Irreducibility can be interpreted as that all the discrete interior points are in contact with each other, which is clearly some discrete interior connectedness property. (C.f. the continuous case.)
- To ensure (s2) (≡(S2)), one possibility is to require

  (s2') $\mathbf{K}_\partial \leq \mathbf{0}$ and at least one non-zero element in every column.

  (s2') can be interpreted that all of the discrete boundary points are in contact with the discrete interior points, which is in some sense some discrete "boundary" connectedness property. (s2') with (s1) implies (s2).
- The only difference between the conditions in Theorem 2.1.4 for the weak and strong maximum principles is the connectedness of the domain Ω. (That theorem gave only sufficient conditions for the different maximum principles.) Now, we have seen that connectedness plays an important role in the discrete case, too.
- (s3) can be replaced by (w3'). This is based on the following. (w3') with (s1) implies (s3). (The converse implication is naturally not true.)

We can conclude that irreducibility is necessary for DsMP and DSMP (but it is not sufficient). Anyway, this would be the key-concept if we want something to emphasize.

- Probably the first paper about strong maximum principles was [30]. In that paper it was proven that (w1'a), (s2'), (w3') imply the DsMP. We note that the same theorem can be found in [34, Thm.1.9].

- In [8] the non-negativity and positivity of the discrete Green function were investigated (and illustrated with interesting numerical examples), which is in close relation with our topic, namely, the non-negativity of the discrete Green function means (w1)–(w2), and the positivity of the discrete Green function means (s1)–(s2).

- Finally, [41] gave necessary and sufficient algebraic conditions for the DsMP and for the DSMP.

3.1.3 Applicability of the framework

We have defined the discrete maximum principles in a natural way for a self-standing discrete operator independently of the original continuous operator. Now we are going to investigate the applicability of these definitions in the light of different discretization methods. This is to be understood as follows. Usually we seek a solution of a given continuous problem $Ku = f$ (where K is in the form (2.1)). But we only look for an approximation of the solution by solving a simpler problem (usually it is a linear algebraic system of equations), because to solve the continuous problem directly is hard or even impossible. In order to construct a simpler problem in the form $\mathbf{Ku} = \mathbf{f}$, a discretization method is applied to the original problem, see the paragraphs "FDM" and "FEM" in the previous chapter.

- In the case of FDM generally no problem occurs since the coordinates of $\mathbf{u}$ represent the values of the approximation at given places. If $\mathbf{u} \leq \mathbf{0}$, then the approximation is also non-positive at the given places, so we can conclude that the DNPP is in harmony with the NPP, in other words, it is applicable (the same can be said about the other discrete maximum principles).

  But if we want a continuous approximation, then we can construct it with some interpolation from $\mathbf{u}$. "Connecting the points" linearly will not cause any problem, however, if, we use a more sophisticated interpolation method, then the obtained function can attain positive values, too, in spite of the fact that $\mathbf{u}$ was non-positive. And this is a problem showing the limits of our definitions.

  Another problem can be caused by the mesh. In a lot of cases (e.g., if the domain is a rectangle) the mesh contains so-called corner points. Corner points

are such boundary points whose all neighbours are boundary points, too, thus these are not connected to interior mesh points and so these have (usually) no effect on the process. We can define their value as we like, independently of the other values, and this makes our definitions meaningless. Naturally, we want to avoid this situation. The easiest way is to omit these points. We will follow this solution, c.f. [34, Thm.1.8], but we mention the paper [30], where the definitions are modified.

- In the case of FEM the approximation is constructed as a linear combination in the form $\langle \mathbf{u}, \mathbf{\Phi} \rangle$, where $\mathbf{\Phi}$ is a vector whose coordinates ϕ_i, $i = 1, \dots, \overline{N}$ are basis functions of some finite-dimensional vector-space. It is clear that if the basis functions are non-negative (e.g., the usual piecewise linear hat functions are of this type) then the non-positivity of $\mathbf{u}$ implies the non-positivity of the approximation. In this case our definition is applicable again.

  But if we use higher order elements, then the usual choice of the basis functions clearly shows us that our definition is not applicable again. In this situation an other approach is needed, the Reader can find information about this in [57, 58], where positive results are obtained (only) for a simple 1D problem and [29], where negative results are obtained for a higher dimensional simple problem.

  Another problem can be if the coordinates of $\mathbf{u}$ do not represent the values of the approximation at the given places (c.f. the FDM). This should be understood as follows. Consider a continuous problem defined on the unit interval. We use a uniform mesh which determinates the sets $\mathcal{P} = \{x_1, x_2, \dots, x_N\}$ and $\mathcal{P}_\partial = \{x_{N+1} = 0, x_{N+2} = 1\}$ containing the vertices in Ω and on $\partial\Omega$, respectively. Now consider the set of the usual hat functions with a small modification: we choose ϕ_{N+1} and ϕ_{N+2} as half of the usual. Then for $\mathbf{u} = (1, \dots, 1)$ $\max \mathbf{u} \leq \max\{0, \mathbf{u}_\partial\}$ clearly holds, on the other hand, $\max_{x_i \in \mathcal{P} \cup \mathcal{P}_\partial} \langle \mathbf{u}, \mathbf{\Phi} \rangle(x_i) \leq \max\{0, \max_{x_i \in \mathcal{P}_\partial} \langle \mathbf{u}, \mathbf{\Phi} \rangle(x_i)\}$ does not hold since the left side is equal to 1 and the right side is equal to $\frac{1}{2}$.

To summarize, we can conclude that the definitions of discrete maximum principles as we introduced them are applicable for FDM (except for the case mentioned above), and for FEM with the usual linear and multilinear elements (since in these cases the basis functions are nonnegative and possess the "value representing condition").

3.2 Numerical examples on the differences between the discrete elliptic weak and strong maximum principles

In this section we present numerical examples, visualized with the help of Matlab in order to show the differences between the discrete elliptic weak and strong maximum principles. In all examples we used linear finite element discretization (because in this case the FDM is less interesting). We focus on the irreducibility property, i.e., we give examples where the discrete domain is not connected from some point of view. This can easily happen when the domain consists of two relatively large areas connected in the middle with a thin "path". In this case the program package COMSOL can produce qualitatively incorrect meshes, too. This section is based on the paper [41].

In the first three examples $K = -\Delta$, in the fourth it is defined as $Ku = -\Delta u + 128u$. In all examples $Ku = 0$. In the first two cases u is defined as 1 on the boundary of the left square, 0 on the boundary of the right square and linearly decreasing from 1 to 0 on the boundary of the middle square. The boundary condition of the third example differs only on the middle part: on the left part of the boundary of it, i.e. on $\{(x, y) : x \in [3, 3.5], y \in \{1, 2\}\}$, u is 1, then linearly decreasing from 1 to 0 on the right part of the boundary of the middle square i.e. on $\{(x, y) : x \in [3.5, 4], y \in \{1, 2\}\}$. The fourth example is similar to the first two.

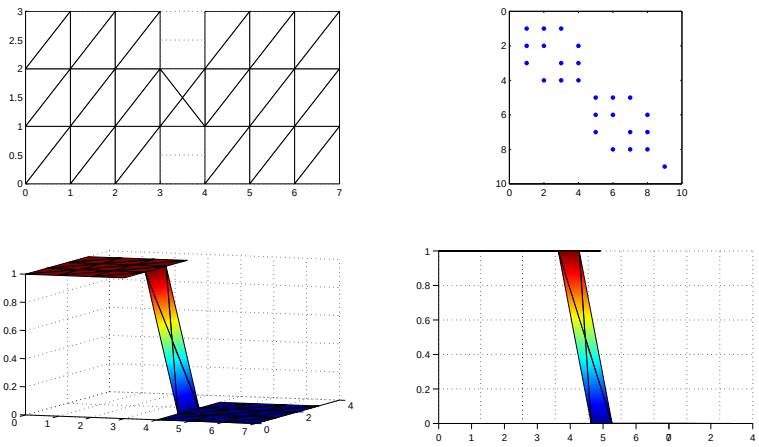

Figure 3.1: 1. Example: The mesh results in a reducible matrix. The DsMP failed, while the DwMP was fulfilled.

The arrangement within the figures is as follows. The top left panel presents the

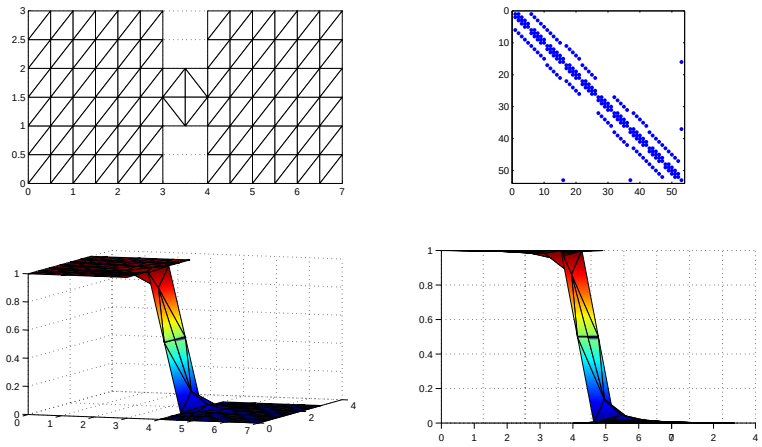

Figure 3.2: 2. Example: The mesh results in an irreducible matrix. Both of the DsMP and DwMP were fulfilled.

mesh, the top right panel presents the nonzero elements of the matrix $\mathbf{K}_0$, and in the bottom panels $\mathbf{u}$ is plotted from two different angles, the right one shows us better where the function is constant.

The first example shows us how an inadequate mesh can result in a reducible matrix and so losing the DSMP (while the DWMP is fulfilled). The second is the "good" example, here both discrete maximum principles are fulfilled. In [8] a mesh is presented, this is the third example here, which seems to be good at first sight, but the two right angles damage the connection of the two seemingly connected points in the middle, c.f. [25], too.

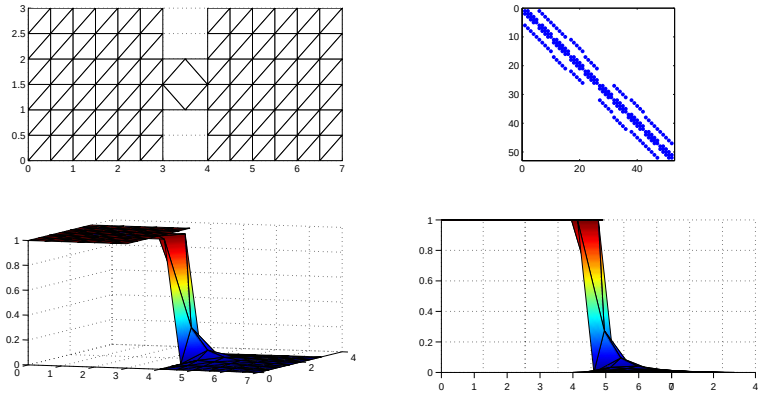

Figure 3.3: 3. Example: The mesh results in a reducible matrix. The DsMP failed, while the DwMP was fulfilled.

The fourth example presents a mesh, which results in losing the DsMP, while the DwMP is fulfilled. It is caused surprisingly by the use of equilateral triangles.

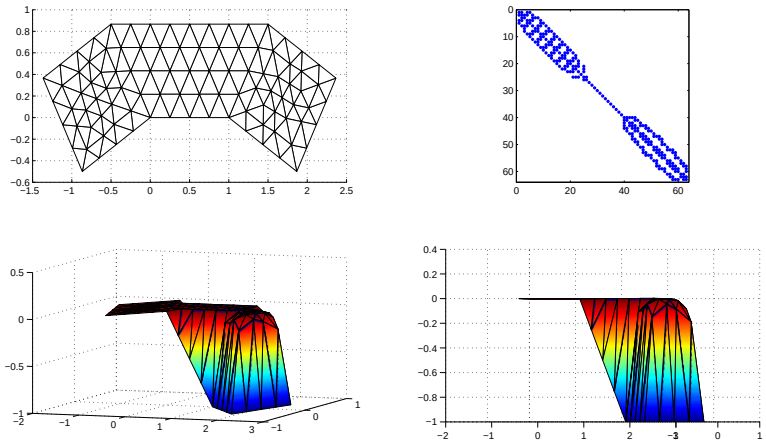

Figure 3.4: 4. Example: The mesh which contains equilateral triangles can result in a reducible matrix, too. The DsMP failed, while the DwMP was fulfilled.

With these examples we demonstrated the usefulness of the algebraic framework.

3.3 Discrete maximum principles for interior penalty discontinuous Galerkin elliptic operators

In the previous sections an algebraic framework was presented with numerical examples. But our job has not been completed yet. The algebraic conditions need to be translated into mesh conditions. There are numerous papers dealing with mesh conditions for FDM and for FEM with linear and continuous elements which guarantee the most popular (and important) maximum principles, the DnP and the DwMP, see e.g., the papers referred to in Section 3.1 besides [46]. Instead of giving an overview of these results here we present how the DnP and DwMP can be guaranteed when interior penalty discontinuous Galerkin method (IPDG) is applied to a 1D elliptic operator (containing diffusion and reaction terms). We formulate the problem and we give the construction of the IPDG operator. After this conditions are derived under which the DnP and DwMP holds. Finally, the sharpness of our conditions is investigated with the help of numerical examples. This section is based on the paper [28].

3.3.1 Interior penalty discontinuous Galerkin elliptic operators

Problem setting. Let us set $\Omega = (0,1)$ and consider the elliptic operator K, defined as

$$Ku = -(pu')' + k^2 u\,, \tag{3.7}$$

where $\operatorname{dom} K = H^1(0,1)$, $p, k \in \mathbb{R}$, $p > 0$.

It is clear that for this operator the nP and wMP holds due to Theorem 2.1.4 and Remark 2.1.3.

There are several sorts of discontinuous Galerkin methods in the literature. Here the interior penalty discontinuous Galerkin method is considered.

Construction of the IPDG elliptic operator. The idea behind the discontinuous Galerkin method in comparison with FEM with piecewise linear and continuous basis functions is to get better approximation and/or to spare computational time by dropping the continuity requirement (even in the case when the solution of the original problem is continuous, which holds for many applications).

As opposed to the standard FEM approach, here the first step to discretize the operator (3.7) with the interior penalty discontinuous Galerkin method is to define a mesh on $(0,1)$. Let us denote it by τ_h and define it in the following way: $0 = x_0 < x_1 < x_2 < \ldots < x_{N-1} < x_N = 1$. We use the notations $I_n = [x_{n-1}, x_n]$, $h_n = |I_n|$, $h_{n-1,n} = \max\{h_{n-1}, h_n\}$, (with $h_{0,1} = h_1$, $h_{N,N+1} = h_N$).

The next step is to define the space $D_l(\tau_h) = \{v : v|_{I_n} \in P_l(I_n), \forall n = 1, 2, \ldots, N\}$ – piecewise polynomials over every interval with maximal degree l. For these functions we introduce the right and left hand side limits $v(x_n^+) = \lim\limits_{t\to 0^+} v(x_n + t)$, $v(x_n^-) = \lim\limits_{t\to 0^+} v(x_n - t)$, and jumps and averages over the mesh nodes as

$$[\![u(x_n)]\!] = u(x_n^-) - u(x_n^+)\,, \quad \{\!\{u(x_n)\}\!\} = \frac{1}{2}(u(x_n^-) + u(x_n^+))\,.$$

At the boundary nodes these are defined as

$$[\![u(x_0)]\!] = -u(x_0^+)\,,\ \{\!\{u(x_0)\}\!\} = u(x_0^+)\,,\ [\![u(x_N)]\!] = u(x_N^-)\,,\ \{\!\{u(x_N)\}\!\} = u(x_N^-)\,.$$

We fix the penalty parameter $\sigma \geq 0$ and ε, which can be any arbitrary number, but it is usually chosen from the set $\{-1, 0, 1\}$. The value $\varepsilon = 1$ gives the nonsymmetric, $\varepsilon = 0$ the incomplete, and $\varepsilon = -1$ the symmetric IPDG.

After these preparations we are ready to define the (discrete) IPDG bilinear form as

$$a_{DG}(u,v) = \sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} pu'(x)v'(x)\,\mathrm{d}x - \sum_{n=0}^{N} \{\!\{pu'(x_n)\}\!\}\,[\![v(x_n)]\!] + \varepsilon \sum_{n=0}^{N} \{\!\{pv'(x_n)\}\!\}\,[\![u(x_n)]\!] + \sum_{n=0}^{N} \frac{\sigma}{h_{n,n+1}}\,[\![v(x_n)]\!]\,[\![u(x_n)]\!] + \int_0^1 k^2 uv\,\mathrm{d}x\,. \tag{3.8}$$

Note that fixing the parameters σ, ε and the mesh τ_h can be done in parallel.

The crucial step is the following. We fix a basis in the space $D_l(\tau_h)$. If we want to use the algebraic framework of Section 3.1, then $l = 1$ needs to be chosen. Moreover, the basis functions need to be non-negative and have to possess the "value representing condition" at least in a generalized sense. This can be done with the following choice, where on the other hand we set aside continuity.

We will use $\Phi_i^1(x)$ for the $(2(i-1)+1)$th basis functions, and $\Phi_i^2(x)$ for the $(2(i-1)+2)$th basis functions, see Figure 3.5. On interval I_i the function $\Phi_i^1(x)$ is the linear function with $\Phi_i^1(x_{i-1}^+) = 1$, $\Phi_i^1(x_i^-) = 0$ and $\Phi_i^2(x)$ is the linear function with $\Phi_i^1(x_{i-1}^+) = 0$, $\Phi_i^1(x_i^-) = 1$, and these functions are zero outside I_i, see Figure 3.5. Thus, here the basis functions can be associated to the subintervals opposed to the standard FEM approach where the basis functions can be associated to the vertices of the mesh.

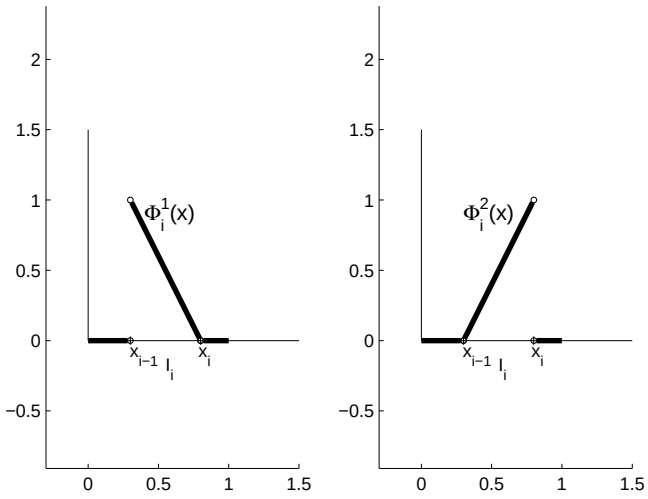

Figure 3.5: $\Phi_i^1(x)$ and $\Phi_i^2(x)$

Finally, we construct the IPDG elliptic operator $\mathbf{K} = (\mathbf{K}_0|\mathbf{K}_\partial)$ similarly to the way as we did in the case of the standard FEM approach. However, there are small

differences since here $\mathbf{K} \in \mathbb{R}^{(2N-2)\times(2N)}$, $\mathbf{K}_0 \in \mathbb{R}^{(2N-2)\times(2N-2)}$, and $\mathbf{K}_\partial \in \mathbb{R}^{(2N-2)\times 2}$. The $2N$ basis function are ordered as follows: the first $2N-2$ are the basis functions that belong to the interior nodes and they are numbered from left to right. The $(2N-1)$th belongs to the left boundary and the $2N$th belongs to the right boundary.

A few words generally about interior penalty discontinuous Galerkin methods. Discontinuous Galerkin methods have been thoroughly investigated in recent years [2, 27, 1]. These methods have several advantages:

- built-in stability for time-dependent advection-convection equations,
- adaptivity can be done easily (the basis function do not have to be continuous over the interfaces),
- the mesh does not have to be regular, hanging-nodes can be handled easily,
- conservation laws could be achieved by the numerical solutions.

There are some disadvantages of this method, too, e.g. there is no guarantee that for a given problem it will work better than the usual FEM approach. Moreover, there are still holes in the theory of the method including questions on the choice of the penalty parameter.

In [2], where several DG methods were examined, the following conditions on the convergence can be found. The nonsymmetric version converges for all $\sigma > 0$, while the two other (symmetric and incomplete) converge only for $\sigma > \sigma^*$, where σ^* is unknown for both methods. The symmetric method is the only one of them that guarantees optimal convergence order.

We note another important difference between the usual FEM and the IPDG method, and it is the treatment of the boundary conditions. In the FEM it is strongly imposed, while in the IPDG case it is imposed only weakly. This means that we need to solve (not trivial) equations to get an approximation for the boundary values. And this is one argument for defining maximum principles for the operator and not for the equation.

For more details about discontinuous Galerkin methods see [7, 9, 45].

The exact form of K. In the following we calculate the elements of the matrix $\mathbf{K}$.

It is easy to check that

$$\partial_x \Phi_i^1(x) = -\frac{1}{h_i}\,, \qquad \partial_x \Phi_i^2(x) = \frac{1}{h_i}\,,$$

which means that the averages are

$$\{\!\{\partial_x \Phi_i^1(x_k)\}\!\} = -\frac{1}{2h_i}\,, \qquad \{\!\{\partial_x \Phi_i^2(x_k)\}\!\} = \frac{1}{2h_i}$$

at both endpoints x_k of I_i, with the exception of the boundary nodes, where there is no division by 2. Similarly, the jumps are

$$[\![\Phi_i^1(x_{i-1})]\!] = -1\,, \qquad [\![\Phi_i^2(x_i)]\!] = 1$$

and zero elsewhere. Using these facts we can calculate the matrix entries.

Summing them up we have the following discretization matrices

$$\mathbf{K}_0 = \begin{pmatrix} d_1 & r_1 & s_2 & & & & & & \\ t_2 & e_2 & q_2 & w_2 & & & & & \\ w_2 & q_2 & d_2 & r_2 & s_3 & & & & \\ & s_2 & t_3 & e_3 & q_3 & w_3 & & & \\ & & & & \ddots & & & & \\ & & & w_i & q_i & d_i & r_i & s_{i+1} & \\ & & & & s_{i-1} & t_i & e_i & q_i & w_i \\ & & & & & & & \ddots & \\ & & & & & & w_{N-1} & q_{N-1} & d_{N-1} & r_{N-1} \\ & & & & & & & s_{N-1} & t_N & e_N \end{pmatrix}, \quad \mathbf{K}_\partial = \begin{pmatrix} v_1 & 0 \\ s_1 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ 0 & 0 \\ 0 & s_N \\ 0 & v_N \end{pmatrix},$$

where

$$\begin{aligned}
d_i &= \frac{p}{2h_i} + \frac{\sigma}{h_{i,i+1}} + \frac{p\varepsilon}{2h_i} + k^2\frac{h_i}{3}\,, \quad i = 1,\dots,N-1\,,\\
e_i &= \frac{p}{2h_i} + \frac{\sigma}{h_{i-1,i}} + \frac{p\varepsilon}{2h_i} + k^2\frac{h_i}{3}\,, \quad i = 2,\dots,N\,,\\
w_i &= \frac{p\varepsilon}{2h_i}\,, \quad i = 2,\dots,N-1\,,\\
q_i &= -\frac{p}{h_i} + \frac{p}{2h_i} - \frac{p\varepsilon}{2h_i} + k^2\frac{h_i}{6}\,, \quad i = 2,\dots,N-1\,,\\
r_i &= \frac{p}{2h_{i+1}} - \frac{\sigma}{h_{i,i+1}} - \frac{p\varepsilon}{2h_i}\,, \quad i = 1,\dots,N-1\,,\\
s_i &= -\frac{p}{2h_i}\,, \quad i = 1,\dots,N\,,\\
t_i &= \frac{p}{2h_{i-1}} - \frac{\sigma}{h_{i-1,i}} - \frac{p\varepsilon}{2h_i}\,, \quad i = 2,\dots,N\,,\\
v_i &= -\frac{p}{h_i} + \frac{p}{2h_i} - \frac{p\varepsilon}{h_i} + k^2\frac{h_i}{6}\,, \quad i = 1,\dots,N
\end{aligned}$$

and zero elsewhere.

3.3.2 Discrete weak non-negativity preservation property and discrete weak maximum principle for interior penalty discontinuous Galerkin elliptic operators

We remark that the space $H^1(0,1)$ consists of continuous functions. Continuity is an important qualitative property, and it cannot be preserved by the discontinuous Galerkin method. This is one reason why we need to be careful, especially with the preservation of some milder qualitative properties which are in connection with the continuity. This leads directly to the investigation of maximum principles for the discontinuous Galerkin method.

Our aim is to get useful mesh conditions that guarantee the DnP and the DwMP. DnP will be guaranteed by the condition (w1'd), while the DwMP will be guaranteed by the conditions (w1'd), (w2') and (w3').

First we deal with the condition (w1'd). Here first we guarantee that the diagonal elements of the matrix $\mathbf{K}_0$ are non-negative and the off-diagonal elements are non-positive resulting in that $\mathbf{K}_0$ is a Z-matrix. This means for the elements

- d_i, e_i:

  we get the following conditions for ε:

  $$\varepsilon \geq -1 - \frac{2\sigma h_i}{p h_{i,i+1}} - \frac{2k^2 h_i^2}{3p}, \quad i = 1, \ldots, N-1$$
  $$\varepsilon \geq -1 - \frac{2\sigma h_i}{p h_{i-1,i}} - \frac{2k^2 h_i^2}{3p}, \quad i = 2, \ldots, N\,.$$

- w_i:

  w_i should be non-positive, which indicates

  $$\varepsilon \leq 0 \tag{3.9}$$

  in the case where we have more than two subintervals. See the third part of Remark 3.3.4 for the degenerate case. This means that for $\varepsilon = 1$ generally we cannot guarantee the DnP and the DwMP.

- q_i:

  because of q_i we need to guarantee $-\frac{p}{2h_i} - \frac{p\varepsilon}{2h_i} + k^2\frac{h_i}{6} \leq 0$, $i = 2, \ldots, N-1$, which means the following for ε:

  $$\varepsilon \geq -1 + \frac{k^2 h_i^2}{3p}, \quad i = 2, \ldots, N-1\,.$$

Or, rephrasing it for the mesh, we have

$$h_i^2 \leq \frac{3(1+\varepsilon)p}{k^2}\,, \quad i = 2, \ldots, N-1$$

in the case where $k \neq 0$. (In the case $k = 0$ we simply have $\varepsilon \geq -1$.)

- s_i:

  Inequality $s_i < 0$ always holds.

- r_i, t_i:

  we need to guarantee $\frac{p}{2h_{i+1}} - \frac{\sigma}{h_{i,i+1}} - \frac{p\varepsilon}{2h_i} \leq 0$ and $\frac{p}{2h_{i-1}} - \frac{\sigma}{h_{i-1,i}} - \frac{p\varepsilon}{2h_i} \leq 0$. After re-indexing t_i and reformulating we have

$$\frac{h_{i,i+1}}{h_{i+1}} - \frac{\varepsilon h_{i,i+1}}{h_i} \leq \frac{2\sigma}{p} \quad \text{and} \quad \frac{h_{i,i+1}}{h_i} - \frac{\varepsilon h_{i,i+1}}{h_{i+1}} \leq \frac{2\sigma}{p}\,, \quad i = 1, \ldots, N-1\,. \tag{3.10}$$

Then we use the "dominant vector" condition to guarantee that $\mathbf{K}_0$ is a non-singular M-matrix, see Theorem 5.0.14 in the Appendix.

Lemma 3.3.1. *There exists* $\mathbf{v} > \mathbf{0}$ *with* $\mathbf{K}_0\mathbf{v} > \mathbf{0}$.

Proof. Fist we consider the case $k = 0$ and $p = 1$.

We choose the dominant vector $\mathbf{v}$ as the piecewise linear interpolation of the function $d(x) = c - x^2$ with the bases of Φ_i^j in the interior nodes and zero at $x = 0, 1$, where $c \geq 1$, see Figure 3.6. We prove that this choice is suitable.

Let us denote this interpolation by $\Pi_d(x)$ and the vector of the coefficients by $\mathbf{v}$, so $\Pi_d(x) = \sum_{(i,j)\in \mathrm{int}(\tau_h)} v_{2(i-1)+j-1}\Phi_i^j(x)$, where the summation goes over all basis functions with the exception of the two that belong to the boundary nodes, ($\Phi_1^1(x)$ and $\Phi_N^2(x)$). It is clear that $\mathbf{v} > \mathbf{0}$, and we need to prove that $\mathbf{K}_0\mathbf{v} > \mathbf{0}$ holds. The meaning of this inequality is that $a_{DG}(\Pi_d(x), \Phi_i^j(x)) > 0$ holds for all basis functions, since e.g. for the first coordinate of $\mathbf{K}_0\mathbf{v}$:

$$(\mathbf{K}_0\mathbf{v})_1 = \sum_{(i,j)\in \mathrm{int}(\tau_h)} v_{2(i-1)+j-1} a_{DG}\left(\Phi_i^j(x), \Phi_1^2(x)\right) =$$

$$a_{DG}\left(\sum_{(i,j)\in \mathrm{int}(\tau_h)} v_{2(i-1)+j-1}\Phi_i^j(x), \Phi_1^2(x)\right) = a_{DG}\left(\Pi_d(x), \Phi_1^2(x)\right)\,.$$

Next we calculate this bilinear form. The function $\Pi_d(x)$ is continuous, therefore its jumps are zero all over the nodes, which means that we have to take into account neither ε, nor the penalty terms.
The derivative of $\Pi_d(x)$ can be calculated on every I_n. It is

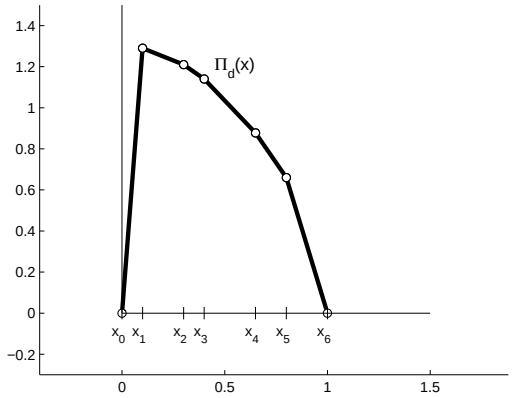

Figure 3.6: $\Pi_d(x)$ for $c = 1.3$

- $\dfrac{c - x_1^2}{x_1}$ on I_1,
- $-\dfrac{x_i^2 - x_{i-1}^2}{x_i - x_{i-1}} = -(x_i + x_{i-1})$ on $I_i \quad i = 2, \ldots, N-1$,
- $\dfrac{x_{N-1}^2 - c}{1 - x_{N-1}}$ on I_N.

This means

$$a_{DG}(\Pi_d(x), \Phi_1^2(x)) = \int_{I_1} \partial_x \Pi_d(x) \partial_x \Phi_1^2(x)\, dx - \{\!\{\partial_x \Pi_d(x_1)\}\!\} \left[\!\left[\Phi_1^2(x_1)\right]\!\right] =$$
$$\left(\frac{c - x_1^2}{x_1}\right) \underbrace{\int_{I_1} \frac{1}{h_1}\, dx}_{=1} - \left(\frac{\frac{c-x_1^2}{x_1} - x_1 - x_2}{2}\right) \cdot 1 = \frac{c - x_1^2}{2x_1} + \frac{x_1 + x_2}{2}\,. \tag{3.11}$$

Similarly,

$$a_{DG}(\Pi_d(x), \Phi_2^1(x)) = \frac{c - x_1^2}{2x_1} + \frac{x_1 + x_2}{2}\,.$$

For $i \neq 1, N-1, N$:

$$a_{DG}(\Pi_d(x), \Phi_i^2(x)) = \int_{I_i} \partial_x \Pi_d(x) \partial_x \Phi_i^2(x)\, dx - \{\!\{\partial_x \Pi_d(x_i)\}\!\} \left[\!\left[\Phi_i^2(x_i)\right]\!\right] =$$
$$-(x_i + x_{i-1}) \int_{I_i} \frac{1}{h_i}\, dx - \left(-\frac{x_i + x_{i-1} + x_i + x_{i+1}}{2}\right) \cdot 1 = \frac{x_{i+1} - x_{i-1}}{2}\,. \tag{3.12}$$

For $i \neq 1, 2, N$:

$$a_{DG}(\Pi_d(x), \Phi_i^1(x)) = \int_{I_i} \partial_x \Pi_d(x) \partial_x \Phi_i^i(x)\, dx - \{\!\{\partial_x \Pi_d(x_{i-1})\}\!\} \left[\!\left[\Phi_i^2(x_{i-1})\right]\!\right] =$$
$$-(x_i + x_{i-1}) \int_{I_i} -\frac{1}{h_i}\, dx - \left(-\frac{x_i + x_{i-1} + x_{i-1} + x_{i-2}}{2}\right) \cdot (-1) = \frac{x_i - x_{i-2}}{2}\,. \tag{3.13}$$

On I_{N-1}:

$$a_{DG}(\Pi_d(x), \Phi_{N-1}^2(x)) =$$
$$\int_{I_{N-1}} \partial_x \Pi_d(x) \partial_x \Phi_{N-1}^2(x)\, dx - \{\!\{\partial_x \Pi_d(x_{N-1})\}\!\} \left[\!\left[\Phi_{N-1}^2(x_{N-1})\right]\!\right] =$$
$$-(x_{N-2} + x_{N-1}) \int_{I_{N-1}} \frac{1}{h_{N-1}}\, dx - \left(\frac{-(x_{N-2} + x_{N-1}) + \frac{x_{N-1}^2 - c}{1 - x_{N-1}}}{2}\right) \cdot 1 =$$
$$-\frac{x_{N-2} + x_{N-1}}{2} + \frac{c - x_{N-1}^2}{2(1 - x_{N-1})}\,. \tag{3.14}$$

Finally,

$$a_{DG}(\Pi_d(x), \Phi_N^1(x)) = -\frac{x_{N-2} + x_{N-1}}{2} + \frac{c - x_{N-1}^2}{2(1 - x_{N-1})}\,.$$

We have to prove that these are positive values. The first three (3.11) – (3.13) are trivial. To prove that (3.14) is positive, some simple calculation is still needed. $-\frac{x_{N-2}+x_{N-1}}{2} + \frac{c-x_{N-1}^2}{2(1-x_{N-1})} > 0\,, \quad \frac{c-x_{N-1}^2}{1-x_{N-1}} > x_{N-2} + x_{N-1}$ and this holds, since $\frac{c-x_{N-1}^2}{1-x_{N-1}} = \frac{(\sqrt{c}-x_{N-1})(\sqrt{c}+x_{N-1})}{1-x_{N-1}} = \frac{\sqrt{c}-x_{N-1}}{1-x_{N-1}}(\sqrt{c} + x_{N-1}) > \sqrt{c} + x_{N-1} > 1 + x_{N-1} > x_{N-2} + x_{N-1}\,.$

When $p \neq 1$, we only have to multiply the matrix $\mathbf{K}_0$ with p, which makes no difference in the sign of the product.

When $k \neq 0$, we have the extra terms $\int_{I_i} k^2 \Phi_i^j(x) \cdot \Phi_i^l(k)$, where $j, l \in \{1, 2\}$. All functions are positive, so these integrals are also positive, hence we have just increased the elements of $\mathbf{K}_0$, consequently increased the coordinates of $\mathbf{K}_0\mathbf{v}$. □

Property (w2') means that v_1 and v_N should be non-positive, i.e.,

$$\varepsilon \geq \frac{-3p + k^2 h_i^2}{6p} = -\frac{1}{2} + \frac{k^2 h_i^2}{6p} \geq -\frac{1}{2}\,, \qquad i = 1, N\,. \tag{3.15}$$

Note that this means that in the case $\varepsilon = -1$ we cannot guarantee the DwMP.

Property (w3') means the condition $\mathbf{0} \leq (\mathbf{K}_0 | \mathbf{K}_\partial)\mathbf{e}$. It is equivalent to the condition $a_{DG}(1, \Phi_i^j) \geq 0$ for $(i, j) \in \mathrm{int}(\tau_h)$, for example, for the first coordinate of $(\mathbf{K}_0 | \mathbf{K}_\partial)\mathbf{e}$

this means the following:

$$((\mathbf{K}_0|\mathbf{K}_\partial)\mathbf{e})_1 = \sum_{i=1}^{N}\sum_{j=1}^{2} 1 \cdot a_{DG}\left(\Phi_i^j(x), \Phi_1^2(x)\right) =$$
$$a_{DG}\left(\sum_{i=1}^{N}\sum_{j=1}^{2} 1 \cdot \Phi_i^j(x), \Phi_1^2(x)\right) = a_{DG}\left(1, \Phi_1^2(x)\right) .$$

The result of this matrix-vector product is

$$\left(\frac{k^2h_1}{2} - \varepsilon\frac{p}{h_1}, \frac{k^2h_2}{2}, \dots, \frac{k^2h_{N-1}}{2}, \frac{k^2h_N}{2} - \varepsilon\frac{p}{h_N}\right)^T ,$$

which is non-negative if

$$\varepsilon \leq \frac{k^2h_i^2}{2p}, \qquad i = 1, N . \tag{3.16}$$

We note that we need to take it into consideration only in the degenerate case, when the interval is divided into two subintervals, since (3.9) is stricter.
Inequalities (3.15) and (3.16) can be pulled together as

$$-\frac{1}{2} + \frac{k^2h_i^2}{6p} \leq \varepsilon \leq \frac{k^2h_i^2}{2p}, \qquad i = 1, N \tag{3.17}$$

or rephrasing it for the mesh,

$$\frac{2p\varepsilon}{k^2} \leq h_i^2 \leq \frac{3p\,(2\varepsilon+1)}{k^2}, \qquad i = 1, N . \tag{3.18}$$

Mesh conditions. We sum up and systematize the conditions we have obtained. Our plan is to give a "recipe" on how we should choose the parameters and the mesh to guarantee the DnP and DwMP. The trick is that we fix the order of the choices.

First we suppose that the interval $(0, 1)$ is divided into more than two subintervals.

Theorem 3.3.2. *Let* $\mathbf{K} = (\mathbf{K}_0|\mathbf{K}_\partial)$ *be the matrix constructed from* (3.7) *by the IPDG method as described earlier. This matrix has the DnP if we choose*

- ε *as*

$$-1 \leq \varepsilon \leq 0, \qquad \text{when} \quad k = 0,$$
$$-1 < \varepsilon \leq 0, \qquad \text{when} \quad k > 0,$$

- σ *as*

$$\frac{p(1-\varepsilon)}{2} \leq \sigma ,$$

- *the mesh* τ_h *as*

$$h_i^2 \leq \frac{3p(\varepsilon+1)}{k^2}\,, \qquad i=2,\dots,N-1\,, \qquad \text{(fineness at the interior)}$$

$$\begin{aligned} &\frac{h_{i,i+1}}{h_{i+1}} - \frac{\varepsilon h_{i,i+1}}{h_i} \leq \frac{2\sigma}{p} \quad \text{and} \\ &\frac{h_{i,i+1}}{h_i} - \frac{\varepsilon h_{i,i+1}}{h_{i+1}} \leq \frac{2\sigma}{p}\,, \quad i=1,\dots,N-1\,. \end{aligned} \qquad \text{(uniformity)}$$

Theorem 3.3.3. *Let* $\mathbf{K} = (\mathbf{K}_0|\mathbf{K}_\partial)$ *be the matrix constructed from* (3.7) *by the IPDG method as described earlier. This matrix possesses the DwMP if we choose*

- ε *as*

$$\begin{aligned} -\frac{1}{2} \leq \varepsilon \leq 0\,, &\qquad \text{when} \quad k=0\,, \\ -\frac{1}{2} < \varepsilon \leq 0\,, &\qquad \text{when} \quad k>0\,, \end{aligned}$$

- σ *as*

$$\frac{p(1-\varepsilon)}{2} \leq \sigma\,,$$

- *the mesh* τ_h *as*

$$h_i^2 \leq \frac{3p(2\varepsilon+1)}{k^2}\,, \qquad i=1,N\,, \qquad \text{(fineness at the boundary)}$$

$$h_i^2 \leq \frac{3p(\varepsilon+1)}{k^2}\,, \qquad i=2,\dots,N-1\,, \qquad \text{(fineness at the interior)}$$

$$\begin{aligned} &\frac{h_{i,i+1}}{h_{i+1}} - \frac{\varepsilon h_{i,i+1}}{h_i} \leq \frac{2\sigma}{p} \quad \text{and} \\ &\frac{h_{i,i+1}}{h_i} - \frac{\varepsilon h_{i,i+1}}{h_{i+1}} \leq \frac{2\sigma}{p}\,, \quad i=1,\dots,N-1\,. \end{aligned} \qquad \text{(uniformity)}$$

Proof (of both theorems). Almost all of the conditions are simple consequences of the above calculations.

The condition for σ can be derived from (3.10) by taking its minimum

$$\frac{2\sigma}{p} \geq \frac{h_{i,i+1}}{h_{i+1}} - \frac{\varepsilon h_{i,i+1}}{h_i} \geq 1-\varepsilon$$

□

Note that we have two types of mesh conditions, one is about the fineness of the mesh and the other is about the uniformity. The first determines the maximum size of the subintervals and it depends on the choice of ε, with $\varepsilon = 0$ being the less restrictive one. The second determines the maximum ratio of the size of the neighbouring subintervals, and it depends on the choice of σ, $\sigma = \frac{p(1-\varepsilon)}{2}$ is the most restrictive.

Remark 3.3.4. We investigate the popular cases: $\varepsilon \in \{-1, 0, 1\}$, too.

- $\varepsilon = -1$:

  We can guarantee only the DnP, and only in the case if $k = 0$ holds. In this case (3.10) simplifies to

  $$\frac{h_{i,i+1}}{h_i} + \frac{h_{i,i+1}}{h_{i+1}} \leq \frac{2\sigma}{p}\,, \quad i = 1, \ldots, N-1\,. \tag{3.19}$$

  This has the consequence that σ needs to be chosen $\geq p$.

- $\varepsilon = 0$:

  We have no additional restrictions in this case. The conditions simplify as

  $\frac{h_{i,i+1}}{h_{i+1}} \leq \frac{2\sigma}{p}$ and $\frac{h_{i,i+1}}{h_i} \leq \frac{2\sigma}{p}\,, \quad i = 1, \ldots, N-1$

  which can be pulled together as

  $$\frac{h_{i,i+1}}{\min\{h_i, h_{i+1}\}} \leq \frac{2\sigma}{p}\,, \quad i = 1, \ldots, N-1 \tag{3.20}$$

  since it is enough to guarantee that the inequality holds for the greater left-hand side. Thus, σ needs to be chosen $\geq p/2$.

- $\varepsilon = 1$:

  We can guarantee the DnP in this case only if $(0, 1)$ is subdivided into two subintervals. Then (3.10) leads to the following conditions

  $\frac{h_{1,2}}{h_1} - \frac{h_{1,2}}{h_2} \leq \frac{2\sigma}{p}$ and $\frac{h_{1,2}}{h_2} - \frac{h_{1,2}}{h_1} \leq \frac{2\sigma}{p}\,.$

  They can be pulled together as

  $$\frac{h_{1,2} - \min\{h_1, h_2\}}{\min\{h_1, h_2\}} \leq \frac{2\sigma}{p}\,. \tag{3.21}$$

  For the DwMP we have more conditions, namely $k > 0$ and

  $$\frac{2p}{k^2} \leq h_i^2 \leq \frac{9p}{k^2}\,, \qquad i = 1, 2\,.$$

Remark 3.3.5. If we choose a different definition for $h_{n-1,n}$, namely, if it is defined as $= \min\{h_{n-1}, h_n\}$ (c.f. [7, Ch.4, Definition 4.5] and [45, Ch.1]), the condition for σ will coincide with the condition that describes the relation between the neighbouring subintervals.

3.3.3 Numerical examples – on the sharpness of the conditions

In this subsection we will investigate the mesh conditions we derived. Naturally, the obtained conditions cannot be sharp since we used practical conditions and these are only sufficient and not necessary. However, we will show that our conditions are sharp in some sense.

Example 3.3.6. Let us set $p = 1$, $\varepsilon = 0$, $\sigma = 5$, $k = 0$. First of all it is clear that condition (3.17) holds for ε and (3.18) is out of view. In this case for the mesh:

$$\tau_h = \{0,\, 0.02,\, 0.22,\, 0.8,\, 1\}$$

the condition (3.20) is sharp in the following sense. Let us modify this mesh as

$$\tau_h^m = \{0,\, 0.02,\, 0.22 + \frac{1}{10^m},\, 0.8,\, 1\}\,.$$

Let us consider the vector $\mathbf{v} = (-1,\, \frac{1}{10^m},\, 0,\, 0,\, 0,\, 0)^T$, see Figure 3.7. The following calculation shows that the resulting right-hand side is non-positive, which means that the maximum principle fails.

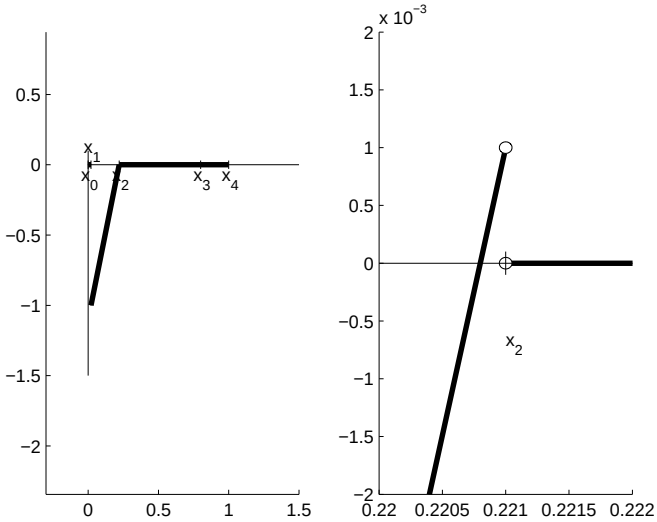

Figure 3.7: Left: the counterexample with $m = 3$. Right: the positive value at the node 0.221

The product $\mathbf{Kv}$ has four non-zero coordinates: $(-d_1 + r_1/10^m, -t_2 + e_2/10^m, -w_2 + q_2/10^m, s_2/10^m, 0, 0, 0, 0)^T$. In this case $h_{1,2} = h_2$. Let us examine these terms.

$$-d_1 + \frac{r_1}{10^m} = -\frac{1}{2h_1} - \frac{5}{h_2} + \frac{1}{10^m}\left(\frac{1}{2h_2} - \frac{5}{h_2}\right) = -\frac{1}{2h_1} - \frac{5}{h_2} - \frac{1}{10^m}\cdot\frac{9}{2h_2} < 0\,.$$

The second one is

$$-t_2+\frac{e_2}{10^m}=-\frac{1}{2h_1}+\frac{5}{h_2}+\frac{1}{10^m}\left(\frac{1}{2h_2}+\frac{5}{2h_2}\right)=-25+\frac{1}{h_2}\left(5+\frac{11}{2\cdot 10^m}\right)=$$
$$-25+\frac{25\cdot 10^m+55/2}{10^m+5}<0\,.$$

The last two terms are easier

$$-w_2+\frac{q_2}{10^m}=0+\frac{q_2}{10^m}=\frac{1}{10^m}\left(-\frac{1}{h_1}+\frac{1}{2h_1}\right)=-\frac{1}{2\cdot 10^m\cdot h_1}<0\,,$$
$$\frac{s_2}{10^m}=\frac{1}{10^m}\cdot\left(-\frac{1}{2h_1}\right)<0\,.$$

Example 3.3.7. Let us set $p=1$, $\varepsilon=1$, $\sigma=5$, $k=0$. In the case that was discussed in the third part of Remark 3.3.4 the mesh

$$\tau_h=\{0,\,1/12,\,1\}$$

is sharp in the same sense as in the last example with respect to (3.21). Similarly as above, we modify the mesh as

$$\tau_h^m=\{0,\,1/12-1/10^m,\,1\}$$

and choose $\mathbf{v}$ as $\mathbf{v}_0=(-1,\frac{1}{10^m})^T$ and 0 elsewhere. This setting breaks the DnP.

$\mathbf{Kv}$ is non-positive since $\mathbf{K}_0\mathbf{v}_0=(-d_1+r_1/10^m,-t_2+e_2/10^m)^T$, where

$$-d_1+\frac{r_1}{10^m}=-\frac{1}{2h_1}-\frac{5}{h_2}-\frac{1}{2h_1}+\frac{1}{10^m}\left(\frac{1}{2h_2}-\frac{5}{h_2}-\frac{1}{2h_1}\right)=$$
$$-\frac{1}{h_1}-\frac{5}{h_2}-\frac{1}{2\cdot 10^m}\left(\frac{9}{h_2}+\frac{1}{h_1}\right)<0$$

and

$$-t_2+\frac{e_2}{10^m}=-\frac{1}{2h_1}+\frac{5}{h_2}+\frac{1}{2h_2}+\frac{1}{10^m}\left(\frac{1}{2h_2}+\frac{5}{h_2}+\frac{1}{2h_2}\right)=$$
$$-\frac{1}{2h_1}+\frac{1}{2h_2}\left(11+\frac{12}{10^m}\right)$$

and similar calculations as before give its negativity $\left(\frac{1}{12}-\frac{1}{10^m}\right)\cdot\left(11+\frac{12}{10^m}\right)<\frac{11}{12}+\frac{1}{10^m}$, and this holds for all $m>0$ since $\frac{11}{12}+\frac{1}{10^m}-\frac{11}{10^m}-\frac{12}{10^{2m}}<\frac{11}{12}+\frac{1}{10^m}$.

Conclusion. First of all, we have shown that it is possible to guarantee the DnP and DwMP when the IPDG discretization is used. However, our conditions are restrictive at the following points:

- the choice of the basis functions,
- $\varepsilon = 1$ is excluded from a practical point of view,
- we can handle $\varepsilon = -1$ only in special cases.

On the other hand, we could state that $\varepsilon = 0$ works very well from the viewpoint of the discrete maximum principle and the conditions suggest that we need to take into consideration a non-integer $\varepsilon \in \left(-\frac{1}{2}, 0\right)$, too.

We have shown with numerical examples that our conditions are sharp in some sense. The numerical examples and computational tests suggest the following points of interest:

- for the symmetric IPDG (3.19) does not seem to be sharp,
- the mesh condition (3.20) seems to be sharp only at the boundary, it could be slightly broken in the interior intervals without losing the DwMP,
- for meshes that consist of more than two subintervals, the condition (3.21) seems to be irrelevant for the neighbouring elements.

Summary of the chapter. In Section 3.1 of this chapter we presented an algebraic framework on discrete maximum principles for matrices. The framework contained sufficient and necessary algebraic conditions (for each introduced discrete maximum principle) including our own results on discrete strong maximum principles, namely, Lemma 3.1.5, Theorem 3.1.10 and Theorem 3.1.9. We gave an overview of the practical conditions ensuring the DwMP, the DsMP and the DSMP by listing the known results and completing them with our own conditions. We investigated the applicability of the framework, too. In Section 3.2 we illustrated the differences between the weak and strong discrete maximum principles with several numerical examples. Section 3.1 and 3.2 were based on the paper [41, Mincsovics and Horváth, 2012].

In Section 3.3, using the algebraic framework we investigated an elliptic problem where the interior penalty discontinuous Galerkin method was applied as discretization. Here we gave sufficient conditions on the parameters ε and σ and on the mesh under which the DnP and the DwMP are fulfilled, see Theorem 3.3.2 and Theorem 3.3.3, respectively. We investigated the sharpness of the necessary conditions of these theorems with numerical examples as well. Section 3.3 was based on the paper [28, Horváth and Mincsovics, 2013].

†

Chapter 4

Discrete parabolic maximum principles

In this chapter first we present an algebraic framework on the important discrete maximum principles defined for a certain class of hyper-matrices. We give algebraic results on discrete maximum principles, both theoretical and practical ones, and we investigate the applicability of the framework as well. Furthermore we apply the framework and present practical conditions when FEM is applied as the spacial discretization and the θ-method for the time integration on a wide class of linear parabolic operators. This is based on the paper [39]. Finally in this chapter we investigate the relation of discrete elliptic and parabolic maximum principles. These results are from the paper [40].

4.1 Algebraic framework

In this section we applied a brief style, since it is very similar to Section 3.1 both in its content and structure.

4.1.1 Discrete parabolic maximum principles

We define maximum principles for a hyper-matrix $\mathcal{L}$ in a special form, acting on a hyper-vector ν

$$\mathcal{L} = \begin{pmatrix} \bar{\mathbf{I}} & \mathbf{0} & \dots & \dots & \mathbf{0} \\ -\mathbf{X}_2 & \mathbf{X}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & -\mathbf{X}_2 & \mathbf{X}_1 & \mathbf{0} & \dots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & -\mathbf{X}_2 & \mathbf{X}_1 \end{pmatrix}, \qquad \nu = \begin{pmatrix} \mathbf{v}^0 \\ \mathbf{v}^1 \\ \vdots \\ \vdots \\ \mathbf{v}^M \end{pmatrix}, \tag{4.1}$$

where $\overline{\mathbf{I}} = (\mathbf{I}_0|\mathbf{0}_\partial)$, $\mathbf{X}_1 = (\mathbf{X}_{10}|\mathbf{X}_{1\partial})$, $\mathbf{X}_2 = (\mathbf{X}_{20}|\mathbf{X}_{2\partial}) \in \mathbb{R}^{N\times\overline{N}}$; $\mathbf{I}_0$, $\mathbf{X}_{10}, \mathbf{X}_{20} \in \mathbb{R}^{N\times N}$; $\mathbf{0}_\partial, \mathbf{X}_{1\partial}, \mathbf{X}_{2\partial} \in \mathbb{R}^{N\times N_\partial}$, $\overline{N} = N+N_\partial$. $(\nu)^n = \mathbf{v}^n = (\mathbf{v}_0^n|\mathbf{v}_\partial^n)^T \in \mathbb{R}^{\overline{N}}$, $\mathbf{v}_0^n \in \mathbb{R}^N$, $\mathbf{v}_\partial^n \in \mathbb{R}^{N_\partial}$.

We mention that the vectors $\mathbf{v}_0^n$ and $\mathbf{v}_\partial^n$ are some approximations of the interior and boundary values of the function v on some time-level, respectively, and $\mathcal{L}$ is the discrete parabolic operator corresponding to L c.f. the paragraph "Problem 2" in 1.2.2.

Thus, we can write $(\mathcal{L}\nu)^0 = \mathbf{v}^0$, $(\mathcal{L}\nu)^n = \mathbf{X}_1\mathbf{v}^n - \mathbf{X}_2\mathbf{v}^{n-1}$, $n = 1, \dots, M$. With this notation and by the assumption that $\mathbf{X}_{10}$ is non-singular, the following iteration form can be created

$$\mathbf{v}_0^n = \mathbf{X}_{10}^{-1}\mathbf{X}_2\mathbf{v}^{n-1} - \mathbf{X}_{10}^{-1}\mathbf{X}_{1\partial}\mathbf{v}_\partial^n + \mathbf{X}_{10}^{-1}(\mathcal{L}\nu)^n\,, \quad n = 1, \dots, M\,, \tag{4.2}$$

which serves to compute (theoretically) $\mathbf{v}_0^n$ if the boundary values $\mathbf{v}_\partial^n$, the initial vector $\mathbf{v}^0$ and $(\mathcal{L}\nu)^n$ are given.

To formalize the discrete maximum principles we introduce the notations $\nu_0(k) = \{\mathbf{v}_0^1, \dots, \mathbf{v}_0^k\}$; $\nu_\partial(k^0) = \{\mathbf{v}_\partial^0, \mathbf{v}_\partial^1, \dots, \mathbf{v}_\partial^k\}$; $(\mathcal{L}\nu)(k) = \{(\mathcal{L}\nu)^1, \dots, (\mathcal{L}\nu)^k\}$ and $(\mathcal{L}\nu)(k^0) = \{(\mathcal{L}\nu)^0, (\mathcal{L}\nu)^1, \dots, (\mathcal{L}\nu)^k\}$.

Then the corresponding maximum principles read as follows.

Definition 4.1.1. We say that the hyper-matrix $\mathcal{L}$ in the form (4.1) possesses

- the *discrete non-negativity preservation property* (DnP) if for all $k = 1, 2, \dots, M$ the following implication holds.

$$\max(\mathcal{L}\nu)(k^0) \le 0\,, \quad \max\nu_\partial(k^0) \le 0 \quad \Rightarrow \quad \max\nu_0(k) \le 0\,;$$

- the *discrete maximum principle* (DmP) if for all $k = 1, 2, \dots, M$ the following implication holds.

$$\max(\mathcal{L}\nu)(k) \le 0 \quad \Rightarrow \quad \max\nu_0(k) \le \max\{0, (\mathcal{L}\nu)^0, \max\nu_\partial(k^0)\}\,.$$

- the *discrete strict maximum principle* (DMP) if for all $k = 1, 2, \dots, M$ the following implication holds.

$$\max(\mathcal{L}\nu)(k) \le 0 \quad \Rightarrow \quad \max\nu_0(k) \le \max\{(\mathcal{L}\nu)^0, \max\nu_\partial(k^0)\}\,.$$

Remark 4.1.2. Even though the discrete parabolic maximum principles are less investigated, there are some important works in this topic. We give a short list of the recommended literature.

- Probably the first paper on a discrete parabolic maximum principle is [32].
- From the early years the paper [24] should be mentioned which was the starting-point for almost every later published work in this topic.
- From the recent years the works [11, 17] contain a detailed investigation of a whole family of discrete (and continuous) parabolic maximum principles.

4.1.2 Algebraic results on discrete parabolic maximum principles

Our aim is to give necessary and sufficient conditions for the above defined discrete maximum principles, moreover, we also touch upon useful practical conditions which can be used from an application point of view.

First, exploiting the iteration form (4.2) we reformulate Definition 4.1.1 into a more suitable form.

Lemma 4.1.3. *The hyper-matrix $\mathcal{L}$ in the form (4.1) possesses*

- *the DnP if and only if (for all $\mathbf{v}^n, \mathbf{v}^{n-1}$) the following implication holds.*

$$(\mathcal{L}\nu)^n \equiv \mathbf{X}_1\mathbf{v}^n - \mathbf{X}_2\mathbf{v}^{n-1} \leq \mathbf{0}\,, \quad \max\{\mathbf{v}^{n-1}, \mathbf{v}^n_\partial\} \leq 0 \quad \Rightarrow \quad \max \mathbf{v}^n \leq 0\,;$$

- *the DmP if and only if (for all $\mathbf{v}^n, \mathbf{v}^{n-1}$) the following implication holds.*

$$(\mathcal{L}\nu)^n \equiv \mathbf{X}_1\mathbf{v}^n - \mathbf{X}_2\mathbf{v}^{n-1} \leq \mathbf{0} \quad \Rightarrow \quad \max \mathbf{v}^n \leq \max\{0, \mathbf{v}^{n-1}, \mathbf{v}^n_\partial\}\,;$$

- *the DMP if and only if (for all $\mathbf{v}^n, \mathbf{v}^{n-1}$) the following implication holds.*

$$(\mathcal{L}\nu)^n \equiv \mathbf{X}_1\mathbf{v}^n - \mathbf{X}_2\mathbf{v}^{n-1} \leq \mathbf{0} \quad \Rightarrow \quad \max \mathbf{v}^n \leq \max\{\mathbf{v}^{n-1}, \mathbf{v}^n_\partial\}\,.$$

Proof. The "if" part follows from induction, the "only if" part is trivial. □

Now, based on Lemma 4.1.3 we can give sufficient and necessary algebraic conditions for the DnP.

Lemma 4.1.4. *[11, L. 2.3.39] The hyper-matrix $\mathcal{L}$ in the form* (4.1) *possesses the DnP if and only if the following three conditions hold.*

(n1) $\mathbf{X}_{10}^{-1} \geq \mathbf{0}\,;$ (n2) $-\mathbf{X}_{10}^{-1}\mathbf{X}_{1\partial} \geq \mathbf{0}\,;$ (n3) $\mathbf{X}_{10}^{-1}\mathbf{X}_2 \geq \mathbf{0}\,.$

Proof. This can be proven exactly in the same way as Lemma 3.1.4, thus we recall only the important steps.

– First, we assume (n1)–(n3). Then the identity (4.2) gives immediately the DnP.

– Second, we assume the DnP. First we can prove that $\mathbf{X}_{10}$ is non-singular, and this means that it is allowed to use the identity (4.2). Then (n1) follows from the setting $\mathbf{v}^{n-1} = \mathbf{0}$, $\mathbf{v}^n_\partial = \mathbf{0}$, (n2) follows from the setting $(\mathcal{L}\nu)^n = \mathbf{0}$, $\mathbf{v}^{n-1} = \mathbf{0}$, and (n3) follows from the setting $(\mathcal{L}\nu)^n = \mathbf{0}$, $\mathbf{v}^n_\partial = \mathbf{0}$.

□

We finish with the DmP and the DMP.

Theorem 4.1.5. *[40, 11] The hyper-matrix $\mathcal{L}$ in the form* (4.1) *possesses*

- *the DmP if and only if the following four conditions hold.*

  (m1) $\mathbf{X}_{10}^{-1} \geq \mathbf{0}$; (m2) $-\mathbf{X}_{10}^{-1}\mathbf{X}_{1\partial} \geq \mathbf{0}$; (m3) $\mathbf{X}_{10}^{-1}\mathbf{X}_2 \geq \mathbf{0}$;
  (m4) $\mathbf{X}_{10}^{-1}\mathbf{X}_2\mathbf{e} - \mathbf{X}_{10}^{-1}\mathbf{X}_{1\partial}\mathbf{e} \leq \mathbf{e}$.

- *the DMP if and only if the following four conditions hold.*

  (M1) $\mathbf{X}_{10}^{-1} \geq \mathbf{0}$; (M2) $-\mathbf{X}_{10}^{-1}\mathbf{X}_{1\partial} \geq \mathbf{0}$; (M3) $\mathbf{X}_{10}^{-1}\mathbf{X}_2 \geq \mathbf{0}$;
  (M4) $\mathbf{X}_{10}^{-1}\mathbf{X}_2\mathbf{e} - \mathbf{X}_{10}^{-1}\mathbf{X}_{1\partial}\mathbf{e} = \mathbf{e}$.

Proof. • The DmP case. Note that (m1)–(m3) are identical with (n1)–(n3). The proof goes in the same way as the proof of Lemma 3.1.6.

– First we assume (m1)–(m4), then

$$(\mathcal{L}\nu)^n \leq \mathbf{0} \quad \Rightarrow \quad \mathbf{v}^n_0 \leq \mathbf{X}_{10}^{-1}\mathbf{X}_2\mathbf{v}^{n-1} - \mathbf{X}_{10}^{-1}\mathbf{X}_{1\partial}\mathbf{v}^n_\partial$$
$$\leq \mathbf{X}_{10}^{-1}\mathbf{X}_2 \max\{0, \mathbf{v}^{n-1}, \mathbf{v}^n_\partial\}\mathbf{e} - \mathbf{X}_{10}^{-1}\mathbf{X}_{1\partial} \max\{0, \mathbf{v}^{n-1}, \mathbf{v}^n_\partial\}\mathbf{e} \leq \max\{0, \mathbf{v}^{n-1}, \mathbf{v}^n_\partial\}\mathbf{e}.$$

– Second, to prove the reverse direction we assume the DMP. DMP implies DnP, and that gives (m1)–(m3) (≡ (n1)–(n3)). (m4) follows from putting $\mathbf{v}^{n-1} = \mathbf{e}$, $\mathbf{v}^n_\partial = \mathbf{e}$, $(\mathcal{L}\nu)^n = \mathbf{0}$ in (4.2).

- The DMP case. Note that (M1)–(M3) are identical with (n1)–(n3). Then we can proceed similarly as in the proof of Lemma 3.1.8, thus we omit the details.

□

Some remarks this theorem:

- Note that (m4) corresponds to $c \leq 0$ (c.f. the continuous case).
- (M4) is equivalent to $\mathbf{Ke} = \mathbf{0}$. This corresponds to $c = 0$ (c.f. the continuous case).
- There are many papers containing some variants of the above lemma and theorem, e.g. [13, 14, 15], but in most cases the discretization method is fixed at the beginning, thus the algebraic framework is not independent. An independent algebraic framework can be found in [11] and in [40].

Practical algebraic conditions for the discrete maximum principles. Lemma 4.1.4 and Theorem 4.1.5 are not applicable directly. From an application point of view it is necessary to give more useful (but only sufficient) conditions in order to guarantee the DnP/DmP/DMP.

- The condition (m2) is usually replaced by the assumption

  (m2') $\mathbf{X}_{1\partial} \leq \mathbf{0}$.

  Then (m2') with (m1) implies (m2), but the converse is not true.

- The condition (m3) is usually replaced by the assumption

  (m3') $\mathbf{X}_2 \geq \mathbf{0}$.

  Then (m3') with (m1) clearly implies (m3), but the converse is not true.

- The condition (m4) is usually replaced by the assumption

  (m4') $\mathbf{Ke} \geq \mathbf{0}$.

  Then (m4') with (m1) implies (m4), but the converse is not true.

- To ensure (m1) is the hardest task here, too, and it is usually replaced by the assumption

  (m1') $\mathbf{X}_{10}$ is an M-matrix

  and we can apply each one from the list that can be found in the paragraph "Practical algebraic conditions for the DwMP" in Subsection 3.1.2.

  These conditions appeared (a little bit hidden) already in [24].

Applicability. Finally, we turn our attention on the applicability to the above defined discrete parabolic maximum principles. Here we defined discrete maximum principles in the natural way, too. Thus, the applicability of this framework depends on the same questions as in the elliptic case, however, with some additional things to consider.

- This framework is designed only for some discretization methods. Namely, only for those when the discretization is done in the following two consecutive steps:

  1. The spatial discretization. This can be done e.g. by FDM or FEM as in the elliptic case. This means that the same applicability problems occur that we explained in details in the paragraph "Applicability" in Subsection 3.1.1.

  2. The time-integration. The special structure of the hyper-matrix $\mathcal{L}$ (4.1) reveals that only one-step methods are allowed. (Naturally, this could be extended to contain multistep methods as well.) This means that L is approximated by the formula

     $$(Lv)(\mathbf{x}_i, n\Delta t) \approx (\mathcal{L}\nu)_i^n = (\mathbf{X}_1\mathbf{v}^n - \mathbf{X}_2\mathbf{v}^{n-1})_i,$$

     where $\mathbf{x}_i \in \mathcal{P}$, $n = 1, 2, \ldots, M$ and $\Delta t = T/M$ is the time-step.

     At the time-integration part the same problems can occur that we investigated at the FDM case in the paragraph "Applicability" of Subsection 3.1.1.

  A typical choice is FEM + θ-method. Then $\mathbf{X}_1 = \frac{1}{\Delta t}\mathbf{M} + \theta\mathbf{K}$, $\mathbf{X}_2 = \frac{1}{\Delta t}\mathbf{M} - (1-\theta)\mathbf{K}$, where $\mathbf{M}$ is the so-called mass matrix, $\mathbf{K}(= \mathbf{X}_1 - \mathbf{X}_2)$ is the so-called stiffness matrix and $\theta \in [0, 1]$ is a parameter.

- Another applicability restriction which comes from the form of the hyper-matrix is that seemingly we can handle only the case where the coefficient functions are time independent. However, this deficiency can be stopped easily (introducing one more index), but that would complicate matters unnecessarily.

- In the discrete elliptic case it was definitively advantageous to define maximum principles for the operator, but here it has some disadvantages, too, c.f. the notions IAP and CAP in [11].

4.2 Discrete maximum principles for some finite element + θ-method parabolic operator

In this section we investigate the way how the DmP and the DMP can be guaranteed for a given (pretty general) linear parabolic operator when the FEM+θ-method is applied as discretization method. The section is organized as follows. First we obtain the hyper-matrix by the discretization applied to the given operator. Then practical conditions are obtained including a mesh condition, restriction to the parameter θ and restriction to the time step Δt under which the DmP/DMP is fulfilled for the hyper-matrix. Finally, numerical examples are presented in order to investigate the sharpness of the conditions. This section is mainly based on the paper [39], which generalizes the results of [18].

4.2.1 Finite element + θ-method parabolic operators

Problem setting. Let $\Omega \subset \mathbb{R}^d$ be an open and bounded domain that can be covered by a regular simplicial mesh $\mathcal{T}_h$ with the property that this mesh is of nonobtuse type, i.e., all the angles made by any faces of each simplex $S \in \mathcal{T}_h$ are not greater than $\pi/2$.

We consider the parabolic operator which is defined for the functions $v(x,t) \in C^{2,1}(Q_T) \cap C(\bar{Q}_T)$ and which can be described as

$$L_{a,b,c}v = \frac{\partial v}{\partial t} - \operatorname{div}(a \operatorname{grad} v) + \langle b, \operatorname{grad} v\rangle + c\, v, \tag{4.3}$$

where $a, c : \Omega \to \mathbb{R}$, $b : \Omega \to \mathbb{R}^d$, $a, b, c \in C(\overline{\Omega})$ and $a \in C^1(\Omega)$. The symbol $\langle \cdot, \cdot \rangle$ stands for the usual scalar product in $\mathbb{R}^d$.

In the sequel we assume that $0 < a_m \le a \le a_M$, $\|b\| \le b_M$ and $0 \le c \le c_M$ holds with the constants a_m, a_M, b_M, c_M. $\|\cdot\|$ denotes the norm of $\mathbb{R}^d$ induced by the scalar product $\langle \cdot, \cdot \rangle$.

Then, by Theorem 2.2.2 the operator $L_{a,b,c}$ satisfies the DnP and the DmP, moreover, the operator $L_{a,b,0}$ satisfies the DMP.

Discretization. We proceed in the same way as in the paragraph "Problem 2" in 1.2.2. Using the FEM+θ-method, where we cover Ω by a regular simplicial mesh $\mathcal{T}_h$ and we use the usual hat functions resulting in the discrete parabolic operator

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

where $n = 1, \dots, M$, $M\Delta t = T$, $\theta \in [0,1]$ and the matrices $\mathbf{M}$, $\mathbf{K}$ are defined by bilinear forms as

$$(\mathbf{M})_{ij} = B_1(\phi_j, \phi_i) = \int_\Omega \phi_j \phi_i \, \mathrm{d}\mathbf{x}\,,$$

$$(\mathbf{K})_{ij} = B_2(\phi_j, \phi_i) = \int_\Omega a \ \langle \mathrm{grad}\phi_j, \mathrm{grad}\phi_i \rangle \ \mathrm{d}\mathbf{x} + \int_\Omega \langle b, \mathrm{grad}\phi_j \rangle \, \phi_i \ \mathrm{d}\mathbf{x} + \int_\Omega c\phi_j\phi_i \ \mathrm{d}\mathbf{x}\,,$$

where $i = 1, \dots, N$, $j = 1, \dots, \overline{N}$. This can be rewritten to the familiar form

$$\mathbf{X}_1 \mathbf{v}^{n+1} - \mathbf{X}_2 \mathbf{v}^n$$

which will be denoted by $\mathcal{L}_{a,b,c}$ or $\mathcal{L}_{a,b,0}$ if our starting point was the operator $L_{a,b,c}$ or $L_{a,b,0}$, with the roles $\mathbf{X}_1 = \frac{1}{\Delta t}\mathbf{M} + \theta\mathbf{K}$, $\mathbf{X}_2 = \frac{1}{\Delta t}\mathbf{M} - (1-\theta)\mathbf{K}$.

4.2.2 Discrete maximum principles for some FEM + θ-method parabolic operator

First we give some useful results.

Lemma 4.2.1. *The earlier described discretization method applied to the operator* (4.3) *results in a hyper-matrix* $\mathcal{L}_{a,b,c}$ *with the properties*

(i) $\mathbf{M} \geq \mathbf{0}$;

(ii) $\mathbf{M}_0\mathbf{e} > \mathbf{0}$;

(iii) $\mathbf{K}\mathbf{e} \geq \mathbf{0}$.

Proof. (i) $(\mathbf{M})_{ij} = B_1(\phi_j, \phi_i) = \int_\Omega \phi_j\phi_i \, \mathrm{d}\mathbf{x} \geq 0$, since the basis functions are non-negative.

(ii) It follows from the previous item, since $(\mathbf{M})_{ii} > 0$.

(iii) $(\mathbf{K}\mathbf{e})_i = \sum_{j=1}^{\overline{N}} B_2(\phi_j, \phi_i) = B_2(\sum_{j=1}^{\overline{N}} \phi_j, \phi_i) = B_2(1, \phi_i) = \int_\Omega c\phi_i \, \mathrm{d}\mathbf{x} \geq 0.$

□

Lemma 4.2.2. *Under the assumptions*

(P1) $\quad (\mathbf{K})_{ij} \leq 0\,, \quad i \neq j,\ i = 1, \dots, N\,,\ j = 1, \dots, \overline{N}\,,$

(P2) $\quad \Delta t(\mathbf{X}_1)_{ij} = (\mathbf{M})_{ij} + \Delta t\theta(\mathbf{K})_{ij} \leq 0\,, \quad i \neq j,\ i = 1, \dots, N\,,\ j = 1, \dots, \overline{N}\,,$

(P3) $\Delta t(\mathbf{X}_2)_{ii} = (\mathbf{M})_{ii} - \Delta t(1-\theta)(\mathbf{K})_{ii} \geq 0\,, \quad i = 1, \ldots, N\,.$

the conditions (m1')–(m4') are satisfied.

Proof. We organize the proof going from simple to difficult.

(m4') is independent of the conditions (P1)–(P3), it is ensured by the choice of the basis functions, see Lemma 4.2.1.

(m2') is ensured by (P2).

(m3') is ensured by (P1), (P3) and the fact that $\mathbf{M} \geq \mathbf{0}$ (see Lemma 4.2.1).

(m1') is ensured by (P1), (P2) and by Lemma 4.2.1. In details: $\mathbf{X}_{10}$ is a Z-matrix by (P2).

$$\mathbf{K}_0\mathbf{e} = \mathbf{K}\mathbf{e}\underbrace{-\mathbf{K}_\partial\mathbf{e}}_{\geq\mathbf{0}} \geq \mathbf{K}\mathbf{e} \geq \mathbf{0}$$

by (P1) and Lemma 4.2.1. Thus,

$$\mathbf{X}_{10}\mathbf{e} = \left(\frac{1}{\Delta t}\mathbf{M}_0 + \theta\mathbf{K}_0\right)\mathbf{e} = \underbrace{\frac{1}{\Delta t}\mathbf{M}_0}_{>\mathbf{0}} + \underbrace{\theta\mathbf{K}_0\mathbf{e}}_{\geq\mathbf{0}} > \mathbf{0}$$

by Lemma 4.2.1, which implies that $\mathbf{X}_{10}$ is SDD. Finally, we recall that a SDD Z-matrix is an M-matrix (see the Appendix).

□

Remark 4.2.3. We list some comments on the conditions of Lemma 4.2.2.

- Since $(\mathbf{M})_{ij} = (\mathbf{K})_{ij} = 0$, $(i \neq j)$ for the index pairs which determine non-neighbouring vertices, we need to investigate only the remainder.
- (P1) is one additional restriction for the mesh, (P2) and (P3) give a lower and an upper bound for the time-step Δt. Naturally, the lower bound must be smaller than the upper bound, this can be attained by the corresponding choice of θ.
- In the case $\theta = 0$ the condition (P2) cannot be fulfilled. Thus, we fix that $\theta \in (0, 1]$. (However, if we use the lumped mass technique (see e.g. in [24]), then $\theta = 0$ is possible, too.) In case $\theta = 1$ (P3) is automatically fulfilled.
- (P2) implies (P1). However, we need to require a strict inequality in (P1) for the index pairs which determine neighbouring vertices, to make (P2) possible. Let us denote this modified condition by (P1'). Since we want to get a usable condition for the mesh, we investigate (P1') instead of (P1) in the following.

Local conditions for the DmP/DMP. We define/estimate the elements of the local mass and stiffness matrices similarly as in [4, 18, 26].

The contributions to the mass matrix $\mathbf{M}$ over the simplex $S \in \mathcal{T}_h$ are

$$M_{ij}|_S = \frac{\mathrm{meas_d S}}{(d+1)(d+2)}\,,\ (i \neq j); \qquad M_{ii}|_S = \frac{2\ \mathrm{meas_d S}}{(d+1)(d+2)}\,. \tag{4.4}$$

We estimate the contribution to the stiffness matrix $\mathbf{K}$ over the simplex S in the following way. If the simplex S is tightened by the $d+1$ piece vertices $\mathbf{x}_i$, and we denote by S_i the $(d-1)$-dimensional face opposite to the vertex $\mathbf{x}_i$, then $\cos\gamma_{ij}$ is the cosine of the interior angle between faces S_i and S_j. Note that $(\mathrm{meas_d S})d = (\mathrm{meas_{d-1} S_i})m_i$, where m_i is the (Euclidean) distance between S_i and $\mathbf{x}_i$.

Let us introduce the notations: $a_m(S) = \min_S a$, $a_M(S) = \max_S a$, $b_M(S) = \max_S \|b\|$, $c_M(S) = \max_S c$. Then,

$$\int\limits_S a\,\langle \mathrm{grad}\phi_j, \mathrm{grad}\phi_i\rangle\ \mathrm{d}\mathbf{x} = -\int\limits_S a\,\|\mathrm{grad}\phi_j\|\,\|\mathrm{grad}\phi_i\|\cos\gamma_{ij}\ \mathrm{d}\mathbf{x} =$$
$$-\frac{\cos\gamma_{ij}}{m_j m_i}\int\limits_S a\ \mathrm{d}\mathbf{x} \le -\frac{a_m(S)(\mathrm{meas_d S})}{m_i m_j}\cos\gamma_{ij}\ (\le 0) \quad \text{in case } i \neq j\,,$$

otherwise

$$\int\limits_S a\,\langle \mathrm{grad}\phi_i, \mathrm{grad}\phi_i\rangle\ \mathrm{d}\mathbf{x} = \int\limits_S a\,\|\mathrm{grad}\phi_i\|^2\ \mathrm{d}\mathbf{x} \le \frac{a_M(S)(\mathrm{meas_d S})}{m_i^2}$$

and

$$\left|\int\limits_S \langle b, \mathrm{grad}\phi_j\rangle\,\phi_i\ \mathrm{d}\mathbf{x}\right| \le \int\limits_S |\langle b, \mathrm{grad}\phi_j\rangle|\,\phi_i\ \mathrm{d}\mathbf{x} \le$$
$$\int\limits_S \|b\|\,\|\mathrm{grad}\phi_j\|\,\phi_i\ \mathrm{d}\mathbf{x} \le \frac{b_M(S)}{m_j}\int\limits_S \phi_i\ \mathrm{d}\mathbf{x} = \frac{b_M(S)(\mathrm{meas_d S})}{m_j(d+1)}$$

hold.

Thus we have the estimation

$$K_{ij}|_S \le (\mathrm{meas_d S})\left[-\frac{a_m(S)\cos\gamma_{ij}}{m_i m_j} + \frac{b_M(S)}{m_j(d+1)} + \frac{c_M(S)}{(d+1)(d+2)}\right] \tag{4.5}$$

for the non-diagonal elements, and

$$K_{ii}|_S \le (\mathrm{meas_d S})\left[\frac{a_M(S)}{m_i^2} + \frac{b_M(S)}{m_i(d+1)} + \frac{2c_M(S)}{(d+1)(d+2)}\right] \tag{4.6}$$

for the diagonal elements.

If we require (P1')–(P3) on every simplex $S \in \mathcal{T}_h$, then we get a sufficient condition to fulfil (P1')–(P3). Thus, one can easily check on the basis of (4.4) – (4.6) that the following lemma is valid.

Lemma 4.2.4. *Let us assume that for the mesh $\mathcal{T}_h$ the geometrical condition*

$$\cos\gamma_{ij} > \frac{b_M(S)}{a_m(S)}\frac{m_i}{d+1} + \frac{c_M(S)}{a_m(S)}\frac{m_i m_j}{(d+1)(d+2)} \tag{4.7}$$

is satisfied. Then, for Δt chosen in accordance with the lower bound

$$\Delta t \geq \frac{1}{\theta}\left[a_m(S)\cos\gamma_{ij}\frac{(d+1)(d+2)}{m_i m_j} - b_M(S)\frac{d+2}{m_j} - c_M(S)\right]^{-1} \tag{4.8}$$

and the upper bound

$$\Delta t \leq \frac{1}{1-\theta}\left[\frac{a_M(S)}{2}\frac{(d+1)(d+2)}{m_i^2} + \frac{b_M(S)}{2}\frac{d+2}{m_i} + c_M(S)\right]^{-1}, \tag{4.9}$$

respectively, $\mathcal{L}_{a,b,c}/\mathcal{L}_{a,b,0}$ satisfies the DmP/DMP.

Global conditions for the DmP/DMP. Lemma 4.2.4 is of little use in practice, since conditions (4.7)-(4.9) should be checked for each $S \in \mathcal{T}_h$, moreover, it does not contain any useful information about the corresponding choice of θ. In the following we deal with getting rid of these problems. The trick is the same as in Theorems 3.3.2 and 3.3.3 that we fix the order of the choices.

In order to formalize the theorem, let us introduce the notations

$$m = \min_{\mathcal{T}_h} m_i\,, \quad M = \max_{\mathcal{T}_h} m_i\,, \quad G = \min_{\mathcal{T}_h} \cos\gamma_{ij}\,,$$

$$\spadesuit = \frac{a_M}{2}\frac{(d+1)(d+2)}{m^2} + \frac{b_M}{2}\frac{d+2}{m} + c_M\,, \quad \heartsuit = a_m G\frac{(d+1)(d+2)}{M^2} - b_M\frac{d+2}{m} - c_M\,.$$

Then, from Lemma 4.2.4 it follows:

Theorem 4.2.5. *Let us assume that for the mesh $\mathcal{T}_h$ the geometrical-fineness condition*

$$0 < \heartsuit \qquad \text{(mesh condition)}$$

holds.

Moreover the condition

$$\frac{\spadesuit}{\spadesuit + \heartsuit} \leq \theta \qquad \text{(restriction for the parameter } \theta\text{)}$$

holds, too. Then under the condition

$$\frac{1}{\theta}\frac{1}{\heartsuit} \leq \Delta t \leq \frac{1}{1-\theta}\frac{1}{\spadesuit} \qquad \text{(restriction for the time step } \Delta t\text{)}$$

$\mathcal{L}_{a,b,c}/\mathcal{L}_{a,b,0}$ satisfies the DmP/DMP.

Remark 4.2.6.

- We remark that the mesh condition can be substituted for the less restrictive condition

$$G > \frac{b_M}{a_m}\frac{M}{d+1} + \frac{c_M}{a_m}\frac{M^2}{(d+1)(d+2)}\,.$$

  However, with this condition we cannot guarantee that the right side of the condition "restriction for the parameter θ" is not greater than one.

- The mesh condition gives an upper bound for the angles, and it depends on how fine the mesh is, i.e., the ratio M^2/m cannot be too large. Note that $G \leq \frac{1}{d}$, and here the equality holds only in the case where $\mathcal{T}_h$ is a uniformly regular simplicial mesh, i.e., that consisting of the congruent regular simplices, see [18]. Naturally, this can be attained if Ω is special. This case allows us the widest choice of the parameters θ, Δt. However even in this case $\spadesuit > \heartsuit$ for $d > 2$, which means by using the condition "restriction for the parameter θ" that the Crank–Nicolson method is excluded for us.

- If $\mathcal{T}_h$ and θ is such that the conditions "mesh condition" and "restriction for the parameter θ" hold, then the lower and upper bounds for Δt determine a non-empty interval, this is condition "restriction for the time step Δt".

- Note that our bounds contain as special case the bounds obtained in [18] – in which the operator $L_{a,0,c}$ with constant coefficients was investigated – if we set the parameters as $a_M = a_m = a$, $b_M = 0$, $c_M = c$.

4.2.3 Numerical examples

As one can see, the conditions collected in the last subsection are sufficient, but not necessary to guarantee the DmP/DMP. Consequently, we need to investigate how sharp our conditions are. This subsection is devoted to illustrate this question with several (extreme) numerical examples.

We fix the dimension $d = 2$ and the parameters $a \equiv 1$, $b \equiv (6, 0)$, $c \equiv 10$. We investigate two operators $L_{a,b,c}$ with homogeneous Dirichlet boundary conditions which differ only in their domains, see Figure 4.1.

In the first case the domain is a rhombus, determined by the vertices $(0, 0)$, $(1, 0)$, $(3/2, \sqrt{3}/2)$, $(1/2, \sqrt{3}/2)$, which allows us to use a uniformly regular simplicial mesh, however, the finer mesh from the two is still relatively coarse.

In the second case the domain is a unit square, here we used a mesh which contains right-angled triangles, which is problematical from the point of view of Theorem 4.2.5.

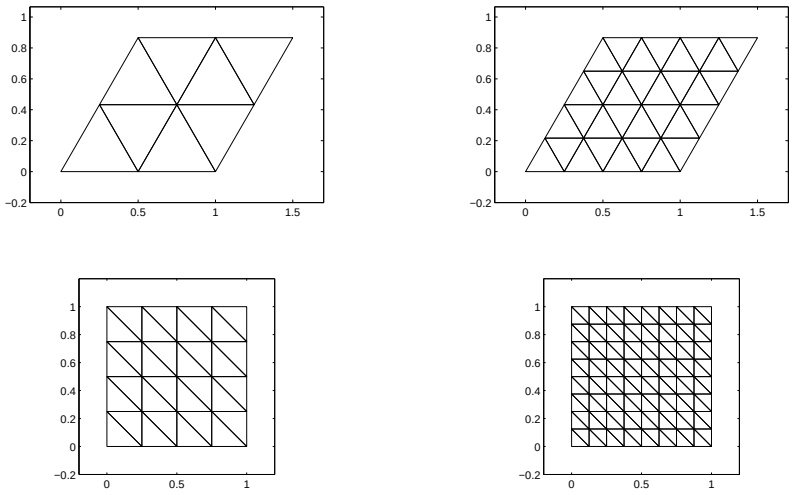

Figure 4.1: Mesh and refined mesh on two different domains Ω

The question is which bounds we obtain from Theorem 4.2.5, see Table 4.1, and how these compare with the real bounds of the DmP, see Table 4.2.

	rhombus (l)	rhombus (r)	square (l)	square (r)
mesh condition	not fulfilled	fulfilled	not fulfilled	not fulfilled
lower bound for θ	–	0.9644	–	–
$\theta = 1/2$, bounds for Δt	–	–	–	–
$\theta = 1$, bound for Δt	–	0.1399	–	–

Table 4.1: Bounds of DmP obtained from Theorem 4.2.5

	rhombus (l)	rhombus (r)	square (l)	square (r)
some mesh condition	fulfilled	fulfilled	not fulfilled	fulfilled
lower bound for θ	0	0.8525	–	0.9809
$\theta = 1/2$, bounds for Δt	0 and 0.0476	–	–	–
$\theta = 1$, bound for Δt	0	0.0415	–	0.0699

Table 4.2: The real bounds of DmP

Giving an explanation on the results showed by Tables 4.1 and 4.2 we note that the symbol "–" means in Table 4.1 that we cannot choose the corresponding parameter to fulfil the DmP by Theorem 4.2.5, and in Table 4.2 that it is not possible to choose the corresponding parameter to fulfil the DmP in fact.

The column corresponding to rhombus (l) is problematic from Theorem 4.2.5 since the mesh is too coarse. The columns corresponding to rhombus (r) are completely comparable, in this case Theorem 4.2.5 works very well.

The square (l) was problematic in reality not only for our theorem, the right angles are intolerable for Theorem 4.2.5 and the situation cannot alter by refining the mesh, however, in reality it helps a little bit, but only a little bit as it is shown in the column corresponding to square (r) of Table 4.2.

Finally we turn our attention to the column corresponding to square (l) of Table 4.2 again, since it seems to be mysterious. After fixing the mesh we cannot choose the parameter θ and the time step Δt to ensure the DmP (in reality). It means that we can spoil the things already with an inadequate choice of the mesh! This is represented by the row "some mesh condition" in Table 4.2. We investigate this (temporarily unknown) property in the following section.

4.3 Relation between discrete elliptic and discrete parabolic maximum principles

In this section we are looking for an answer to the problem the numerical results give rise to. This section is based on the paper [40].

4.3.1 Discrete stabilization property, discrete elliptic and discrete parabolic maximum principles

We return to the algebraic framework, thus we investigate the hyper-matrix $\mathcal{L}$, which can be defined by the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$. Let $\mathbf{K}$ be defined as $\mathbf{K} = \mathbf{X}_1 - \mathbf{X}_2$. We recall that if $\mathbf{X}_{10}$ is non-singular, then we can introduce the notation

$$\mathbf{T} = \mathbf{X}_{10}^{-1}\mathbf{X}_{20} \tag{4.10}$$

and with that the iteration form (4.2) can be rewritten as

$$\mathbf{v}_0^n = \mathbf{T}\mathbf{v}_0^{n-1} + \mathbf{X}_{10}^{-1}\mathbf{X}_2\mathbf{v}_\partial^{n-1} - \mathbf{X}_{10}^{-1}\mathbf{X}_{1\partial}\mathbf{v}_\partial^n + \mathbf{X}_{10}^{-1}(\mathcal{L}\nu)^n\,, \quad n = 1, \dots\,. \tag{4.11}$$

The discrete stabilization property.

Definition 4.3.1. The hyper-matrix $\mathcal{L}$ possesses the *discrete stabilization property* (DSP) if $\mathbf{K}_0$ is non-singular and for all $\mathbf{u}$, $\mathbf{v}_0^0$ the iteration

$$\mathbf{X}_1\mathbf{v}^n - \mathbf{X}_2\mathbf{v}^{n-1} = \mathbf{K}\mathbf{u}\,, \quad \mathbf{v}_\partial^{n-1} = \mathbf{u}_\partial\,, \quad n = 1, \dots$$

is convergent, moreover

$$\mathbf{v}^n \to \mathbf{u}$$

holds.

Remark 4.3.2. DSP is related to some continuous property, which is called in various ways in the literature, we favour the name stabilization property (SP), but absolute stability is used, e.g., in [31, Ch.10.1] instead of SP, where it is explained in a simple way for the Laplace operator.

To characterise the DSP we need to recall some notions of the matrix splitting theory. This can be found in the Appendix. We collected there the basic results of that topic at the Reader's convenience, too.

Lemma 4.3.3. *The hyper-matrix $\mathcal{L}$ possesses the DSP if and only if $\mathbf{K}_0 = \mathbf{X}_{10} - \mathbf{X}_{20}$ defines a convergent splitting.*

Proof. This is a trivial consequence of the corresponding part of the Appendix, namely of Definition 5.0.24, of Remark 5.0.25 and of the iteration form (4.11). □

The relation of the DSP, the DwMP and the DmP. Here we show the connection between the discrete stabilization property and the discrete elliptic and parabolic maximum principles. Note that the notation DnP is used for matrices (elliptic case) and for hyper-matrices (parabolic case) too, in this case we tried to make it clear which property we are talking about.

Theorem 4.3.4. *We assume that the hyper-matrix $\mathcal{L}$ possesses the DnP property. Then the DnP property of $\mathbf{K}$ is equivalent to the DSP of $\mathcal{L}$.*

Proof. – We assume the DnP property of $\mathbf{K}$ besides the DnP property of $\mathcal{L}$. It means that $\mathbf{K}_0^{-1} \geq \mathbf{0}$ holds, and this implies the DSP of $\mathcal{L}$ by Lemma 4.3.3 and Theorem 5.0.26.

– We assume the DSP and the DnP property of $\mathcal{L}$.

Then $\mathbf{K}_0^{-1} \geq \mathbf{0}$ holds by Lemma 4.3.3 and Theorem 5.0.26.

$$-\mathbf{K}_0^{-1}\mathbf{K}_\partial = (\mathbf{I}-\mathbf{T})^{-1}(-\mathbf{X}_{10}^{-1}\mathbf{K}_\partial) = \underbrace{\left(\sum_{k=0}^{\infty}\mathbf{T}^k\right)}_{\geq \mathbf{0}}(\underbrace{\mathbf{X}_{10}^{-1}\mathbf{X}_{2\partial}}_{\geq \mathbf{0}}\underbrace{-\mathbf{X}_{10}^{-1}\mathbf{X}_{1\partial}}_{\geq \mathbf{0}}) \geq \mathbf{0}\,, \quad (4.12)$$

due to (m2), (m3) and Lemma 5.0.27 (see Appendix).

□

The main theorem of this chapter comes, which will give the theoretical answer on our open problem from the last chapter.

Theorem 4.3.5. *We assume that the hyper-matrix $\mathcal{L}$ defines a non-singular matrix $\mathbf{K}_0$. Then the DmP of $\mathcal{L}$ implies the DSP for $\mathcal{L}$ and the DwMP for $\mathbf{K}$.*

In order to prove this theorem, first we give several useful results.

Lemma 4.3.6. *The DmP of $\mathcal{L}$ implies $\|\mathbf{T}\|_\infty \leq 1$.*

Proof. From (m4) we have

$$\mathbf{e} \geq \mathbf{X}_{10}^{-1}\mathbf{X}_2\mathbf{e} - \mathbf{X}_{10}^{-1}\mathbf{X}_{1\partial}\mathbf{e} = \mathbf{T}\mathbf{e} + \underbrace{\mathbf{X}_{10}^{-1}\mathbf{X}_{2\partial}}_{\geq \mathbf{0}}\mathbf{e}\underbrace{-\mathbf{X}_{10}^{-1}\mathbf{X}_{1\partial}}_{\geq \mathbf{0}}\mathbf{e} \geq \mathbf{T}\mathbf{e}\,,$$

due to (m3) and (m2), respectively. Finally, the claimed result follows from the non-negativity of $\mathbf{T}$, which is guaranteed by (m3). □

Remark 4.3.7. Note that Lemma 4.3.6 has a simple consequence, namely, the DmP of $\mathcal{L}$ implies $\rho(\mathbf{T}) \leq 1$. If for the matrix $\mathbf{T}$ in the form (4.10) the property $\|\mathbf{T}\|_\infty \leq 1$ holds, then in [14] $\mathcal{L}$ is said to possesses the discrete maximum norm contractivity (DMNC), however, the name non-expansivity would be certainly more accurate.

Lemma 4.3.8. *Let us fix that $\mathbf{K} = \mathbf{X}_1 - \mathbf{X}_2$. If $\mathbf{K}_0$ and $\mathbf{X}_{10}$ are non-singular, then $\mathbf{I} - \mathbf{T}$ is nonsingular, too. Thus, one is not an eigenvalue of $\mathbf{T}$.*

Proof.

$$\mathbf{X}_{10}^{-1}\mathbf{K}_0 = \mathbf{I} - \mathbf{T} \tag{4.13}$$

and the left side is invertible. □

Now we are ready with the preparations.

Proof. (of Theorem 4.3.5)

We assume the DmP of $\mathcal{L}$ and that $\mathbf{K}_0$ is non-singular.

- First, we prove that the DSP holds. To show that, we need to prove that $\rho(\mathbf{T}) < 1$, according to Lemma 4.3.3. We already know from Lemma 4.3.6 and Remark 4.3.7 that the DPMP implies $\rho(\mathbf{T}) \leq 1$. We suppose that $\rho(\mathbf{T}) = 1$. Then one is an eigenvalue of $\mathbf{T}$, due to the non-negativity of $\mathbf{T}$ and Theorem 5.0.28 (consequence of the Perron-Frobenius theorem, see Appendix). On the other hand, using Lemma 4.3.8 contradicts to that. Thus, we proved that $\rho(\mathbf{T}) < 1$.

- Second, we prove the DwMP.

  (w1) and (w2) follows from Theorem 4.3.4.

  (w3) results from (m4), which is equivalent to

  $$(\mathbf{I}-\mathbf{T})\mathbf{e} \geq -\mathbf{X}_{10}^{-1}\mathbf{K}_{\partial}\mathbf{e}\,.$$

  Multiplying with $(\mathbf{I}-\mathbf{T})^{-1} \geq \mathbf{0}$ and using the first part of the identity (4.12) gives the desired result.

□

Remark 4.3.9. The conclusion is that the DSP of $\mathcal{L}$ and the DwMP of $\mathbf{K}$ are necessary to fulfil the DmP, but not sufficient, as the next example (constructed by using [38, Ex. 4.1]) shows us:

$$\mathbf{K}_0 = \begin{pmatrix} 4 & -1 \\ -1 & 1 \end{pmatrix} \quad \mathbf{X}_{10} = \begin{pmatrix} 14 & 4 \\ 0 & 2 \end{pmatrix} \quad \mathbf{K}_{\partial} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

since in this case

$$\mathbf{K}_0^{-1} = \begin{pmatrix} 1/3 & 1/3 \\ 1/3 & 4/3 \end{pmatrix} \quad \mathbf{X}_{10}^{-1} = \begin{pmatrix} 1/14 & -2/14 \\ 0 & 1/2 \end{pmatrix} \ngeq \mathbf{0} \quad \mathbf{T} = \begin{pmatrix} 4/7 & 3/14 \\ 1/2 & 1/2 \end{pmatrix},$$

thus $\rho(\mathbf{T}) < 1$, and it can be seen that (w1)–(w3) hold, on the other hand, (m1) fails.

4.3.2 Numerical examples revisited

Now, based on Theorem 4.3.5 we can define that unknown "some mesh condition" of Table 4.2. It is the DwMP for the matrix $\mathbf{K}$. One can check that for square (l) already the condition $\mathbf{K}_0^{-1} \geq \mathbf{0}$ fails, thus the DmP fails, too.

In the following more numerical examples are investigated from the sufficiency point of view c.f. Remark 4.3.9. We assume that $K = -\triangle$ and $L = \frac{\partial}{\partial t} - \triangle$, where $\triangle$ denotes the Laplacian operator. For these operators the corresponding maximum principles hold. We choose different domains, methods and parameter settings. We focus on the conditions $\mathbf{K}_0^{-1} \geq \mathbf{0}$, $\mathbf{T} \geq \mathbf{0}$ and $\rho(\mathbf{T}) < 1$.

Example 4.3.10. In this case we set $\Omega = (0, 1)$. We use a FDM with uniform mesh to the space discretization – we denote the mesh parameter by h – and the θ-method to the time discretization. The usual calculation gives

$$\mathbf{X}_{10} = \text{tridiag}\left[-\frac{\theta}{h^2},\ \frac{1}{\Delta t} + \frac{2\theta}{h^2},\ -\frac{\theta}{h^2}\right],$$

$$\mathbf{X}_{20} = \text{tridiag}\left[-\frac{1-\theta}{h^2},\ \ \frac{1}{\Delta t} - \frac{2(1-\theta)}{h^2},\ \ -\frac{1-\theta}{h^2}\right],$$

where the matrices are of the size $n \times n$, and $n = \frac{1}{h} - 1$. We set $n = 4$, $\theta = 1/2$ (Crank-Nicolson scheme) and $\Delta t = 0.05$. Then one can check that the conditions $\mathbf{K}_0^{-1} \geq \mathbf{0}$ and $\rho(\mathbf{T}) < 1$ hold, while the condition $\mathbf{T} \geq \mathbf{0}$ fails. Refining the time step to $\Delta t = 0.04$ (and keeping the other parameters) we find that all the three conditions hold.

Example 4.3.11. In this case we set $\Omega = (0,1)^2$. We use a FEM with a uniform triangle mesh – see Figure 4.1, square – to the space discretization – we denote the mesh parameter by h – and the θ-method to the time discretization. The usual calculation gives

$$\mathbf{M}_0 = \frac{h^2}{2}\text{tridiag}\left[\text{tridiag}\left[0, 1/6, 1/6\right], \text{tridiag}\left[1/6, 1, 1/6\right], \text{tridiag}\left[1/6, 1/6, 0\right]\right],$$

$$\mathbf{K}_0 = \text{tridiag}\left[-\mathbf{I},\ \ \text{tridiag}\left[-1, 4, -1\right], -\mathbf{I}\right],$$

where the matrices $\mathbf{M}_0$ and $\mathbf{K}_0$ are of the size $n^2 \times n^2$, and $n = \frac{1}{h} - 1$.

First we set $n = 3$ (Figure 4.1, square (l)), $\theta = 0.9$ and $\Delta t = 0.1$. Then one can check that the conditions $\mathbf{K}_0^{-1} \geq \mathbf{0}$ and $\rho(\mathbf{T}) < 1$ hold, while the condition $\mathbf{T} \geq \mathbf{0}$ fails. Choosing the time step as $\Delta t = 0.05$ (and keeping the other parameters) we find that the conditions $\mathbf{K}_0^{-1} \geq \mathbf{0}$ and $\rho(\mathbf{T}) < 1$ hold, as well as the condition $\mathbf{T} \geq \mathbf{0}$.

Second we set $n = 7$ (Figure 4.1, square (r)), $\theta = 0.9$ and $\Delta t = 0.05$. Then one can check that the conditions $\mathbf{K}_0^{-1} \geq \mathbf{0}$ and $\rho(\mathbf{T}) < 1$ hold, while the condition $\mathbf{T} \geq \mathbf{0}$ fails. Choosing the time step as $\Delta t = 0.01$ (and keeping the other parameters) we find that the conditions $\mathbf{K}_0^{-1} \geq \mathbf{0}$ and $\rho(\mathbf{T}) < 1$ hold, as well as the condition $\mathbf{T} \geq \mathbf{0}$.

The above examples demonstrate that the DmP implies the DwMP and the DSP, but the converse implication fails.

Summary of the chapter. In this chapter in Section 4.1 an algebraic framework was presented on discrete maximum principles for hyper-matrices. Both theoretical and practical conditions were listed on discrete maximum principles besides investigating the applicability of the framework.

In Section 4.2 by using this framework we investigated some parabolic operator when the FEM + θ-method was applied as a discretization. We gave sufficient conditions on the mesh, on the time step and on the parameter θ to fulfil the DmP and the DMP, see Theorem 4.2.5. We investigated the sharpness of the conditions of this theorem with several numerical examples. Section 4.2 was based on the paper [39, Mincsovics, 2010].

In Section 4.3 the relation of discrete elliptic and parabolic maximum principles was investigated. We introduced the notion of DSP. In Theorem 4.3.4 we stated that under the parabolic DnP property the elliptic DnP property is equivalent to the DSP. In Theorem 4.3.5 we stated that DmP implies DwMP and DSP. The practical conclusion of these theorems is that with an inadequate mesh (independently of the choice of the time step and parameter θ) the DmP can be spoiled. To illustrate this we added some numerical examples. Section 4.3 was based on the paper [40, Mincsovics, 2010].

†

Chapter 5

Appendix

In this chapter we collected the definitions and results we used in the other chapters.

1 Basic notions in numerical analysis

proof of Lemma 1.1.22. It is enough to show that $B_{R/S}(G(v)) \subset G(B_R(v))$, due to Corollary 1.1.21. We assume indirectly that there exists $w \in B_{R/S}(G(v))$ such that $w \notin G(B_R(v))$. We define the line $w(\lambda) = (1-\lambda)G(v) + \lambda w$ for $\lambda \geq 0$, and introduce the number $\hat{\lambda}$ as follows:

$$\hat{\lambda} := \begin{cases} \sup\{\lambda' > 0 \,|\, w(\lambda) \in G(B_R(v)) \,\forall \lambda \in [0, \lambda')\} \text{ , if it exists,} \\ 0 \text{ , else.} \end{cases}$$

Then clearly the inequality $\hat{\lambda} \leq 1$ holds. We will show that $\hat{w} =: w(\hat{\lambda}) \in G(B_R(v))$.

For $\hat{\lambda} = 0$ this trivially holds. For $\hat{\lambda} > 0$ we observe that G is invertible on $w(\hat{\lambda} - \varepsilon)$, (i.e., the elements $G^{-1}(w(\hat{\lambda} - \varepsilon)) \in B_R(v)$ exist) for all $\varepsilon : \ \hat{\lambda} \geq \varepsilon > 0$. Thus, we can use the stability estimate (1.14)

$$\begin{aligned} \left\| G^{-1}(w(\hat{\lambda} - \varepsilon)) - v \right\|_{\mathcal{V}} \leq S \left\| w(\hat{\lambda} - \varepsilon) - G(v) \right\|_{\mathcal{W}} = \\ S(\hat{\lambda} - \varepsilon) \underbrace{\|w - G(v)\|_{\mathcal{W}}}_{= \frac{R}{S} - \frac{\delta}{S}} < \hat{\lambda}(R - \delta) \leq R - \delta \,, \end{aligned}$$

for some $\delta > 0$, and using again the stability estimate we can conclude that the function $h(\varepsilon) = G^{-1}(w(\hat{\lambda} - \varepsilon))$ is uniformly continuous at $\varepsilon \in (0, \hat{\lambda}]$. Thus, there exists $\lim_{\varepsilon \searrow 0} h(\varepsilon) =: z \in B_R(v)$. Using the continuity of G, we get $G(z) = \hat{w}$.

Now we can choose a closed ball $\bar{B}_r(z) \subset B_R(v)$, $(r > 0)$ whose image $G(\bar{B}_r(z))$ contains a neighbourhood of $\hat{w}$, due to Brouwer's invariance domain theorem. This results in a contradiction.

Finally, the Lipschitz continuity with the constant S is a simple consequence of (1.14). □

Definition 5.0.12. A real square matrix is said to be a *Z-matrix* if its off-diagonal entries are nonpositive.

Definition 5.0.13. We call a real square matrix *M-matrix* if it can be represented as $s\mathbf{I}-\mathbf{B}$, where $\mathbf{I}$ is the identity matrix and $\mathbf{B} \leq \mathbf{0}$ (i.e. each entries of the matrix $\mathbf{B}$ are nonpositive), moreover $s \geq \varrho(\mathbf{B})$, where ϱ denotes the spectral radius of a matrix.

It is obvious that an M-matrix is a Z-matrix, too.

Theorem 5.0.14. *[3, Ch.6, Th.2.3] We assume that the matrix* $\mathbf{A}$ *is a Z-matrix. Then the following statements are equivalent.*

1. $\mathbf{A}$ *is a nonsingular M-matrix.*
2. *There exists* $\mathbf{d} > \mathbf{0}$ *with* $\mathbf{Ad} > \mathbf{0}$.
3. *There exists* $\mathbf{A}^{-1}$, *and* $\mathbf{A}^{-1} \geq \mathbf{0}$.

The following lemma (which can be found e.g. in [51, I/Lemma 1.8.]) provides a tool to estimate the norm of the inverse of an M-matrix.

Lemma 5.0.15. *We assume that the matrix* $\mathbf{A}$ *is a nonsingular M-matrix with the dominant vector* $\mathbf{d}$. *Then*

$$\|\mathbf{A}^{-1}\|_\infty \leq \frac{\|\mathbf{d}\|_\infty}{\min(\mathbf{Ad})_i}\,. \tag{5.1}$$

†

2 Maximum principles

Definition 5.0.16. We say that K, defined in (2.1), is *uniformly elliptic* if there exists a constant $m > 0$ such that

$$\sum_{i,j=1}^{d} a_{ij}(\mathbf{x})\xi_i\xi_j \geq m\,|\xi|^2$$

holds for all $\mathbf{x} \in \Omega$, $\xi = (\xi_1, \xi_2, \ldots, \xi_d) \in \mathbb{R}^d$.

Definition 5.0.17. We say that L, defined in (2.4), is *uniformly parabolic* if there exists a constant $m > 0$ such that

$$\sum_{i,j=1}^{d} a_{ij}(\mathbf{x}, t)\xi_i\xi_j \geq m\,|\xi|^2$$

holds for all $(\mathbf{x}, t) \in \Omega \times (0, T]$, $\xi \in \mathbb{R}^d$.

†

3 Discrete elliptic maximum principles

Definition 5.0.18.

- $\mathbf{A} \in \mathbb{R}^{n\times n}$ is *cogredient* to $\mathbf{E} \in \mathbb{R}^{n\times n}$ if for some permutation matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$, $\mathbf{PAP}^T = \mathbf{E}$.
- $\mathbf{A}$ is *reducible* if it is cogredient to

$$\begin{pmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{C} & \mathbf{D} \end{pmatrix},$$

where $\mathbf{B}$ and $\mathbf{D}$ are square matrices, or if $n = 1$ and $\mathbf{A} = \mathbf{0}$. Otherwise, $\mathbf{A}$ is *irreducible*.

Definition 5.0.19.

- $\mathbf{A} \in \mathbb{R}^{n\times n}$ is *diagonally dominant* (DD) if

$$|A_{ii}| \geq \sum_{j\neq i} |A_{ij}| \tag{5.2}$$

holds for all $1 \leq i \leq n$.

- $\mathbf{A} \in \mathbb{R}^{n\times n}$ is *strictly diagonally dominant* (SDD) if strict inequality is valid for all $1 \leq i \leq n$ in (5.2).
- $\mathbf{A} \in \mathbb{R}^{n\times n}$ is *irreducibly diagonally dominant* (IDD) if it is irreducible and DD, moreover, strict inequality is valid for at least one i in (5.2).

Definition 5.0.20. A Z-matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is a *Stieltjes matrix* if it is symmetric and positive definite.

The above definitions can be found in almost every textbook on the theory of matrices e.g. in [3], or in [55]. In the following some basic results are presented on the introduced notions, based also on the aforementioned books.

Lemma 5.0.21. *[55, Cor. 3.20.] If* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *is an IDD Z-matrix with positive diagonal entries, then* $\mathbf{A}^{-1} > \mathbf{0}$.

Theorem 5.0.22. *[3, part of Thm. 2.7. in Ch. 6.2.] We assume that* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *is an irreducible Z-matrix. Then the following two statements are equivalent.*

(i) $\mathbf{A}$ *is a nonsingular M-matrix;*

(ii) $\mathbf{A}^{-1} > \mathbf{0}$.

Theorem 5.0.23. *[3, part of Thm. 2.3. in Ch. 6.2.] We assume that* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *is a Z-matrix. Then the following two statements are equivalent.*

(i) $\mathbf{A}$ *is a nonsingular M-matrix;*

(ii) $\mathbf{A}^{-1} \geq \mathbf{0}$.

†

4 Discrete parabolic maximum principles Matrix splitting theory plays a fundamental role in solving large system of linear equations. Here we give only a short introduction into the basic definitions and results which will be important for us. The Reader can find more about this topic in [3, 55, 59].

Definition 5.0.24. For the non-singular matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ the decomposition $\mathbf{A} = \mathbf{M} - \mathbf{N}$ represents a *splitting of* $\mathbf{A}$, where $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{N\times N}$.

- The splitting is called *convergent splitting* if $\mathbf{M}$ is non-singular with $\rho(\mathbf{M}^{-1}\mathbf{N}) < 1$.
- The splitting is called *weak regular splitting* if $\mathbf{M}$ is non-singular with $\mathbf{M}^{-1} \geq \mathbf{0}$ and $\mathbf{M}^{-1}\mathbf{N} \geq \mathbf{0}$.

Remark 5.0.25. The idea behind the notion of convergent splitting can be explained as follows. Consider the linear system of equations $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is non-singular. Then for the iteration $\mathbf{My}^n - \mathbf{Ny}^{n-1} = \mathbf{b}$, $\mathbf{y}^n \to \mathbf{x}$ for every initial vector $\mathbf{y}^0$ if and only if $\mathbf{A} = \mathbf{M} - \mathbf{N}$ defines a convergent splitting.

The following theorem summarizes the essence of the relation of the above given matrix splitting types.

Theorem 5.0.26. *[3, Ch.6, Th.2.3] For the non-singular matrix* $\mathbf{A} \in \mathbb{R}^{N\times N}$ *the following statements are equivalent.*

(a) $\mathbf{A}^{-1} \geq \mathbf{0}$.

(b1) *There exists a convergent weak regular splitting of* $\mathbf{A}$.

(b2) *There exists a weak regular splitting of* $\mathbf{A}$ *and every weak regular splitting of* $\mathbf{A}$ *is a convergent splitting.*

The following two results are used in the proofs of Theorems 4.3.4 and 4.3.5.

Lemma 5.0.27. *[55, Th.3.15] If for an arbitrary matrix* $\mathbf{T} \in \mathbb{R}^{N\times N}$ $\rho(\mathbf{T}) < 1$ *holds, then* $\mathbf{I} - \mathbf{T}$ *is non-singular and*

$$(\mathbf{I} - \mathbf{T})^{-1} = \sum_{k=0}^{\infty} \mathbf{T}^k = \mathbf{I} + \mathbf{T} + \mathbf{T}^2 + \dots . \tag{5.3}$$

The following theorem is a consequence of the Perron–Frobenius theorem, see e.g. in [59, Th.2.2].

Theorem 5.0.28. *If* $\mathbf{T} \geq \mathbf{0}$*, then* $\rho(\mathbf{T})$ *is an eigenvalue of the matrix* $\mathbf{T}$.

†

Conclusions

This dissertation consisted of two parts. The topic of the first part was the Lax theory of the numerical solution of linear and nonlinear equations. The second part dealt with discrete elliptic and parabolic maximum principles.

To approximate the solution of some equation, usually a numerical method is used, the success of which depends on its convergence. The definition of convergence is theoretical since it contains the unknown solution, however, this problem can be solved with the following idea. The directly unverifiable notion of convergence can be substituted with the notions of consistency and stability. In the linear case stability and convergence are equivalent under the consistency assumption, this is the Lax equivalence theorem.

In the first part of the dissertation, our goal was to present a framework that unifies the known results, completes the theory and clarifies the relations between the basic notions of consistency, stability and convergence. These goals were realized in the following way.

- We reformulated the results of Stetter in order to fit into our framework (since we used a different stability notion).

- We completed the nonlinear theory by adding our own results in Subsection 1.1.3, i.e., we introduced the notion of dense consistency, see Definition 1.1.28 and we proved that dense consistency together with stability together implies convergence, see Theorem 1.1.36. Moreover, we stated that stability "near to the solution" implies stability, see Lemma 1.1.37. These results together provide the opportunity for using our nonlinear framework in applications.

- We gave numerous examples in order to shed some light on the relation of the basic notions in the nonlinear case, see Subsection 1.1.4. We proceeded in the same way in the linear case, too, see Subsection 1.2.2.

When choosing a numerical method to approximate the solution of a continuous mathematical problem, the first thing to consider is which method results in a good approximation from a quantitative point of view. This was investigated in the first part of the thesis. However, in most of the cases it is not enough. The original problem (which is usually some model of a phenomenon) possesses important qualitative properties, and a natural requirement from the numerical solution is to preserve these qualitative properties. E.g., when we seek an approximation of the Laplace's equation where the boundary condition is defined to be nonnegative then the solution is nonnegative, too and a good approximation should be nonnegative as well. For linear elliptic and parabolic problems the main qualitative properties are the various maximum principles.

In Chapter 3, which dealt with discrete elliptic maximum principles, our aim was twofold. Firstly, we wanted to present a unified algebraic framework giving the known results and completing the theory with our results on discrete strong maximum principles. Secondly, we wanted to apply this framework on a certain problem. These were realized in the following way.

- In Section 3.1, which is based on the paper [41, Mincsovics and Horváth, 2012], we investigated six different types of maximum principles including the most known ones, like the discrete weak non-positivity preservation property (DnP) and the discrete weak maximum principle (DwMP). We presented sufficient and necessary conditions for each of these discrete maximum principles, including our own results on the strong maximum principles. See the discrete strong non-positivity preservation property in Lemma 3.1.5, the discrete strong maximum principle (DsMP)in Theorem 3.1.10 and the discrete strictly strong maximum principle (DSMP) in Theorem 3.1.9.

- In the same section, we gave an overview on practical conditions ensuring the DwMP, the DsMP and the DSMP listing the known results and completing with our own conditions.

- We also investigated the applicability of our algebraic framework. See Subsection 3.1.3.

- We illustrated the differences between the weak and strong discrete maximum principles with several numerical examples. See Section 3.2, which is also based on [41, Mincsovics and Horváth, 2012].

- In Section 3.3, based on [28, Horváth and Mincsovics, 2013], using the algebraic framework we investigated some elliptic problem where an interior penalty

discontinuous Galerkin method is applied as discretization. We gave sufficient conditions on the parameters ε and σ and on the mesh fulfilling the DnP and the DwMP, see Theorem 3.3.2 and Theorem 3.3.3, respectively. We investigated the sharpness of the necessary conditions of these theorems with numerical examples as well.

In Chapter 4, which dealt with discrete parabolic maximum principles, our aim was the following. Firstly, to present an algebraic framework on discrete parabolic maximum principles collecting the known results. Next, we wanted to apply this framework on a certain practical problem. Finally, we also wanted to find some connection between discrete elliptic and discrete parabolic maximum principles. These were realized in the following steps.

- In Section 4.1 we presented an algebraic framework on discrete parabolic maximum principles. We studied three types of maximum principles, listing the known sufficient and necessary conditions for each type. We also investigated the applicability of the framework.

- In Section 4.2, based on [39, Mincsovics, 2010], we investigated a parabolic problem when some FEM + θ-method discretization is used and we derived practical conditions under which the most important discrete parabolic maximum principles can be preserved, see Theorem 4.2.5.

  In Subsection 4.2.3 we presented numerical examples showing that a not carefully chosen mesh can already hinder to fulfil discrete parabolic maximum principles.

- In Section 4.3, based on [40, Mincsovics, 2010], we introduced a new notion, the discrete stabilization property (DSP), see Definition 4.3.1. We gave sufficient and necessary condition to fulfil this property in Lemma 4.3.3. Additionally, we presented our results on the relation of the DSP and the discrete elliptic and discrete parabolic maximum principles, see Theorems 4.3.4 and 4.3.5.

  These results explain the earlier mentioned property, namely, that a non-adequate mesh can already hinder to fulfil discrete parabolic maximum principles.

Bibliography

[1] Ainsworth, M., and Rankin, R.: Technical Note: A note on the selection of the penalty parameter for discontinuous Galerkin finite element schemes. Numerical Methods for Partial Differential Equations, 28, (3), 1099–1104 (2012)

[2] Arnold, D. N., Brezzi, F., Cockburn, B., Marini, D.: Unified analysis of discontinuous Galerkin methods for elliptic problems. SIAM Journal on Numerical Analysis, 39, 1749–1779 (2002)

[3] Berman, A. and Plemmons, R. J.: Nonnegative matrices in the mathematical sciences. Academic Press, New York, (1979)

[4] Brandts, J., Korotov, S., Krizek, M.: Simplicial finite elements in higher dimensions. Applications of Mathematics 52, 251–265 (2006)

[5] Ciarlet, P. G.: Discrete maximum principle for finite-difference operators. Aequationes Math. 4, 338–352, (1970)

[6] Ciarlet, P. G., Raviart, P.-A.: Maximum principle and uniform convergence for the finite element method. Comput. Methods Appl. Mech. Engrg., 2, 17–31 (1973)

[7] Di Pierto, D. A. and Ern, A.: Mathematical Aspects of Discontinuous Galerkin Methods. Springer-Verlag, New York, (2012)

[8] Draganescu, A., Dupont, T. F., Scott, L. R.: Failure of the discrete maximum principle for an elliptic finite element problem. Math. Comp., 74, n. 249, 1–23 (2005)

[9] Ern, A. and Guermond, J.-L.: Theory and practice of finite elements. Springer-Verlag, New York, (2004)

[10] Evans, L. C.: Partial Differential Equations. Graduate Studies in Mathematics Vol. 19, AMS (1997)

[11] Faragó, I.: Numerical Treatment of Linear Parabolic Problems. Dissertation for the degree MTA Doktora (2008)

[12] Faragó, I.: Matrix and Discrete Maximum Principles. LSSC 2009, LNCS 5910, 563–570 (2010)

[13] Faragó, I., Horváth, R., Korotov, S.: Discrete maximum principle for linear parabolic problems solved on hybrid meshes. Appl. Num. Math., 53, 249–264 (2005)

[14] Faragó, I., Horváth, R.: Discrete maximum principle and adequate discretizations of linear parabolic problems. SIAM Sci. Comput., 28, 2313–2336 (2006)

[15] Faragó, I., Horváth, R.: A review of reliable numerical models for three-dimensional linear parabolic problems. Int. J. Numer. Meth. Engng., 70, 25–45 (2007)

[16] Faragó, I., Horváth, R.: A Review of Reliable Numerical Models for Three-Dimensional Linear Parabolic Problems. Int. J. Numer. Meth. Engng., 70, 25–45 (2007)

[17] Faragó, I., Horváth, R.: Continuous and discrete parabolic operators and their qualitative properties. IMA Journal of Numerical Analysis 29, 606–631 (2009)

[18] Faragó, I.: Discrete maximum principle for finite element parabolic models in higher dimensions. Math. Comp. Sim., 80, 1601–1611 (2010)

[19] Faragó, I., Horváth, R.: Qualitative Properties of Monotone Linear Parabolic Operators. E. J. of Qualitative Theory of Differential Equations, Proc. 8th Coll. QTDE, 2008, No. 8, 1–15 (2009)

[20] Faragó, I., Korotov, S. and Szabó, T.: On modifications of continuous and discrete maximum principles for reaction-diffusion problems. Advances in Applied Mathematics and Mechanics, 3(1), 109–120 (2011)

[21] Faragó, I., Korotov, S. and Szabó, T.: On sharpness of two-sided discrete maximum principles for reaction-diffusion problems. In: Proc. of the Int. Conf. APLIMAT-2011, 247–254 (2011)

[22] Faragó, I., Korotov, S. and Szabó, T.: On continuous and discrete maximum principles for elliptic problems with the third boundary condition. Applied Mathematics and Computation, 219, 7215–7224 (2013)

[23] Faragó, I., Mincsovics, M. E., Fekete, I.: Notes on the Basic Notions in Nonlinear Numerical Analysis. E. J. of Qualitative Theory of Differential Equations, Proc. 9'th Coll. QTDE, 2011, No. 6, 1–22 (2012)

[24] Fujii, H.: Some remarks on finite element analysis of time-dependent field problems. in: Theory and Practice in Finite Element Structural Analysis. (Y. Yamada and R. H. Gallagher eds.), Tokyo: University of Tokyo Press, 91–106. (1973)

[25] Hannukainen, A., Korotov, S., Vejchodský, T.: On weakening conditions for discrete maximum principles for linear finite element schemes. NAA 2008, LNCS 5434, 297–304 (2009)

[26] Holand, I., Bell, K.: Finite element methods in stress analysis. Tapir, Trondheim (1996)

[27] Houston, P., Suli, E., Wihler, T. P.: A posteriori error analysis of hp-version discontinuous Galerkin finite-element methods for second-order quasi-linear elliptic PDEs. IMA Journal of Numerical Analysis, 28, (2), 245–273 (2008)

[28] Horváth, T. L. and Mincsovics, M. E.: Discrete maximum principle for interior penalty discontinuous Galerkin methods. CEJM, 11 no.4, 664–679 (2013)

[29] Höhn, W. and Mittelmann H.-D.: Some remarks on the discrete maximum principle for finite elements of higher order. Computing, 27, 145–154 (1981)

[30] Ishihara, K.: Strong and weak discrete maximum principles for matrices associated with elliptic problems. Linear Algebra Appl., 88/89, 431–448 (1987)

[31] Kan, van J., Segal, A., Vermolen, F.: Numerical methods in scientific computing. VSSD (2005)

[32] Keller, H. B.: The numerical solution of parabolic partial differential equations. in: Mathematical Methods for Digital Computers ed. A Raelston, H. S. Wilf, New York, 135–143 (1960)

[33] Keller, H. B.: Approximation Methods for Nonlinear Problems with Application to Two-Point Boundary Value Problems. Math. Comput., 130, 464–474 (1975)

[34] Knabner, P. and Angermann, L.: Numerical Methods for Elliptic and Parabolic Partial Differential Equations. Springer-Verlag, New York (2003)

[35] Lax, P. D.: Functional Analysis. John Wiley and Sons, Inc., New York, NY (2002)

[36] Lax, P. D. and Richtmyer, R. D.: Survey of Stability of Linear Finite Difference Equations. Comm. Pure Appl. Math., 9, 267–293 (1956)

[37] López-Marcos, J. C. and Sanz-Serna, J. M.: Stability and Convergence in Numerical Analysis III: Linear Investigation of Nonlinear Stability. IMA J. Numer. Anal., 8, 71–84 (1988)

[38] Marek, I. and Szyld, D. B.: Comparison theorems for weak splittings of bounded operators. Numer. Math. 58, 387–397 (1990)

[39] Mincsovics, M. E.: Discrete maximum principle for finite element parabolic operators. LSSC 2009, LNCS 5910, 604–612 (2010)

[40] Mincsovics, M. E.: Discrete and continuous maximum principles for parabolic and elliptic operators. JCAM 235, 470–477 (2010)

[41] Mincsovics, M. E. and Horváth, T. L.: On the differences of the discrete weak and strong maximum principles for elliptic operators. LSSC 2011, LNCS 7116, 614–621 (2012)

[42] Palencia, C. and Sanz-Serna, J. M.: An Extension of the Lax-Richtmyer Theory. Numer. Math., 44, 279–283 (1984)

[43] Palencia, C. and Sanz-Serna, J. M.: Equivalence Theorems for Incomplete Spaces: an Appraisal. IMA J. Numer. Anal., 4, 109–115 (1984)

[44] Palencia, C. and Sanz-Serna, J. M.: A General Equivalence Theorem in the Theory of Discretization Methods. Math. of Comp., 45/171, 143–152 (1985)

[45] Rivière, B.: Discontinuous Galerkin methods for solving elliptic and parabolic equations. SIAM, (2008)

[46] Ruas Santos, V.: On the strong maximum principle for some piecewise linear finite element approximate problems of nonpositive type. J. Fac. Sci. Univ. Tokyo Sect. IA Math., 29, 473–491 (1982)

[47] Samarskii, A. A., Matus, P. P., Vabishchevich, P. N.: Difference Schemes with Operator Factors. Springer Science+Business Media, Dordrecht (2002)

[48] Stetter, H. J.: Analysis of Discretization Methods for Ordinary Differential Equations. Springer, Berlin, (1973)

[49] Stoyan, G.: On a maximum principle for matrices and on conservation of monotonicity with applications to discretization methods. Z. Angew. Math. Mech. 62, 375–381 (1982)

[50] Stoyan, G.: On maximum principles for monotone matrices. Lin. Alg. Appl. 78, 147–161 (1986)

[51] Stoyan, G. and Takó, G.: Numerikus Módszerek I–III. (Hungarian) ELTE-Typotex, Bp. (1993)

[52] Szabó, T.: Qualitative Properties of some Discretized Partial Differential Equations and Reliable Fuel Cell Modelling. Ph.D thesis (2011)

[53] Temam, R.: Navier-Stokes Equations, Theory and Numerical Analysis. North-Holland, Amsterdam (1977)

[54] Trenogin, V. A.: Functional Analysis. Nauka, Moscow, (1980) (in Russian)

[55] Varga, R. S.: Matrix Iterative Analysis. (Second Revised and Expanded Edition) Springer-Verlag, Berlin Heidelberg (2000)

[56] Varga, R.: On discrete maximum principle. J. SIAM Numer. Anal. 3, 355–359, (1966)

[57] Vejchodský, T.: Discrete Maximum Principles. Habilitation thesis, Institute of Mathematics of the Academy of Sciences and Faculty of Mathematics and Physics, Charles University, Prague (2011)

[58] Vejchodský, T., Solin, P.: Discrete maximum principle for higher-order finite elements in 1D. Math. Comput., 76, 1833–1846 (2007)

[59] Woźnicki, Z. I.: Nonnegative splitting theory. Japan J. Indust. Appl. Math., 11, 289–342 (1994)

Köszönetnyilvánítás

Szeretnék köszönetet mondani témavezetõmnek, *Faragó István*nak, akitõl rengeteg támogatást és sok türelmet kaptam az elmúlt években, illetve számtalan lehetõséget a fejlõdésre. Köszönöm *Havasi Ági*nak, hogy bármikor is fordultam hozzá, azonnal segített. Köszönet az Alkalmazott Analízis és Számításmatematikai Tanszék munkatársainak, akik körében hasznos éveket tölthettem el.

Köszönet azoknak a közösségeknek, melyeknek tagja lehettem, barátaimnak és a családomnak, akik így együtt biztos hátteret nyújtottak számomra.

Végül, köszönet mindenkinek, aki valamilyen hatással volt az életemre az utóbbi években, legyen az akár jó, akár rossz.

‡

Summary

This dissertation consisted of two parts. The first part addressed the Lax theory of numerical methods. The second part dealt with discrete elliptic and parabolic maximum principles.

To approximate the solution of some equation, usually a numerical method is used which success depends on its convergence. The definition of convergence is theoretical since it contains the unknown solution, however, this problem can be solved with the following idea. The directly unverifiable notion of convergence can be substituted with the notions consistency and stability. In the linear case stability and convergence are equivalent under the consistency assumption, this is the Lax equivalence theorem.

We presented an abstract framework which is useful for application. We showed that it is sufficient to check consistency for a set of elements whose image is dense in some neighbourhood of the zero, which can be done parallel. Moreover, it is enough to check stability "near to the solution". We investigated the relation of the basic notions (consistency, stability and convergence) providing numerous examples both in the linear and nonlinear case.

When choosing a numerical method to approximate the solution of a continuous mathematical problem, the first thing to consider is which method results in an good approximation from a quantitative point of view. This was investigated in the first part of the dissertation. However, in most of the cases it is not enough. Usually the original problem possesses important qualitative properties and a natural requirement is from the numerical solution to keep possessing these qualitative properties. For linear elliptic and parabolic problems the main qualitative properties are the various maximum principles.

We gave an algebraic framework both on discrete elliptic and discrete parabolic maximum principles. At the elliptic case we focused on the differences between the weak and strong discrete maximum principles. We investigated some elliptic problem where interior penalty discontinuous Galerkin method is applied as discretization. We gave sufficient conditions on the discretization parameters and on the mesh fulfilling the most important discrete elliptic maximum principles. We investigated a parabolic problem where some FEM + θ-method discretization is used and we derived practical conditions under which the most important discrete parabolic maximum principles can be preserved. We introduced a new notion, the discrete stabilization property (DSP), and we presented our results on the relation of the DSP and the discrete elliptic and discrete parabolic maximum principles. These results explain the property that a non-adequate mesh can already hinder to fulfil discrete parabolic maximum principles.

Összefoglalás

Ez a disszertáció két részre oszlik. Az elsõ rész numerikus módszerek Lax-féle elméletét tartalmazza, míg a második rész a diszkrét elliptikus és parabolikus maximumelvvel foglalkozik.

Egy egyenlet megoldásának approximációjához általában valamilyen numerikus módszert használunk, melynek sikerességét a konvergencia fogalmának segítségével mérhetjük. Ezen fogalom definíciója viszont tartalmazza az ismeretlen megoldást. Vagyis a konvergencia direkt úton nem ellenõrizhetõ. Ugyanakkor a konzisztencia és a stabilitás fogalmainak bevezetésével kiküszöbölhetjük ezt a problémát. Lineáris esetben a stabilitás és a konvergencia ekvivalensek, ha feltesszük a konzisztenciát, ez a Lax-féle ekvivalencia tétel.

Kidolgoztunk egy, alkalmazásoknak is megfelelõ absztrakt felépítését a témakörnek. Megmutattuk, hogy a konzisztenciát elég megvizsgálni egy halmazon, melynek képe sûrû a nulla egy környezetében. Ennek ellenõrzése párhuzamosítható. Továbbá, a stabilitást elegendõ "a megoldáshoz közel" megvizsgálni. Számos példán keresztül tárgyaltuk az alapfogalmak (konzisztencia, stabilitás és konvergencia) kapcsolatát mind a lineáris, mind a nemlineáris esetben.

Numerikus módszer használata estén az elsõ kérdés az, hogy kvantitatív szempontból megfelelõ-e. Ezt a disszertáció elsõ része tartalmazta. Ugyanakkor ez sok esetben nem elégséges. Általában a kiindulási feladat fontos kvalitatív tulajdonságokkal rendelkezik, és természetes elvárás egy numerikus módszertõl, hogy ezen tulajdonságokat õrizze meg. Elliptikus és parabolikus parciális differenciálegyenletek esetében a legfontosabb kvalitatív tulajdonságok a különbözõ maximumelvek.

Tárgyaltuk a diszkrét elliptikus és parabolikus maximumelveket algebrai keretben, ahol az elliptikus esetben az erõs és gyenge maximumelvek különbségeire fókuszáltunk. Megvizsgáltunk egy elliptikus problémát, ahol "interior penalty discontinuous Galerkin" módszert alkalmaztunk. Elégséges feltételeket adtunk a diszkretizációs paraméterekre és a rácsháló́ra, amelyek mellett megõrzõdnek a fontosabb diszkrét maximumelvek. Megvizsgáltunk egy parabolikus problémát, ahol végeselem + θ-módszert alkalmaztunk és a gyakorlatban használható feltételeket adtunk, amelyek mellett a fontos maximumelvek megõrzõdnek. Bevezettünk egy új fogalmat, a "discrete stabilization property"-t (DSP). Megmutattuk, hogy milyen kapcsolatban állnak egymással a DSP és a diszkrét elliptikus és diszkrét parabolikus maximumelvek. Ezek az eredmények mutatják, hogy nem megfelelõ rácsháló́ választása egymagában is meg tudja akadályozni a diszkrét parabolikus maximumelvek teljesülését.