Numerical Treatment of Linear Parabolic Problems

Dissertation submitted to
The Hungarian Academy of Sciences
for the degree "MTA Doktora"

István Faragó

Eötvös Loránd University

Budapest

2008

Contents

Chapter 1

Preface

Many phenomena in nature can be described by mathematical models which consist of functions of a certain number of independent variables and parameters. In particular, if some phenomenon is given by a function of spatial positions and time, then its description gives a handle to a wealth of (mathematical) models, which often consist of equations, usually containing a large variety of derivatives with respect to the variables. Apart from the spatial variable(s), which are essential in the problems to be considered, the time variable plays a special role. Indeed, many processes exhibit gradual or rapid changes as time proceeds. They are said to have an evolutionary character and an essential part of their modelling is therefore based on causality; i.e., at any time the situation is dependent of the past. Mathematical modelling of such phenomena leads to the so-called time-dependent partial differential equations, i.e., to equations that involve time t as a variable. The analysis of mathematical models of this kind is the topic of this dissertation.

Since we are not typically able to give the solution of the mathematical model in a closed (analytical) form, we construct some numerical and computer models that are useful for practical purposes. The ever-increasing advances in computer technology has enabled us to apply numerical methods to simulate plenty of physical and mechanical phenomena in science and engineering. As a result, numerical methods do not usually give the exact solution to the given problem, they can only provide approximations, getting closer and closer to the solution with each computational step. Numerical methods are generally useful only when they are implemented on computer using a computer programming language.

The study of the performance of numerical methods is called numerical analysis. This is a mathematical subject that considers estimating/controlling of the error in the processing of numerical methods and the subsequent re-design of the methods.

We note that applied mathematics started in the 17th century. Numerical aspects found a natural place in the analysis but the expression "numerical mathematics" did not exist at that time. However, numerical methods invented by Newton, Euler, and at a later stage by Gauss, still play an important role even today. In that time fundamental laws were formulated for various sub-domains of physics, like mechanics and hydrodynamics. These took the form of simple looking mathematical equations. To the disappointment of the many, these equations could be solved analytically in a few special cases only. For this reason the technological development was only loosely connected with mathematics. The appearance and availability of the modern digital computer has changed this situation. Using a computer, it is possible to gain quantitative (and later qualitative) information with detailed and realistic mathematical models and numerical methods for a multitude of phenomena and processes in physics and technology. Application of computers and

numerical methods has become ubiquitous. Computations are often cheaper than experiments; experiments can be expensive, dangerous or downright impossible. Real-life experiments can often be performed on a small scale only and that makes their results less reliable.

The present dissertation was motivated by two main objectives.

- The above modelling process of real-life phenomena

  $$\boxed{\begin{array}{l}\text{real-life problem}\\ \text{+ physical model}\end{array}} \Rightarrow \boxed{\begin{array}{l}\text{mathematical}\\ \text{model}\end{array}} \Rightarrow \boxed{\text{numerical model}} \Rightarrow \boxed{\text{computer model}}$$

  can qualitatively deform the models: those qualitative properties which are inherent in the original real-life process are not preserved for the other models. Therefore, the first goal is to guarantee quality preservation during all the above steps. (We note that the first and the last step in this modelling process are out of the scope of the dissertation.)

- It is almost obvious that the complexity of a model defines its tractableness: for structurally simple models, usually, it is easier to give qualitative characterization and/or define its solution. (For complex problems, in general, it is even impossible.) The operator splitting method is a powerful tool to decompose a complex time-dependent problem into a sequence of simpler sub-problems. The construction of such methods, their thorough analysis and application to different real-life problems are important issues of the applied mathematics.

The structure of the dissertation follows the above formulated aims.

The first chapter deals with the qualitative properties of linear parabolic problems. After a short overview, we analyze the qualitative properties in continuous models. Then we define the discrete analogues of the basic continuous properties. In both cases, the connections between the different basic qualitative properties are shown. We examine the two-level discretizations in detail, and review the finite difference and linear finite element schemes in different space dimensions. For the heat equation we examine the special scheme, known as the Crank-Nicolson method. We analyze its qualitative properties and we point out those exact bounds for the time-step under which the Crank-Nicolson scheme is qualitatively adequate. However, with a suitable modification of the Crank-Nicolson method we suggest a method which allows us to get rid of such a barrier.

The second chapter gives a systematic analysis of the operator splitting theory. We discuss the traditional methods and we present new results for their behaviour. We formulate some new operator splitting methods and analyze them. We investigate the error analysis of the operator splitting both in the cases where the split sub-problems are solved exactly and where we apply different numerical methods for the time integration of the split sub-problems.

The theoretical results are confirmed by several numerical (computer) results, a part of which is related to real-life applications.

The dissertation is based on the author's several decades of work in the field of numerical analysis. A major part of the results has already been published. Nevertheless, the other part is still unpublished, either because it has just been submitted or because it is under preparation. The author is grateful to everyone who contributed to the achievement of the results and the preparation of the dissertation. It would be a hopeless attempt to

list all the names, therefore I only mention some of them. From my teachers my first supervisor, Igor Nikolayevich Molchanov and later László Czách are those who have been motivating me during all my career as a mathematician. I am also grateful to those outstanding scientists with whom I had the opportunity to work as a co-author, especially Owe Axelsson, Cesar Palencia and Zahari Zlatev. However, my greatest thanks are to those young people, in whose careers I could actively participate in the beginning, and with whom I could work together, with some of them up to now. To my pride, several of them are recognized scientists not only in Hungary, but also abroad. I am glad and lucky to have such a long list that I would not be able to present it here. I do not want to mention any name because this could rightly hurt the remaining ones.

I thank Ágnes Havasi, Róbert Horváth and Sergey Korotov for their assistance in the preparation of the dissertation.

My last thanks are of course to my family. It is not only this dissertation that could not have been written without them: their presence and constant assistance gave me the power to achieve the results of the dissertation.

Budapest, February 2008

Chapter 2

Qualitative properties of linear parabolic problems - reliable models

Time-dependent partial differential equations are involved into mathematical models of phenomena, like heat conduction or diffusion processes, reaction-diffusion problems (such as air pollution models, e.g., [157]), problems of electrodynamics (Maxwell equations, see e.g., [138]), option pricing models (Black-Scholes models [13, 103]), and many others arising in different fields of biology, chemistry, economy, sociology, etc. It is true that the state of the art in the solution of partial differential equations has not been advanced to the level that allows the researchers to obtain close-form analytic solutions of a large number of systems. This involves the need of using numerical approach.

When we construct mathematical and/or numerical models in order to model or solve a real-life problem, these models should have different qualitative properties, which typically arise from some basic principles of the modelled phenomena. In other words, it is important to preserve characteristic properties of the original process, i.e., the models have to possess the natural equivalents of these properties. E.g., many processes, varying in time, have such properties as the monotonicity, the non-negativity preservation and the maximum principles. We will examine these qualitative properties in this part of the dissertation. We note that, even if we consider most simple problems, like the so-called heat equation, they can be viewed as a sub-problem, obtained by using the operator splitting for a more complex reaction-diffusion-advection equation. (This is the topic of the next chapter.) Hence, for such a simple problem, the conditions of the preservation of the main qualitative properties of the continuous problem play an important role, too.

2.1 History, motivation

The classical theory of partial differential equations investigates general issues such as the analytical form, existence and uniqueness of the solutions, and also propose some methods which can produce exact solutions, see, e.g., [55, 59, 83, 124]. Qualitative investigations came into being from the mid-fifties. Researchers assumed that the solution of the problem is at hand and tried to answer the questions: What kind of special properties does the solution have? What class of functions does the solution belong to? The most representative result in this field is the well-known maximum principle. A comprehensive survey of the qualitative properties of the second order linear partial differential equations can be found, e.g., in [34, 55, 110, 128, 149].

Real-life phenomena possess a number of characteristic properties. For instance, let

us consider the non-stationary heat conduction process in a physical body. When we increase the strength of the heat sources inside the body, and also the temperature on the boundary and the temperature in the initial state, then it is physically natural that the temperature does not have to decrease inside the body. Such a general property is called *monotonicity*. Clearly, when there is a certain heat source inside the body, the temperature on the boundary and the temperature in the initial state are non-negative, then the temperature inside the body is also non-negative at any fixed time. This property is called *non-negativity*. *Maximum principles* express the fact of existence of natural lower and upper bounds for the magnitude of temperature in the body. These bounds are defined by the (known) values of the temperature at the boundary, the initial state and the source. The simplest form of them maximum principle states that, if there are no heat sources and sinks present inside the body, then the maximum temperature appears also on the boundary of the body or in the initial state.

As an illustration, we present several simple numerical examples for the source-free heat conduction problem.

In the first one we solve two-dimensional heat equation with a homogeneous boundary condition in the unit square. The material parameters are set to be constant one. We apply the finite element method with bilinear elements on a rectangular mesh with mash-spacing $\Delta x = 1/10$ and $\Delta y = 1/12$. For the time discretization, the so-called Crank-Nicolson method is used with a fixed time-step Δt. A non-negative discretization of a non-negative initial function is depicted in Figure 2.1.1.

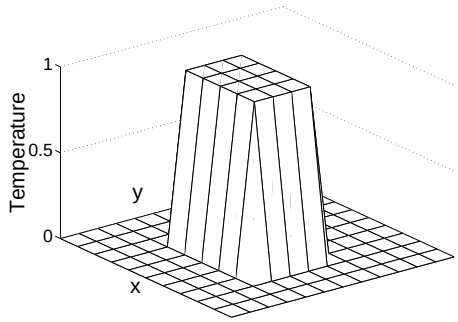

Figure 2.1.1: *Approximation of the initial function.*

Let us choose the time-step $\Delta t = 0.1$ and compute the approximation of the temperature at the fixed time level $t = 1$, i.e., at the 10-th time level. The result is shown on the left-hand side of Figure 2.1.2. The full time history of the approximation to the temperature at the fixed spatial point $(1/2, 1/6)$ on the interval $[0, 2.5]$ (i.e., during the first 25 timesteps) is displayed on the right-hand side of the same figure. We can observe that the non-negativity property of the initial temperature is not preserved. Naturally, negative values are impossible from the physical point of view, because both the initial temperature and the boundary temperature are non-negative. Moreover, the solution produces strange spurious oscillations, which are not present in the real physical process. Thus, the time-step $\Delta t = 0.1$ results in a qualitatively incorrect numerical solution. This observation can lead to the thought that the time-step has to be decreased.

Let us choose the time-step $\Delta t = 0.005$ and execute the same calculations like above. The result can be seen in Figure 2.1.3. The numerical solution seems to be qualitatively correct and we can be led to the false conclusion that small time-steps make the numerical solution better from the qualitative point of view.

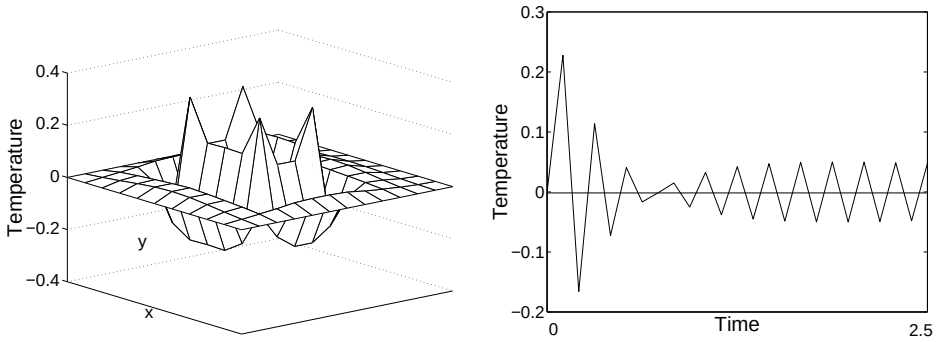

Figure 2.1.2: *Approximation by the Crank-Nicolson method of the temperature at the 10th time level with* $\Delta t = 0.1$ *and the time history of the temperature at the spatial point* $(1/2, 1/6)$.

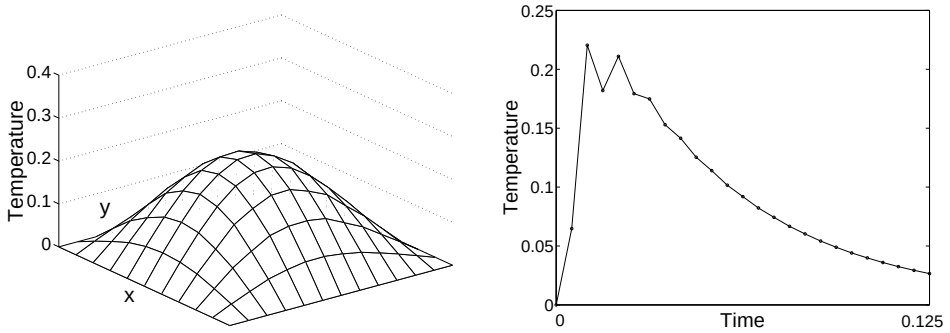

Figure 2.1.3: *Approximation by the Crank-Nicolson method of the temperature at the 10th time level with* $\Delta t = 0.005$ *and the time history of the temperature at the spatial point* $(1/2, 1/6)$.

In order to demonstrate that it is not correct in general, let us choose even smaller time-step $\Delta t = 0.0001$. The obtained result is shown in Figure 2.1.4. This numerical solution has again negative values, and it breaks the so-called maximum-minimum principle and the maximum norm contractivity property, too. This indicates that, most probably, the time-step has to be chosen within a certain interval, that is it should be neither too small, nor too large.

In the second numerical example, we solve the same problem with the implicit Euler method. Choosing the same time-steps, the time histories of the temperature are displayed in Figure 2.1.5. As we can see, in the case of the implicit Euler method, only small time-steps produce qualitative deficiency (negative values).

Let us turn now to the finite difference methods. We solve the problem considered above with the implicit Euler method using finite difference spatial discretization. Calculating with the same time-steps as in the previous two examples we obtain the time histories depicted in Figure 2.1.6. The results obtained demonstrate that there seems that no restrictions on the time-step are needed when the finite difference spatial discretization is combined with the implicit Euler time discretization.

Finally, we consider an example in three dimensions. We show that the suitable choice of the time-step is essential in this case, too. Thus, let us consider the three-dimensional heat equation in the unit cube. The material parameters are set to be constant one again.

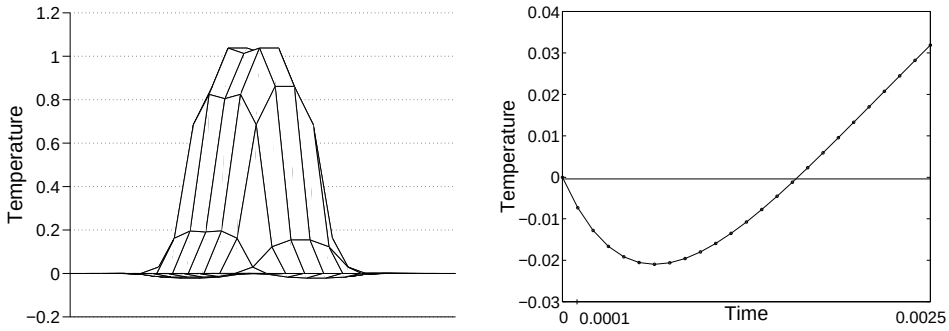

Figure 2.1.4: *Approximation of the temperature at the 10th time level with* $\Delta t = 0.0001$ *and the time history of the temperature at the spatial point* $(1/2, 1/6)$.

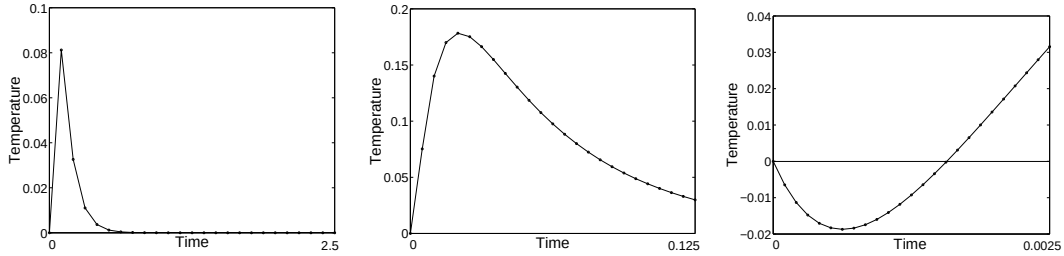

Figure 2.1.5: *Time histories of the approximated temperature at the point* $(1/2, 1/6)$ *with the time-steps* $\Delta t = 0.1$, $\Delta t = 0.005$ *and* $\Delta t = 0.0001$, *respectively, using the implicit Euler method and finite element spatial discretization.*

The boundary points are at constant temperature zero. We apply the finite difference method with the equidistant step sizes $\Delta x = \Delta y = \Delta z = 1/10$ combined with the Crank-Nicolson time discretization method. Let us suppose that an approximation of a continuous non-negative initial function is zero in every grid point except for 27 grid points in the middle of the region, where the temperature is approximated by one. The time history of the temperature at the point $(4/10, 4/10, 4/10)$ using the time-step $\Delta t = 0.05$ can be seen in Figure 2.1.7. The time-step $\Delta t = 0.05$ results in both positive and negative temperatures, which contradicts to the non-negativity preservation property. Choosing the time-step to be $\Delta t = 0.003$, we obtain a qualitatively adequate time history indicated on the right-hand side of Figure 2.1.7.

The above examples illustrate the fact often observed in real calculations that certain time-steps of some numerical schemes result in qualitatively adequate numerical models, while the others do not (e.g., [48, 65]). It is apparent that not only relatively large time-steps cause problems but small ones too. Moreover, unconditionally stable schemes, like the Crank-Nicolson or the implicit Euler scheme, can also produce qualitative deficiencies. These observations rise the demand for figuring out such time-step choices that result in numerical models that mirror the characteristic nature of the original phenomenon.

The above examples show that when we construct a mathematical model of a phenomenon, it is important to investigate whether the mathematical model (continuous/discrete) possesses the same properties as the modelled process. In the sequel we investigate the subject for the second order linear parabolic partial differential operator and for its discretizations, and reveal the connections between the various qualitative properties. The results of the qualitative theory of differential equations and their discrete analogues,

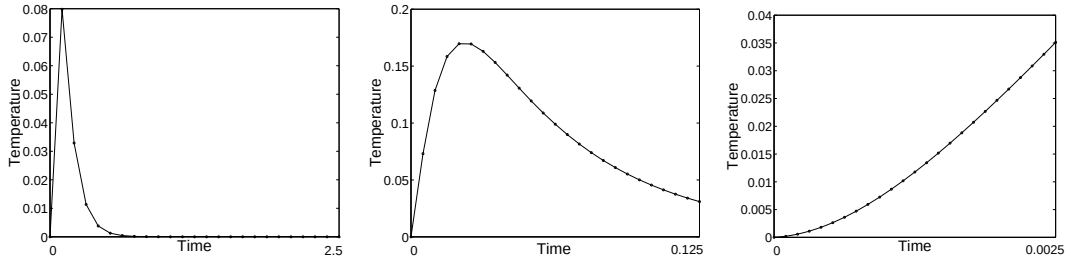

Figure 2.1.6: *Time histories of the temperature at the point* $(1/2, 1/6)$ *with the time-steps* $\Delta t = 0.1$, $\Delta t = 0.005$ *and* $\Delta t = 0.0001$, *respectively, using the implicit Euler method and finite difference spatial discretization.*

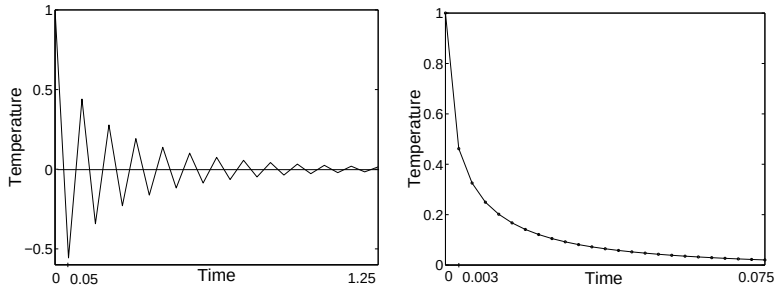

Figure 2.1.7: *Time history of the temperature at the point* $(4/10, 4/10, 4/10)$ *using the time-steps* $\Delta t = 0.05$ *and* $\Delta t = 0.003$, *respectively.*

albeit they have the importance on their own, help us to show that the qualitative properties of a mathematical model correspond to the qualitative properties of the modelled phenomenon.

The discrete version of the maximum-minimum principle is commonly called the *discrete maximum-minimum principle* (or DMP in short). The topic of construction and preserving the validity of various discrete maximum principles arose already 40 years ago and was first investigated for elliptic problems (see, e.g., [23, 24, 77, 118]). Sufficient conditions for the validity of the DMP were given in [144] in terms of the matrix appearing in the finite difference discretization. Recently, this question for the elliptic problems is intensively investigated in many works, see, e.g., [77, 81, 125, 146]. The paper [76] investigates nonlinear problems. The discrete maximum principle is generally guaranteed by some geometrical conditions for the meshes. The discrete maximum principle for parabolic problems was originally discussed discussed about 25 years ago, see, e.g., [57, 83, 131]. In [57], based on the acuteness of the tetrahedral meshes, a sufficient condition of the DMP was obtained for the Galerkin finite element solution of certain parabolic problems, including both the lumped and the non-lumped approaches. The lumped mass method and some hyperbolic prolems are considered in [10]. Actually this topic is considered in the works [36, 46, 47, 48]. In paper [46], a necessary and sufficient condition of the DMP was derived for Galerkin finite element methods and sufficient conditions were given for hybrid meshes. A comprehensive survey on DMPs can be found in papers [18, 19].

The conditions of the *discrete non-negativity preservation* was discussed in [43, 64] for linear finite elements in one, two and three dimensions, and in [38] in one dimensional case with the combination of the finite difference and finite element methods. The discrete non-negativity preservation is investigated for nonlinear problems in [145].

The *discrete maximum norm contractivity* was analyzed for one-dimensional parabolic problems in [69, 80, 131, 132]. In the papers [69, 80] the necessary and sufficient conditions were given. In the first one, the dependence on the spatial discretization was also discussed. In the papers [131, 132] sufficient conditions were given.

For one-dimensional problems, we can deduce some other remarkable qualitative properties such as the preservation of the shape and the monotonicity of the initial function, and the sign-stability (see, e.g., [52, 70, 71, 72, 107]).

2.2 Qualitative properties of the continuous models - reliable continuous models

In this part we define the main qualitative properties for the continuous models, namely, the maximum-minimum principle, the monotonicity, and the maximum norm contractivity. First we consider the general setting, then we analyze the second order linear partial differential operator. We also demonstrate various interrelations between these properties.

Let Ω denote a bounded, simply connected domain in $\mathbb{R}^d$ ($d \in \mathbb{N}^+$) with a Lipschitz-continuous boundary $\partial\Omega$. We introduce the following sets

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

for any arbitrary positive number τ. The set Γ_τ is usually called *parabolic boundary*. For some fixed number $T > 0$, we consider the linear partial differential operator

$$L \equiv \frac{\partial}{\partial t} - \sum_{0 \le |\varsigma| \le \delta} a_\varsigma \frac{\partial^{|\varsigma|}}{\partial^{\varsigma_1} x_1 \dots \partial^{\varsigma_d} x_d} \equiv \frac{\partial}{\partial t} - \sum_{0 \le |\varsigma| \le \delta} a_\varsigma D^\varsigma, \tag{2.2.1}$$

where δ is the order of the operator, $\varsigma_1, \dots, \varsigma_d$ denote non-negative integers, $|\varsigma|$ is defined as $|\varsigma| = \varsigma_1 + \dots + \varsigma_d$ for the multi-index $\varsigma = (\varsigma_1, \dots, \varsigma_d)$, and the coefficient functions $a_\varsigma : Q_T \to \mathbb{R}$ are bounded in the set Q_T. For the sake of simplicity, in what follows, the coefficient function $a_{(0,\dots,0)}$ will be simply denoted by a_0. We define the domain of the operator L, denoted by $\operatorname{dom} L$, as the space of functions $v \in C(\bar{Q}_T)$, for which all the partial derivatives $D^\varsigma v$ ($0 < |\varsigma| \le \delta$) and $\partial v/\partial t$ exist in Q_T and they are bounded. It can be seen easily that Lv is bounded in $Q_{\bar{t}^\star}$ for each $v \in \operatorname{dom} L$ and $t^\star \in (0,T)$, which means that $\inf_{Q_{\bar{t}^\star}} Lv$ and $\sup_{Q_{\bar{t}^\star}} Lv$ are finite values.

2.2.1 Qualitative properties of the linear operators for the continuous models

Operator (2.2.1) appears in the mathematical models of many physical phenomena ([73, 82]). In these phenomena, the following quantities, often called *input data*, can be observed and measured, and hence they are supposed to be known (or easily computable):

- the values of the unknown investigated physical quantities on the parabolic boundary of the solution domain,
- the source density of the quantities inside the solution domain.

Our task is to determine the physical quantities inside the given domain. It can be usually observed in practice that the increase of the input data implies the increase of the quantities inside the solution domain for the physical phenomena described by (2.2.1).

In the mathematical models of the physical phenomena, the function $v \in \operatorname{dom} L$ describes the values of the physical quantity in the domain $\bar{Q}_T$, that is the dependence of the quantity on place and time. The above mentioned physical property can be connected by the following definition.

Definition 2.2.1 *Operator* (2.2.1) *is said to be monotone if for all* $t^\star \in (0,T)$ *and* $v_1, v_2 \in \operatorname{dom} L$ *such that* $v_1|_{\Gamma_{t^\star}} \geq v_2|_{\Gamma_{t^\star}}$ *and* $(Lv_1)|_{Q_{\bar{t}^\star}} \geq (Lv_2)|_{Q_{\bar{t}^\star}}$, *the relation* $v_1|_{Q_{\bar{t}^\star}} \geq v_2|_{Q_{\bar{t}^\star}}$ *holds.*[1]

Clearly, the monotonicity property of the linear operator (2.2.1) is equivalent (due to its linearity) to the widely used non-negativity preservation property.

Definition 2.2.2 *The operator* L *is called non-negativity preserving (NP) when for any* $v \in \operatorname{dom} L$ *and* $t^\star \in (0,T)$ *such that* $v|_{\Gamma_{t^\star}} \geq 0$ *and* $(Lv)|_{Q_{\bar{t}^\star}} \geq 0$, *the relation* $v|_{Q_{\bar{t}^\star}} \geq 0$ *holds.*

The physical quantities inside the solution domain can be obtained by computation of the function v with given initial data. Often we may need only certain characterization of v, which does not require the knowledge of v in the whole domain. It is typical that we are interested in range(v) over $\bar{Q}_T$. From the practical point of view, only such estimates are suitable which include only the known initial data. This kind of estimations is called *maximum-minimum principles*.

For different operators different maximum-minimum principles are valid. These are widely used in literature, because they well characterize the operator L itself (cf. [34, 55, 83, 110, 124, 128] and references therein). Now we list four possible variants of the maximum-minimum principles.

Definition 2.2.3 *We say that the operator* L *satisfies the weak maximum-minimum principle (WMP) if for any function* $v \in \operatorname{dom} L$ *and any* $t^\star \in (0,T)$ *the inequalities*

$$\min\{0, \min_{\Gamma_{t^\star}} v\} + t^\star \cdot \min\{0, \inf_{Q_{\bar{t}^\star}} Lv\} \leq \min_{\bar{Q}_{t^\star}} v \leq \max_{\bar{Q}_{t^\star}} v \leq \max\{0, \max_{\Gamma_{t^\star}} v\} + t^\star \cdot \max\{0, \sup_{Q_{\bar{t}^\star}} Lv\} \tag{2.2.2}$$

are valid.

Definition 2.2.4 *We say that the operator* L *satisfies the strong maximum-minimum principle (SMP) if for any function* $v \in \operatorname{dom} L$ *and any* $t^\star \in (0,T)$ *the inequalities*

$$\min_{\Gamma_{t^\star}} v + t^\star \cdot \min\{0, \inf_{Q_{\bar{t}^\star}} Lv\} \leq \min_{\bar{Q}_{t^\star}} v \leq \max_{\bar{Q}_{t^\star}} v \leq \max_{\Gamma_{t^\star}} v + t^\star \cdot \max\{0, \sup_{Q_{\bar{t}^\star}} Lv\} \tag{2.2.3}$$

are satisfied.

When the sign of Lv is known, then it is possible that the estimates involve only the known values of v on the parabolic boundary. These types of maximum-minimum principles are called *boundary maximum-minimum principles*. (Boundary maximum-minimum principles are frequently used in proofs of the uniqueness theorems.)

[1] This property is also known as the comparion principle.

Definition 2.2.5 *We say that the operator L satisfies the weak boundary maximum-minimum principle (WBMP) if for any function $v \in \operatorname{dom} L$ and any $t^\star \in (0,T)$ such that $Lv|_{Q_{\bar{t}^\star}} \geq 0$ the inequalityies*

$$\min\{0, \min_{\Gamma_{t^\star}} v\} \leq \min_{\bar{Q}_{t^\star}} v \leq \max_{\bar{Q}_{t^\star}} v \leq \max\{0, \max_{\Gamma_{t^\star}} v\} \tag{2.2.4}$$

hold.

Definition 2.2.6 *We say that the operator L satisfies the strong boundary maximum-minimum principle (SBMP) if for any function $v \in \operatorname{dom} L$ and any $t^\star \in (0,T)$ such that $Lv|_{Q_{\bar{t}^\star}} \geq 0$ the realtions*

$$\min_{\Gamma_{t^\star}} v = \min_{\bar{Q}_{t^\star}} v \leq \max_{\bar{Q}_{t^\star}} v = \max_{\Gamma_{t^\star}} v \tag{2.2.5}$$

hold.

Remark 2.2.7 *To show the validity of the relations* (2.2.4) *and* (2.2.5)*, it is enough to show only one relation in each of them: the relation either for the minimum or for the maximum. This is true, because $v \in \operatorname{dom} L$ implies $-v \in \operatorname{dom} L$ and the maximum of a real valued function v is minus one times the minimum of $-v$, we obtain that if an operator L satisfies the WBMP, then $Lv|_{Q_{\bar{t}^\star}} \leq 0$ implies $\max\{0, \max_{\Gamma_{t^\star}} v\} \geq \max_{\bar{Q}_{t^\star}} v$. Similarly, if an operator L satisfies the SBMP, then $\max_{\Gamma_{t^\star}} v = \max_{\bar{Q}_{t^\star}} v$ whenever $Lv|_{Q_{\bar{t}^\star}} \leq 0$.*

Although the left-hand side inequalities in (2.2.2) and (2.2.3) also imply the inequalities on the right-hand side, for practical reasons, we wrote out both the upper and the lower estimates for the function v.

The WMP and the SMP generally do not disclose the place of the maximum or minimum values of v. The WBMP (resp. SBMP) implies that the non-negative maximum (resp. maximum) and the non-positive minimum (resp. minimum) taken over the set $\bar{Q}_{t^\star}$ of the functions $v \in \operatorname{dom} L$ for which $Lv|_{Q_{\bar{t}^\star}} \leq 0$ (or $Lv|_{Q_{\bar{t}^\star}} \geq 0$), can be found also on the parabolic boundary $\Gamma_{t^\star}$.

Remark 2.2.8 *We could pose the natural question of whether it is possible to define another maximum-minimum principle that is somewhat stronger than the SMP. This could be done in the form*

$$\min_{\Gamma_{t^\star}} v + t^\star \cdot \inf_{Q_{\bar{t}^\star}} Lv \leq \min_{\bar{Q}_{t^\star}} v \leq \max_{\bar{Q}_{t^\star}} v \leq \max_{\Gamma_{t^\star}} v + t^\star \cdot \sup_{Q_{\bar{t}^\star}} Lv, \tag{2.2.6}$$

i.e., without the zero values in (2.2.3)*. It is easy to see that there is no sense in defining such a maximum-minimum principle because the simplest one-dimensional heat conduction operator*

$$L = \frac{\partial}{\partial t} - \frac{\partial^2}{\partial x^2}$$

on $Q_T \equiv (0,\pi) \times (0,T)$ does not possess this property. To show this, let us consider the function $v(x,t) = e^{-t}(\sin x - 2) \in \operatorname{dom} L$, for which

$$(Lv)(x,t) = \frac{\partial v}{\partial t}(x,t) - \frac{\partial^2 v}{\partial x^2}(x,t) = 2e^{-t}.$$

For a fixed $t^\star \in (0,T)$, we have

$$\min_{\Gamma_{t^\star}} v + t^\star \cdot \inf_{Q_{\bar{t}^\star}} Lv = -2 + 2t^\star e^{-t^\star} > -2.$$

On the other hand, we have

$$\min_{\bar{Q}_{t^\star}} v = \left(\min_{[0,\pi]}(\sin x - 2)\right) \cdot (\max_{[0,t^\star]} e^{-t}) = -2,$$

which shows the uselessness of such a definition. The explanation of this phenomena is the following. The different maximum principles based on the comparison of the unknown solution with a function, about which we a priori know that it takes bigger values on the parabolic boundary. Then we use the monotonicity property. (We investigate the relation between the different qualitative properties in the next section in more details.) If we choose

$$v_1 = \min\{0, \min_{\Gamma_t} v\} + t \cdot \min\{0, \inf_{Q_{\bar{t}}} Lv\},$$

which stands on the left side of (2.2.2) *at $t = t^\star$, and $v_2 = v$, then the conditions of the monotonicity in Definition 2.2.1 are valid. However, with the choice*

$$v_1 = \min\{0, \min_{\Gamma_t} v\} + t \cdot \inf_{Q_{\bar{t}}} Lv$$

it is not true anymore.

The maximum-minimum principles are in close connection with the maximum norm contractivity, which can be formulated as follows.

Definition 2.2.9 *The operator L is called contractive in the maximum norm (MNC) if for any two functions $\hat{v}$, $\tilde{v} \in \operatorname{dom} L$ and any $t^\star \in (0, T)$ such that $L\hat{v}|_{Q_{\bar{t}^\star}} = L\tilde{v}|_{Q_{\bar{t}^\star}}$ and $\hat{v}|_{\partial\Omega\times[0,t^\star]} = \tilde{v}|_{\partial\Omega\times[0,t^\star]}$, the property*

$$\max_{\mathbf{x}\in\bar{\Omega}} |\hat{v}(\mathbf{x}, t^\star) - \tilde{v}(\mathbf{x}, t^\star)| \le \max_{\mathbf{x}\in\bar{\Omega}} |\hat{v}(\mathbf{x}, 0) - \tilde{v}(\mathbf{x}, 0)|$$

is valid.

2.2.2 Connections between the qualitative properties

In the next theorem, the logical implications between the qualitative properties defined in Section 2.2.1 are proved. In order to see the analogy between the qualitative properties of operator (2.2.1) and its discrete versions, the conditions of the theorem are formulated for the function $L\mathtt{1}$, where $\mathtt{1} : (\mathbf{x}, t) \mapsto 1$ is the identically one function. Naturally, for operator (2.2.1), $L\mathtt{1} = -a_0$.

Theorem 2.2.10 *The implications between the qualitative properties are shown in Figure 2.2.1. The solid arrows mean the implications without any additional condition, while the dashed ones are true under the indicated assumptions on the sign of a_0.*

Proof.

Implications I and II: These implications follow from the relations $\min\{0, \min_{\Gamma_{t^\star}} v\} \le \min_{\Gamma_{t^\star}} v$ and $\max\{0, \max_{\Gamma_{t^\star}} v\} \ge \max_{\Gamma_{t^\star}} v$.

Implication III: Due to the inclusion $\Gamma_{t^\star} \subset \bar{Q}_{t^\star}$, the trivial relation $\min_{\Gamma_{t^\star}} v \ge \min_{\bar{Q}_{t^\star}} v$ holds. The reverse relation follows from the left-hand side relation of (2.2.3) and the non-negativity of Lv in $Q_{\bar{t}^\star}$.

Implication IV: For functions v with $Lv|_{Q_{\bar{t}^\star}} \ge 0$, the left-hand side relation of (2.2.2) ensures the required relation (2.2.4).

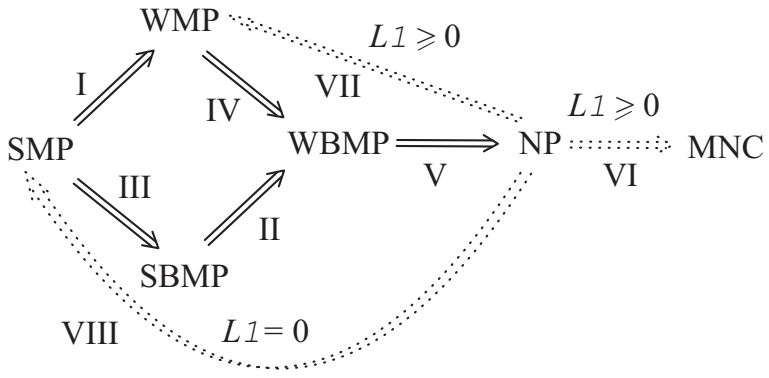

Figure 2.2.1: *Implications between the qualitative properties.*

Implication V: This statement is a direct consequence of the definition of the WBMP.

Implication VI: Let $\hat{v}$ and $\tilde{v} \in \operatorname{dom} L$ be two arbitrary functions with $L\hat{v}|_{Q_{\bar{t}^\star}} = L\tilde{v}|_{Q_{\bar{t}^\star}}$ and $\hat{v}|_{\partial\Omega\times[0,t^\star]} = \tilde{v}|_{\partial\Omega\times[0,t^\star]}$. We consider the functions $v_\pm = \zeta \pm (\hat{v} - \tilde{v})$ with $\zeta = \max_{\mathbf{x}\in\bar{\Omega}} |\hat{v}(\mathbf{x},0) - \tilde{v}(\mathbf{x},0)|$. For these functions, in view of the non-positivity of a_0 and non-negativity of ζ, the estimations $Lv_\pm|_{Q_{\bar{t}^\star}} = (-a_0\zeta)|_{Q_{\bar{t}^\star}} \geq 0$ and $\min_{\Gamma_{t^\star}} v_\pm \geq 0$ are true, which implies the non-negativity of $v_\pm$ on $Q_{\bar{t}^\star}$. Thus, we have

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

Implication VII: We suppose that $a_0 \leq 0$. We choose an arbitrary function $v \in \operatorname{dom} L$ and apply the operator L to the function $\bar{v} = v - \min\{0, \min_{\Gamma_{t^\star}} v\} - t \cdot \min\{0, \inf_{Q_{\bar{t}^\star}} Lv\}$. Clearly, $\bar{v}|_{\Gamma_{t^\star}} \geq 0$. Moreover, we obtain that

$$L\bar{v}|_{Q_{\bar{t}^\star}} = (Lv - \min\{0, \inf_{Q_{\bar{t}^\star}} Lv\} + a_0 \cdot \min\{0, \min_{\Gamma_{t^\star}} v\} + a_0 \cdot t \cdot \min\{0, \inf_{Q_{\bar{t}^\star}} Lv\})|_{Q_{\bar{t}^\star}} \geq 0,$$

which implies that $\bar{v}$ is non-negative on $Q_{\bar{t}^\star}$ by virtue of the non-negativity preservation assumption. Thus

$$\min\{0, \min_{\Gamma_{t^\star}} v\} + t^\star \cdot \min\{0, \inf_{Q_{\bar{t}^\star}} Lv\} \leq \min\{0, \min_{\Gamma_{t^\star}} v\} + t \cdot \min\{0, \inf_{Q_{\bar{t}^\star}} Lv\} \leq v(\mathbf{x},t)$$

for all $\mathbf{x} \in \bar{\Omega}$ and $t \in [0, t^\star]$.

Implication VIII: We suppose that $a_0 = 0$. We choose an arbitrary function $v \in \operatorname{dom} L$ and apply the operator L to the function $\bar{v} = v - \min_{\Gamma_{t^\star}} v - t \cdot \min\{0, \inf_{Q_{\bar{t}^\star}} Lv\}$. Clearly, $\bar{v}|_{\Gamma_{t^\star}} \geq 0$. Moreover, we obtain that

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

which implies that $\bar{v}$ is non-negative on $Q_{\bar{t}^\star}$ by virtue of the non-negativity preservation assumption. Thus

$$\min_{\Gamma_{t^\star}} v + t^\star \cdot \min\{0, \inf_{Q_{\bar{t}^\star}} Lv\} \leq \min_{\Gamma_{t^\star}} v + t \cdot \min\{0, \inf_{Q_{\bar{t}^\star}} Lv\} \leq v(\mathbf{x},t)$$

for all $\mathbf{x} \in \bar{\Omega}$ and $t \in [0, t^\star]$. ■

An important and direct consequence of the above theorem can be formulated for non-negativity preserving operators as follows.

Theorem 2.2.11 *For a non-negativity preserving operator* (2.2.1) *with* $a_0 \leq 0$*, the weak maximum-minimum principles and the maximum norm contractivity properties are also satisfied. If, in addition* $a_0 = 0$*, then the non-negativity preserving operator possesses all the other defined qualitative properties.*

2.2.3 Qualitative properties of the second order linear operators

The second order linear operators (i.e., $\delta = 2$) of the form (2.2.1) have a great practical importance. Such a type of operators appears in parabolic partial differential equations, which serve as mathematical models of several important real-life problems such as heat conduction, advection-diffusion, option pricing, etc. Based on the results of the previous section, we investigate qualitative properties of the following operator

$$L \equiv \frac{\partial}{\partial t} - \sum_{m,k=1}^{d} a_{m,k} \frac{\partial^2}{\partial x_m \partial x_k} - \sum_{m=1}^{d} a_m \frac{\partial}{\partial x_m} - a_0, \tag{2.2.7}$$

where the coefficient functions and $\operatorname{dom} L$ are defined as before (cf. introduction of this Section). The maximum principle for some special case is investigated e.g., in [34, 35, 83, 110, 128]. Let $\mathbf{S}(\mathbf{x}, t)$ be the matrix of the coefficients of the second derivative terms at the point $(\mathbf{x}, t)$, i.e.,

$$\mathbf{S}(\mathbf{x}, t) := [a_{m,k}(\mathbf{x}, t)]_{m,k=1}^{d}. \tag{2.2.8}$$

A sufficient condition for the operator (2.2.7) being non-negativity preserving can be formulated as follows. (We note that, with a similar approach, analogical results are proved for some more special cases in [83] and [35].)

Theorem 2.2.12 *Assume that the matrix $\mathbf{S}(\mathbf{x}, t)$ is positive semi-definite at each point of Q_T. Then the operator* (2.2.7) *is non-negativity preserving.*

Proof. First we prove a lower estimation for the functions $v \in \operatorname{dom} L$, which will show the non-negativity preservation of the operator immediately. Thus, let $v \in \operatorname{dom} L$ an arbitrary fixed function. Then the function

$$\hat{v}(\mathbf{x}, t) \equiv v(\mathbf{x}, t) e^{-\lambda t} \tag{2.2.9}$$

also belongs to $\operatorname{dom} L$ for any real parameter λ. Expressing v from (2.2.9) and applying operator (2.2.7) to it, we get

$$Lv = L(e^{\lambda t}\hat{v}) = e^{\lambda t} \left[\frac{\partial \hat{v}}{\partial t} - \sum_{m,k=1}^{d} a_{m,k} \frac{\partial^2 \hat{v}}{\partial x_m \partial x_k} - \sum_{m=1}^{d} a_m \frac{\partial \hat{v}}{\partial x_m} + (\lambda - a_0)\hat{v} \right]. \tag{2.2.10}$$

Let us fix the parameter $t^\star \in (0, T)$. Since $\hat{v}$ is a continuous function on $\bar{Q}_{t^\star}$, its minimum exists on $\bar{Q}_{t^\star}$ and it is taken at some point $(\mathbf{x}^0, t^0) \in \bar{Q}_{t^\star}$.

- First we assume that this point belongs to the parabolic boundary, i.e., $(\mathbf{x}^0, t^0) \in \Gamma_{t^\star}$. Then, due to the obvious relation

  $$\hat{v}(\mathbf{x}, t) \geq \hat{v}(\mathbf{x}^0, t^0) = \min_{\Gamma_{t^\star}} \hat{v}$$

  for all $(\mathbf{x}, t) \in \bar{Q}_{t^\star}$, we get the estimation

  $$\inf_{Q_{\bar{t}^\star}} \hat{v} \geq \min_{\Gamma_{t^\star}} \hat{v}. \tag{2.2.11}$$

- Assume now that $(\mathbf{x}^0, t^0) \in Q_{\bar{t}^\star}$. Then we get the relations

$$\frac{\partial \hat{v}}{\partial t}(\mathbf{x}^0, t^0) \le 0, \quad \frac{\partial \hat{v}}{\partial x_m}(\mathbf{x}^0, t^0) = 0, \tag{2.2.12}$$

and, because $(\mathbf{x}^0, t^0)$ is a minimum point, the second derivative matrix

$$\hat{\mathbf{V}}(\mathbf{x}^0, t^0) := \left[\frac{\partial^2 \hat{v}}{\partial x_m \partial x_k}(\mathbf{x}^0, t^0)\right]_{m,k=1}^{d}$$

is positive semi-definite.

Let us denote by $\mathbf{S}(\mathbf{x}^0, t^0) \circ \hat{\mathbf{V}}(\mathbf{x}^0, t^0) \in \mathbb{R}^{d\times d}$ the Hadamard product

$$\Big[\mathbf{S}(\mathbf{x}^0, t^0) \circ \hat{\mathbf{V}}(\mathbf{x}^0, t^0)\Big]_{m,k} = a_{m,k}(\mathbf{x}^0, t^0) \cdot \frac{\partial^2 \hat{v}}{\partial x_m \partial x_k}(\mathbf{x}^0, t^0). \tag{2.2.13}$$

Due to the assumptions, both the matrices $\mathbf{S}(\mathbf{x}^0, t^0)$ and $\hat{\mathbf{V}}(\mathbf{x}^0, t^0)$ are positive semi-definite, hence, according to the Schur theorem (e.g., Theorem 7.5.3 in [68]), the matrix $\mathbf{S}(\mathbf{x}^0, t^0) \circ \hat{\mathbf{V}}(\mathbf{x}^0, t^0)$ is also positive semi-definite.

We investigate (2.2.10) in the rearranged form

$$e^{-\lambda t} Lv + \sum_{m=1}^{d} a_m \frac{\partial \hat{v}}{\partial x_m} - (\lambda - a_0)\hat{v} = \frac{\partial \hat{v}}{\partial t} - \sum_{m,k=1}^{d} a_{m,k} \frac{\partial^2 \hat{v}}{\partial x_m \partial x_k}. \tag{2.2.14}$$

Using the notation $\mathbf{e} = [1, 1, \ldots, 1]^\top \in \mathbb{R}^d$, the relation

$$\sum_{m,k=1}^{d} a_{m,k}(\mathbf{x}^0, t^0) \frac{\partial^2 \hat{v}}{\partial x_m \partial x_k}(\mathbf{x}^0, t^0) = ((\mathbf{S}(\mathbf{x}^0, t^0) \circ \hat{\mathbf{V}}(\mathbf{x}^0, t^0))\mathbf{e}, \mathbf{e}) \ge 0. \tag{2.2.15}$$

is valid. On the base of (2.2.12) and (2.2.15), the right-hand side of (2.2.14) is nonpositive at the point $(\mathbf{x}^0, t^0)$. Hence, the inequality

$$e^{-\lambda t^0}(Lv)(\mathbf{x}^0, t^0) - (\lambda - a_0(\mathbf{x}^0, t^0))\hat{v}(\mathbf{x}^0, t^0) \le 0 \tag{2.2.16}$$

holds. Let us introduce the notations $a_{\inf} := \inf_{Q_T} a_0$ and $a_{\sup} := \sup_{Q_T} a_0$, which are well-defined because of the boundedness of the coefficient function a_0. For any $\lambda > a_{\sup}$, we have

$$\begin{aligned} \hat{v}(\mathbf{x}^0, t^0) &\ge \frac{e^{-\lambda t^0}(Lv)(\mathbf{x}^0, t^0)}{\lambda - a_0(\mathbf{x}^0, t^0)} \ge \frac{e^{-\lambda t^0}(Lv)(\mathbf{x}^0, t^0)}{\lambda - a_{\inf}} \ge \\ &\ge \frac{1}{\lambda - a_{\inf}} \inf_{Q_{\bar{t}^\star}}(e^{-\lambda t}(Lv)(\mathbf{x}, t)). \end{aligned} \tag{2.2.17}$$

Since the function $\hat{v}$ takes its minimum at the point $(\mathbf{x}^0, t^0)$, therefore estimation (2.2.17) implies the inequality

$$\inf_{Q_{\bar{t}^\star}} \hat{v} \ge \frac{1}{\lambda - a_{\inf}} \inf_{Q_{\bar{t}^\star}}(e^{-\lambda t}(Lv)(\mathbf{x}, t)). \tag{2.2.18}$$

Clearly, the estimates of the two different cases, namely (2.2.11) and (2.2.18) together, imply that

$$\inf_{\bar{Q}_{t^\star}} \hat{v} \geq \min\{\inf_{\Gamma_{t^\star}} \hat{v};\ \frac{1}{\lambda - a_{\inf}} \inf_{Q_{\bar{t}^\star}} \left(e^{-\lambda t}(Lv)(\mathbf{x},t)\right)\}. \tag{2.2.19}$$

From (2.2.19) and from the definition of the function $\hat{v}$ in (2.2.9), we obtain that

$$v(\mathbf{x},t^\star) \geq \sup_{\lambda > a_{\sup}} \left\{ e^{\lambda t^\star} \min\left\{ \min_{\Gamma_{t^\star}}(ve^{-\lambda t}), \frac{1}{\lambda - a_{\inf}} \inf_{Q_{\bar{t}^\star}}(e^{-\lambda t}(Lv)(\mathbf{x},t)) \right\} \right\}. \tag{2.2.20}$$

The statement of the theorem follows from the definition of the non-negativity preservation and the estimation (2.2.20). ■

Remark 2.2.13 *Let us consider the d-dimensional heat conduction operator*

$$L \equiv \frac{\partial}{\partial t} - \sum_{m=1}^{d} \frac{\partial^2}{\partial x_m^2}. \tag{2.2.21}$$

In this case $\mathbf{S}(\mathbf{x},t) = \mathbf{I}$*, where* $\mathbf{I}$ *denotes the* $d \times d$ *unit matrix, which is obviously positive definite. Thus, for this operator the NP property holds and, according to Theorem 2.2.11, if* $a_0 = 0$ *it satisfies all the above discussed qualitative properties.*
For the more general operator,

$$L \equiv \frac{\partial}{\partial t} - \sum_{m=1}^{d} \frac{\partial}{\partial x_m}(k_m(\mathbf{x},t)\frac{\partial}{\partial x_m}) - a_0(\mathbf{x},t) \tag{2.2.22}$$

the condition of the positive semi-definitness reads, obviously, as

$$k_m(\mathbf{x},t) \geq 0 \text{ for all } (\mathbf{x},t) \in Q_T \text{ and } m = 1,2,\ldots,d. \tag{2.2.23}$$

2.2.4 On necessity of the conditions

In this section we show that certain implications in Theorem 2.2.10 are strong in the sense that they cannot be reversed or sharpened. Namely, let us investigate whether the condition $-a_0 = L1 \geq 0$ can be changed to $-a_0 = L1 \geq \gamma$, with some constant $\gamma \neq 0$ such that the implications indicated in Figure 2.2.1 remain valid.

Theorem 2.2.14 *The infimum of those values* γ *for which Implications VI and VII in Theorem 2.2.10 are valid, under the condition* $-a_0 = L1 \geq \gamma$*, is zero. Similarly, for the Implication VIII the given condition of is also necessary in the same sense.*

Proof. Let γ be an arbitrary negative number (i.e., $a_0 \equiv -\gamma > 0$) and we consider the one-dimensional operator

$$L \equiv \frac{\partial}{\partial t} + \frac{\gamma}{2}\frac{\partial^2}{\partial x^2} + \gamma, \tag{2.2.24}$$

where dom L is defined similarly as for operator (2.2.1) and $Q_T = (0,\pi)\times(0,T)$. Naturally, based on Theorem 2.2.12, operator (2.2.24) is non-negativity preserving as $a_{11} = -\gamma/2 > 0$. Moreover $L1 = \gamma = -a_0$ and hence $a_0 > 0$.

We show that operator (2.2.24) does not possess the WMP and the MNC. Let us choose the function $v(x,t) = (-\gamma/2)e^{-\gamma t/2}\sin x$, for which function the relation $Lv(x,t) = 0$ is true. Thus, we have

$$-\frac{\gamma}{2}e^{-\gamma t^\star/2} = \max_{\bar{Q}_{t^\star}} v > \max\{0, \max_{\Gamma_{t^\star}} v\} + t^\star \cdot \max\{0, \sup_{Q_{\bar{t}^\star}} Lv\} = -\frac{\gamma}{2}$$

for any $t^\star \in (0,T)$. This shows that the WMP does not hold for the operator defined by (2.2.24). Due to Implication I, the SMP cannot be valid, either. Let us set $\hat{v} = v$ and $\tilde{v} = 0$. The relations $L\hat{v}|_{Q_{\bar{t}^\star}} = L\tilde{v}|_{Q_{\bar{t}^\star}} = 0$ and $\hat{v}|_{\partial\Omega\times[0,t^\star]} = \tilde{v}|_{\partial\Omega\times[0,t^\star]} = 0$ obviously hold for these functions and we have

$$-\frac{\gamma}{2}e^{-\gamma t^\star/2} = \max_{x\in\bar{\Omega}} |\hat{v}(x,t^\star) - \tilde{v}(x,t^\star)| > \max_{x\in\bar{\Omega}} |\hat{v}(x,0) - \tilde{v}(x,0)| = -\frac{\gamma}{2}.$$

This shows that the operator (2.2.24) does not have the MNC property.

Now let γ be an arbitrary positive number and consider the non-negativity preserving operator

$$L \equiv \frac{\partial}{\partial t} - \frac{\gamma}{2}\frac{\partial^2}{\partial x^2} + \gamma. \tag{2.2.25}$$

We set $v(x,t) = 0.5\,\gamma e^{-\gamma t/2}(\sin x - 2)$ and $Q_T = (0,\pi)\times(0,T)$, for which $Lv(x,t) = 0.5\,\gamma^2 e^{-\gamma t/2}(\sin x - 1) \le 0$. For this function v, we get the relation

$$\max_{\bar{Q}_{t^\star}} v = -\frac{\gamma}{2}e^{-\gamma t^\star/2} > \max\{-\gamma e^{-\gamma t^\star/2}, -\gamma/2\} = \max_{\Gamma_{t^\star}} v = \max_{\Gamma_{t^\star}} v + t^\star\cdot\max\{0, \sup_{Q_{\bar{t}^\star}} Lv\}$$

for any $t^\star \in (0,T)$, showing that the SMP is not satisfied in this case. This completes the proof. ■

Remark 2.2.15 *The operator* (2.2.24) *with the function* $v(x,t) = a_0 e^{a_0 t}\sin x$ *also demonstrates that Implication V cannot be reversed for* $a_0 > 0$. *Namely,* $\max\{0, \max_{\Gamma_{t^\star}} v\} = a_0 < a_0 e^{a_0 t^\star} = \max_{\bar{Q}_{t^\star}} v$. *Similarly, operator* (2.2.25) *and the function* $v(x,t) = -a_0 e^{a_0 t}(\sin x - 2)$ *show that Implications I and II are not reversible, provided* $a_0 \neq 0$.

Finally, we note that a more general setting of the problem and other counterexamples can be found in [49].

2.3 Discrete analogs of the qualitative properties - reliable discrete models

In this part we present the natural discrete analogs of the qualitative properties formulated in Section 2.2 for the continuous models.

2.3.1 Qualitative properties of discrete mesh operators

Let us assume that the sets $\mathcal{P} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and $\mathcal{P}_\partial = \{\mathbf{x}_{N+1}, \mathbf{x}_{N+2}, \ldots, \mathbf{x}_{N+N_\partial}\}$ consist of different vertices in Ω and on $\partial\Omega$, respectively. (The bounded domain Ω and its boundary $\partial\Omega$ were defined in Section 2.2.)

We set $\bar{N} = N + N_\partial$ and $\bar{\mathcal{P}} = \mathcal{P}\cup\mathcal{P}_\partial$. Let T and $\Delta t < T$ be two arbitrary positive numbers. Moreover, let us suppose that the natural number M satisfies the condition $M\Delta t \le T < (M+1)\Delta t$ and introduce the set $\mathcal{R} = \{t_n = n\Delta t \,|\, n = 0,1,\ldots,M\}$. For any values τ from the set $\mathcal{R}$ we introduce the notations

$$\begin{aligned} \mathcal{R}_\tau &= \{t\in\mathcal{R}\,|\,0 < t < \tau\},\\ \mathcal{R}_{\bar{\tau}} &= \{t\in\mathcal{R}\,|\,0 < t \le \tau\},\\ \mathcal{R}^0_{\bar{\tau}} &= \{t\in\mathcal{R}\,|\,0 \le t \le \tau\}, \end{aligned} \tag{2.3.1}$$

and the sets

$$\mathcal{Q}_\tau = \mathcal{P}\times\mathcal{R}_\tau, \qquad \bar{\mathcal{Q}}_\tau = \bar{\mathcal{P}}\times\mathcal{R}^0_{\bar{\tau}}, \qquad \mathcal{Q}_{\bar{\tau}} = \mathcal{P}\times\mathcal{R}_{\bar{\tau}}, \qquad \mathcal{G}_\tau = (\mathcal{P}_\partial\times\mathcal{R}^0_{\bar{\tau}})\cup(\mathcal{P}\times\{0\}).$$

Definition 2.3.1 *Linear mappings that map from the space of real-valued functions defined on $\bar{\mathcal{Q}}_{t_M}$ to the space of real-valued functions defined on $\mathcal{Q}_{t_M}$ are called discrete (linear) mesh operators.*

As we will see later, finite difference or finite element solution methods for time-dependent (parabolic) partial differential equations can be written in a discrete mesh operator form.

The domain of a discrete mesh operator $\mathcal{L}$, that is the space of real valued functions defined on $\bar{\mathcal{Q}}_{t_M}$, is denoted by $\operatorname{dom}\mathcal{L}$. We define the qualitative properties of the discrete mesh operators in an analogous way to those in the linear partial differential operator case in Section 2.2.1.

Definition 2.3.2 *We say that the discrete mesh operator $\mathcal{L}$ is monotone if for all $t^\star \in \mathcal{R}_{t_M}$ and $\nu_1, \nu_2 \in \operatorname{dom}\mathcal{L}$ such that $\nu_1|_{\mathcal{G}_{t^\star}} \geq \nu_2|_{\mathcal{G}_{t^\star}}$ and $(\mathcal{L}\nu_1)|_{\mathcal{Q}_{\bar{t}^\star}} \geq (\mathcal{L}\nu_2)|_{\mathcal{Q}_{\bar{t}^\star}}$, the relation $\nu_1|_{\mathcal{Q}_{\bar{t}^\star}} \geq \nu_2|_{\mathcal{Q}_{\bar{t}^\star}}$ holds.*[2]

Similarly to the continuous case, the monotonicity of a discrete mesh operator is equivalent to the non-negativity property, defined below.

Definition 2.3.3 *The discrete mesh operator $\mathcal{L}$ is called non-negativity preserving (DNP) if for any $\nu \in \operatorname{dom}\mathcal{L}$ and any $t^\star \in \mathcal{R}_{t_M}$ such that $\min_{\mathcal{G}_{t^\star}} \nu \geq 0$ and $\mathcal{L}\nu|_{\mathcal{Q}_{\bar{t}^\star}} \geq 0$, the relation $\nu|_{\mathcal{Q}_{\bar{t}^\star}} \geq 0$ holds.*

The discrete maximum-minimum principles can be formulated as follows.

Definition 2.3.4 *We say that a discrete mesh operator $\mathcal{L}$ satisfies the discrete weak maximum-minimum principle (DWMP) if for any function $\nu \in \operatorname{dom}\mathcal{L}$ and $t^\star \in \mathcal{R}_{t_M}$ the inequalities*

$$\min\{0, \min_{\mathcal{G}_{t^\star}} \nu\} + t^\star \cdot \min\{0, \min_{\mathcal{Q}_{\bar{t}^\star}} \mathcal{L}\nu\} \leq \min_{\bar{\mathcal{Q}}_{t_M}} \nu \leq \max_{\bar{\mathcal{Q}}_{t_M}} \nu \leq \max\{0, \max_{\mathcal{G}_{t^\star}} \nu\} + t^\star \cdot \max\{0, \max_{\mathcal{Q}_{\bar{t}^\star}} \mathcal{L}\nu\} \tag{2.3.2}$$

hold.

Definition 2.3.5 *We say that a discrete mesh operator $\mathcal{L}$ satisfies the discrete strong maximum-minimum principle (DSMP) if for any function $\nu \in \operatorname{dom}\mathcal{L}$ and $t^\star \in \mathcal{R}_{t_M}$ the inequalities*

$$\min_{\mathcal{G}_{t^\star}} \nu + t^\star \cdot \min\{0, \min_{\mathcal{Q}_{\bar{t}^\star}} \mathcal{L}\nu\} \leq \min_{\bar{\mathcal{Q}}_{t_M}} \nu \leq \max_{\bar{\mathcal{Q}}_{t_M}} \nu \leq \max_{\mathcal{G}_{t^\star}} \nu + t^\star \cdot \max\{0, \max_{\mathcal{Q}_{\bar{t}^\star}} \mathcal{L}\nu\} \tag{2.3.3}$$

hold.

Definition 2.3.6 *We say that the discrete mesh operator $\mathcal{L}$ satisfies the discrete weak boundary maximum-minimum principle (DWBMP) if for any function $\nu \in \operatorname{dom}\mathcal{L}$ and $t^\star \in \mathcal{R}_{t_M}$ such that $\mathcal{L}\nu|_{\mathcal{Q}_{\bar{t}^\star}} \geq 0$ the inequalities*

$$\min\{0, \min_{\mathcal{G}_{t^\star}} \nu\} \leq \min_{\bar{\mathcal{Q}}_{t^\star}} \nu \leq \max_{\bar{\mathcal{Q}}_{t^\star}} \nu \leq \max\{0, \max_{\mathcal{G}_{t^\star}} \nu\} \tag{2.3.4}$$

hold.

[2] The ordering relations for vectors and matrices are always meant elementwise.

Definition 2.3.7 *We say that the discrete mesh operator $\mathcal{L}$ satisfies the discrete strong boundary maximum-minimum principle (DSBMP) if for any function $\nu \in \operatorname{dom}\mathcal{L}$ and $t^\star \in \mathcal{R}_{t_M}$ such that $\mathcal{L}\nu|_{\mathcal{Q}_{\bar{t}^\star}} \geq 0$ the relations*

$$\min_{\mathcal{G}_{t^\star}} \nu = \min_{\bar{\mathcal{Q}}_{t^\star}} \nu \leq \max_{\bar{\mathcal{Q}}_{t^\star}} \nu = \max_{\mathcal{G}_{t^\star}} \nu \tag{2.3.5}$$

hold.

Remark 2.3.8 *If a discrete mesh operator $\mathcal{L}$ satisfies DWBMP (or DSBMP) then, for the vector $\nu \in \operatorname{dom}\mathcal{L}$ with the reverse inequality, i.e., under the condition $\mathcal{L}\nu|_{\mathcal{Q}_{\bar{t}^\star}} \leq 0$, the analogous implications for the maximum are valid. Obviously, these two properties are equivalent and hence it is enough to assume only one of them.*

The definition of the discrete maximum norm contractivity is as follows.

Definition 2.3.9 *The discrete mesh operator $\mathcal{L}$ is called contractive in maximum norm (DMNC) when for any arbitrary two functions $\hat{\nu}$, $\tilde{\nu} \in \operatorname{dom}\mathcal{L}$ and $t^\star \in \mathcal{R}_{t_M}$ such that $\mathcal{L}\hat{\nu}|_{\mathcal{Q}_{\bar{t}^\star}} = \mathcal{L}\tilde{\nu}|_{\mathcal{Q}_{\bar{t}^\star}}$ and $\hat{\nu}|_{\mathcal{P}_\partial \times \mathcal{R}^0_{\bar{t}^\star}} = \tilde{\nu}|_{\mathcal{P}_\partial \times \mathcal{R}^0_{\bar{t}^\star}}$, the relation*

$$\max_{\mathbf{x}\in\bar{\mathcal{P}}} |\hat{\nu}(\mathbf{x}, t^\star) - \tilde{\nu}(\mathbf{x}, t^\star)| \leq \max_{\mathbf{x}\in\bar{\mathcal{P}}} |\hat{\nu}(\mathbf{x}, 0) - \tilde{\nu}(\mathbf{x}, 0)|$$

is valid.

2.3.2 Connections between the qualitative properties for the discrete operators

Let us introduce two special mesh functions, $\mathbb{1}$ and $\sharp$, defined on $\bar{\mathcal{Q}}_{t_M}$ with the following equalities $\mathbb{1}(\mathbf{x}_i, t_n) = 1$, $\sharp(\mathbf{x}_i, t_n) = n\Delta t$ for all $(\mathbf{x}_i, t_n) \in \bar{\mathcal{Q}}_{t_M}$. These mesh functions are the discrete analogue of the continuous functions $v(\mathbf{x}, t) = 1$ and $v(\mathbf{x}, t) = t$, respectively, associated with the mesh $\bar{\mathcal{Q}}_{t_M}$.

Now we are able to formulate the discrete analogue of Theorem 2.2.10.

Theorem 2.3.10 *The implications between the discrete qualitative properties are shown in Figure 2.3.1. The solid arrows mean the implications without any condition, while the*

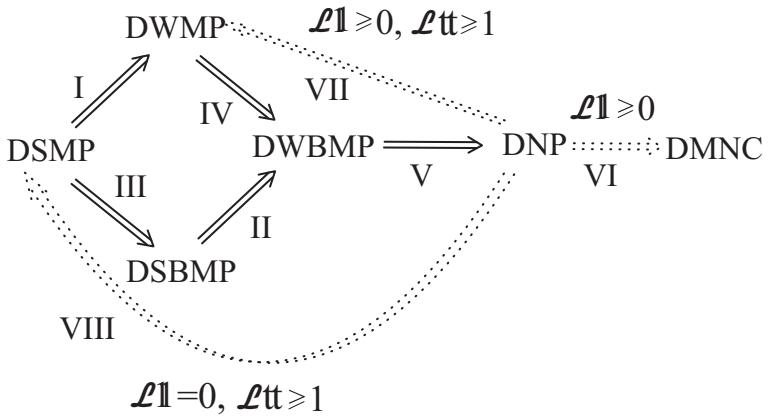

Figure 2.3.1: *Implications between the various discrete qualitative properties.*

dashed ones are true only under the indicated assumptions.

PROOF. Implications I-V can be proved similarly as in the continuous case in Theorem 2.2.10.

Implication VI: Let $\hat{\nu}$ and $\tilde{\nu} \in \operatorname{dom} \mathcal{L}$ be two arbitrary functions with $\mathcal{L}\hat{\nu}|_{\mathcal{Q}_{\bar{t}^\star}} = \mathcal{L}\tilde{\nu}|_{\mathcal{Q}_{\bar{t}^\star}}$ and $\hat{\nu}|_{\mathcal{P}_\partial \times \mathcal{R}^0_{\bar{t}^\star}} = \tilde{\nu}|_{\mathcal{P}_\partial \times \mathcal{R}^0_{\bar{t}^\star}}$. We consider the functions $\nu_\pm = \zeta \pm (\hat{\nu} - \tilde{\nu})$ with $\zeta = \max_{\mathbf{x} \in \bar{\mathcal{P}}} |\hat{\nu}(\mathbf{x}, 0) - \tilde{\nu}(\mathbf{x}, 0)|$. For these functions, the estimations $\mathcal{L}\nu_\pm|_{\mathcal{Q}_{\bar{t}^\star}} = (\zeta(\mathcal{L}\mathbb{1}))|_{\mathcal{Q}_{\bar{t}^\star}} \geq 0$ and $\min_{\mathcal{G}_{t^\star}} \nu_\pm \geq 0$ are true, which implies the non-negativity of $\nu_\pm$ on $\mathcal{Q}_{\bar{t}^\star}$. Thus, we have

$$\max_{\mathbf{x} \in \bar{\mathcal{P}}} |\hat{\nu}(\mathbf{x}, t^\star) - \tilde{\nu}(\mathbf{x}, t^\star)| \leq \max_{\mathbf{x} \in \bar{\mathcal{P}}} |\hat{\nu}(\mathbf{x}, 0) - \tilde{\nu}(\mathbf{x}, 0)|.$$

Implication VII: We choose an arbitrary function $\nu \in \operatorname{dom} \mathcal{L}$ and apply the operator $\mathcal{L}$ to the function $\bar{\nu}(\mathbf{x}_i, t_n) = \nu(\mathbf{x}_i, t_n) - \min\{0, \min_{\mathcal{G}_{t^\star}} \nu\} - n\Delta t \cdot \min\{0, \inf_{\mathcal{Q}_{\bar{t}^\star}} \mathcal{L}\nu\}$. Clearly, $\bar{\nu}|_{\mathcal{G}_{t^\star}} \geq 0$. Moreover, we obtain that

$$\begin{aligned}\mathcal{L}\bar{\nu}|_{\mathcal{Q}_{\bar{t}^\star}} &= (\mathcal{L}\nu - \min\{0, \min_{\mathcal{G}_{t^\star}} \nu\}(\mathcal{L}\mathbb{1}) - \min\{0, \inf_{\mathcal{Q}_{\bar{t}^\star}} \mathcal{L}\nu\}(\mathcal{L}t))|_{\mathcal{Q}_{\bar{t}^\star}} \geq \\ &\geq (\mathcal{L}\nu - \min\{0, \inf_{\mathcal{Q}_{\bar{t}^\star}} \mathcal{L}\nu\})|_{\mathcal{Q}_{\bar{t}^\star}} \geq 0,\end{aligned}$$

which implies that $\bar{\nu}$ is non-negative on $\mathcal{Q}_{\bar{t}^\star}$ by virtue of the non-negativity preservation assumption. Thus

$$\min\{0, \min_{\mathcal{G}_{t^\star}} \nu\} + t^\star \cdot \min\{0, \inf_{\mathcal{Q}_{\bar{t}^\star}} \mathcal{L}\nu\} \leq \min\{0, \min_{\mathcal{G}_{t^\star}} \nu\} + t \cdot \min\{0, \inf_{\mathcal{Q}_{\bar{t}^\star}} \mathcal{L}\nu\} \leq \nu(\mathbf{x}, t)$$

for all $\mathbf{x} \in \bar{\mathcal{P}}$ and $t \in \mathcal{R}^0_{\bar{t}^\star}$.

Implication VIII: We choose an arbitrary function $\nu \in \operatorname{dom} \mathcal{L}$ and apply the operator $\mathcal{L}$ to the function $\bar{\nu}(\mathbf{x}_i, t_n) = \nu(\mathbf{x}_i, t_n) - \min_{\mathcal{G}_{t^\star}} \nu - n\Delta t \cdot \min\{0, \inf_{\mathcal{Q}_{\bar{t}^\star}} \mathcal{L}\nu\}$. Clearly, $\bar{\nu}|_{\mathcal{G}_{t^\star}} \geq 0$. Moreover, we obtain that

$$\begin{aligned}\mathcal{L}\bar{\nu}|_{\mathcal{Q}_{\bar{t}^\star}} &= (\mathcal{L}\nu - \min_{\mathcal{G}_{t^\star}} \nu \cdot (\mathcal{L}\mathbb{1}) - \min\{0, \inf_{\mathcal{Q}_{\bar{t}^\star}} \mathcal{L}\nu\}(\mathcal{L}t))|_{\mathcal{Q}_{\bar{t}^\star}} \geq \\ &\geq (\mathcal{L}\nu - \min\{0, \inf_{\mathcal{Q}_{\bar{t}^\star}} \mathcal{L}\nu\})|_{\mathcal{Q}_{\bar{t}^\star}} \geq 0,\end{aligned}$$

which implies that $\bar{\nu}$ is non-negative on $\mathcal{Q}_{\bar{t}^\star}$ by virtue of the non-negativity preservation assumption. Thus

$$\min_{\mathcal{G}_{t^\star}} \nu + t^\star \cdot \min\{0, \inf_{\mathcal{Q}_{\bar{t}^\star}} \mathcal{L}\nu\} \leq \min_{\mathcal{G}_{t^\star}} \nu + t \cdot \min\{0, \inf_{\mathcal{Q}_{\bar{t}^\star}} \mathcal{L}\nu\} \leq \nu(\mathbf{x}, t)$$

for all $\mathbf{x} \in \bar{\mathcal{P}}$ and $t \in \mathcal{R}^0_{\bar{t}^\star}$. ■

Remark 2.3.11 *Let us consider operator* (2.2.1). *If* $L1 \geq 0$, *then the relation* $Lt = 1 - a_0 t \geq 1$ *is valid. Comparing Figure 2.3.1 with Figure 2.2.1, we observe that the relations between the qualitative properties of the continuous and the discrete operators have the same structure.*

It is worth mentioning that the DWBMP and the DSBMP qualitative properties are defined only for mesh functions with the properties $\mathcal{L}\nu|_{\mathcal{Q}_{\bar{t}^\star}} \leq 0$ and $\mathcal{L}\nu|_{\mathcal{Q}_{\bar{t}^\star}} = 0$, respectively. This information relaxes the required conditions in the implications VII and VIII. Namely, introducing the notation

$$H_0 = \{\nu \in \operatorname{dom} \mathcal{L}; \quad \mathcal{L}\nu|_{\mathcal{Q}_{\bar{t}^\star}} \leq 0\}, \tag{2.3.6}$$

for the mesh function from H_0 the condition $\mathcal{L}\sharp \geq 1$ is not necessary, because for such mesh functions $\min\{0, \inf_{\mathcal{Q}_{\bar{t}^\star}} \mathcal{L}\nu\} = 0$. We also note, that, on the subset

$$H_1 = \{\nu \in \operatorname{dom}\mathcal{L};\ \ \nu|_{\mathcal{G}_{\bar{t}^\star}} \geq 0;\ \mathcal{L}\nu|_{\mathcal{Q}_{\bar{t}^\star}} \leq 0\} \tag{2.3.7}$$

the properties DSBMP and DSMP are equivalent.

The above considerations can be summarized in the following theorem.

Theorem 2.3.12 *Assume that the mesh operator $\mathcal{L}$ is DNP and the relation $\mathcal{L}\mathbb{1} \geq 0$ is satisfied. Then, on the set H_0 the operator $\mathcal{L}$ has the DWMP, and hence the DWBMP properties. If $\mathcal{L}\mathbb{1} = 0$, then for the functions $\nu \in H_1$ the operator $\mathcal{L}$ has the DSMP, and hence the DWMP, DSBMP and DWBMP properties, too.*

Finally we note that similar statements can be formulated for the minimum on the subsets

$$H_0^1 = \{\nu \in \operatorname{dom}\mathcal{L};\ \mathcal{L}\nu|_{\mathcal{Q}_{\bar{t}^\star}} \geq 0\} \tag{2.3.8}$$

and

$$H_1^1 = \{\nu \in \operatorname{dom}\mathcal{L}; \nu|_{\mathcal{G}_{\bar{t}^\star}} \leq 0;\ \mathcal{L}\nu|_{\mathcal{Q}_{\bar{t}^\star}} \geq 0\}, \tag{2.3.9}$$

respectively.

2.3.3 Two-level discrete mesh operators

In the sequel, the values $\nu(\mathbf{x}_i, n\Delta t)$ of the function ν defined in $\bar{\mathcal{Q}}_{t_M}$ will be denoted by ν_i^n. Similar notation is applied to the function $\mathcal{L}\nu$. We introduce the vectors

$$\boldsymbol{\nu}^n = [\nu_1^n, \ldots, \nu_{\bar{N}}^n], \quad \boldsymbol{\nu}_0^n = [\nu_1^n, \ldots, \nu_N^n], \quad \boldsymbol{\nu}_\partial^n = [\nu_{N+1}^n, \ldots, \nu_{\bar{N}}^n].$$

In many numerical methods, the discrete mesh operators have a special form, namely, they are defined as

$$(\mathcal{L}\nu)_i^n = (\mathbf{X}_1^{(n)}\boldsymbol{\nu}^n - \mathbf{X}_2^{(n)}\boldsymbol{\nu}^{n-1})_i, \quad i = 1, \ldots, N, \quad n = 1, \ldots, M, \tag{2.3.10}$$

where $\mathbf{X}_1^{(n)}, \mathbf{X}_2^{(n)} \in \mathbb{R}^{N\times\bar{N}}$ are some given matrices.[3] In order to give the connections between the qualitative properties of such a type of mesh operators, we reformulate the conditions in Theorem 2.3.10, see Figure 2.3.1. We have already introduced the notation $\mathbf{e} = [1, \ldots, 1] \in \mathbb{R}^{\bar{N}}$. The N-element and the $(\bar{N} - N)$-element version of this vector will be denoted by $\mathbf{e}_0$ and $\mathbf{e}_\partial$, respectively, i.e., $\mathbf{e} = [\mathbf{e}_0|\, \mathbf{e}_\partial]$. Then the conditions $\mathcal{L}\mathbb{1} = 0$ and $\mathcal{L}\mathbb{1} \geq 0$ read as

$$\mathbf{X}_1^{(n)}\mathbf{e} - \mathbf{X}_2^{(n)}\mathbf{e} = (\mathbf{X}_1^{(n)} - \mathbf{X}_2^{(n)})\mathbf{e} = \mathbf{0} \quad \text{and} \quad (\mathbf{X}_1^{(n)} - \mathbf{X}_2^{(n)})\mathbf{e} \geq \mathbf{0}\ \ (n = 1, \ldots, M),$$

respectively, while condition $\mathcal{L}\sharp \geq 1$ means that

$$\mathbf{X}_1^{(n)}(\Delta t n\mathbf{e}) - \mathbf{X}_2^{(n)}(\Delta t(n-1)\mathbf{e}) = \Delta t(n(\mathbf{X}_1^{(n)} - \mathbf{X}_2^{(n)})\mathbf{e} + \mathbf{X}_2^{(n)}\mathbf{e}) \geq \mathbf{e}_0.$$

If $(\mathbf{X}_1^{(n)} - \mathbf{X}_2^{(n)})\mathbf{e} = \mathbf{0}$ $(n = 1, \ldots, M)$, then the above condition reduces to $\Delta t\mathbf{X}_2^{(n)}\mathbf{e} \geq \mathbf{e}_0$. Hence, we have

[3]The term "two-level method" refers to the fact that two discrete time levels are involved into the definition of the mesh operator. Sometimes such a method is also called "one-step method".

Theorem 2.3.13 *If a non-negativity preserving discrete mesh operator of type* (2.3.10) *has such a structure that the conditions* $(\mathbf{X}_1^{(n)} - \mathbf{X}_2^{(n)})\,\mathbf{e} \geq \mathbf{0}$ *and* $\Delta t(n(\mathbf{X}_1^{(n)} - \mathbf{X}_2^{(n)})\,\mathbf{e} + \mathbf{X}_2^{(n)}\,\mathbf{e}) \geq \mathbf{e}_0$ *hold, then the discrete weak maximum-minimum principles and the discrete maximum norm contractivity properties are always satisfied. If, in addition,* $(\mathbf{X}_1^{(n)} - \mathbf{X}_2^{(n)})\,\mathbf{e} = \mathbf{0}$ *and* $\Delta t\mathbf{X}_2^{(n)}\,\mathbf{e} \geq \mathbf{e}_0$*, then the operator possesses all the discrete qualitative properties introduced in Section 2.3.1.*

As we can see from (2.3.10), the values $(\mathcal{L}\nu)(\mathbf{x}_i, t_n)$ $(i = 1, \ldots, N)$ depend only on the values of the function ν taken from the sets $\bar{\mathcal{P}} \times \{t_n\}$ and $\bar{\mathcal{P}} \times \{t_{n-1}\}$. This suggests that the discrete qualitative properties can be written in such a form where only two levels in t are involved instead of all the levels from 0 to $t^\star$. In order to define the qualitative properties in such a two-level form, we introduce the vector $\boldsymbol{\lambda}_0^n = [(\mathcal{L}\nu)_1^n, \ldots, (\mathcal{L}\nu)_N^n]$.

Let us consider the following property (denoted by DNP2) of a discrete mesh operator $\mathcal{L}$: For any $\nu \in \operatorname{dom}\mathcal{L}$ and $n \in \{1, \ldots, M\}$ such that

$$\boldsymbol{\nu}_0^{n-1}, \boldsymbol{\nu}_\partial^{n-1}, \boldsymbol{\nu}_\partial^n, \boldsymbol{\lambda}_0^n \geq \mathbf{0},$$

the relation $\boldsymbol{\nu}_0^n \geq \mathbf{0}$ holds.

The two-level forms of the maximum-minimum principles can be formulated similarly as this is done in [46, 47, 57].

Theorem 2.3.14 *For a discrete mesh operator* $\mathcal{L}$ *in the form* (2.3.10)*, the DNP property is equivalent to the DNP2 property.*

Proof. First we prove that the DNP2 implies the DNP. Let $\nu \in \operatorname{dom}\mathcal{L}$ and $t^\star = n^\star\Delta t \in \mathcal{R}_{t_M}$ such that $\min_{\mathcal{G}_{t^\star}} \nu \geq 0$ and $\mathcal{L}\nu|_{\mathcal{Q}_{\bar{t}^\star}} \geq 0$. Thus we have the relations

$$\boldsymbol{\nu}_0^0, \boldsymbol{\nu}_\partial^0, \boldsymbol{\nu}_\partial^1, \ldots, \boldsymbol{\nu}_\partial^{n^\star}, \boldsymbol{\lambda}_0^1, \boldsymbol{\lambda}_0^2, \ldots, \boldsymbol{\lambda}_0^{n^\star} \geq \mathbf{0}.$$

We need to show that $\nu|_{\mathcal{Q}_{\bar{t}^\star}} \geq 0$. Based on the property DNP2, $\boldsymbol{\nu}_0^0, \boldsymbol{\nu}_\partial^0, \boldsymbol{\nu}_\partial^1, \boldsymbol{\lambda}_0^1 \geq \mathbf{0}$ imply $\boldsymbol{\nu}_0^1 \geq \mathbf{0}$. Similarly, $\boldsymbol{\nu}_0^1, \boldsymbol{\nu}_\partial^1, \boldsymbol{\nu}_\partial^2, \boldsymbol{\lambda}_0^2 \geq \mathbf{0}$ imply $\boldsymbol{\nu}_0^2 \geq \mathbf{0}$, and so on. At last, $\boldsymbol{\nu}_0^{n^\star-1}, \boldsymbol{\nu}_\partial^{n^\star-1}, \boldsymbol{\nu}_\partial^{n^\star}, \boldsymbol{\lambda}_0^{n^\star} \geq \mathbf{0}$ imply $\boldsymbol{\nu}_0^{n^\star} \geq \mathbf{0}$. Thus, $\boldsymbol{\nu}_0^1, \ldots, \boldsymbol{\nu}_0^{n^\star} \geq \mathbf{0}$ and $\nu|_{\mathcal{Q}_{\bar{t}^\star}} \geq 0$.

Now let us suppose that the DNP property holds, and for a function $\nu \in \operatorname{dom}\mathcal{L}$ and $n^\star \in \{1, \ldots, M\}$ the conditions $\boldsymbol{\nu}_0^{n^\star-1}, \boldsymbol{\nu}_\partial^{n^\star-1}, \boldsymbol{\nu}_\partial^{n^\star}, \boldsymbol{\lambda}_0^{n^\star} \geq \mathbf{0}$ are valid. Let us define the function $\bar{\nu}$ as follows: $\bar{\nu}_i^0 = \nu_i^{n^\star-1}$, $\bar{\nu}_i^1 = \nu_i^{n^\star}$ and $\bar{\nu}_i^n$ is chosen arbitrarily if $1 < n \leq M$ $(i = 1, \ldots, \bar{N})$. Applying the DNP property for the function $\bar{\nu}$ with $t^\star = \Delta t$, we obtain that $\bar{\boldsymbol{\nu}}_0^1 = \boldsymbol{\nu}_0^{n^\star} \geq \mathbf{0}$. This completes the proof. ■

The two-level form of the discrete non-negativity preservation property makes possible the formulation of its necessary and sufficient conditions. In order to give this condition in a linear algebraic form, we introduce the following convenient partitions of the matrices $\mathbf{X}_1^{(n)}$ and $\mathbf{X}_2^{(n)}$:

$$\mathbf{X}_1^{(n)} = [\mathbf{X}_{10}^{(n)} |\, \mathbf{X}_{1\partial}^{(n)}], \quad \mathbf{X}_2^{(n)} = [\mathbf{X}_{20}^{(n)} |\, \mathbf{X}_{2\partial}^{(n)}],$$

where $\mathbf{X}_{10}^{(n)}$ and $\mathbf{X}_{20}^{(n)}$ are square matrices from $\mathbb{R}^{N\times N}$, and $\mathbf{X}_{1\partial}^{(n)}, \mathbf{X}_{2\partial}^{(n)} \in \mathbb{R}^{N\times N_\partial}$.

Theorem 2.3.15 *Let us suppose that the matrices* $\mathbf{X}_{10}^{(n)}$ $(n = 1, \ldots, M)$ *of the discrete mesh operator* $\mathcal{L}$ *defined in* (2.3.10) *are regular. Then* $\mathcal{L}$ *possesses the discrete non-negativity preservation property (DNP or DNP2) if and only if the following relations hold for all* $n = 1, \ldots, M$*,*

(P1) $(\mathbf{X}_{10}^{(n)})^{-1} \geq \mathbf{0}$*,*
(P2) $-(\mathbf{X}_{10}^{(n)})^{-1}\mathbf{X}_{1\partial}^{(n)} \geq \mathbf{0}$*,*
(P3) $(\mathbf{X}_{10}^{(n)})^{-1}\mathbf{X}_2^{(n)} \geq \mathbf{0}$*.*

PROOF. With the above notations and based on (2.3.10), we can write an identity in the following linear algebraic form

$$\mathbf{X}_{10}^{(n)}\boldsymbol{\nu}_0^n = -\mathbf{X}_{1\partial}^{(n)}\boldsymbol{\nu}_\partial^n + \mathbf{X}_2^{(n)}\boldsymbol{\nu}^{n-1} + \boldsymbol{\lambda}_0^n, \quad (n = 1, \dots, M). \tag{2.3.11}$$

Supposing the regularity of the matrix $\mathbf{X}_{10}$, we arrive at the iteration form

$$\boldsymbol{\nu}_0^n = -(\mathbf{X}_{10}^{(n)})^{-1}\mathbf{X}_{1\partial}^{(n)}\boldsymbol{\nu}_\partial^n + (\mathbf{X}_{10}^{(n)})^{-1}\mathbf{X}_2^{(n)}\boldsymbol{\nu}^{n-1} + (\mathbf{X}_{10}^{(n)})^{-1}\boldsymbol{\lambda}_0^n, \quad (n = 1, \dots, M).$$

According to the DNP2 property (which is equivalent to the DNP), the vector $\boldsymbol{\nu}_0^n$ is non-negative for all non-negative vectors $\boldsymbol{\nu}^{n-1}$, $\boldsymbol{\nu}_\partial^n$ and $\boldsymbol{\lambda}_0^n$ if and only if the coefficient matrices $(\mathbf{X}_{10}^{(n)})^{-1}$, $-(\mathbf{X}_{10}^{(n)})^{-1}\mathbf{X}_{1\partial}^{(n)}$ and $(\mathbf{X}_{10}^{(n)})^{-1}\mathbf{X}_2^{(n)}$ are non-negative matrices. This completes the proof. ■

Summarizing the results of the above three theorems, we can conclude the following.

Theorem 2.3.16 *Let us assume that the non-negativity assumption (conditions (P1)-(P3)) is satisfied. Then, besides the DNP property,*

- *under the conditions*

$$(\mathbf{X}_1^{(n)} - \mathbf{X}_2^{(n)})\,\mathbf{e} \geq \mathbf{0}; \quad \text{and} \quad \Delta t(n(\mathbf{X}_1^{(n)} - \mathbf{X}_2^{(n)})\,\mathbf{e} + \mathbf{X}_2^{(n)}\,\mathbf{e}) \geq \mathbf{e}_0 \tag{2.3.12}$$

  the qualitative properties DWMP, DWBMP and DMNC;

- *under the conditions*

$$(\mathbf{X}_1^{(n)} - \mathbf{X}_2^{(n)})\,\mathbf{e} = \mathbf{0} \quad \text{and} \quad \Delta t\mathbf{X}_2^{(n)}\,\mathbf{e} \geq \mathbf{e}_0 \ (\text{or } \Delta t\mathbf{X}_1^{(n)}\,\mathbf{e} \geq \mathbf{e}_0) \tag{2.3.13}$$

  the qualitative properties DWMP, DSMP, DWBMP, DSBMP and DMNC

are valid.

According to Theorem 2.3.12, for the mesh functions from H_0 and H_1 the conditions can be relaxed, namely, the second condition in (2.3.12) and (2.3.13) can be neglected. Hence, we get

Theorem 2.3.17 *Let us assume that the non-negativity assumption (conditions (P1)-(P3)) is satisfied. Then, besides the DNP property,*

- *under the conditions*

$$(\mathbf{X}_1^{(n)} - \mathbf{X}_2^{(n)})\mathbf{e} \geq \mathbf{0} \tag{2.3.14}$$

  the qualitative properties DWMP, DWBMP and DMNC;

- *under the conditions*

$$(\mathbf{X}_1^{(n)} - \mathbf{X}_2^{(n)})\mathbf{e} = \mathbf{0} \tag{2.3.15}$$

  the qualitative properties DWMP, DSMP, DWBMP, DSBMP and DMNC

are also valid for any mesh function from H_0 and H_1.

The operator $\mathcal{L}$ on the n-th time level is completely defined by the matrices $\mathbf{X}_1^{(n)}$ and $\mathbf{X}_2^{(n)}$, see (2.3.10). In the typical numerical applications, as also in our work, (cf. Section 2.3.5), they are derived from the approximation of the concrete continuous operator L. When we use a one-parameter family of the approximation (which is called the θ-method), these matrices are defined by the matrices $\mathbf{M}^{(n)}$, $\mathbf{K}^{(n)}$, (called mass and stiffness matrices, respectively), and a real parameter θ, as follows

$$\begin{aligned} \mathbf{X}_1^{(n)} &= \frac{1}{\Delta t}\mathbf{M}^{(n)} + \theta\mathbf{K}^{(n)}, \\ \mathbf{X}_2^{(n)} &= \frac{1}{\Delta t}\mathbf{M}^{(n)} - (1-\theta)\mathbf{K}^{(n)}. \end{aligned} \tag{2.3.16}$$

The matrices $\mathbf{M}^{(n)}$ and $\mathbf{K}^{(n)}$ have the size $N \times \bar{N}$. Hence, the discrete mesh operator $\mathcal{L}$ in (2.3.10) can be written in the following (so-called canonical) form:

$$(\mathcal{L}\nu)_i^n = (\mathbf{M}^{(n)}\frac{\boldsymbol{\nu}^n - \boldsymbol{\nu}^{n-1}}{\Delta t} + \theta\mathbf{K}^{(n)}\boldsymbol{\nu}^n + (1-\theta)\mathbf{K}^{(n)}\boldsymbol{\nu}^{n-1})_i. \tag{2.3.17}$$

Therefore, conditions (2.3.12) and (2.3.13) in Theorem 2.3.16 can be formulated as follows:

$$\mathbf{K}^{(n)}\mathbf{e} \geq \mathbf{0} \quad \text{and} \quad \Delta t(n-1+\theta)\mathbf{K}^{(n)}\mathbf{e} + \mathbf{M}^{(n)}\mathbf{e} \geq \mathbf{e}_0; \tag{2.3.12$'$}$$

and

$$\mathbf{K}^{(n)}\mathbf{e} \ = \mathbf{0} \quad \text{and} \quad \mathbf{M}^{(n)}\mathbf{e} \geq \mathbf{e}_0. \tag{2.3.13$'$}$$

Thus, we have

Theorem 2.3.18 *Let us assume that the discrete mesh operator $\mathcal{L}$, defined by* (2.3.10) *and* (2.3.16)*, is non-negativity preserving and the relation*

$$\mathbf{M}^{(n)}\mathbf{e} \geq \mathbf{e}_0 \tag{2.3.18}$$

holds. Then, beyond the DNP property, under the condition $\mathbf{K}^{(n)}\mathbf{e} \ \geq \mathbf{0}$, the operator $\mathcal{L}$ is DWMP, DWBMP and DMNC; while in the case $\mathbf{K}^{(n)}\mathbf{e} \ = \mathbf{0}$ it obeys each of the DWMP, DSMP, DWBMP, DSBMP and DMNC properties.

Remark 2.3.19 *Let us note that, according to Theorem 2.3.12, the assumption is simpler for the mesh functions from H_1: if $\mathcal{L}$ is a discrete non-negativity preserving operator and the condition*

$$\mathbf{K}^{(n)}\mathbf{e} \geq \mathbf{0} \tag{2.3.19}$$

is satisfied, then it possesses all the other qualitative properties, too.

Remark 2.3.20 *The above consideration is a special case of the following general approach. For the pair of the matrices $(\mathbf{X}_1^{(n)}, \mathbf{X}_2^{(n)})$ (which define the two-level operator) we define a mapping $\varphi : \mathbb{R}^{N\times\bar{N}} \times \mathbb{R}^{N\times\bar{N}} \to \mathbb{R}^{N\times\bar{N}} \times \mathbb{R}^{N\times\bar{N}}$ which is assumed to be a bijection. Then, we write the conditions, obtained for the matrices $\mathbf{X}_1^{(n)}$ and $\mathbf{X}_2^{(n)}$, for the matrices $\varphi^{-1}(\mathbf{X}_1^{(n)}, \mathbf{X}_2^{(n)})$, denoted by $(\mathbf{M}^{(n)}, \mathbf{K}^{(n)})$.*

For the θ-method, as one can see from (2.3.16)*, the inverse of this bijection is*

$$\varphi_\theta^{-1}(\mathbf{A}, \mathbf{B}) := \left(\frac{1}{\Delta t}\mathbf{A} + \theta\mathbf{B}, \frac{1}{\Delta t}\mathbf{A} - (1-\theta)\mathbf{B}\right), \tag{2.3.20}$$

where θ is any fixed number. Hence,

$$\varphi_\theta(\mathbf{A},\mathbf{B}) = (\Delta t((1-\theta)\mathbf{A}+\theta\mathbf{B}), \mathbf{A}-\mathbf{B}). \tag{2.3.21}$$

This means the following. The above approach (knowing $\mathbf{M}^{(n)}$ and $\mathbf{K}^{(n)}$ and selecting the mapping, i.e., by fixing θ, we define $\mathbf{X}_1^{(n)}$ and $\mathbf{X}_2^{(n)}$) can be reversed: when the matrices $\mathbf{X}_1^{(n)}$ and $\mathbf{X}_2^{(n)}$ are a priori given, and we know that the discretization was obtained by use of the θ-method, we proceed as follows. We introduce the matrices

$$\mathbf{M}^{(n)} = \Delta t\left((1-\theta)\mathbf{X}_1^{(n)}+\theta\mathbf{X}_2^{(n)}\right), \quad \mathbf{K}^{(n)} = \mathbf{X}_1^{(n)}-\mathbf{X}_2^{(n)}, \tag{2.3.22}$$

and use the same conditions. However, as we will see later, the first approach is more natural, because, as it was already mentioned, we define the matrices $\mathbf{M}^{(n)}$ and $\mathbf{K}^{(n)}$ a priori from the approximation of the continuous operator.

2.3.4 Matrix maximum principles and their relations

In the literature, for partitioned matrices with a certain structure some qualitative properties have been introduced (e.g., [23, 57, 132]). In this section we analyze their relation to our notions.

We consider the block-matrix $\mathbf{H} \in \mathbb{R}^{k\times k}$ and the block-vector $\mathbf{y} \in \mathbb{R}^k$ in the form

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_1 & \mathbf{H}_2 \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}, \tag{2.3.23}$$

where submatrices $\mathbf{H}_1 \in \mathbb{R}^{k_1\times k_1}$, $\mathbf{I} \in \mathbb{R}^{k_2\times k_2}$, $\mathbf{H}_2 \in \mathbb{R}^{k_1\times k_2}$, $\mathbf{0} \in \mathbb{R}^{k_2\times k_1}$, $\mathbf{y}_1 \in \mathbb{R}^{k_1}$ and $\mathbf{y}_2 \in \mathbb{R}^{k_2}$ with $k = k_1 + k_2$. In the sequel, for arbitrary vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^k$, we will use the following notations:

$$\begin{gathered} \max\{\mathbf{v}\} := \max\{v_1, v_2, \ldots, v_k\}, \quad \max\{0, \mathbf{v}\} := \max\{0, \max\{\mathbf{v}\}\}, \\ \max\{\mathbf{v}, \mathbf{w}\} := \max\{\max\{\mathbf{v}\}, \max\{\mathbf{w}\}\}. \end{gathered} \tag{2.3.24}$$

According to Ciarlet's and Stoyan's works (see [23], [131], [132]) we introduce the following definitions.

Definition 2.3.21 *We say that a matrix $\mathbf{H}$ satisfies the Ciarlet matrix maximum principle (CMMP) if for arbitrary vectors $\mathbf{y}_1 \in \mathbb{R}^{k_1}$ and $\mathbf{y}_2 \in \mathbb{R}^{k_2}$, such that $\mathbf{H}_1\mathbf{y}_1 + \mathbf{H}_2\mathbf{y}_2 \le \mathbf{0}$, the inequality $\max\{\mathbf{y}_1\} \le \max\{0, \mathbf{y}_2\}$ holds.*

Definition 2.3.22 *We say that a matrix $\mathbf{H}$ satisfies the Stoyan matrix maximum principle (SMMP) if for arbitrary vectors $\mathbf{y}_1 \in \mathbb{R}^{k_1}$ and $\mathbf{y}_2 \in \mathbb{R}^{k_2}$, such that $\mathbf{H}_1\mathbf{y}_1 + \mathbf{H}_2\mathbf{y}_2 = \mathbf{0}$ and $\mathbf{y}_2 \ge \mathbf{0}$, the inequality $\max\{\mathbf{y}_1\} \le \max\{\mathbf{y}_2\}$ holds.*

Remark 2.3.23 *We note that in [131] the SMMP was originally formulated for general, un-partitioned matrices as follows: a quadratic matrix $\mathbf{H}$ is said to satisfy the maximum principle if the relation $\mathbf{Hy} \ge \mathbf{0}$ implies that $\mathbf{y} \ge \mathbf{0}$, moreover, when $\max\{\mathbf{y}\} = y_{i_0}$, (i.e., the maximum is taken on the i_0-th component) then $(\mathbf{Hy})_{i_0} > 0$. An application of this principle to the structured matrix $\mathbf{H}$ of the form (2.3.23) yields the definition of the SMMP in Definition 2.3.22.*

The above definitions give information about the location of the maximum components of the unknown vector $\mathbf{y} \in \mathbb{R}^k$, using some a priori information: for the CMMP the non-negative maximum, while for the SMMP the maximum is taken over the last indices $i = k_1 + 1, k_1 + 2, \ldots, k$, i.e., on the sub-vector $\mathbf{y}_2$.

Remark 2.3.24 *If a matrix* $\mathbf{H}$ *satisfies the CMMP, then it is necessarily regular. To show this, it is enough to prove that the relation* $\mathbf{H}_1\mathbf{y}_1 = \mathbf{0}$ *implies that* $\mathbf{y}_1 = \mathbf{0}$. *By the choice* $\mathbf{y}_2 = \mathbf{0}$, *the application of the CMMP implies the inequality* $\mathbf{y}_1 \leq \mathbf{0}$. *Repeating this argument for* $-\mathbf{y}_1$, *we obtain* $-\mathbf{y}_1 \leq \mathbf{0}$, *which shows the validity of the required equality. The similar statement holds for the SMMP, too.*

The CMMP and SMMP properties can be guaranteed in the following way.

Theorem 2.3.25 ([23]) *The matrix* $\mathbf{H}$ *satisfies the CMMP if and only if the following two matrix conditions hold:*

(C1): $\mathbf{H}$ *is monotone, i.e.,*

$$\mathbf{H}^{-1} = \begin{pmatrix} \mathbf{H}_1^{-1} & -\mathbf{H}_1^{-1}\mathbf{H}_2 \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \geq \mathbf{0}; \tag{2.3.25}$$

(C2): as before, using the notation $\mathbf{e}_{k_m} \in \mathbb{R}^{k_m}$ $(m = 1, 2)$ *for the vectors with all coordinates equal to one, we have*

$$-\mathbf{H}_1^{-1}\mathbf{H}_2\mathbf{e}_{k_2} \leq \mathbf{e}_{k_1}. \tag{2.3.26}$$

The condition (C2) can be relaxed by the following sufficient condition

(C2'): the row sums of the matrix $\mathbf{H}$ *are all non-negative, i.e.,*

$$\mathbf{H}_1\mathbf{e}_{k_1} + \mathbf{H}_2\mathbf{e}_{k_2} \geq \mathbf{0}. \tag{2.3.27}$$

The following statement gives an equivalent condition for the CMMP.

Lemma 2.3.26 *A matrix* $\mathbf{H}$ *satisfies the CMMP if and only if the implications*

$$\mathbf{H}_1\mathbf{y}_1 + \mathbf{H}_2\mathbf{y}_2 \leq \mathbf{0}, \text{ and } \mathbf{y}_2 \leq \mathbf{0} \quad \Rightarrow \quad \max\{\mathbf{y}_1\} \leq \mathbf{0}; \tag{2.3.28}$$

$$\mathbf{H}_1\mathbf{y}_1 + \mathbf{H}_2\mathbf{y}_2 \leq \mathbf{0}, \text{ and } \mathbf{y}_2 \geq \mathbf{0} \quad \Rightarrow \quad \max\{\mathbf{y}_1\} \leq \max\{\mathbf{y}_2\} \tag{2.3.29}$$

are valid.

Proof. It is obvious that CMMP implies both (2.3.28) and (2.3.29). Therefore have to show only the converse implication. From the assumption (2.3.28) it follows that the vector $\mathbf{Hy}$ is also non-positive for any non-positive $\mathbf{y}$. This yields the monotonicity of $\mathbf{H}$, i.e., $\mathbf{H}^{-1} \geq \mathbf{0}$ is valid. On the other side, let us choose $\mathbf{y}_1 = -\mathbf{H}_1^{-1}\mathbf{H}_2\mathbf{e}_{k_2}$ and $\mathbf{y}_2 = \mathbf{e}_{k_2}$. Then $\mathbf{H}_1\mathbf{y}_1 + \mathbf{H}_2\mathbf{y}_2 = \mathbf{0}$ and $\mathbf{y}_2 \geq \mathbf{0}$. For these vectors we can use (2.3.29) and obtain the relation $\max\{-\mathbf{H}_1^{-1}\mathbf{H}_2\mathbf{e}_{k_2}\} \leq \max\{\mathbf{e}_{k_2}\} = 1$. Hence, $-\mathbf{H}_1^{-1}\mathbf{H}_2\mathbf{e}_{k_2} \leq \mathbf{e}_{k_1}$. Hence, according to the Theorem 2.3.25, we have showed the CMMP property for the matrix $\mathbf{H}$. ■

Remark 2.3.27 *The CMMP obviously implies the SMMP. Therefore, the above conditions also guarantee the SMMP property. It is worth mentioning that for the un-partitioned matrix* $\mathbf{H}$ *the monotonicity and the condition* $\mathbf{He} \geq \mathbf{0}$ *(which is, in fact, the analogue of the condition* (2.3.27)*) are necessary conditions for validity of the SMMP. When* $\mathbf{H}$ *is an M-matrix*[4]*, then these conditions are necessary and sufficient ([131]).*

We can combine the CMMP and the SMPP as follows: we require that under the CMMP condition the implication in the SMMP is true, i.e., we introduce

Definition 2.3.28 *We say that a matrix* $\mathbf{H}$ *satisfies the Ciarlet-Stoyan matrix maximum principle (CSMMP) if for arbitrary vectors* $\mathbf{y}_1 \in \mathbb{R}^{k_1}$ *and* $\mathbf{y}_2 \in \mathbb{R}^{k_2}$*, such that* $\mathbf{H}_1\mathbf{y}_1 + \mathbf{H}_2\mathbf{y}_2 \leq \mathbf{0}$*, the relation* $\max\{\mathbf{y}_1\} \leq \max\{\mathbf{y}_2\}$ *holds.*

Obviously, the CSMMP implies both the CMMP and SMMP properties. This property can be guaranteed by the following statement (cf. [76]).

Lemma 2.3.29 *Assume that* $\mathbf{H}$ *is monotone and the condition*

$$\mathbf{H}_1\mathbf{e}_{k_1} + \mathbf{H}_2\mathbf{e}_{k_2} = \mathbf{0} \tag{2.3.30}$$

holds. Then $\mathbf{H}$ *has the CSMMP property.*

Proof. Let $\mathbf{y}_1 \in \mathbb{R}^{k_1}$ and $\mathbf{y}_2 \in \mathbb{R}^{k_2}$ be arbitrary vectors with the property $\mathbf{H}_1\mathbf{y}_1 + \mathbf{H}_2\mathbf{y}_2 \leq \mathbf{0}$. Since $\mathbf{H}$ is monotone, therefore $\mathbf{H}_1^{-1} \geq \mathbf{0}$ and $-\mathbf{H}_1^{-1}\mathbf{H}_2 \geq \mathbf{0}$ (cf. (2.3.25)). Therefore we have

$$\mathbf{y}_1 \leq -\mathbf{H}_1^{-1}\mathbf{H}_2\mathbf{y}_2 \leq -\mathbf{H}_1^{-1}\mathbf{H}_2(\max\{\mathbf{y}_2\}\mathbf{e}_{k_2}) = -(\max\{\mathbf{y}_2\})\mathbf{H}_1^{-1}\mathbf{H}_2\mathbf{e}_{k_2}. \tag{2.3.31}$$

Due to the assumption (2.3.30), the relation (2.3.31) implies that

$$\mathbf{y}_1 \leq (\max\{\mathbf{y}_2\})\mathbf{e}_{k_1}, \tag{2.3.32}$$

which proves the statement. ■

In the next statement, we show that the conditions in the Lemma 2.3.29 are not only sufficient but they are necessary, too.

Lemma 2.3.30 *Assume that* $\mathbf{H}$ *has the CSMMP property. Then* $\mathbf{H}$ *is monotone and the relation*

$$\mathbf{H}_1\mathbf{e}_{k_1} + \mathbf{H}_2\mathbf{e}_{k_2} = \mathbf{0} \tag{2.3.33}$$

holds.

Proof. Since the CSMMP property implies the CMMP property, therefore, due to the Theorem 2.3.25, $\mathbf{H}$ is monotone.

In order to show the second condition, first, let us put $\mathbf{y}_1 = -\mathbf{H}_1^{-1}\mathbf{H}_2\mathbf{e}_{k_2}$ and $\mathbf{y}_2 = \mathbf{e}_{k_2}$ (as in the proof of the Lemma 2.3.26). Since for this choice the CSMMP is applicable, we get the estimation $\max\{-\mathbf{H}_1^{-1}\mathbf{H}_2\mathbf{e}_{k_2}\} \leq \max\{\mathbf{e}_{k_2}\} = 1$. Let us put now $\mathbf{y}_1 = \mathbf{H}_1^{-1}\mathbf{H}_2\mathbf{e}_{k_2}$ and $\mathbf{y}_2 = -\mathbf{e}_{k_2}$. The CSMMP is again applicable and we get the estimation $\max\{\mathbf{H}_1^{-1}\mathbf{H}_2\mathbf{e}_{k_2}\} \leq \max\{-\mathbf{e}_{k_2}\} = -1$.

[4]A Z-matrix (a square matrix with all off-diagonal entries are less than or equal to zero) A is called an M-matrix if the relation$Av \geq 0$ implies that $v \geq 0$. There are many equivalent definitions, see e.g., [9].

The above two estimations clearly result in the equality $-\mathbf{H}_1^{-1}\mathbf{H}_2\mathbf{e}_{k_2} = \mathbf{e}_{k_1}$, which yields the required (2.3.33). ■

In the sequel we apply the above theory to a linear algebraic system of a special form. Namely, we use the notation $\mathbf{b}$ for the vector $\mathbf{Hy} \in \mathbb{R}^k$, i.e., we consider the system

$$\mathbf{Hy} = \mathbf{b}. \tag{2.3.34}$$

Hence, $\mathbf{b}$ has the partitioning

$$\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix}, \tag{2.3.35}$$

where $\mathbf{b}_1 \in \mathbb{R}^{k_1}$ and $\mathbf{b}_2 \in \mathbb{R}^{k_2}$, respectively. Moreover, let

$$\mathbf{H} = \begin{pmatrix} \mathbf{A} & -\mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} \mathbf{u}^{(n)} \\ \mathbf{u}^{(n-1)} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \bar{\mathbf{f}}^{(n)} \\ \mathbf{u}^{(n-1)} \end{pmatrix}, \tag{2.3.36}$$

with

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_0 & \mathbf{A}_\partial \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{B}_0 & \mathbf{B}_\partial \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \quad \mathbf{u}^{(n)} = \begin{pmatrix} \mathbf{u}_0^{(n)} \\ \mathbf{u}_\partial^{(n)} \end{pmatrix}, \quad \bar{\mathbf{f}}^{(n)} = \begin{pmatrix} \mathbf{f}^{(n)} \\ \mathbf{f}_\partial^{(n)} \end{pmatrix}. \tag{2.3.37}$$

Here $\mathbf{A}_0, \mathbf{B}_0 \in \mathbb{R}^{N\times N}$, $\mathbf{A}_\partial, \mathbf{B}_\partial \in \mathbb{R}^{N\times N_\partial}$, $\mathbf{u}_0^{(n)}, \mathbf{f}^{(n)} \in \mathbb{R}^N$ and $\mathbf{u}_\partial^{(n)}, \mathbf{f}_\partial^{(n)} \in \mathbb{R}^{N_\partial}$. Then in the problem (2.3.34),(2.3.36),(2.3.37) $\mathbf{H} \in \mathbb{R}^{2\bar{N}\times 2\bar{N}}$, $\mathbf{y}, \mathbf{b} \in \mathbb{R}^{2\bar{N}}$. (As before, $\bar{N} = N + N_\partial$.) Let us notice that the problem (2.3.34),(2.3.36),(2.3.37) is equivalent to the system of linear algebraic equations of the form

$$\mathbf{Au}^{(n)} = \mathbf{Bu}^{(n-1)} + \bar{\mathbf{f}}^{(n)}. \tag{2.3.38}$$

(Notice that $\mathbf{f}_\partial^{(n)} = \mathbf{u}_\partial^{(n)} - \mathbf{u}_\partial^{(n-1)}$.) We will refer to this problem as the canonical algebraic problem (CAP). The qualitative properties of this problem can be defined as follows.

Definition 2.3.31 *We say that the CAP is non-negativity preserving, when* $\mathbf{b} \geq \mathbf{0}$ *results in the relation* $\mathbf{y} \geq \mathbf{0}$*, that is, the implication*

$$\mathbf{f}^{(n)} \geq \mathbf{0};\ \mathbf{u}_\partial^{(n)} - \mathbf{u}_\partial^{(n-1)} \geq \mathbf{0};\ \mathbf{u}_0^{(n-1)} \geq \mathbf{0} \text{ and } \mathbf{u}_\partial^{(n-1)} \geq \mathbf{0} \ \Rightarrow\ \mathbf{u}_0^{(n)} \geq \mathbf{0} \text{ and } \mathbf{u}_\partial^{(n)} \geq \mathbf{0}, \tag{2.3.39}$$

which is the same as

$$\mathbf{f}^{(n)} \geq \mathbf{0};\ \mathbf{u}_0^{(n-1)} \geq \mathbf{0} \text{ and } \mathbf{u}_\partial^{(n)} \geq \mathbf{u}_\partial^{(n-1)} \geq \mathbf{0} \ \Rightarrow\ \mathbf{u}_0^{(n)} \geq \mathbf{0} \tag{2.3.40}$$

is true.

Definition 2.3.32 *We say that the CAP satisfies the Ciarlet maximum principle, when the corresponding matrix* $\mathbf{H}$ *defined in* (2.3.36),(2.3.37) *has the CMMP property, i.e., the implication*

$$\mathbf{f}^{(n)} \leq \mathbf{0} \text{ and } \mathbf{u}_\partial^{(n)} - \mathbf{u}_\partial^{(n-1)} \leq \mathbf{0} \ \Rightarrow\ \max\{\mathbf{u}_0^{(n)}, \mathbf{u}_\partial^{(n)}\} \leq \max\{0, \mathbf{u}_0^{(n-1)}, \mathbf{u}_\partial^{(n-1)}\}, \tag{2.3.41}$$

or, equivalently, the implication

$$\mathbf{f}^{(n)} \leq \mathbf{0} \text{ and } \mathbf{u}_\partial^{(n)} \leq \mathbf{u}_\partial^{(n-1)} \ \Rightarrow\ \max\{\mathbf{u}_0^{(n)}\} \leq \max\{0, \mathbf{u}_0^{(n-1)}, \mathbf{u}_\partial^{(n-1)}, \mathbf{u}_\partial^{(n)}\} \tag{2.3.42}$$

is true.

Analogically, we introduce the following

Definition 2.3.33 *We say that the CAP* (2.3.34),(2.3.36),(2.3.37)*, (or equivalently, the problem* (2.3.37)- (2.3.38)*), satisfies the Stoyan maximum principle, when the correspondig matrix* $\mathbf{H}$ *by* (2.3.36),(2.3.37)*, has the SMMP property, which yields that the implication*

$$\mathbf{f}^{(n)} = \mathbf{0},\ \mathbf{u}_0^{(n-1)} \geq \mathbf{0},\ \mathbf{u}_\partial^{(n)} = \mathbf{u}_\partial^{(n-1)} \geq \mathbf{0}\ \Rightarrow\ \max\{\mathbf{u}_0^{(n)}, \mathbf{u}_\partial^{(n)}\} \leq \max\{\mathbf{u}_0^{(n-1)}, \mathbf{u}_\partial^{(n-1)}\}, \tag{2.3.43}$$

i.e.,

$$\mathbf{f}^{(n)} = \mathbf{0},\ \mathbf{u}_0^{(n-1)} \geq \mathbf{0},\ \mathbf{u}_\partial^{(n)} = \mathbf{u}_\partial^{(n-1)} \geq \mathbf{0}\ \Rightarrow\ \max\{\mathbf{u}_0^{(n)}\} \leq \max\{\mathbf{u}_0^{(n-1)}, \mathbf{u}_\partial^{(n-1)}, \mathbf{u}_\partial^{(n)}\} \tag{2.3.44}$$

is true.

It follows from the definitions that the validity of the Ciarlet maximum principle implies the validity of the Stoyan maximum principle.

First we investigate the non-negativity preservation property of the CAP. Since, according to (2.3.25),

$$\mathbf{H}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} & \mathbf{A}^{-1}\mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \tag{2.3.45}$$

we need the monotonicity of the matrix $\mathbf{H}$, which is valid only under the conditions

$$\mathbf{A}^{-1} \geq \mathbf{0}; \quad \mathbf{A}^{-1}\mathbf{B} \geq \mathbf{0}. \tag{2.3.46}$$

(This yields that the matrices $\mathbf{A}$ and $\mathbf{B}$ must form a weak regular splitting of the matrix $\mathbf{A} - \mathbf{B}$.) Using (2.3.37), we get that (2.3.46) is valid if and only if the relations

$$\mathbf{A}_0^{-1} \geq \mathbf{0}, \quad -\mathbf{A}_0^{-1}\mathbf{A}_\partial \geq \mathbf{0}, \quad \mathbf{A}_0^{-1}\mathbf{B}_0 \geq \mathbf{0}, \quad \mathbf{A}_0^{-1}(\mathbf{B}_\partial - \mathbf{A}_\partial) \geq \mathbf{0} \tag{2.3.47}$$

are true. Hence, the following statement is true.

Lemma 2.3.34 *The CAP is non-negativity preserving if and only if the conditions in* (2.3.47) *are satisfied.*

We pass to the investigation of the Ciarlet maximum principle property of the CAP. Due to Theorem 2.3.25, it is sufficient to require the monotonicity of the matrix $\mathbf{H}$ and the relation

$$(\mathbf{A} - \mathbf{B})\mathbf{e} \geq \mathbf{0}. \tag{2.3.48}$$

Substituting (2.3.37), the condition (2.3.48) yields the condition

$$(\mathbf{A}_0 - \mathbf{B}_0)\mathbf{e}_0 + (\mathbf{A}_\partial - \mathbf{B}_\partial)\mathbf{e}_\partial \geq \mathbf{0}. \tag{2.3.49}$$

Hence we get

Lemma 2.3.35 *The CAP satisfies both the Ciarlet and Stoyan maximum principle properties if the conditions* (2.3.47) *and* (2.3.49) *are satisfied.*

In typical applications we a priori know the vectors $\mathbf{u}_0^{(n-1)}$, $\mathbf{u}_\partial^{(n-1)}$, $\mathbf{u}_\partial^{(n)}$ and $\mathbf{f}^{(n)}$ and we want to guarantee some qualitative properties of the only unknown vector $\mathbf{u}_0^{(n)}$. (We recall the relation $\mathbf{f}_\partial^{(n)} = \mathbf{u}_\partial^{(n)} - \mathbf{u}_\partial^{(n-1)}$, which means that $\mathbf{f}_\partial$ is not a free parameter in the CAP.) This means that we investigate the problem

$$\mathbf{A}_0\mathbf{u}_0^{(n)} = \mathbf{B}_0\mathbf{u}_0^{(n-1)} + \mathbf{B}_\partial\mathbf{u}_\partial^{(n-1)} - \mathbf{A}_\partial\mathbf{u}_\partial^{(n)} + \mathbf{f}^{(n)}, \tag{2.3.50}$$

where the "input vectors" on the right-hand side are arbitrary, given vectors, i.e., they are chosen from the set

$$H = \{\mathbf{u}_0^{(n-1)}, \mathbf{f}^{(n)} \in \mathbb{R}^N, \quad \mathbf{u}_\partial^{(n-1)}, \mathbf{u}_\partial^{(n)} \in \mathbb{R}^{N_\partial}\}. \tag{2.3.51}$$

We will refer to the problem (2.3.50) as iterative algebraic problem (IAP). If we define $\mathbf{X}_1^{(n)} = \mathbf{A}$ and $\mathbf{X}_2^{(n)} = \mathbf{B}$ in the two-level mesh operator $\mathcal{L}$ defined in (2.3.10), we can establish a direct connection between the qualitative properties of $\mathcal{L}$ and the IAP. (We note that, due to the above choice, $\mathbf{X}_{10}^{(n)} = \mathbf{A}_0$, $\mathbf{X}_{1\partial}^{(n)} = \mathbf{A}_\partial$, $\mathbf{X}_{20}^{(n)} = \mathbf{B}_0$ and $\mathbf{X}_{2\partial}^{(n)} = \mathbf{B}_\partial$.) The following definitions are straightforward.

Definition 2.3.36 *We say that the IAP* (2.3.50) *is non-negativity preserving if the implication*

$$\mathbf{f}^{(n)} \geq \mathbf{0};\ \ \mathbf{u}_\partial^{(n-1)} \geq \mathbf{0};\ \ \mathbf{u}_\partial^{(n-1)} \geq \mathbf{0} \text{ and } \mathbf{u}_0^{(n)} \geq \mathbf{0} \ \ \Rightarrow \ \ \mathbf{u}_0^{(n)} \geq \mathbf{0} \tag{2.3.52}$$

is true.

Definition 2.3.37 *We say that the IAP* (2.3.50) *satisfies the discrete weak boundary maximum principle (DWBMP) when the implication*

$$\mathbf{f}^{(n)} \leq \mathbf{0} \ \ \Rightarrow \ \ \max \mathbf{u}_0^{(n)} \leq \max\{0, \mathbf{u}_0^{(n-1)}, \mathbf{u}_\partial^{(n-1)}, \mathbf{u}_\partial^{(n)}\} \tag{2.3.53}$$

is true.

Definition 2.3.38 *We say that the IAP* (2.3.50) *satisfies the discrete strong boundary maximum principle (DSBMP), when the implication*

$$\mathbf{f}^{(n)} \leq \mathbf{0} \ \ \Rightarrow \ \ \max \mathbf{u}_0^{(n)} \leq \max\{\mathbf{u}_0^{(n-1)}, \mathbf{u}_\partial^{(n-1)}, \mathbf{u}_\partial^{(n)}\} \tag{2.3.54}$$

is true.

In the sequel, we analyze the relation between the qualitative properties of the CAP and the IAP.

First of all, we introduce some subsets in H, defined in (2.3.51). Namely, we define

$$\begin{aligned} H_+ &= H \bigcap \{\mathbf{f}^{(n)} \geq \mathbf{0},\ \mathbf{u}_0^{(n-1)} \geq \mathbf{0},\ \mathbf{u}_\partial^{(n-1)} \geq \mathbf{0},\ \mathbf{u}_\partial^{(n)} \geq \mathbf{0}\}, \\ H_+^M &= H \bigcap \{\mathbf{f}^{(n)} \geq \mathbf{0},\ \mathbf{u}_0^{(n-1)} \geq \mathbf{0},\ \mathbf{u}_\partial^{(n)} \geq \mathbf{u}_\partial^{(n-1)} \geq \mathbf{0}\}, \end{aligned} \tag{2.3.55}$$

$$H_{DBMP} = H \bigcap \{\mathbf{f}^{(n)} \leq \mathbf{0}\},\ \ H_{DWBMP}^C = H_{DBMP} \bigcap \{\ \mathbf{u}_\partial^{(n-1)} \geq \mathbf{u}_\partial^{(n)}\}, \tag{2.3.56}$$

$$H_{DSBMP}^S = H \bigcap \{\mathbf{f}^{(n)} = \mathbf{0},\ \mathbf{u}_0^{(n-1)} \geq \mathbf{0},\ \mathbf{u}_\partial^{(n-1)} = \mathbf{u}_\partial^{(n)} \geq \mathbf{0}\}. \tag{2.3.57}$$

The inclusions

$$H_+ \supset H_+^M; \quad H_{DBMP} \supset H_{DWBMP}^C; \quad H_{DBMP} \supset H_{DSBMP}^S \tag{2.3.58}$$

are obvious. For the CAP the different qualitative properties (non-negativity preservation, Ciarlet maximum principle, Stoyan maximum principle) are defined on the sub-sets H_+^M, H_{DSBMP}^S and H_{DSBMP}^C, respectively. For the IAP, the corresponding qualitative properties (non-negativity preservation, DWBMP, DSBMP) are defined on the wider subsets H_+ and H_{DBMP}, respectively. Moreover, if some qualitative property is guaranteed for the IAP, then the corresponding CAP also possesses this qualitative property on the smaller subset, where it is defined. In the following we compare those conditions that guarantee these qualitative properties.

We start with the non-negativity preservation property. Based on Theorem 2.3.15, the following lemma holds.

Lemma 2.3.39 *The IAP* (2.3.50) *is non-negativity preserving if and only if the conditions*

$$\mathbf{A}_0^{-1} \geq \mathbf{0}; \quad \mathbf{A}_0^{-1}\mathbf{B}_\partial \geq \mathbf{0}, \quad \mathbf{A}_0^{-1}\mathbf{B}_0 \geq \mathbf{0}, \quad -\mathbf{A}_0^{-1}\mathbf{A}_\partial \geq \mathbf{0} \tag{2.3.59}$$

are satisfied.

We can see that the conditions (2.3.59) imply the conditions (2.3.47), i.e., the non-negativity preservation property of the IAP implies the non-negativity preservation property of the CAP. To analyze the validity of the converse implication, we consider an example.

Example 2.3.40 *We choose* $N = N_\partial$, $\mathbf{A}_0 = k\mathbf{I}$, $\mathbf{B}_0 = \mathbf{I}$, $\mathbf{A}_\partial = -2\mathbf{I}$ *and* $\mathbf{B}_\partial = -\mathbf{I}$, *where* $k > 0$ *is an arbitrary number. Then the conditions in* (2.3.59) *are not satisfied, while the conditions* (2.3.47) *are true. For this case we have*

$$\mathbf{H} = \begin{pmatrix} k\mathbf{I} & -2\mathbf{I} & -\mathbf{I} & \mathbf{I} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & -\mathbf{I} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix}; \quad \mathbf{H}^{-1} = \begin{pmatrix} \frac{1}{k}\mathbf{I} & \frac{2}{k}\mathbf{I} & \frac{1}{k}\mathbf{I} & \frac{1}{k}\mathbf{I} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix}. \tag{2.3.60}$$

This shows that the conditions in (2.3.59) *are not necessary for the conditions* (2.3.47), *i.e., the non-negativity preservation property of the CAP does not imply automatically the non-negativity preservation property of the IAP. However, in the case* $\mathbf{u}_\partial^{(n)} = \mathbf{u}_\partial^{(n-1)}$ *the conditions* (2.3.47) *and* (2.3.59) *are the same, i.e., the non-negativity preservation properties of the CAP and IAP are equivalent.*

We pass to the investigation of the maximum principles.

Comparing the implications (2.3.42) and (2.3.53), the implication "DWBMP $\Rightarrow$ Ciarlet maximum principle" is obviously true. Similarly, based on (2.3.44) and (2.3.54), the implication "DSBMP $\Rightarrow$ Stoyan maximum principle" is also valid. Due to Theorem 2.3.17, under the non-negativity preservation property, the condition (2.3.14) guarantees the DWBMP property and the condition (2.3.15) results in the DSBMP for the IAP. As we have seen, the condition (2.3.59) implies the conditions (2.3.47), and the condition (2.3.14) coincides with (2.3.48). Therefore, the conditions of DWBMP guaranties the Ciarlet maximum principle. However, as the following example shows, the relaxed conditions (2.3.47) and (2.3.48) cannot guarantee the DWBMP property on the whole H_{DWBMP}.

(This shows that the Ciarlet maximum principle is guaranteed not by the exact condition of the DWBMP but by a sufficient condition of it.)

EXAMPLE 2.3.41 *We consider Example 2.3.40 with the choice $k = 2$. This means that* (2.3.47)*, i.e., the monotonicity of* $\mathbf{H}$ *holds. (However, since* (2.3.59) *is not valid, the IAP method is not non-negativity preserving.) Moreover, the row sums of the matrix* $\mathbf{H}$ *are non-negative. Hence, the corresponding IAP satisfies the Ciarlet maximum principle, i.e., the DWBMP on* H^C_{DWBMP}*. However, for arbitrary element from* H_{DWBMP} *the DWBMP does not hold. To show this, let us choose*

$$\mathbf{u}_\partial^{(n-1)} = -2\mathbf{e}, \quad \mathbf{u}_0^{(n-1)} = \mathbf{u}_\partial^{(n)} = \mathbf{f}^{(n)} = \mathbf{0}. \tag{2.3.61}$$

Since the equation has the form

$$2\mathbf{u}_0^{(n)} = \mathbf{u}_0^{(n-1)} - \mathbf{u}_\partial^{(n-1)} + 2\mathbf{u}_\partial^{(n)} + \mathbf{f}^{(n)}, \tag{2.3.62}$$

we get $\mathbf{u}_0^{(n)} = \mathbf{e}$*. Hence, for this choice the DWBMP is not true, i.e., the implication "Ciarlet maximum principle* $\Rightarrow$ *DWBMP" is not valid.*

As we have seen, the Stoyan maximum principle of the CAP can be guaranteed by the conditions (2.3.47) and (2.3.48). On the other hand, the DSBMP of IAP follows from (2.3.59) and from the condition

$$(\mathbf{A} - \mathbf{B})\mathbf{e} = \mathbf{0}, \tag{2.3.63}$$

which follows from (2.3.15) in Theorem 2.3.17. Let us notice that on the subset H^S_{DSBMP} the DSBMP and DWBMP properties are equivalent, thus the conditions of DWBMP are sufficient for the DSBMP property on H^S_{DSBMP}. Therefore, the condition (2.3.63) can be relaxed by (2.3.48). However, as Example 2.3.41 shows, the implication "Stoyan maximum principle $\Rightarrow$ DSBMP" is not true.

In conclusion, in Table 2.3.1 we give the conditions of the applicability of the different qualitative properties on an initial first boundary value problem for a time dependent linear PDE. (We assume that the discretization preserves the qualitative properties of the continuous functions.) We use the following notations: $f(\mathbf{x}, t)$ is the source (forcing) function, $u_0(\mathbf{x})$ is the initial function and $u_\partial(\mathbf{x}, t)$ the boundary function. We may observe that the applicability of the Ciarlet and Stoyan maximum principles is rather restrictive: the first one can be applied only for a problem with sign-determined source function and to boundary conditions decreasing in time (e.g., time-independent). The Stoyan maximum principle gives some information about the maximum (minimum) only for a homogeneous equation with non-negative initial and time-independent, non-negative boundary conditions. If one of the above conditions does not hold, we must apply another principle.

2.3.5 Basic conditions for the finite difference and finite element approximations

As it was mentioned in Section 2.3.3, the discrete mesh operators are derived via the discretization of the partial differential operators. Therefore, $\mathcal{L}$ is usually some *approximation* to L, which means the following. Let $\mathbf{P}$ denote the projection operator from the space $\mathrm{dom}L$ to $\mathrm{dom}\,\mathcal{L}$ defined as follows. For $v \in \mathrm{dom}\,L$, $(\mathbf{P}v)(\mathbf{x}_i, t_n) = v(\mathbf{x}_i, t_n)$ $(i = 1, \ldots, \bar{N};\ n = 0, \ldots, M)$. (We note that $\mathrm{dom}\,\mathcal{L}$, and hence $\mathcal{L}$, depends on the choice

	f	u_0	u_∂
DNP	non-negative	non-negative	non-negative
NPCAP	non-negative	non-negative	non-negative and time-decreasing
DWMP	any	any	any
DSMP	any	any	any
DWBMP	non-positive	any	any
Ciarlet	non-positive	any	time-decreasing
DSBMP	non-positive	any	any
Stoyan	zero	non-negative	non-negative and time-independent

Table 2.3.1: Conditions for the given data of the continuous problem providing different qualitative properties.

of the mesh $\bar{\mathcal{Q}}_{t_M}$, i.e., on the discretization parameters h and Δt. Therefore, for refined meshes they denote a family of the operators.)

Definition 2.3.42 *We say that $\mathcal{L}$ locally approximates the operator L if for all functions $v \in \operatorname{dom} L$ and for all points $(\mathbf{x}^\star, t^\star) \in Q_T$ we have*

$$(Lv)(\mathbf{x}^\star, t^\star) - (\mathcal{L}(\mathbf{P}v))(\mathbf{x}^\star_h, t^\star_{\Delta t}) \to 0, \tag{2.3.64}$$

when $\mathbf{x}^\star_h \to \mathbf{x}^\star$ and $t^\star_{\Delta t} \to t^\star$ as $h, \Delta t \to 0$.

The expression on the left-hand side in (2.3.64) is called *local approximation error* and the rate of its convergence to zero defines the order of the approximation. This will be denoted by the symbol $\mathcal{O}(g(\Delta t, h))$, where g is some function (typically polynom) of Δt and h. [5]

Aiming at preserving the qualitative properties, we want to use Theorem 2.3.18. Therefore, first we should analyze validity of the conditions (2.3.12′) and (2.3.13′). In what follows, we consider the differential operator in the standard form

$$L \equiv \frac{\partial}{\partial t} - \sum_{m=1}^{d} \frac{\partial}{\partial x_m}(k_m(\mathbf{x}, t)\frac{\partial}{\partial x_m}) - \sum_{m=1}^{d} a_m(\mathbf{x}, t)\frac{\partial}{\partial x_m} - a_0(\mathbf{x}, t), \tag{2.3.65}$$

which will be discretized by two popular numerical techniques - the finite difference and finite element methods. Henceforward we assume that k_m are positive functions and a_0 is a non-positive function.

a. The finite difference discretization

In the following we approximate the operator L in (2.3.65) with sufficiently smooth coefficient functions on a rectangular mesh by the usual finite difference method according to Figure 2.3.2. (See e.g., [73, 116, 133]). The interior points of the mesh are denoted

[5] The "Big Oh notation" (it is also called as Landau notation, Bachmann-Landau notation, asymptotic notation) first appeared in the second volume of Bachmann's treatise on number theory [2], and Landau obtained this notation in Bachmann's book [85]. This symbol means the following. Let $\mathbf{g}_\tau$ be a vector function defined on an interval $\mathcal{I} \subset \mathbb{R}$, $\mathbf{g}_\tau : \mathcal{I} \to \mathbb{R}^n$, with τ being a scalar parameter. We write $\mathbf{g}_\tau(t) = \mathcal{O}(\tau^p)$ if there exists a constant $C_0 > 0$ such that for sufficiently small values of $|\tau|$ the inequality

$$\|\mathbf{g}_\tau(t)\| \le C_0|\tau|^p$$

holds uniformly with respect to $t \in \mathcal{I}$ and $\|\cdot\|$ is any vector norm on $\mathbb{R}^n$.

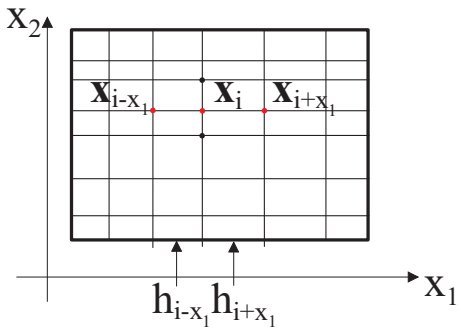

Figure 2.3.2: *A grid of a two-dimensional rectangular domain.*

by $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and the boundary ones by $\mathbf{x}_{N+1}, \ldots, \mathbf{x}_{\bar{N}}$. For the sake of simplicity, we also use the notation $\mathbf{x}_{i+x_m}$ ($\mathbf{x}_{i-x_m}$) for the grid point adjoint to $\mathbf{x}_i$ in positive (negative) x_m-direction. The lengths of the segments $[\mathbf{x}_i, \mathbf{x}_{i+x_m}]$ and $[\mathbf{x}_{i-x_m}, \mathbf{x}_i]$ are denoted by h_{i+x_m} and h_{i-x_m}, respectively. Furthermore, let us denote the uniform temporal discretization step size with $\Delta t > 0$, and we will use the notation $\bar{t}_n = t_n - 0.5\Delta t$. The integer number M is defined by the property $M\Delta t \le T < (M+1)\Delta t$.

Using the notation ν_i^n for the value of $v(\mathbf{x}_i, t_n)$, the finite difference approximation results in the discrete mesh operator $\mathcal{L}$ defined in the canonical form (2.3.17), where for the entries of $\mathbf{M}^{(n)}(= \mathbf{M})$ we have

$$M_{i,j}^{(n)} = M_{i,j} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j, \quad i = 1, \ldots, N;\ j = 1, \ldots, \bar{N}. \end{cases} \tag{2.3.66}$$

Applying the central difference approximation for the first order derivatives, the nonzero elements of the i-th row of $\mathbf{K}^{(n)}$ are $K_{i,i-x_m}^{(n)}, K_{i,i+x_m}^{(n)}$ $(m = 1, \ldots, d)$ and $K_{i,i}^{(n)}$, where

$$K_{i,i-x_m}^{(n)} = \frac{-2(k_m)_{i-0.5}^{(n)}}{h_{i-x_m}(h_{i-x_m} + h_{i+x_m})} + \frac{(a_m)_i^{(n)}}{h_{i-x_m} + h_{i+x_m}} \tag{2.3.67}$$

$$K_{i,i+x_m}^{(n)} = \frac{-2(k_m)_{i+0.5}^{(n)}}{h_{i+x_m}(h_{i-x_m} + h_{i+x_m})} - \frac{(a_m)_i^{(n)}}{h_{i-x_m} + h_{i+x_m}} \tag{2.3.68}$$

and

$$K_{i,i}^{(n)} = \sum_{m=1}^{d} \frac{2}{h_{i-x_m} + h_{i+x_m}} \left(\frac{(k_m)_{i+0.5}^{(n)}}{h_{i+x_m}} + \frac{(k_m)_{i-0.5}^{(n)}}{h_{i-x_m}} \right) - (a_0)_i^{(n)}, \tag{2.3.69}$$

where $(k_m)_{i\pm0.5}^{(n)} = 0.5(k_m(x_i, \bar{t}_n) + k_m(x_{i\pm1}, \bar{t}_n))$, $(a_m)_i^{(n)} = a_m(x_i, \bar{t}_n)$ and $(a_0)_i^{(n)} = a_0(x_i, \bar{t}_n)$.

Hence, for the finite difference discrete mesh operators $\mathcal{L}$, defined by (2.3.17) and (2.3.66)–(2.3.69), the relations

$$\mathbf{Me} = \mathbf{e}_0 \tag{2.3.70}$$

and

$$\mathbf{K}^{(n)}\mathbf{e} \begin{cases} \ge \mathbf{0}, & \text{if } a_0 \le 0; \\ = \mathbf{0}, & \text{if } a_0 = 0 \end{cases} \tag{2.3.71}$$

hold. Hence, we have

Theorem 2.3.43 *Let us assume that the finite difference discrete mesh operator $\mathcal{L}$, defined by* (2.3.17) *and* (2.3.66)-(2.3.69)*, is non-negativity preserving. Then, beyond the NP property, when $a_0 \le 0$, the operator $\mathcal{L}$ is DWMP, DWBMP and DMNC, too; while in the case $a_0 = 0$ it has each of the DWMP, DSMP, DWBMP, DSBMP and DMNC properties.*

Let us replace the central difference approximation with the upwind (upstream) approximation. In this case the matrix $\mathbf{M}$ does not change and the elements of $\mathbf{K}$ have the following form

$$K^{(n)}_{i,i-x_m} = \frac{-2(k_m)^{(n)}_{i-0.5}}{h_{i-x_m}(h_{i-x_m}+h_{i+x_m})} + \frac{(a_m)^{(n)}_i - |(a_m)^{(n)}_i|}{2h_{i-x_m}} \tag{2.3.72}$$

$$K^{(n)}_{i,i+x_m} = \frac{-2(k_m)^{(n)}_{i+0.5}}{h_{i+x_m}(h_{i-x_m}+h_{i+x_m})} - \frac{(a_m)^{(n)}_i + |(a_m)^{(n)}_i|}{2h_{i+x_m}} \tag{2.3.73}$$

and

$$\begin{aligned} K^{(n)}_{i,i} &= \sum_{m=1}^{d}\left[\frac{2}{h_{i-x_m}+h_{i+x_m}}\left(\frac{(k_m)^{(n)}_{i+0.5}}{h_{i+x_m}} + \frac{(k_m)^{(n)}_{i-0.5}}{h_{i-x_m}}\right) + \right. \\ &\left. + \frac{(a_m)^{(n)}_i + |(a_m)^{(n)}_i|}{2h_{i+x_m}} - \frac{(a_m)^{(n)}_i - |(a_m)^{(n)}_i|}{2h_{i-x_m}}\right] - (a_0)^{(n)}_i = \\ &= \sum_{m=1}^{d}\left[\frac{2}{h_{i-x_m}+h_{i+x_m}}\left(\frac{(k_m)^{(n)}_{i+0.5}}{h_{i+x_m}} + \frac{(k_m)^{(n)}_{i-0.5}}{h_{i-x_m}}\right) + \right. \\ &\left. + \frac{|(a_m)^{(n)}_i|}{h_{i+sign((a_m)^{(n)}_i)x_m}}\right] - (a_0)^{(n)}_i, \end{aligned} \tag{2.3.74}$$

respectively. One can directly check that for the finite difference discrete mesh operators $\mathcal{L}$, defined by (2.3.17) and (2.3.72)-(2.3.74), the relations (2.3.70) and (2.3.71) are satisfied and, hence, all the results of Theorem 2.3.43 remain valid.

b. The finite element discretization

We consider again the operator L (with homogenous first boundary condition) in (2.3.65) with sufficiently smooth coefficient functions. Then L can be written in the weak form as follows

$$\int_\Omega (Lv)(\mathbf{x},t)w(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_\Omega \frac{\partial v}{\partial t}(\mathbf{x},t)w(\mathbf{x})\,\mathrm{d}\mathbf{x} +$$

$$\int_\Omega \left[\sum_{m=1}^{d}\left(k_m(\mathbf{x},t)\frac{\partial v}{\partial x_m}(\mathbf{x},t)\frac{\partial w}{\partial x_m}(\mathbf{x}) - a_m(\mathbf{x},t)\frac{\partial v}{\partial x_m}(\mathbf{x},t)w(\mathbf{x})\right) - a_0(\mathbf{x},t)v(\mathbf{x},t)w(\mathbf{x})\right]\mathrm{d}\mathbf{x},$$

where $w(\mathbf{x}) \in H^1_0(\Omega)$ are the test functions and the solution $v(\mathbf{x},t)$ is assumed to be continuously differentiable w.r.t. t and belongs to $H^1(\Omega)$ for any fixed t.

In order to define a discrete finite element mesh operator, let $\phi_1,\dots,\phi_{\bar{N}}$ be finite element basis functions from $H^1(\Omega)$ with the property

$$\sum_{i=1}^{\bar{N}} \phi_i(\mathbf{x}) \equiv 1 \tag{2.3.75}$$

in $\bar{\Omega}$. Applying these functions to the space discretization and the θ-method to the time discretization, we arrive again at the discrete mesh operator in the canonical form (2.3.17). Now the matrices $\mathbf{M} \in \mathbb{R}^{N\times\bar{N}}$ and $\mathbf{K}^{(n)} \in \mathbb{R}^{N\times\bar{N}}$, respectively, have the elements

$$M_{i,j} = (M_\star)_{i,j}\frac{1}{\int_\Omega \phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}}, \qquad K^{(n)}_{i,j} = (K^{(n)}_\star)_{i,j}\frac{1}{\int_\Omega \phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}}, \tag{2.3.76}$$

where $\mathbf{M}_\star$ and $\mathbf{K}^{(n)}_\star$ are, respectively, the so-called mass and stiffness matrices with the entries

$$(M_\star)_{i,j} = \int_\Omega \phi_j(\mathbf{x})\phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}, \tag{2.3.77}$$

$$\begin{aligned}(K^{(n)}_\star)_{i,j} = \int_\Omega \left(\sum_{m=1}^{d} k_m(\mathbf{x},t_n)\frac{\partial\phi_j}{\partial x_m}(\mathbf{x})\frac{\partial\phi_i}{\partial x_m}(\mathbf{x})\right)\mathrm{d}\mathbf{x}-\\ -\int_\Omega \left(\sum_{m=1}^{d} a_m(\mathbf{x},t_n)\frac{\partial\phi_j}{\partial x_m}(\mathbf{x})\phi_i(\mathbf{x}) + a_0(\mathbf{x},t_n)\phi_j(\mathbf{x})\phi_i(\mathbf{x})\right)\mathrm{d}\mathbf{x}.\end{aligned} \tag{2.3.78}$$

Therefore, we can use Theorem 2.3.18.

For the row-sums of the matrix $\mathbf{M}$, by using the relation (2.3.75), we get:

$$\begin{aligned}(\mathbf{M}\mathbf{e})_i = \sum_{j=1}^{\bar{N}} M_{i,j} = \frac{1}{\int_\Omega \phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}}\sum_{j=1}^{\bar{N}}\left(\int_\Omega \phi_j(\mathbf{x})\phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) =\\ = \frac{1}{\int_\Omega \phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}}\Big(\int_\Omega \Big(\overbrace{\sum_{j=1}^{\bar{N}}\phi_j(\mathbf{x})}^{=1}\Big)\phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}\Big) = 1\end{aligned} \tag{2.3.79}$$

for all $i = 1,\ldots,N$. Hence, (2.3.18) is satisfied.

For the row-sums of the matrix $\mathbf{K}^{(n)}$, we get

$$(\mathbf{K}^{(n)}\mathbf{e})_i = \sum_{j=1}^{\bar{N}} K^{(n)}_{i,j} = \frac{1}{\int_\Omega \phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}}\sum_{j=1}^{\bar{N}}(K^{(n)}_\star)_{i,j} =$$

$$= \frac{1}{\int_\Omega \phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}}\sum_{j=1}^{\bar{N}}\int_\Omega\sum_{m=1}^{d} k_m(\mathbf{x},t_n)\frac{\partial\phi_j}{\partial x_m}(\mathbf{x})\frac{\partial\phi_i}{\partial x_m}(\mathbf{x})\,\mathrm{d}\mathbf{x}-$$

$$-\frac{1}{\int_\Omega \phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}}\sum_{j=1}^{\bar{N}}\int_\Omega\left(\sum_{m=1}^{d} a_m(\mathbf{x},t_n)\frac{\partial\phi_j}{\partial x_m}(\mathbf{x})\phi_i(\mathbf{x}) + a_0\phi_j(\mathbf{x})\phi_i(\mathbf{x})\right)\mathrm{d}\mathbf{x} =$$

$$= \frac{1}{\int_\Omega \phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}}\int_\Omega\Big(\sum_{m=1}^{d} k_m(\mathbf{x},t_n)\frac{\partial\overbrace{\big(\sum_{j=1}^{\bar{N}}\phi_j\big)}^{=1}}{\partial x_m}(\mathbf{x})\frac{\partial\phi_i}{\partial x_m}(\mathbf{x})-$$

$$-\frac{1}{\int_\Omega \phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}}\int_\Omega \Big(\sum_{m=1}^{d} a_m(\mathbf{x},t_n)\frac{\partial \overbrace{(\sum_{j=1}^{\bar N}\phi_j)}^{=1}}{\partial x_m}(\mathbf{x})\phi_i(\mathbf{x}) + a_0(\mathbf{x},t_n)\overbrace{(\sum_{j=1}^{\bar N}\phi_j)}^{=1}\phi_i(\mathbf{x})\Big)\,\mathrm{d}\mathbf{x} =$$

$$= -\frac{1}{\int_\Omega \phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}}\int_\Omega a_0(\mathbf{x},t_n)\phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

for all $i = 1,\dots,N$. Therefore, when $a_0 = 0$ then $\mathbf{K}^{(n)}\mathbf{e} = \mathbf{0}$. When a_0 is non-positive and it is independent of $\mathbf{x}$, then $\mathbf{K}^{(n)}\mathbf{e} \geq \mathbf{0}$. If a_0 is non-positive and it depends on $\mathbf{x}$, then additionally we assume that the finite element basis functions are non-negative, i.e., the condition

$$\phi_i(\mathbf{x}) \geq 0 \tag{2.3.80}$$

is satisfied. Then $\int_\Omega -a_0(\mathbf{x},t_n)\phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x} \geq \inf(-a_0)\int_\Omega \phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}$. Hence, for this case $\mathbf{K}^{(n)}\mathbf{e} \geq \mathbf{0}$. We can summarize our results as follows.

Theorem 2.3.44 *Let us assume that the finite element discrete mesh operator $\mathcal{L}$, defined by* (2.3.17) *and* (2.3.76)-(2.3.78) *for arbitrary finite element basis functions, is non-negativity preserving. For $a_0 = 0$ it has each of the DNP, DWMP, DSMP, DWBMP, DSBMP and DMNC properties. When a_0 is a non-positive, independent of $\mathbf{x}$, function, or, when it varies in $\mathbf{x}$ and non-positive and for the basis functions the condition* (2.3.80) *is satisfied, then $\mathcal{L}$ has the DNP, DWMP, DWBMP and DMNC properties.*

In the sequel we deal with the problem of how the non-negativity of the discrete mesh operator can be guaranteed for the above cases.

2.3.6 The non-negativity preservation of the discrete heat conduction mesh operator in 1D case

We start with the investigation of the one-dimensional heat conduction operator with a constant coefficient, which is assumed, for simplicity, to be equal to one, i.e.,

$$L \equiv \frac{\partial}{\partial t} - \frac{\partial^2}{\partial x^2}. \tag{2.3.81}$$

On a fixed uniform mesh we define a one-step discrete mesh operator $\mathcal{L}$ in the form (2.3.10) with $N_\partial = 2$, $\bar N = N + 2$ and matrices

- for the finite difference method

$$\mathbf{X}_{10} = \frac{1}{\Delta t}\mathbf{I}_0 + \theta\mathbf{K}_0 = \mathrm{tridiag}\left[-\frac{\theta}{h^2}, \frac{1}{\Delta t} + 2\frac{\theta}{h^2}, -\frac{\theta}{h^2}\right] \in \mathbb{R}^{N\times N},$$

$$\mathbf{X}_{20} = \frac{1}{\Delta t}\mathbf{I}_0 - (1-\theta)\mathbf{K}_0 = \mathrm{tridiag}\left[\frac{1-\theta}{h^2}, \frac{1}{\Delta t} - 2\frac{1-\theta}{h^2}, \frac{1-\theta}{h^2}\right] \in \mathbb{R}^{N\times N},$$

$$\mathbf{X}_{1\partial} = -\frac{\theta}{h^2}\mathbf{E};\quad \mathbf{X}_{2\partial} = \frac{1-\theta}{h^2}\mathbf{E},\quad \text{where}\quad \mathbf{E} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}^T \in \mathbb{R}^{N\times 2}. \tag{2.3.82}$$

- for the linear finite element method

$$\mathbf{X}_{10} = \frac{1}{\Delta t}\mathbf{I}_0 + \theta\mathbf{K}_0 = \text{tridiag}\left[\frac{1}{6\Delta t} - \frac{\theta}{h^2}, \frac{2}{3\Delta t} + 2\frac{\theta}{h^2}, \frac{1}{6\Delta t} - \frac{\theta}{h^2}\right]$$

$$\mathbf{X}_{20} = \frac{1}{\Delta t}\mathbf{I}_0 - (1-\theta)\mathbf{K}_0 = \text{tridiag}\left[\frac{1}{6\Delta t} + \frac{1-\theta}{h^2}, \frac{2}{3\Delta t} - 2\frac{1-\theta}{h^2}, \frac{1}{6\Delta t} + \frac{1-\theta}{h^2}\right]$$

$$\mathbf{X}_{1\partial} = \left(\frac{1}{6\Delta t} - \frac{\theta}{h^2}\right)\mathbf{E}; \quad \mathbf{X}_{2\partial} = \left(\frac{1}{6\Delta t} + \frac{1-\theta}{h^2}\right)\mathbf{E}. \tag{2.3.83}$$

We have to guarantee the conditions (P1)-(P3) in Theorem 2.3.15.

Lemma 2.3.45 *For the finite difference scheme* (2.3.82) *the conditions (P1)-(P3) in Theorem 2.3.15 are satisfied if and only if*

$$\mathbf{X}_{10}^{-1} \geq \mathbf{0} \text{ and } \mathbf{X}_{pr} := \mathbf{X}_{10}^{-1}\mathbf{X}_{20} \geq \mathbf{0}. \tag{2.3.84}$$

Under the additional assumption

$$\frac{\Delta t}{h^2} \geq \frac{1}{6\theta}, \tag{2.3.85}$$

the assumptions (2.3.84) *are also necessary and sufficient for the validity of (P1)-(P3) for the finite element scheme* (2.3.83).

The proof follows directly from the non-negativity of the matrices $-\mathbf{X}_{1\partial}$ and $\mathbf{X}_{2\partial}$.

Remark 2.3.46 *Lemma 2.3.45 yields that $\mathcal{L}$ is non-negativity preserving if and only if the matrices $\mathbf{X}_{10}$ and $\mathbf{X}_{20}$ form a weak regular splitting for the matrix $\mathbf{X}_{10} - \mathbf{X}_{20} = \mathbf{K}_0$.*

Let us notice that the matrices in (2.3.82) and (2.3.83) have special structure: only the entries of the main-, super- and sub-diagonals differ from zero and the elements standing on the same diagonal are equal. Moreover, these matrices are symmetric, too. This kind of matrices, i.e., a matrix of the form tridiag$[a, b, a]$ is called *uniformly continuant, symmetrical tridiagonal matrix*, [114]. (Clearly, these matrices are symmetric, tridiagonal Toeplitz matrices.) These matrices have some special qualitative propertiess, which will be considered in the sequel.

a. Non-negativity of uniformly continuant, symmetrical tridiagonal matrices

Hereafter we investigate the conditions (2.3.84) for special matrices, namely we consider the real, uniformly continuant, symmetrical tridiagonal matrices

$$\mathbf{X}_{10} = \text{tridiag}[-z, 2\tilde{w}, -z]; \quad \mathbf{X}_{20} = \text{tridiag}[s, \tilde{p}, s]. \tag{2.3.86}$$

We may assume that $z \geq 0$ and $s \geq 0$.

The simplest way to satisfy the conditions (2.3.84) is to guarantee the monotonicity of $\mathbf{X}_{10}$ and the non-negativity of $\mathbf{X}_{20}$. (I.e., $\mathbf{X}_{10}$ and $\mathbf{X}_{20}$ form a regular splitting for the matrix $\mathbf{K}_0$.) However, as it is known, in several cases the condition of the non-negativity of $\mathbf{X}_{20}$ can be relaxed. As it was showed in [91] and [131], the condition $\mathbf{X}_{pr} \geq \mathbf{0}$ in (2.3.84) is valid for certain negative $\tilde{p}$'s, too. Our aim is to give the exact bounds.

When $z = 0$ then $\mathbf{X}_{10} = 2\tilde{w}\ \mathbf{I}_0$. Hence, for this case the exact conditions are

$$\tilde{w} > 0 \ \text{ and } \ \mathbf{X}_{20} \geq \mathbf{0}. \tag{2.3.87}$$

When $s = 0$, then $\mathbf{X}_{20} = \tilde{p}\ \mathbf{I}_0$. Hence, for this case the exact conditions are

$$\tilde{p} \geq 0 \ \text{ and } \ \mathbf{X}_{10} \ \text{ is a monotone matrix.} \tag{2.3.88}$$

We pass to the case when $z > 0$ and $s > 0$. Then, we can consider the equivalent form of the matrices,

$$\mathbf{X}_{10} = z \cdot \text{tridiag}[-1, 2w, -1]; \ \ \mathbf{X}_{20} = s \cdot \text{tridiag}[1, p, 1], \tag{2.3.89}$$

where $w = \tilde{w}/z$ and $p = \tilde{p}/s$. First we introduce the following one-pair matrix $\mathbf{G} = (G_{ij})$,[6] depending on the parameter w:

$$G_{i,j} = \begin{cases} \gamma_{i,j}, & \text{if } \ i \leq j \\ \gamma_{j,i}, & \text{if } \ j \leq i \end{cases} \tag{2.3.90}$$

$(i, j = 1, 2, \ldots, N)$, where

$$\gamma_{i,j} = \begin{cases} \dfrac{\text{sh}(i\vartheta)\text{sh}(N+1-j)\vartheta}{\text{sh}\vartheta\text{sh}(N+1)\vartheta}, & \vartheta = \text{arch}(w), \qquad \text{if } \ w > 1; \\ \dfrac{i(N+1-j)}{N+1}, & \qquad\qquad\quad \text{if } \ w = 1; \\ \dfrac{\sin(i\vartheta)\sin(N+1-j)\vartheta}{\sin\vartheta\sin(N+1)\vartheta}, & \vartheta = \arccos(w), \qquad \text{if } \ |w| < 1; \\ (-1)^{i+j-1}\dfrac{i(N+1-j)}{N+1}, & \qquad\qquad\quad \text{if } \ w = -1; \\ (-1)^{i+j-1}\dfrac{\text{sh}(i\vartheta)\text{sh}(N+1-j)\vartheta}{\text{sh}\vartheta\text{sh}(N+1)\vartheta}, & \vartheta = \text{arch}(w), \qquad \text{if } \ w < -1. \end{cases} \tag{2.3.91}$$

In case $z \neq 0$ we have the relation $\mathbf{X}_{10}^{-1} = (1/z)\mathbf{G}$ (see [114]), thus a direct computation verifies the validity of the following

Lemma 2.3.47 *For the matrices* $\mathbf{X}_{10}$ *and* $\mathbf{X}_{20}$ *of the form* (2.3.89) *the matrix* $\mathbf{X}_{pr} = \mathbf{X}_{10}^{-1}\mathbf{X}_{20}$ *can be expressed as*

$$\mathbf{X}_{pr} = \frac{s}{z}\left[(2w + p)\mathbf{G} - \mathbf{I}_0\right]. \tag{2.3.92}$$

Hence, one of the conditions $\text{offdiag}(\mathbf{G}) \geq \mathbf{0}$ and $\text{offdiag}(\mathbf{G}) \leq \mathbf{0}$ is necessary[7]. Using the formula (2.3.91) for the elements of the matrix $\mathbf{G}$, this implies the following

[6] A matrix is called one-pair matrix if its (i, j)-th elements have the representation $s_i \cdot t_j$, for $i \leq j$, and $s_j \cdot t_i$, for $j \leq i$. See [58], [114].

[7] The symbol $\text{offdiag}(\mathbf{G})$ denotes the matrix with zeros in the main diagonal and with the off-diagonal elements of the matrix $\mathbf{G}$, i.e., $\text{offdiag}(\mathbf{G}) = \mathbf{G} - \text{diag}(\mathbf{G})$.

Lemma 2.3.48 *The condition* $\mathbf{X}_{pr} \geq \mathbf{0}$ *may be satisfied only if the relation*

$$\begin{array}{l} w \geq 1, \quad \text{or} \\ |w| < 1 \quad \text{and } \arccos(w) < (\pi/N) \end{array} \tag{2.3.93}$$

holds.

Since the condition, imposed for $|w| < 1$, can be guaranteed only for small values N, and in our applications N becomes arbitrarily large, this case is negligible. Hence, if we intend to get the result for arbitrary large dimension, then it can be achieved for $w \geq 1$, only. That is, the matrix $\mathbf{X}_{pr}$ may be non-negative for arbitrary dimension only in the case $\mathbf{X}_{10}$ is a positive definite M-matrix.

We can summarize our results as

Lemma 2.3.49 *Let us suppose that* $w \geq 1$. *Then,* $\mathbf{X}_{pr} \in \mathbb{R}^{N\times N}$ *is non-negative for an arbitrary fixed* N *if and only if the conditions*

$$2w + p > 0 \tag{2.3.94}$$

and

$$\gamma_{i,i} \geq \frac{1}{2w+p}, \;\; i = 1, 2, \ldots, N \tag{2.3.95}$$

are fulfilled.

We note, that, for the case $ws + p = 0$, the non-negativity of $\mathbf{X}$ implies the relation $\mathbf{X_2} = \mathbf{0}$, which is out of our interest.

Now we analyze the expression on the left-hand side in condition (2.3.95).

Lemma 2.3.50 *When* $w \geq 1$, *then for the diagonal elements of the matrix* $\mathbf{X}_{pr}$ *the relation*

$$\min\{\gamma_{i,i}, \;\; i = 1, 2, \ldots, N\} = \gamma_{1,1} = \gamma_{N,N} \tag{2.3.96}$$

holds.

Proof. Introducing the functions $h_1(y) = K_1 \mathrm{sh}(Cy)\ \mathrm{sh}(C(N+1-y))$ and $h_2(y) = K_2 y(N+1-y)$ on the interval $[1, N]$, (where K_1 , K_2 and C are some positive constants), one can check that both functions take their maxima at the same point $y = (N+1)/2$. Moreover, on the interval $[1, (N+1)/2)$ they are monotonically increasing, while on the interval $((N+1)/2, N]$ they are monotonically decreasing. Using this fact and the expressions for $\gamma_{i,i}$, we get the statement. ■

Combining Lemma 2.3.49 and Lemma 2.3.50, we obtain

Theorem 2.3.51 *Assume that* $z > 0$, $s > 0$ *and* $w > 1$. *Then,* $\mathbf{X}_{pr} \in \mathbb{R}^{N\times N}$ *is non-negative for arbitrary fixed* N *if and only if the conditions* (2.3.94) *and*

$$a(N) := \frac{\mathrm{sh}(N\vartheta)}{\mathrm{sh}((N+1)\vartheta)} \geq \frac{1}{2w+p} \tag{2.3.97}$$

are satisfied.

Obviously, (2.3.94) and (2.3.97) are necessary and sufficient conditions of the non-negativity for some fixed dimension N. Let us turn to the examination of a varying N. Due to the relations

$$\frac{\mathrm{sh}(N\vartheta)}{\mathrm{sh}((N+1)\vartheta)} = \mathrm{ch}(\vartheta) - \mathrm{coth}((N+1)\vartheta)\mathrm{sh}(\vartheta), \tag{2.3.98}$$

we have

$$\sup\left\{\frac{\mathrm{sh}(N\vartheta)}{\mathrm{sh}((N+1)\vartheta)};\ N \in \mathbb{N}\right\} = \mathrm{ch}(\vartheta) - \mathrm{sh}(\vartheta) = \exp(-\vartheta). \tag{2.3.99}$$

Since the sequence $a(N)$ is monotonically increasing, it converges to its limit (which is its supremum) monotonically. Thus, the conditions (2.3.94) and (2.3.97), that is, the necessary and sufficient conditions for some fixed N, serve as sufficient condition for the non-negativity of the matrices $\mathbf{X}_{pr} \in \mathbb{R}^{N_1 \times N_1}$ for all $N_1 \geq N$.

Let us observe that

$$\begin{aligned}\exp(-\vartheta) = \exp(-\mathrm{arch}(w)) &= \exp\left(\ln\left[w + \sqrt{w^2-1}\right]^{-1}\right) \\ &= \left[w + \sqrt{w^2-1}\right]^{-1}.\end{aligned} \tag{2.3.100}$$

Therefore, from some sufficiently large $N_0 \in \mathbb{N}$ the relation $\mathbf{X}_{pr} \geq \mathbf{0}$ may be true only if the condition

$$\left[w + \sqrt{w^2-1}\right]^{-1} > \frac{1}{2w+p}, \tag{2.3.101}$$

i.e., the condition

$$p > -w + \sqrt{w^2-1} \tag{2.3.102}$$

is fulfilled. This proves the following

Theorem 2.3.52 *Assume that $z > 0$, $s > 0$ and $w > 1$. If, for some number $N_0 \in \mathbb{N}$, the conditions* (2.3.94) *and* (2.3.97) *are satisfied, then, all matrices $\mathbf{X}_{pr} \in \mathbb{R}^{N\times N}$ with $N \geq N_0$, are non-negative. Moreover, there exists such a number N_0, if and only if the condition* (2.3.94) (2.3.102) *holds.*

Remark 2.3.53 *Since*

$$a(1) = \frac{\mathrm{sh}\vartheta}{\mathrm{sh}(2\vartheta)} = \frac{1}{2\mathrm{ch}\vartheta} = \frac{1}{2w},$$

therefore, (2.3.97) *results in the condition*

$$p \geq 0. \tag{2.3.103}$$

Hence, the matrix $\mathbf{X}_{pr} \in \mathbb{R}^{N\times N}$ is non-negative for all $N = 1, 2, \ldots,$ if and only if $\mathbf{X}_{10}$ is an M-matrix and $\mathbf{X}_{10} - \mathbf{X}_{20}$ is a regular splitting of the matrix $\mathbf{K}_0$.

Remark 2.3.54 *Due to the relation*

$$a(2) = \frac{\mathrm{sh}(2\vartheta)}{\mathrm{sh}(3\vartheta)} = \frac{2\mathrm{ch}(\vartheta)}{4\mathrm{ch}^2(\vartheta) - 1} = \frac{2w}{4w^2-1},$$

condition (2.3.97) *results in the assumption*

$$p \geq -\frac{1}{2w}. \tag{2.3.104}$$

That is, $\mathbf{X}_{pr} \in \mathbb{R}^{N\times N}$ is non-negative for all $N = 2, 3, \ldots,$ if and only if $\mathbf{X}_{10}$ is an M-matrix and (2.3.104) *is valid.*

Remark 2.3.55 *The conditions* (2.3.103) *and* (2.3.104) *(corresponding to the cases* $N = 1$ *and* $N = 2$*, respectively) are sufficient conditions for the non-negativity of the matrix* $\mathbf{X}_{pr}$ *in any larger dimension. For increasing* N*, the new obtained conditions are approaching the necessary condition of the non-negativity. Using* (2.3.98) *and* (2.3.99) *we can characterize the rate of the convergence: it is equal to the rate of convergence of the sequence* $\{\coth(N\vartheta),\ N = 1, 2, \dots\}$ *to one. Clearly,*

$$\coth(N\vartheta) = 1 + \frac{2}{[\exp(\vartheta)]^{2N} - 1}.$$

Therefore, introducing the notation in (2.3.100) *as*

$$\exp(\vartheta) = w + \sqrt{w^2 - 1} =: \beta, \tag{2.3.105}$$

we get that the sequence of the bounds of the sufficient conditions converges linearly with the ratio $1/\beta^2$ *to the bound of the necessary condition.*

Remark 2.3.56 *The consideration of the case* $w = 1$ *is obvious. As an easy computation shows, in this case* $a(N) = N/(N + 1)$ *and, hence,* $\mathbf{X}_{pr} \geq \mathbf{0}$ *for all* $N = 1, 2, \dots$ *if and only if* $p \geq 0$*. (I.e., Remark 2.3.53 remains true for this case.) The above relation holds for all* $N = 2, 3, \dots$*, if and only if* $p \geq -1/2$*. The necessary condition of the existence of some dimension for the non-negativity is* $p > -1$*.*

b. Non-negativity of difference schemes

The results of the previous part can be used in the qualitative analysis of the finite difference and linear finite element mesh operators in 1D, given by the formula (2.3.82) and (2.3.83), respectively. We will use the notation $q = \Delta t/h^2$.

First we investigate the finite difference mesh operator. According to (2.3.82), the corresponding matrices are uniformly continuant, they can be written in the form (2.3.86) with the choice

$$z = \frac{\theta}{h^2}, \quad s = \frac{1-\theta}{h^2} \quad \tilde{w} = \frac{1}{2\Delta t} + \frac{\theta}{h^2}, \quad \tilde{p} = \frac{1}{\Delta t} - 2\frac{1-\theta}{h^2}. \tag{2.3.106}$$

Therefore, in case $\theta = 0$, due to (2.3.87), the condition is $\tilde{p} \geq 0$, which results in the bound

$$q \leq \frac{1}{2}. \tag{2.3.107}$$

For the case $\theta = 1$ the condition (2.3.88) must be satisfied. Since in our case $\tilde{w} > z$, therefore $\mathbf{X}_{10}$ is a diagonally dominant Z-matrix,[8] and hence it is an M-matrix. This yields that for this case we do not have any condition for the choice of the parameters h and Δt.

In what follows, we pass to the analysis of the case when $z > 0$ and $s > 0$. Then we can use the form (2.3.89) with the choice

$$z = \frac{\theta q}{\Delta t}, \quad s = \frac{(1-\theta)q}{\Delta t} \quad w = \frac{1 + 2\theta q}{2\theta q}, \quad p = \frac{1 - 2(1-\theta)q}{(1-\theta)q}. \tag{2.3.108}$$

[8] A matrix is called Z-matrix if its off-diagonal elements are non-positive [9], [56]. (Cf. footnote on p.29).

The positivity of z and s means that $\theta \in (0,1)$. Let us notice that under the choice (2.3.108) $2w + p = 1/\theta(1-\theta)q$, hence the condition (2.3.94) is always satisfied.

Using (2.3.103), we directly get that the condition of the non-negativity preservation for all $N = 1, 2, \ldots$ is the condition

$$q \leq \frac{1}{2(1-\theta)}. \tag{2.3.109}$$

However, the non-negativity preservation for all $N = 2, 3, \ldots$ should be guaranteed by the weaker condition (2.3.104), which, in our case, yields the inequality

$$\frac{1-2(1-\theta)q}{(1-\theta)q} \geq -\frac{\theta q}{1+2\theta q}. \tag{2.3.110}$$

Solving this problem, we get the upper bound

$$q \leq \frac{-1+2\theta+\sqrt{1-\theta(1-\theta)}}{3\theta(1-\theta)}, \tag{2.3.111}$$

which is larger than the bound in (2.3.107).

Our aim is to get the largest value for q under which the non-negativity preservation for sufficiently large values N still holds. Therefore we put the values w and p from (2.3.108) into the necessary condition (2.3.102). Then we should solve the inequality

$$\frac{1-2(1-\theta)q}{(1-\theta)q} \geq -\frac{1+2\theta q}{2\theta q} + \frac{\sqrt{1+4\theta q}}{2\theta q}. \tag{2.3.112}$$

The solution of (2.3.112) results in the bound

$$q \leq \frac{1-\sqrt{1-\theta}}{\theta(1-\theta)}. \tag{2.3.113}$$

We can summarize our results as follows.

Theorem 2.3.57 *The finite difference discrete mesh operator $\mathcal{L}$, which is defined by* (2.3.82)*, is non-negativity preserving for each $N \geq 1$ if and only if the condition* (2.3.109) *holds. It is non-negativity preserving for each $N \geq 2$ only under the condition* (2.3.111)*. There exists a number $N_0 \in \mathbb{N}$ such that $\mathcal{L}$ is non-negativity preservation for each $N \geq N_0$ if and only if the weaker condition* (2.3.113) *is satisfied.*

Example 2.3.58 *We demonstrate our results on some special choices of θ. Namely, we define upper bounds for*

- *explicit Euler method ($\theta = 0$);*
- *fourth order method $\theta = 1/2 - 1/(12q)$, $q > 1/6$;*
- *Crank-Nicolson method ($\theta = 0.5$);*
- *implicit Euler method ($\theta = 1$).*

θ	$N=1$	$N=2$	$N=\infty$
0	0.5	0.5	0.5
$0.5-(12q)^{-1}$	0.8333	0.9574	0.9661
0.5	1	$2\sqrt{3}/3$	$2(2-\sqrt{2})$
1	∞	∞	∞

Table 2.3.2: Non-negativity providing upper bounds for q in the different finite difference mesh operators.

The results are shown in Table 2.3.2.

We pass to the investigation of the linear finite element mesh operator. According to (2.3.83), the corresponding matrices are also symmetrical, uniformly continuant, tridiagonal and they can be written in the form (2.3.86) with the choice

$$z=\frac{1}{6\Delta t}-\frac{\theta}{h^2},\quad s=\frac{1}{6\Delta t}+\frac{1-\theta}{h^2},\quad \tilde{w}=\frac{1}{3\Delta t}+\frac{\theta}{h^2},\quad \tilde{p}=\frac{2}{3\Delta t}-2\frac{1-\theta}{h^2}. \tag{2.3.114}$$

First we consider the special choices $\theta=0$ and $\theta=1$.
For $\theta=0$ we get $\mathbf{X}_{10}=(1/6\Delta t)\text{tridiag}[1,4,1]$, i.e., $\mathbf{X}_{10}\geq\mathbf{0}$. Therefore we cannot guarantee the monotonicity of $\mathbf{X}_{10}$. When $\theta=1$, then $\mathbf{X}_{20}=(1/6\Delta t)\text{tridiag}[1,4,1]$, hence, the monotonicity of $\mathbf{X}_{10}$ is the necessary and sufficient condition of the non-negativity preservation of the mesh operator $\mathcal{L}$. For this $q\geq 1/6$ is the condition.
In the sequel we consider the case $\theta\in(0,1)$.
When $q=1/(6\theta)$, then $\mathbf{X}_{10}=(1/\Delta t)\mathbf{I}_0$, hence the only condition of the non-negativity preservation is $\mathbf{X}_{20}\geq\mathbf{0}$. This can be guaranteed only by the condition $q\leq(3(1-\theta))^{-1}$.
When $q=(3(1-\theta))^{-1}$, then $\mathbf{X}_{20}=(1/6\Delta t)\text{tridiag}[1,4,1]$, hence the only condition is the monotonicity of $\mathbf{X}_{10}$. As we can see, for this case this matrix is an M-matrix, therefore, there is no additional condition for the non-negativity preservation.

In what follows we assume

$$\frac{1}{6\theta}<q<\frac{1}{3(1-\theta)}, \tag{2.3.115}$$

i.e., $\theta\in(1/3,1)$. Then we can use the form (2.3.89) with the choice

$$\begin{gathered}z=\frac{1}{6\Delta t}-\frac{\theta}{h^2},\quad s=\frac{1}{6\Delta t}+\frac{1-\theta}{h^2},\quad w=\frac{\frac{1}{3\Delta t}+\frac{\theta}{h^2}}{\frac{1}{6\Delta t}-\frac{\theta}{h^2}}=\frac{\frac{1}{3}+\theta q}{\theta q-\frac{1}{6}},\\ p=\frac{\frac{2}{3\Delta t}-2\frac{1-\theta}{h^2}}{\frac{1}{6\Delta t}+\frac{1-\theta}{h^2}}=\frac{\frac{2}{3}-2(1-\theta)q}{(1-\theta)q+\frac{1}{6}}.\end{gathered} \tag{2.3.116}$$

For this choice $2w+p=[(\theta q-1/6)((1-\theta)q+1/6)]^{-1}>0$, therefore (2.3.94) is always satisfied. Let us notice that under the condition (2.3.115) the condition $z>0$ is also satisfied.

The condition of the non-negativity preservation for all $N=1,2,\ldots$ is (2.3.103). Therefore, using (2.3.116), we obtain the upper bound

$$q\leq\frac{1}{3(1-\theta)}. \tag{2.3.117}$$

θ	$N=1$	$N=2$	$N=\infty$
0	not allowed	not allowed	not allowed
0.5	$1/3 \le q \le 2/3$	$1/3 \le q \le \sqrt{5}/3$	$1/3 \le q \le 0.748$
1	$1/6 \le q$	$1/6 \le q$	$1/6 \le q$

Table 2.3.3: Non-negativity providing upper and lower bounds for q in the different finite element mesh operators.

The non-negativity preservation for all $N = 2, 3, \ldots$ should be guaranteed by the weaker condition (2.3.104), which, in our case, yields the upper bound

$$q \le \frac{3(-1+2\theta) + \sqrt{9 - 16\theta(1-\theta)}}{12\theta(1-\theta)}, \tag{2.3.118}$$

which is larger than the bound in (2.3.117).

Our aim is to get the largest value for q under which the non-negativity preservation for sufficiently large values N is still valid. Therefore we put the values w and p from (2.3.116) into the necessary condition (2.3.102). Hence, we obtain that for any fixed $\theta \in (0,1)$ the suitable q are the positive solutions of the inequality

$$\begin{gathered} \theta(1-\theta)q^2 - 1/6(\theta+4)q + A \le 0; \\ A = \sqrt{q\theta + 1/12}[1/6 + (1-\theta)q]. \end{gathered} \tag{2.3.119}$$

We can summarize our results as follows.

Theorem 2.3.59 *The linear finite element discrete mesh operator $\mathcal{L}$, which is defined by* (2.3.83)*, is non-negativity preserving for any $\theta \in [0,1]$*

- *for each $N \ge 1$ if and only if the condition*

$$\frac{1}{6\theta} \le q \le \frac{1}{3(1-\theta)}; \tag{2.3.120}$$

- *for each $N \ge 2$ if and only if the condition*

$$\frac{1}{6\theta} \le q \le \frac{3(-1+2\theta) + \sqrt{9 - 16\theta(1-\theta)}}{12\theta(1-\theta)} \tag{2.3.121}$$

holds. There exists a number $N_0 \in \mathbb{N}$ such that $\mathcal{L}$ is non-negativity preservation for each $N \ge N_0$ if and only if the condition (2.3.119) *is satisfied.*

(In the bounds (2.3.120) and (2.3.121) we mean $1/0 := \infty$.) The bound (2.3.120) shows that the non-negativity preservation property can be guaranteed only for the values $\theta \in [1/3, 1]$.

EXAMPLE 2.3.60 *We demonstrate our results again on some special choice of θ. The results are shown in Table 2.3.3.*

Finally we note that the above results can be successfully applied to the qualitative analysis of several iterative methods for solving a system of linear algebraic equations with special structure (with tridiagonal and block tridiagonal Stieltjes-Toeplitz matrices) [53, 54]. The investigation is based on the matrix splitting method, and for a particular case it is proven that only those SOR methods are qualitatively good that are based on regular splittings.

2.3.7 Non-negativity preservation for more general discrete mesh operators

In this section we analyze the non-negativity preservation of discrete mesh operators obtained by the discretization of the partial differential operators in a general form. We should guarantee the validity of conditions (P1)-(P3) in Theorem 2.3.15. The following statement gives a sufficient condition for this in terms of the matrices $\mathbf{M}^{(n)}$ and $\mathbf{K}^{(n)}$.

Theorem 2.3.61 *Under the following conditions*

$$\begin{array}{ll} (P1') & K_{ij}^{(n)} \le 0, \quad i \ne j,\ i = 1, ..., N,\ j = 1, ..., \bar{N}, \\ (P2') & (\mathbf{X}_1^{(n)})_{ij} = M_{ij}^{(n)} + \theta \Delta t\ K_{ij}^{(n)} \le 0, \quad i \ne j,\ i = 1, ..., N,\ j = 1, ..., \bar{N}, \\ (P3') & (\mathbf{X}_2^{(n)})_{ii} = M_{ii}^{(n)} - (1-\theta)\Delta t\ K_{ii}^{(n)} \ge 0, \quad i = 1, ..., N, \end{array} \tag{2.3.122}$$

the non-negativity assumptions (P1)-(P3) in Theorem 2.3.15 are satisfied.

Proof. According to $(P1')$, the elements of $(\mathbf{X}_1^{(n)})_{i,j}$ for each $i \ne j$ are non-positive. Moreover, by using the basic estimations in Section 2.3.5, we obtain

$$\mathbf{X}_{10}^{(n)}\mathbf{e}_0 \ge \mathbf{X}_{10}^{(n)}\mathbf{e}_0 + \underbrace{\mathbf{X}_{1\partial}^{(n)}\mathbf{e}_\partial}_{\le \mathbf{0}} = \mathbf{X}_1^{(n)}\mathbf{e} = \left(\frac{1}{\Delta t}\mathbf{M} + \theta\mathbf{K}^{(n)}\right)\mathbf{e} > \mathbf{0}.$$

Thus, the matrices $\mathbf{X}_{10}^{(n)}$ are regular M-matrices and as such they are regular and their inverses are non-negative ([9]). Hence, (P1) is satisfied. Due to $(P1')$, $\mathbf{X}_{1\partial}^{(n)} \le \mathbf{0}$, hence (P2) is obvious. Finally, $(P3')$ guarantees the non-negativity of $\mathbf{X}_2^{(n)}$, which, together with (P1), is sufficient for (P3). This completes the proof. ■

First we investigate the finite difference approximations, and then the linear/bilinear finite element discretization.

a. The non-negativity preservation property of the finite difference mesh operators

Let us consider the finite difference approximation of the differential operator in the general form (2.3.65). In the next theorem an a priori sufficient condition of the non-negativity preservation is given for the upwind finite difference operators.

Theorem 2.3.62 *The upwind finite difference discrete mesh operator $\mathcal{L}$, defined by* (2.3.17) *with* (2.3.72)-(2.3.74) *possesses the discrete non-negativity preservation property if the condition*

$$\Delta t \le \frac{1}{(1-\theta)\max_{1\le i\le N} K_{i,i}^{(n)}} \tag{2.3.123}$$

is satisfied.

Proof. The non-negativity preservation can be guaranteed by the Theorem 2.3.61. The conditions $(P1')$ and $(P2')$ are obviously true. The assumption (2.3.123) guarantees $(P3')$. ■

Remark 2.3.63 *By using the coefficient functions in the continuous operator L in* (2.3.65)*, the condition* (2.3.123) *can be guaranteed by the assumption*

$$\Delta t \leq \frac{1}{(1-\theta)\left(\frac{2k^*}{h^2_{\min}} + \frac{a^*}{h_{\min}} - a_{\inf}\right)} = \frac{h^2_{\min}}{(1-\theta)\,(2k^* + a^* h_{\min} - a_{\inf} h^2_{\min})}, \tag{2.3.124}$$

where $k^* = \sup_{(\mathbf{x},t)\in Q_T}\{\sum_{m=1}^d k_m(\mathbf{x},t)\}$*; and* $a^* = \sup_{(\mathbf{x},t)\in Q_T}\{\sum_{m=1}^d |a_m(\mathbf{x},t)|\}$.[9]

Next we pass to the consideration of the finite difference discrete mesh operator $\mathcal{L}$ defined by (2.3.17) with (2.3.66)-(2.3.69) (central difference approximation). For such an approximation the non-positivity of the elements $(\mathbf{X}_1^{(n)})_{i,j}$ for each $i \neq j$ is not automatically satisfied, only if the condition

$$h_{i+sign((a_m)_i^{(n)})x_m} \leq \frac{2(k_m)_{i+0.5sign((a_m)_i^{(n)})}}{|(a_m)_i^{(n)}|} \tag{2.3.125}$$

holds. Then, by repeating the proof of Theorem 2.3.62, we arrive at the following statement.

Theorem 2.3.64 *The finite difference discrete mesh operator* $\mathcal{L}$*, defined by* (2.3.17) *and* (2.3.66)-(2.3.69)*, under the condition* (2.3.125)*, possesses the discrete non-negativity preservation property if the condition* (2.3.123) *is satisfied.*

Remark 2.3.65 *By using the coefficient functions in the continuous operator L in* (2.3.65)*, the condition* (2.3.125) *can be guaranteed by the assumption*

$$h_{max} \leq \frac{2\min_{1\leq m\leq d}\inf_{Q_T} k_m(\mathbf{x},t)}{\max_{1\leq m\leq d}\sup_{Q_T}|a_m(\mathbf{x},t)|}. \tag{2.3.126}$$

The condition (2.3.123) *can be guaranteed by the assumption*

$$\Delta t \leq \frac{1}{(1-\theta)\left(\frac{2k^*}{h^2_{\min}} - a_{\inf}\right)}. \tag{2.3.127}$$

It is worth characterizing the sharpness of the estimates (2.3.124) and (2.3.127) on the heat conduction operator, for which we know the exact bounds. For this operator $k_{st} = 1$, $a_{st} = 0$ and $a_{\inf} = 0$, hence, both the estimations (2.3.124) and (2.3.127) result in the bound (2.3.109), which is the exact bound for all $N = 1, 2, \ldots$.

b. The non-negativity preservation property of the linear finite element mesh operators

In the sequel we analyze the discrete mesh operator obtained by linear finite element space and the θ-method time dicretization. First we investigate the one dimensional case, and then the higher dimensional case will be considered.

b1. Non-negativity preservation property in 1D case

In this part we give the conditions of the non-negativity preservation for the linear finite element discretization on uniform mesh of the operator

$$L \equiv \frac{\partial}{\partial t} - \frac{\partial}{\partial x}(k(\mathbf{x},t)\frac{\partial}{\partial x}) - a(\mathbf{x},t)\frac{\partial}{\partial x} \; - a_0(\mathbf{x},t), \tag{2.3.128}$$

[9]We recall that $a_{\inf} = \inf_{Q_T} a_0$, according to the proof of Theorem 2.2.12.

We assume again that the bounds $k^* \geq k(\mathbf{x},t) \geq k_* > 0$; $a^* \geq |a(\mathbf{x},t)|$ and $a_{inf} \leq a_0(\mathbf{x},t) \leq 0$ are finite.

The non-negativity preservation can be guaranteed again by the use of Theorem 2.3.61. Due to (2.3.78), for the operator L in (2.3.128) the entries of the stiffness matrix are defined from the relation

$$(K_\star^{(n)})_{i,j} = \int_\Omega \left(k(x,t_n)\frac{\mathrm{d}\phi_j}{\mathrm{d}x}(x)\frac{\mathrm{d}\phi_i}{\mathrm{d}x}(x) \right) \mathrm{d}x - \\ - \int_\Omega \left(a(x,t_n)\frac{\mathrm{d}\phi_j}{\mathrm{d}x}(x)\phi_i(x) + a_0(x,t_n)\phi_j(x)\phi_i(x) \right) \mathrm{d}x. \tag{2.3.129}$$

Since in case $|i-j|>1$ we have $supp\,(\phi_j) \bigcap supp\,(\phi_i) = \emptyset$ therefore, for these indices we get $(K_\star^{(n)})_{i,j} = 0$. For the values $x \in supp\,(\phi_i)\backslash\{x_i\}$ we have

$$\frac{\mathrm{d}\phi_i}{\mathrm{d}x}(x) = \pm\frac{1}{h}. \tag{2.3.130}$$

Hence, for the values $j = i \pm 1$, due to the non-positivity of a_0, we have the upper estimation

$$K_{i,j}^{(n)} \leq -\frac{k_\star}{h^2} + \frac{a^\star}{2h} - \frac{a_{inf}}{6}. \tag{2.3.131}$$

Hence, under the assumption

$$-\frac{k_\star}{h^2} + \frac{a^\star}{2h} - \frac{a_{inf}}{6} \leq 0 \tag{2.3.132}$$

$(P1')$ is satisfied. Due to the positivity of $k_\star$, for sufficiently small h, this condition can always be satisfied. Then, $(P2')$ can be satisfied by the condition

$$\Delta t \geq \frac{1}{6\theta\left(\frac{k_\star}{h^2} - \frac{a^\star}{2h} + \frac{a_{inf}}{6}\right)} = \frac{h^2}{\theta\left(6k_\star - 3ha^\star + h^2 a_{inf}\right)}. \tag{2.3.133}$$

In order to satisfy $(P3')$, we need an upper estimation for $K_{ii}^{(n)}$:

$$K_{i,i}^{(n)} \leq \frac{2k^\star}{h^2} + \frac{a^\star}{h} - \frac{2a_{inf}}{3}. \tag{2.3.134}$$

Hence, for $(P3')$ the condition reads as

$$\Delta t \leq \frac{2h^2}{(1-\theta)\left(6k^\star + 3ha^\star - 2h^2 a_{inf}\right)}. \tag{2.3.135}$$

Hence, we can summarize our results as follows.

Theorem 2.3.66 *The discrete mesh operator $\mathcal{L}$, obtained by linear finite element discretization of the operator L in* (2.3.128)*, possesses the discrete non-negativity preservation property if, according to* (2.3.132)*, the space discretization parameter h is sufficiently small, and the time discretization parameter Δt satisfies both the lower and upper bounds, given in* (2.3.133) *and* (2.3.135)*, respectively.*

Remark 2.3.67 *We note that for a special case, namely, for the heat conduction operator L in* (2.3.81)*, the bounds* (2.3.133) *and* (2.3.135) *turn into the exact bound* (2.3.120)*, which is valid for all N.*

Remark 2.3.68 *Naturally, the lower bound should not exceed the upper bound. This results in certain restrictions for the possible choice of the parameter θ. Namely, if we introduce the notation $\mu = \mu(k) := k^\star/k_\star$ for the oscillation of the function $k(x,t)$, then, under the condition*

$$\theta \geq \theta_0(\mu) = \frac{\mu}{2+\mu}, \tag{2.3.136}$$

for sufficiently small h we can always find suitable values for Δt. Since $\mu \geq 1$, therefore $\theta_0 \geq 1/3$ and $\theta_0 = 1/3$ only for $a(x,t) = a_0(x,t) = 0$ and $k(x,t) = constant$. Since $\theta_0(\mu)$ in (2.3.136) *tends to one monotonically as μ tends to infinity, therefore, for any linear finite element discretization with sufficiently small h, there always exists suitable θ_0, such that for any $\theta \in (\theta_0, 1]$, under the conditions* (2.3.133) *and* (2.3.135)*, the discrete mesh operator is non-negativity preserving. It is worth mentioning that the Crank-Nicolson scheme ($\theta = 0.5$) belongs to this interval only when $\mu \leq 2$, which corresponds to the condition $k^\star \leq 2k_\star$.*

b2. Non-negativity preservation property in higher dimensions

To give the exact condition for the non-negativity preservation property of the discrete mesh operator, obtained by the finite element method, for dimensions $d \geq 2$ is a difficult task, even the simplest case $L = \Delta_d$, where Δ_d denotes the d-dimensional Laplace operator. (The problem is the inversion of the block tridiagonal matrices.) The problem turns into more complex task when we consider the discretization of the operator L in the general form (2.3.65). In this part we analyze this problem for a special case: $a_m(\mathbf{x},t) = a_0(\mathbf{x},t) = 0$ and $k_m(\mathbf{x},t) = constant$. (Without loss of generality we assume that this constant equals one.) We give sufficient condition for the non-negativity preservation property.

Hence, we consider the operator

$$L \equiv \frac{\partial}{\partial t} - \sum_{m=1}^{d} \frac{\partial^2}{\partial x_m^2} = \frac{\partial}{\partial t} - \Delta_d. \tag{2.3.137}$$

First we assume that $d = 2$ and Ω is a polygonal domain in $\mathbb{R}^2$ with a boundary $\partial\Omega$, $T > 0$. Let Ω be covered by a hybrid mesh $\mathcal{T}_h$ (see Figure 2.3.3), where h stands for the discretization parameter. (We note that in extremal cases the hybrid mesh covers two special types of meshes - a triangular one and a rectangular one.) Let $P_1, ..., P_N$ denote the interior nodes, and $P_{N+1}, \ldots, P_{\bar{N}}$ the boundary ones in $\mathcal{T}_h$. We also define $N_\partial := \bar{N} - N$.

Let $\phi_1, \ldots, \phi_{\bar{N}}$ be basis functions defined as follows: each ϕ_i is required to be continuous piecewise linear (over triangular elements) and bilinear (over rectangular elements) such that $\phi_i(P_j) = \delta_{ij}$, $i, j = 1, \ldots, \bar{N}$, where δ_{ij} is the Kronecker symbol. For these basis functions we have

$$\phi_i \geq 0, \ i = 1, ..., \bar{N}, \tag{2.3.138}$$

$$\sum_{i=1}^{\bar{N}} \phi_i \equiv 1 \ \text{ in } \overline{\Omega}, \tag{2.3.139}$$

i.e., the assumptions (2.3.75) and (2.3.80) are satisfied. As before, we will apply Theorem 2.3.61 to prove the non-negativity preservation property of such a linear/bilinear finite element discrete mesh operator.

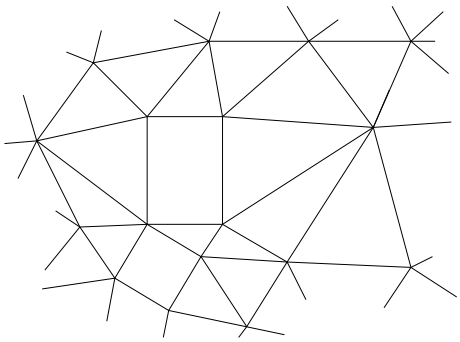

Figure 2.3.3: An example of a hybrid mesh.

To do this, we give some preliminaries. The hybrid unstructured mesh $\mathcal{T}_h$ consists of triangles T_i and rectangles R_j, which together cover the solution domain Ω (see Figure 2.3.3).

Definition 2.3.69 *Let T be a triangle with interior angles α_1, α_2, and α_3. Let the number σ_T be defined as $\sigma_T = \min\{\text{ctg } \alpha_1, \text{ ctg } \alpha_2, \text{ ctg } \alpha_3\}$. We say that the triangle is non-obtuse if $\sigma_T \geq 0$. We say that the triangle is of acute type if $\sigma_T > 0$. We also introduce the following parameters:*

$$\sigma = \min_{T \in \mathcal{T}_h} \sigma_T, \ \lambda^{\triangle}_{min} = \min_{T \in \mathcal{T}_h} \text{meas}_2 T, \ \lambda^{\triangle}_{max} = \max_{T \in \mathcal{T}_h} \text{meas}_2 T, \tag{2.3.140}$$

where $\text{meas}_2 T$ *denotes the area of the triangle T.*

Definition 2.3.70 *Let R be a rectangle with the length of edges a and b. Let us define the number*

$$\mu_R = \frac{2\min^2\{a,b\} - \max^2\{a,b\}}{ab}. \tag{2.3.141}$$

We say that the rectangle is non-narrow if $\mu_R \geq 0$. We call the rectangle strictly non-narrow type if $\mu_R > 0$. We also introduce the following parameters:

$$\mu = \min_{R \in \mathcal{T}_h} \mu_R, \ \lambda^{\#}_{min} = \min_{R \in \mathcal{T}_h} \text{meas}_2 R, \ \lambda^{\#}_{max} = \max_{R \in \mathcal{T}_h} \text{meas}_2 R, \tag{2.3.142}$$

where $\text{meas}_2 R$ *denotes the area of the rectangle R.*

Remark 2.3.71 *A rectangle is non-narrow if its longest edge is not greater than $\sqrt{2}$ times the shortest one.*

Definition 2.3.72 *We say that the hybrid mesh $\mathcal{T}_h$ is of compact type if both $\sigma \geq 0$ and $\mu \geq 0$. We say that the hybrid mesh $\mathcal{T}_h$ is of strictly compact type if $\sigma > 0$ and $\mu > 0$.*

In fact, any nonzero entry K_{ij} of the stiffness matrix $\mathbf{K}$, and any nonzero entry M_{ij} of the mass matrix $\mathbf{M}$, presents a sum of several contributions calculated over triangles and rectangles forming the intersection of the supports of basis functions ϕ_i and ϕ_j. In what follows, we find what these contributions are equal to.

The contributions to the mass matrix $\mathbf{M}$ over the triangle T:

$$M_{ij}|_T = \frac{\text{meas}_2\text{T}}{12} \ (i \neq j), \quad M_{ii}|_T = \frac{\text{meas}_2\text{T}}{6}. \tag{2.3.143}$$

The contribution to the stiffness matrix $\mathbf{K}$ over T (with the vertices denoted by e.g., P_i, P_j, and P_k) is equal to

$$K_{ij}|_T = -\frac{1}{2}\text{ctg } \alpha_{ij} \ (i \neq j), \tag{2.3.144}$$

where α_{ij} is the interior angle opposite the edge P_iP_j. If the triangle T is non-obtuse, then, obviously, $K_{ij}|_T$ is non-positive. Further,

$$K_{ii}|_T = \frac{l_i^2}{4\,\text{meas}_2 T}, \tag{2.3.145}$$

where l_i is the length of the edge of the triangle T opposite the vertex P_i.

Now we pass to the investigation of the contributions from the rectangular elements, i.e., form the rectangular element R with the edges a_R and b_R. We can easily show that

$$M_{ij}|_R \in \left\{\frac{\text{meas}_2 R}{18},\ \frac{\text{meas}_2 R}{36}\right\} (i \neq j), \quad M_{ii}|_R = \frac{\text{meas}_2 R}{9}. \tag{2.3.146}$$

Further,

$$K_{ij}|_R \in \left\{-\frac{2b_R^2 - a_R^2}{6\,\text{meas}_2 R},\ -\frac{a_R^2 + b_R^2}{6\,\text{meas}_2 R},\ -\frac{2a_R^2 - b_R^2}{6\,\text{meas}_2 R}\right\}\ (i \neq j)\,. \tag{2.3.147}$$

If the rectangle R is non-narrow, then $K_{ij}|_R$ is non-positive. Also,

$$K_{ii}|_R = \frac{a_R^2 + b_R^2}{3\,\text{meas}_2 R}. \tag{2.3.148}$$

In the following we formulate three lemmas which give sufficient conditions for the requirements $(P1')$-$(P3')$ in Theorem 2.3.61, respectively.

Lemma 2.3.73 *Let the hybrid mesh $\mathcal{T}_h$ be of compact type, then $K_{ij} \leq 0$ for $i \neq j,\ i = 1, ..., N,\ j = 1, ..., \bar{N}$.*

Proof. We denote $\text{supp}\phi_i \cap \text{supp}\phi_j$ by S, then for $K_{ij},\ i \neq j$, we have

$$\begin{aligned} K_{ij} &= \int\limits_\Omega \text{grad}\phi_j \cdot \text{grad}\phi_i \, dx = \\ &= \sum_{R \subseteq S} \int\limits_R \text{grad}\phi_j \cdot \text{grad}\phi_i \, dx + \sum_{T \subseteq S} \int\limits_T \text{grad}\phi_j \cdot \text{grad}\phi_i \, dx \\ &= \sum_{R \subseteq S} K_{ij}|_R + \sum_{T \subseteq S} K_{ij}|_T \leq 0, \end{aligned} \tag{2.3.149}$$

because the values $K_{ij}|_R$ and $K_{ij}|_T$ are non-positive for any non-narrow rectangle and any non-obtuse triangle, respectively. ■

Lemma 2.3.74 *Let the hybrid mesh $\mathcal{T}_h$ be of strictly compact type, i.e., $\mu > 0$ and $\sigma > 0$, then $A_{ij} = M_{ij} + \theta\Delta t\ K_{ij} \leq 0,\ i \neq j,\ i = 1, ..., N,\ j = 1, ..., \bar{N}$, provided that*

$$\Delta t \geq \frac{1}{6\,\theta\ \min\{\frac{\mu}{3\lambda_{max}^{\#}}, \frac{\sigma}{\lambda_{max}^{\triangle}}\}}. \tag{2.3.150}$$

PROOF. We denote $\mathrm{supp}\phi_i \cap \mathrm{supp}\phi_j$ by S, then

$$M_{ij} + \theta\Delta t K_{ij} = \sum_{R\subseteq S}(M_{ij}|_R + \theta\Delta t K_{ij}|_R) + \sum_{T\subseteq S}(M_{ij}|_T + \theta\Delta t K_{ij}|_T) \leq$$

$$\leq \sum_{R\subseteq S}\left(\frac{\mathrm{meas}_2 R}{18} - \theta\Delta t\frac{\mu_R}{6}\right) + \sum_{T\subseteq S}\left(\frac{\mathrm{meas}_2 T}{12} - \theta\Delta t\frac{\sigma_T}{2}\right) \leq$$

$$\leq \sum_{R\subseteq S}\left(\frac{\mathrm{meas}_2 R}{18} - \theta\Delta t\frac{\mu}{6}\right) + \sum_{T\subseteq S}\left(\frac{\mathrm{meas}_2 T}{12} - \theta\Delta t\frac{\sigma}{2}\right) \leq$$

$$\leq \sum_{R\subseteq S}\left(\frac{\mathrm{meas}_2 R}{12} - \theta\Delta t\frac{\mu\,\mathrm{meas}_2 R}{6\,\mathrm{meas}_2 R}\right) + \sum_{T\subseteq S}\left(\frac{\mathrm{meas}_2 T}{12} - \theta\Delta t\frac{\sigma\,\mathrm{meas}_2 T}{2\,\mathrm{meas}_2 T}\right) \leq \tag{2.3.151}$$

$$\leq \sum_{R\subseteq S}\left(\frac{\mathrm{meas}_2 R}{12} - \theta\Delta t\frac{\mu\,\mathrm{meas}_2 R}{6\,\lambda^{\#}_{max}}\right) + \sum_{T\subseteq S}\left(\frac{\mathrm{meas}_2 T}{12} - \theta\Delta t\frac{\sigma\,\mathrm{meas}_2 T}{2\,\lambda^{\triangle}_{max}}\right) \leq$$

$$\leq \mathrm{meas}_2 S\left(\frac{1}{12} - \theta\frac{\Delta t}{2}\ \min\{\frac{\mu}{3\lambda^{\#}_{max}}, \frac{\sigma}{\lambda^{\triangle}_{max}}\}\right) \leq 0.$$

This completes the proof. ■

Lemma 2.3.75 *Let the hybrid mesh $\mathcal{T}_h$ be of strictly compact type, i.e., $\mu > 0$ and $\sigma > 0$, then $B_{ii} = M_{ii} - (1-\theta)\Delta t\ K_{ii} \geq 0, \quad i = 1, ..., N,$ provided that*

$$\Delta t \leq \frac{1}{9\,(1-\theta)\ \max\{\frac{\gamma^{\#}_{max}}{3\lambda^{\#}_{min}}, \frac{\gamma^{\triangle}_{max}}{4\,\lambda^{\triangle}_{min}}\}}, \tag{2.3.152}$$

where $\gamma^{\triangle}_{max} = \max\limits_{T\in\mathcal{T}_h}\{\frac{l^2_{max}}{\mathrm{meas}_2 T}\}$.

In the formulation of the lemma we have used the notation l_{max} for the length of the longest edge in T and

$$\gamma^{\#}_{max} = \max_{R\in\mathcal{T}_h}\left\{\frac{a^2_R + b^2_R}{\mathrm{meas}_2 R}\right\}.$$

PROOF. We have the following lower estimations:

$$M_{ii} - (1-\theta)\Delta t K_{ii} =$$

$$= \sum_{R\subseteq S}(M_{ii}|_R - (1-\theta)\Delta t K_{ii}|_R) + \sum_{T\subseteq S}(M_{ii}|_T - (1-\theta)\Delta t K_{ii}|_T) \geq$$

$$\geq \sum_{R\subseteq S}\left(\frac{\mathrm{meas}_2 R}{9} - (1-\theta)\Delta t\frac{a^2_R + b^2_R}{3\mathrm{meas}_2 R}\right) + \sum_{T\subseteq S}\left(\frac{\mathrm{meas}_2 T}{6} - (1-\theta)\Delta t\frac{l^2_{max}}{4\mathrm{meas}_2 \mathrm{T}}\right) =$$

$$\geq \sum_{R\subseteq S}\left(\frac{\mathrm{meas}_2 R}{9} - (1-\theta)\Delta t\,\frac{\gamma^{\#}_{max}}{3}\right) + \sum_{T\subseteq S}\left(\frac{\mathrm{meas}_2 T}{9} - (1-\theta)\Delta t\frac{\gamma^{\triangle}_{max}}{4}\right) \geq$$

$$\geq \mathrm{meas}_2 S\left(\frac{1}{9} - (1-\theta)\Delta t\ \max\{\frac{\gamma^{\#}_{max}}{3\lambda^{\#}_{min}}, \frac{\gamma^{\triangle}_{max}}{4\,\lambda^{\triangle}_{min}}\}\right) \geq 0.$$

This completes the proof. ■

From Lemmas 2.3.73–2.3.75 it follows immediately

Theorem 2.3.76 *The linear / bilinear finite element heat conduction discrete mesh operator on a hybrid mesh of strictly compact type is non-negativity preserving if the conditions*

$$\Delta t \geq \frac{1}{6\,\theta\,\min\{\frac{\mu}{3\lambda^{\#}_{max}}, \frac{\sigma}{\lambda^{\triangle}_{max}}\}} \tag{2.3.153}$$

and

$$\Delta t \leq \frac{1}{9\,(1-\theta)\,\max\{\frac{\gamma^{\#}_{max}}{3\lambda^{\#}_{min}}, \frac{\gamma^{\triangle}_{max}}{4\,\lambda^{\triangle}_{min}}\}} \tag{2.3.154}$$

are fulfilled.

Remark 2.3.77 *For pure rectangular meshes we have the weaker lower bound for Δt in the form*

$$\Delta t \geq \frac{\lambda^{\#}_{max}}{3\,\theta\,\mu}, \tag{2.3.155}$$

because in (2.3.151) we can apply a weaker estimation [44]. For a square mesh with the step-size h the sufficient condition of the non-negativity preservation is

$$\Delta t \geq \frac{h^2}{3\theta} \tag{2.3.156}$$

and

$$\Delta t \leq \frac{h^2}{6(1-\theta)}, \tag{2.3.157}$$

respectively. This shows that in this case the non-negativity preservation can be guaranteed only (with our sufficient condition) for methods with $\theta \geq 2/3$, i.e., the Crank-Nicolson scheme is not included.

Now we consider the operator L, defined by (2.3.137), in higher dimensions i.e., for $d \geq 3$. We assume again that $\Omega \subset \mathbb{R}^d$ is a polytopic domain with a boundary $\partial\Omega$.[10] Let Ω be covered by a simplical mesh $\mathcal{T}_h$. For this case we are also able to define the elements of the local mass and stiffness matrices [20].

The contributions to the mass matrix $\mathbf{M}$ over the simplex T:

$$M_{ij}|_T = \frac{1}{(d+1)(d+2)}\mathrm{meas_d T},\ (i \neq j), \quad M_{ii}|_T = \frac{2}{(d+1)(d+2)}\mathrm{meas_d T}. \tag{2.3.158}$$

The contribution to the stiffness matrix $\mathbf{K}$ over the simplex T is equal to

$$K_{ij}|_T = -\frac{(\mathrm{meas_{d-1}S_i})(\mathrm{meas_{d-1}S_j})}{d^2\mathrm{meas_d T}}\cos\gamma_{ij}, \quad (i \neq j),\ K_{ii}|_T = \frac{(\mathrm{meas_{d-1}S_i})^2}{d^2\mathrm{meas_d T}}. \tag{2.3.159}$$

Here T is a simplex with vertices $P_1, \ldots, P_{d+1}$, S_i is the $(d-1)$-dimensional face opposite to the vertex P_i, and $\cos\gamma_{ij}$ is the cosine of the interior angle between faces S_i and S_j.

In the following we formulate those conditions which guarantee the requirements in Theorem 2.3.61.

[10] Mostly we assume that Ω is a d-dimensional rectangle, which is also called as orthotope, hyperrectangle, or d-dimensional box.

$(P1')$ When the condition

$$\gamma_{ij} \in (0, \pi/2) \tag{2.3.160}$$

holds, then clearly $K_{ij}|_T \leq 0$.

$(P2')$ Based on (2.3.158) and (2.3.159), we have

$$M_{ij} + \theta \Delta t K_{ij} = \sum_{T \subseteq \mathcal{T}_h} (M_{ij}|_T + \theta \Delta t K_{ij}|_T) =$$

$$= \sum_{T \subseteq \mathcal{T}_h} \left[\frac{1}{(d+1)(d+2)} \mathrm{meas_d T} - \theta\, \Delta t\, \frac{(\mathrm{meas_{d-1} S_i})(\mathrm{meas_{d-1} S_j})}{d^2 \mathrm{meas_d T}} \cos \gamma_{ij} \right]. \tag{2.3.161}$$

Hence, a sufficient condition of $(P2')$ is that each term in the above sum is non-positive, i.e.,

$$\Delta t \geq \frac{d^2}{(d+1)(d+2)}\ \frac{1}{\theta\, \cos \gamma_{ij}}\ \frac{(\mathrm{meas_d T})^2}{(\mathrm{meas_{d-1} S_i})(\mathrm{meas_{d-1} S_j})}. \tag{2.3.162}$$

$(P3')$ Based also on (2.3.158) and (2.3.159), we also get

$$M_{ii} - (1-\theta)\, \Delta t K_{ii} = \sum_{T \subseteq \mathcal{T}_h} (M_{ii} - (1-\theta)\, \Delta t K_{ii}) =$$

$$= \sum_{T \subseteq \mathcal{T}_h} \left[\frac{2}{(d+1)(d+2)} \mathrm{meas_d T} - (1-\theta)\, \Delta t\, \frac{(\mathrm{meas_{d-1} S_i})^2}{d^2 \mathrm{meas_d T}} \right]. \tag{2.3.163}$$

A sufficient condition of $(P3')$ is that each term in the above sum is non-negative, i.e.,

$$\Delta t \leq \frac{2\, d^2}{(d+1)(d+2)}\ \frac{1}{1-\theta}\ \frac{(\mathrm{meas_d T})^2}{(\mathrm{meas_{d-1} S_i})^2}. \tag{2.3.164}$$

Let us introduce the notations:

$$S_\star = \min_{T \subseteq \mathcal{T}_h} (\mathrm{meas_{d-1} S_i}), \quad S^\star = \max_{T \subseteq \mathcal{T}_h} (\mathrm{meas_{d-1} S_i}), \tag{2.3.165}$$

$$T_\star = \min_{T \subseteq \mathcal{T}_h} (\mathrm{meas_d T_i}), \quad T^\star = \max_{T \subseteq \mathcal{T}_h} (\mathrm{meas_d T_i}), \tag{2.3.166}$$

$$\gamma_\star^{(d)} = \min \cos \gamma_{ij}. \tag{2.3.167}$$

We can summarize our results as follows.

Theorem 2.3.78 *Let us assume that the simplical mesh $\mathcal{T}_h$ is of the strictly acute type, i.e., the geometrical condition* (2.3.160) *is satisfied. Then, for Δt chosen in accordance with the upper and lower bounds* (2.3.162) *and* (2.3.164)*, the linear finite element discrete mesh operator $\mathcal{L}$ is non-negativity preserving.*

Corollary 2.3.79 *Under the condition*

$$\frac{d^2}{(d+1)(d+2)}\ \frac{1}{\theta\, \gamma_\star^{(d)}} \left(\frac{T^\star}{S_\star} \right)^2 \leq \Delta t \leq \frac{2\, d^2}{(d+1)(d+2)}\ \frac{1}{1-\theta} \left(\frac{T_\star}{S^\star} \right)^2 \tag{2.3.168}$$

the linear finite element discrete mesh operator on the strictly acute simplical mesh is non-negativity preserving.

In practice, we apply the condition (2.3.168). Therefore, it is worth analyzing when it is applicable.

Definition 2.3.80 *For an acute simplical mesh $\mathcal{T}_h$ the number*

$$\mu_d(\mathcal{T}_h) = \frac{1}{\gamma_\star^{(d)}} \left(\frac{S^\star}{S_\star}\right)^2 \left(\frac{T^\star}{T_\star}\right)^2 \tag{2.3.169}$$

is called the uniformity number of the partition.

Clearly, $\mu_d(\mathcal{T}_h) > 1$ and the sufficient condition (2.3.168) can be applied only when

$$\mu_d(\mathcal{T}_h) \leq \frac{2\,\theta}{1-\theta} \tag{2.3.170}$$

is true. Let us notice that (2.3.170) can be written as

$$\theta \geq \theta_0^{(d)} := \frac{2\,\mu_d(\mathcal{T}_h)}{2+\mu_d(\mathcal{T}_h)}. \tag{2.3.171}$$

Remark 2.3.81 *We note that on the uniform mesh $\mathcal{T}_h$ we have $S^\star = S_\star$ and $T^\star = T_\star$. Hence, for this case*

$$\mu_d(\mathcal{T}_h) = \frac{1}{\gamma_\star^{(d)}} \text{ and } \theta_0^{(d)} = \frac{1}{2\gamma_\star^{(d)}+1}. \tag{2.3.172}$$

In the 2D case this implies the following. Since $\min(\max\gamma_{ij}) = \pi/3$*, therefore* $\gamma_\star^{(2)} = 0.5$*. Therefore, on the uniform mesh, according to* (2.3.172)*, we get* $\theta_0^{(2)} = 0.5$*, which means that the Crank-Nicolson scheme is applicable. However, as one can see, for the other cases* $\theta_0^{(2)} > 0.5$*, i.e., the Crank-Nicolson scheme is not included. In the 3D case for the uniform partition we get* $\gamma_\star^{(3)} = \cos\gamma_{ij} = 1/3$*. Therefore, on the uniform mesh, according to* (2.3.172)*, we get* $\theta^{(3)} = 0.6$*. This means that the Crank-Nicolson scheme is not applicable. Finally we remark that with the increase of the uniformity number of the partition, the interval of the admissible values θ is getting shorter and it is approaching the only possible choice $\theta = 1$.*

2.4 The Crank-Nicolson scheme to the heat equation

In what follows we apply our results to solving the homogeneous heat equation with some prescribed initial condition. (We assume that the boundary conditions, if they exist, are included into $\operatorname{dom} L$.) Applying some approximation to the operator L, we obtain the discrete mesh operator $\mathcal{L}$, and hence we get the equation

$$\mathcal{L}\nu = \mathbf{0}. \tag{2.4.1}$$

This yields the following algebraic problem: knowing the values of the mesh function ν at the points of the discrete parabolic boundary, using (2.4.1), we seek its values at the interior mesh points. When $\mathcal{L}$ is a two-level discrete mesh operator, then the above algebraic problem leads to the one-step iterative method of the form

$$\boldsymbol{\nu}^n = \widetilde{r_{stab}}(q, \mathbf{M}_0^{(n)}, \mathbf{K}_0^{(n)})\boldsymbol{\nu}^{n-1}, \tag{2.4.2}$$

for $n = 1, 2, \ldots$ with given $\boldsymbol{\nu}^0$, which is defined from the initial condition.

Remark 2.4.1 *For simplicity, for bounded Ω, we assume that the boundary conditions are homogeneous. When Ω is unbounded, then $\mathbf{M}_0^{(n)}$ and $\mathbf{K}_0^{(n)}$ are infinite matrices.*

The function $\widetilde{r_{stab}}(q, \mathbf{M}_0^{(n)}, \mathbf{K}_0^{(n)})$ is called the *stability function of the method* and it is derived in the following way: for the numerical solution of the semidiscretized problem

$$\mathbf{M}_0^{(n)} y'(t) = \frac{1}{h^2}\mathbf{K}_0^{(n)} y(t); \quad \mathbf{M}_0^{(n)} y(0) = y_0, \tag{2.4.3}$$

we apply some one-step numerical integration method, which is based on the rational approximation $r(z)$ of the exponential function $\exp(z)$.[11] When

$$r(z) = r_\theta(z) = \frac{1 + (1-\theta)z}{1 - \theta z}, \tag{2.4.4}$$

with $\theta \in [0, 1]$, then the corresponding stability function is

$$\widetilde{r_\theta}(q, \mathbf{M}_0^{(n)}, \mathbf{K}_0^{(n)}) = r_\theta\left(q\left(\mathbf{M}_0^{(n)}\right)^{-1}\mathbf{K}_0^{(n)}\right) = (\mathbf{M}_0^{(n)} - \theta\mathbf{K}_0^{(n)})^{-1}(\mathbf{M}_0^{(n)} + (1-\theta)\mathbf{K}_0^{(n)}) \tag{2.4.5}$$

for $n = 1, 2, \ldots$. Let us notice that (2.4.2) and (2.4.5) are equivalent to (2.4.1) with the choice $\mathcal{L}$ as in (2.3.17).

The Crank-Nicolson method[12] corresponds to the choice $\theta = 0.5$, i.e, for the finite difference approximation its stability function reads as

$$r_{CN}(q\mathbf{K}_0) = (\mathbf{I} - 0.5q\mathbf{K}_0)^{-1}(\mathbf{I} + 0.5q\mathbf{K}_0), \tag{2.4.6}$$

where, for simplicity, we omit the notation of the dependence of the matrix $\mathbf{K}_0$ on n.

In what follows we consider this method in more details. As it is known, due to its higher time accuracy, it is of a special interest. It is well known that this method is absolute stable, therefore it is convergent for any choice of the discretization parameters. In this part we analyze the usage of this method for the simplest heat equation operator (2.3.81) in 1D.

According to Table 2.3.2, the condition $q \leq 1$ is a necessary and sufficient condition of the discrete non-negativity preservation, for any number of space partition. Due to Theorem 2.3.10, this condition also guarantees the discrete maximum norm contractivity property. However, this is not a necessary condition for it. As it is shown in [80] and [69], the exact condition in this case is $q \leq 1.5$.

Corollary 2.4.2 *According to the above observation, the implication $DNP \Rightarrow DMNC$ in Figure 2.3.1 cannot be reversed. We note that the necessary condition of the discrete non-negativity preservation for all numbers of partition ($q \leq 2(2-\sqrt{2})$) is more restrictive than the above exact condition of the discrete maximum norm contractivity.*

Hence, we can summarize the results as follows.

[11] We attract attention, that, inspite of the previous notation in (2.3.67) and later, here the matrix $\mathbf{K}_0$ does not depend on Δt and h.

[12] John Crank (1915 - 2006) originally worked in industry on the modelling and numerical solution of diffusion in polymers. In 1943, working with Phyllis Nicolson (1917-1968) on finite difference methods for the time dependent heat equation, he proposed the Crank-Nicolson method which has been incorporated universally in the solving of time-dependent problems since then. Their first result on this method was published in [29].

Theorem 2.4.3 *The finite difference Crank-Nicolson mesh operator, which corresponds to the one-dimensional continuous differential operator L in* (2.3.81)*, has the following qualitative properties:*

- *It satisfies the discrete strong maximum principle (and hence all other qualitative properties) for any number of the uniform space partition if and only if the condition $q \in (0, 1]$ holds.*
- *It satisfies the discrete strong maximum principle (and hence all other qualitative properties) for a sufficiently large number of the uniform space partition if and only if the condition $q \in (0, 2(2 - \sqrt{2})]$ holds.*
- *For the values $q \in (0, 1.5]$ it is contractive in the maximum norm.*

In the sequel, we analyze the qualitative behaviour of the operator "after the death", i.e., in the case $q > 1.5$.

2.4.1 Some preliminaries for the Crank-Nicolson scheme

To our theoretical investigation, we consider the initial value problem for the heat equation in an infinite space domain. Then the numerical solution of the one-dimensional heat equation

$$Lu \equiv \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0; \;\; x \in \mathbb{R}, \;\; t \geq 0, \quad u(x, 0) = u_0(x), \;\; x \in \mathbb{R}, \tag{2.4.7}$$

by the Crank-Nicolson scheme leads to the one-step iterative method

$$\boldsymbol{\nu}^n = r_{CN}(q\mathbf{K}_0)\boldsymbol{\nu}^{n-1}, \;\; n = 1, 2, \ldots. \tag{2.4.8}$$

We recall that here $\mathbf{K}_0$ is the infinite tridiagonal matrix

$$\mathbf{K}_0 = \text{tridiag}[1, -2, 1]. \tag{2.4.9}$$

According to the usual theoretical investigations, the approximations are sequences $\boldsymbol{\nu}^n = \{\nu_j^n\}_{j=-\infty}^{+\infty} \simeq \{u(jh, n\Delta t)\}_{j=-\infty}^{+\infty}$ of complex numbers and $\boldsymbol{\nu}^0$ is defined from the initial function u_0.

For $1 \leq p \leq +\infty$, let l_p be the Banach space of all the p-summable sequences of complex numbers $\mathbf{z} = \{z_j\}_{j=-\infty}^{+\infty}$, endowed with the standard norm

$$\| \mathbf{z} \|_p = h^{1/p} \left(\sum_{j=-\infty}^{+\infty} |z_j|^p \right)^{1/p}, \qquad 1 \leq p < +\infty,$$

and

$$\| \mathbf{z} \|_\infty = \sup_{-\infty < j < +\infty} |z_j|.$$

It is also well known (see, e.g., [139]) that for all $1 \leq p \leq +\infty$ the numerical method is stable independently of the step-sizes, i.e., independently of q. This means that there exists $C_p > 0$ such that

$$\| r_{CN}^n(q\mathbf{K}_0) \|_p \leq C_p \tag{2.4.10}$$

for all $q > 0$. As a consequence of the stability, it turns out that for sufficiently smooth initial data u_0 the Crank-Nicolson discretization attains the second order accuracy without any restriction on q.

Remark 2.4.4 *Let us point out that if u_0 is a non-smooth initial data, i.e., u_0 is merely in L_p, the above accuracy requires certain restrictions on q [140], which spoils the use of an implicit method. However, if we use the implicit Euler method for the first two steps, $\boldsymbol{\nu}^1$ and $\boldsymbol{\nu}^2$, and the Crank-Nicolson scheme (2.4.8) for $n \geq 3$, then the usual second order is also attained for non-smooth initial data, without any restriction on q [63, 92]. Further analysis of this combined method can be found in Section 2.4.4 and in [50].*

Henceforth C_p denotes the smallest possible constant fulfilling (2.4.10), which is called *the stability constant* in the corresponding norm. Since (2.4.8) is absolute contractive in the l_2-norm, we have $C_2 = 1$. In [120] (see also [139]) it is proved that $C_\infty < 23$ (see [121] for a related result). Due to Lemma 2.4.3, the method (2.4.8) is contractive in the maximum norm only for the values $0 < q \leq 1.5$. Therefore, clearly $C_\infty > 1$.

Our aim is to sharpen the upper bound for the constant C_∞. (In the sequel we follow the analysis of the paper [51].)

In our consideration we make use of a resolvent estimate for the operator $\mathbf{K}_0$. Let us set $\varphi = \arg(z)$, then, for $|\varphi| < \pi$, the resolvent

$$R(z, \mathbf{K}_0) = (z - \mathbf{K}_0)^{-1}$$

exists[13] and the following estimate (see [37])

$$\| R(z, \mathbf{K}_0) \|_\infty \leq \frac{\sec(\varphi/2)}{|z|} \tag{2.4.11}$$

holds. moreover,we note that the factor $\sec(\varphi/2)$ in the above estimate is sharp (see [4] for related results). Consequently, the theory of sectorial operators can be employed.

Remark 2.4.5 *In practice, very often we have to consider the initial-value problem for the heat equation, posed in a bounded interval, along with either Dirichlet or Neumann boundary conditions. Notice that after a suitable extension of the initial function (i.e., a periodic one that is either odd or even with respect to the extremes of the interval), the problem is transformed into a pure initial value problem on the whole $\mathbb{R}$. Then it is straightforward to prove that the constant C_∞ is an upper bound of the stability constant for the original initial boundary value problem.*

2.4.2 Lower and upper bounds for C_∞

In this section, first we derive the exact value of $\| r_{CN}(q\mathbf{K}_0) \|_\infty$ for any $q > 0$. Clearly, this value serves then as the lower bound for C_∞. (For some related results we refer to [80] or [139].)

Theorem 2.4.6 *The equality*

$$\| r_{CN}(q\mathbf{K}_0) \|_\infty = \begin{cases} 1 & \text{for } 0 < q \leq 3/2 \ , \\ 3 - \frac{4}{\sqrt{1+2q}} & \text{for } q > 3/2 \ , \end{cases} \tag{2.4.12}$$

holds.[14]

[13] For the resolvent we use the conventional notation in the semigroup theory, omitting the notation of the identity operator.

[14] The relation (2.4.12) shows that the Crank-Nicolson method is contractive in the maximum norm really only for $q \leq 1.5$.

PROOF. Let now $\mathbf{E} : l_\infty \to l_\infty$ denote the shift operator $\{x_j\}_{j=-\infty}^{\infty} \to \{x_{j+1}\}_{j=-\infty}^{\infty}$. Since $\mathbf{K}_0 = \mathbf{E}^{-1} - 2\mathbf{I} + \mathbf{E}$ (where $\mathbf{I}$, as before, denotes the identity operator), then, at least formally, we have

$$r_{CN}(q\mathbf{K}_0) = \sum_{j=-\infty}^{\infty} c_j \mathbf{E}^j,$$

where the coefficients $c_n j$ are the ones of the Laurent expansion of the function

$$\varphi(z) = r_{CN}(q(z^{-1} - 2 + z)) = \sum_{j=-\infty}^{\infty} c_j z^j.$$

Moreover, the relation

$$\| r_{CN}(q\mathbf{K}_0) \|_\infty = \sum_{j=-\infty}^{\infty} |c_j| \tag{2.4.13}$$

holds [139]. Thus, we have to develop the rational function

$$\varphi(z) = -1 - \frac{4z}{qz^2 - 2(1+q)z + q}.$$

Obviously, the roots of the denominator are α and β, where $\beta = 1/\alpha$ and α can be expressed as $\alpha = (1+q-\sqrt{1+2q})/q < 1$. Using the decomposition into simple fractions, an easy computation results in

$$\varphi(z) = -1 + \frac{4\alpha}{q(1-\alpha^2)} + \frac{4\alpha}{q(1-\alpha^2)} \sum_{j=1}^{\infty} \alpha^j (z^j + z^{-j}).$$

Since

$$\frac{\alpha}{1-\alpha^2} = \frac{1}{\beta - \alpha} = \frac{q}{2\sqrt{1+2q}},$$

and

$$\sum_{j=1}^{+\infty} |\alpha^j| = \frac{|\alpha|}{1-|\alpha|},$$

using (2.4.13) leads to

$$\| r_{CN}(q\mathbf{K}_0) \|_\infty = \left| -1 + \frac{2}{\sqrt{1+2q}} \right| + \frac{4(1+q-\sqrt{1+2q})}{1+2q-\sqrt{1+2q}}. \tag{2.4.14}$$

If the expression in modulus is non-negative, that is if $q \le 3/2$, then the right-hand side is equal to one and the first part of the statement in (2.4.12) is proved. For $q > 3/2$ the second equality in (2.4.12) follows easily after writing

$$\left| -1 + \frac{2}{\sqrt{1+2q}} \right| = 1 - \frac{2}{\sqrt{1+2q}}.$$

■

Corollary 2.4.7 *Since the right-hand side of (2.4.12) is an increasing function of q, the relation*

$$\lim_{q \to \infty} \| r_{CN}(q\mathbf{K}_0) \|_\infty = 3 \tag{2.4.15}$$

yields that $\| r_{CN}(q\mathbf{K}_0) \|_\infty \le 3$ *and the number* 3 *is the smallest one with this property.*

In the sequel, using the resolvent estimate (2.4.11), we give upper bounds for $\| r_{CN}^n(q\mathbf{K}_0) \|_\infty$, $n \geq 1$. To this aim the basic point is the Cauchy integral representation (see [106])

$$r_{CN}^n(q\mathbf{K}_0) = r_{CN}^n(\infty) + \frac{1}{2i\pi} \int_{\Gamma_n(a,\varphi)} r_{CN}^n(z) R(z, q\mathbf{K}_0)\, \mathrm{d}z, \tag{2.4.16}$$

where $\Gamma_n(a,\varphi)$ is the positively oriented boundary of the domain

$$\{\, z \in \mathbf{C} \,:\, a/n \leq |z| \leq 4n/a, \quad |\arg(z)| \leq \varphi \,\}.$$

This path Γ_n depends on n and on the two parameters $0 < \varphi < \pi$ and $0 < a < 2$. From this representation and by the resolvent estimate (2.4.11) we get,

$$\|r_{CN}^n(q\mathbf{K}_0)\|_\infty \leq 1 + \frac{1}{2\pi} \int_{\Gamma_n(a,\varphi)} |r_{CN}^n(z)| \frac{1}{|z| \cos(\arg(z)/2)}\, |dz|.$$

Noticing that the integrand is symmetric with respect to conjugation, the above integral is twice the contribution due to the part of $\Gamma_n(a,\varphi)$ lying in the upper half plane. Moreover, for $\eta = 4/z$ we have

$$|r_{CN}(\eta)|^n \frac{1}{|\eta| \cos(\arg(\eta)/2)}\, |d\eta| = |r_{CN}(z)|^n \frac{1}{|z| \cos(\arg(z)/2)}\, |dz|.$$

Therefore,

$$\|r_{CN}^n(q\mathbf{K}_0)\|_\infty \leq 1 + \frac{2}{\pi} \left(I_1^{(n)}(a,\varphi) + I_2^{(n)}(a,\varphi) \right),$$

where

$$I_1^{(n)}(a,\varphi) = \sec(\varphi/2) \int_{a/n}^{2} |r_{CN}(\rho e^{i\varphi})|^n \, \frac{d\rho}{\rho}$$

and

$$I_2^{(n)}(a,\varphi) = \int_0^{\varphi} |r_{CN}(a e^{i\theta}/n)|^n \sec(\theta/2)\, d\theta.$$

On the other hand, for $\rho > 0$ and $0 \leq \theta < \pi$, we have

$$|r_{CN}(\rho e^{i\theta})|^2 = 1 + \frac{8\rho \cos\theta}{4 + \rho^2 - 4\rho\cos\theta}, \tag{2.4.17}$$

and therefore

(a) $\partial |r_{CN}(\rho e^{i\theta})| / \partial\theta \leq 0$, if $0 \leq \theta < \pi$,
(b) $\partial |r_{CN}(\rho e^{i\theta})| / \partial\rho \geq 0$, if $0 \leq \theta < \pi/2$,
(c) $\partial |r_{CN}(\rho e^{i\theta})| / \partial\rho \leq 0$, if $\pi/2 \leq \theta < \pi$.

Hence, the integrand in $I_1^{(n)}(a,\varphi)$ is the product of two monotonic mappings w.r.t. the variable ρ, while the one in $I_2^{(n)}(a,\varphi)$ is the product of two monotonic functions w.r.t. the variable θ. Thus, the upper Riemann sums of these integrals can easily be estimated. By means of a simple MATLAB program, for a given n we can minimize the estimations of the upper Riemann sums corresponding to a grid of values of a and φ. We do not take these numerical estimates till the limit, since with this approach we do not expect to achieve an optimal estimate of C_∞, but rather bounds of a reasonable size. In this way we can construct Table 2.4.1. (The upper Riemann sums corresponding to the indicated values of the parameters a and φ are overestimated by using the package INTLAB [115]. For this reason the displayed bounds for $\|r_{CN}(q\mathbf{K}_0)^n\|_\infty$ are fully reliable.)

n	a	φ	bound for $\|r_{CN}(q\mathbf{K}_0)^n\|_\infty$
1	1.9100	0.0460	3.24470834246243
2	0.7396	1.8538	4.25470131010134
3	0.7630	2.0537	4.32262088486189
4	0.7702	2.0901	4.31834433976687
5	0.7702	2.0901	4.30073678468435
6	0.7720	2.0940	4.28502942256316
7	0.7720	2.0940	4.27311136432081
8	0.7720	2.0901	4.26439687442123

Table 2.4.1: Bounds of stability constants for the fixed time levels.

It is worth noticing the accurate estimate for $n = 1$ (recall Corollary 2.4.7). Clearly, from Table 2.4.1 we obtain

$$\|r_{CN}^n(q\mathbf{K}_0)\|_\infty < 4.323, \qquad 1 \le n \le 8. \tag{2.4.18}$$

Next we are going to show that (2.4.18) is valid for arbitrary values of n. To this end we first derive upper bounds for $I_1^{(n)}(a,\varphi)$ and $I_2^{(n)}(a,\varphi)$ that are independent of $n \ge 9$. For $0 \le \theta < \pi$ and $0 < \rho < 2$, (2.4.17) implies

$$|r_{CN}(\rho e^{i\theta})| \le \exp\left(\frac{4\rho\cos\theta}{4+\rho^2-4\rho\cos\theta}\right), \tag{2.4.19}$$

and hence we also get

$$|r_{CN}(\rho e^{i\theta})|^n \le \exp\left(\frac{4n\rho\cos\theta}{(2-\rho)^2}\right), \qquad 0 \le \theta \le \pi/2, \tag{2.4.20}$$

and

$$|r_{CN}(\rho e^{i\theta})|^n \le \exp\left(-\frac{4n\rho|\cos\theta|}{4+(a/n)^2+4(a/n)|\cos\theta|}\right), \qquad \rho \ge a/n, \qquad \pi/2 \le \theta < \pi. \tag{2.4.21}$$

Therefore, for $n \ge 1$, $0 < a < 2$, and $\pi/2 < \varphi < \pi$, by (2.4.21) we have

$$\begin{aligned} I_1^{(n)}(a,\varphi) &= \sec(\varphi/2)\int_{a/n}^2 |r_{CN}(\rho e^{i\varphi})|^n \frac{d\rho}{\rho} \\ &\le \sec(\varphi/2)\int_{a/n}^2 \exp\left(-\frac{4n\rho|\cos\varphi|}{4+(a/n)^2+4(a/n)|\cos\varphi|}\right)\frac{d\rho}{\rho} \\ &\le \sec(\varphi/2)\int_{\frac{4a|\cos\varphi|}{4+(a/n)^2+4(a/n)|\cos\varphi|}}^{+\infty} \frac{e^{-u}}{u}\,du. \end{aligned}$$

Now, recalling the formula (see, e.g., [105])

$$\int_p^\infty \frac{e^{-\sigma}}{\sigma}\,\mathrm{d}\sigma = -\ln p - \gamma + \sum_{k=1}^{+\infty}\frac{(-1)^{k-1}}{k}\frac{p^k}{k!}, \qquad p > 0, \tag{2.4.22}$$

where γ denotes the Euler constant ($\gamma = 0,57721...$), we deduce the following bound for $I_1^{(n)}(a,\varphi)$

$$I_1^{(n)}(a,\varphi) \leq J_1^{(n)}(a,\varphi) := -\ln p - \gamma + \sum_{k=1}^{3} \frac{(-1)^{k-1}}{k} \frac{p^k}{k!}, \tag{2.4.23}$$

with

$$p = \frac{4a|\cos\varphi|}{4 + (a/n)^2 + 4(a/n)|\cos\varphi|}.$$

On the other hand, again by (2.4.19)-(2.4.20), we can bound

$$I_2^{(n)}(a,\varphi) \leq J_2^{(n)}(a,\varphi),$$

where

$$\begin{aligned} J_2^{(n)}(a,\varphi) \;:=\; & \int_0^{\pi/2} \sec(\theta/2) \exp\left(\frac{4a\cos\theta}{(2-(a/n))^2}\right) d\theta \\ & + \int_{\pi/2}^{\varphi} \sec(\theta/2) \exp\left(-\frac{4a|\cos\theta|}{4 + (a/n)^2 + 4(a/n)|\cos\theta|}\right) d\theta. \end{aligned}$$

Noticing that both $I_1^{(n)}(a,\varphi)$ and $J_2^{(n)}(a,\varphi)$ are decreasing in n, we finally conclude that for $n \geq n_0 \geq 1$

$$\begin{aligned} \|r_{CN}^n(q\mathbf{K}_0)\|_\infty &\leq 1 + (2/\pi)\left(I_1^{(n)}(a,\varphi) + I_2^{(n)}(a,\varphi)\right) \\ &\leq 1 + (2/\pi)\left(I_1^{(n_0)}(a,\varphi) + J_2^{(n)}(a,\varphi)\right) \\ &\leq 1 + (2/\pi)\left(J_1^{(n_0)}(a,\varphi) + J_2^{(n_0)}(a,\varphi)\right). \end{aligned}$$

Now we take $n_0 = 9$. A simple calculation shows that each integrand in the two integrals defining $J_2^{(9)}(a,\varphi)$ is a product of two monotonic functions. Therefore, $J_2^{(9)}(a,\varphi)$ can be easily estimated by an appropriate upper Riemann sum. Thus, we can proceed as we did before when obtaining the table: $J_1^{(9)}(a,\varphi) + J_2^{(9)}(a,\varphi)$ is estimated over a discrete grid of values of a and φ, and then we minimize these estimations in a and φ over the grid. In this way, by using again the INTLAB package, we can prove that for $a = 0.6675$ and $\varphi = 2.0200$ we have

$$1 + \frac{2}{\pi}\left(J_1^{(9)}(a,\varphi) + J_2^{(9)}(a,\varphi)\right) \leq 4.32360575826713,$$

which establishes the validity of (2.4.18) for $n \geq 9$.

Finally, we can summarize our results as follows.

Theorem 2.4.8 *The finite difference Crank-Nicolson scheme is stable and contractive in the l_2-norm for any choice of the mesh sizes. In l_∞-norm (the maximum norm) it is also stable for any step sizes. However, in this space it is contractive only for the values $q \in (0, 1.5]$. However, for any choice of the step sizes it is not necessarily contractive, and the maximum norm of the initial function can increase by a factor of C_∞ at most, where $C_\infty \in [3, 4.324)$.*

2.4.3 Maximum norm contractivity and accuracy of the Crank-Nicolson scheme

Applying the Crank-Nicolson discretization to the heat conduction problem, our aim is to get reliable numerical model. This means that we have to guarantee the maximum norm contractivity, too. However, as we have already seen, this requires a bound for the discretization step-sizes. In this section we consider this condition and analyze its effect to the numerical accuracy.

For one-step finite difference methods which are based on some rational approximation of the exponential function, Spijker has shown in [129] that there is an order barrier: only methods with first order of accuracy can be contractive in the maximum norm for all $q > 0$. (Such a method is the backward Euler method with the stability function $r_{BE}(q\mathbf{K}_0) = (I - q\mathbf{K}_0)^{-1}$.) For higher order methods it is necessary to restrict the choice of q (from above) in order to preserve the maximum norm contractivity. As we have seen in the previous section, in Theorem 2.4.8, for the second order Crank-Nicolson scheme, when $r_{CN}(q\mathbf{K}_0) = (I + \frac{q}{2}\mathbf{K}_0)(I - \frac{q}{2}\mathbf{K}_0)^{-1}$ is applied to the one-dimensional heat equation, then the sharp restriction is $q \leq 1.5$. Therefore, the use of the Crank-Nicolson scheme requires the choice of a very small time step τ in the case of a small space discretization parameter h if we want to preserve the maximum norm contractivity. One of the main problems when using the Crank-Nicolson scheme is that it preserves the maximum norm contractivity only for $q \leq 1.5$. This means that if we have a fine mesh for the space variable, we must choose the time step $\tau \leq 1.5h^2$ in order to have maximum norm contractivity. For large values of N (where N denotes the number of the partition in space, i.e., $h = 1/N$) it would mean very small (even useless) τ and it requires considerable computational efforts. Moreover, this also results in an essential loss of accuracy: the computational error, as the result of the large number of iteration steps, might cause an essential loss in the accuracy, i.e., the scheme may loose its second order accuracy.

We illustrate this problem on the numerical solution of the initial-boundary value problem

$$\begin{gathered} \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0, \quad t > 0, \quad x \in (0,1), \\ u(0,t) = u(1,t) = 0, \quad t \geq 0, \\ u(x,0) = u_0(x), \quad x \in [0,1]. \end{gathered} \tag{2.4.24}$$

We demonstrate the behaviour of the numerical solution on two examples, namely, one with a smooth initial function and one with a non-smooth initial function. In the examples we compare the errors at the same fixed time level $T = 1$.

Example 2.4.9 *The first model problem is* (2.4.24) *with the smooth initial function* $u_0(x) = \sin(\pi x)$. *The exact solution is* $u(x,t) = \sin(\pi x)\exp(-\pi^2 t)$.

Example 2.4.10 *The second model problem is* (2.4.24) *with the non-smooth initial function*

$$u_0(x) = \begin{cases} 1 & \text{if } x \in [0.25, 0.75] \\ 0 & \text{otherwise .} \end{cases}$$

q	100	50	40	20	10	8
CN-error	6.35(−5)	4.14(−5)	3.10(−5)	8.79(−6)	1.53(−6)	6.04(−7)
BE-error	6.90(−3)	1.60(−3)	1.00(−3)	2.81(−4)	9.83(−5)	7.29(−5)
q	7	6	5	4	2	1.5
CN-error	2.22(−7)	1.147(−7)	4.05(−7)	6.40(−7)	9.54(−7)	9.89(−7)
BE-error	6.27(−5)	4.86(−5)	4.07(−5)	3.15(−5)	1.50(−5)	1.20(−5)
q	1	0.4	0.1	0.05	0.01	0.005
CN-error	1.03(−6)	1.06E(−6)	1.06(−6)	1.06(−6)	1.06(−6)	1.06E(−6)
BE-error	7.75E(−6)	3.67(−6)	1.70(−6)	1.38(−6)	1.12(−6)	1.09(−6)

Table 2.4.2: The maximum norm error for $h = 0.05$ for the CN and BE methods.

q	4000	2000	1500	1000	500
CN-error	3.16(−5)	9.70(−6)	5.12(−6)	2.54(−6)	6.35(−7)
BE-error	9.91(−4)	2.76(−4)	1.58(−4)	9.59(−5)	3.91(−5)
q	400	200	100	40	20
CN-error	4.03(−7)	9.30(−8)	1.54(−8)	6.35(−9)	9.46(−9)
BE-error	3.00(−5)	1.38(−5)	6.59(−6)	2.58(−6)	1.28E(−6)
q	10	4	2	1.5	1
CN-error	1.02(−8)	1.05(−8)	1.05(−8)	1.05(−8)	1.05(−8)
BE-error	6.43(−7)	2.63(−7)	1.37(−7)	1.05(−7)	7.35(−8)

Table 2.4.3: The maximum norm error for $h = 0.005$ for the CN and BE methods.

The exact solution is

$$u(x,t) = \frac{2}{\pi}\sum_{m=1}^{\infty}\frac{1}{m}\left(\cos\frac{m\pi}{4} - \cos\frac{3m\pi}{4}\right)\sin(m\pi x)\exp(-m^2\pi^2 t).$$

First we consider the maximum norm error of the Crank-Nicolson (CN) and the backward Euler (BE) methods, applied to the numerical solution of Example 2.4.9, on different meshes. Table 2.4.2 and 2.4.3 show that by refining the mesh under the maximum norm contractivity condition, the Crank-Nicolson scheme loses its higher accuracy with respect to the backward Euler method and in limit they result in the same accuracy.
The following Tables 2.4.4 - 2.4.6 serve to demonstrate the behavior of the maximum norm error of the Crank-Nicolson scheme with further different discretization parameters. These results show that the optimal accuracy of the Crank-Nicolson scheme is attained at some value $q_{opt} = q_{opt}(h)$ which is greater than 1.5. Moreover, by decreasing h (that

q	10	5	2	1.5	1
error	2.93(−5)	5.93(−6)	2.62(−6)	3.23(−6)	3.92(−6)
q	0.5	0.1	0.05	0.01	0.005
error	4.25(−6)	4.35(−6)	4.36(−6)	4.36(−6)	4.36(−6)

Table 2.4.4: The maximum norm error of the Crank-Nicolson scheme for $h = 0.1$.

q	1000	500	250	160	80	40	20
error	0.245	0.014	9.54(−5)	3.14(−5)	9.48(−6)	2.29(−6)	3.84(−7)
q	16	4.8	1.6	1.5	1	0.8	0.16
error	1.52(−7)	1.57(−7)	2.59(−7)	2.59(−7)	2.61(−7)	1.85(−7)	1.86(−7)

Table 2.4.5: The maximum norm error of the Crank-Nicolson scheme for $h = 0.025$.

q	1000	500	100	50	40	30
error	$3.16E(-5)$	9.67(−6)	3.72(−7)	6.16(−8)	2.43(−8)	4.61(−9)
q	20	10	5	1.5	1	0.5
error	2.54(−8)	3.79(−8)	4.10(−8)	2.96(−8)	2.97(−8)	2.97(−8)

Table 2.4.6: The maximum norm error of the Crank-Nicolson scheme for $h = 0.01$.

is, by refining the mesh), the values q_{opt} are increasing. Table 2.4.7 shows the loss in the accuracy. The fifth column in this table shows how much more CPU-time is used to obtain the less accurate result (with $q = 1.5$). The accuracy with the choice $q = 1.5$ is attained with some $q_{big} > 1.5$, too. The approximate values of these parameters and the corresponding CPU ratios are included in the last two columns.

Let us compare the above numerical methods on the non-smooth problem, i.e., on the Example 2.4.10 by using the Crank-Nicolson and the backward Euler methods.

Table 2.4.8 summarizes the errors in maximum norm for the Crank-Nicolson and backward Euler methods. The behaviour of the Crank-Nicolson scheme is similar as for the smooth initial function. However, the smoothing property of the backward Euler method is considerable. We remark that the same conclusions can be made for the other choices of h.

The Crank-Nicolson scheme has a local error $\mathcal{O}(\tau^2 + h^2)$. Therefore, the optimal accuracy (i.e., second order), is achieved when $\tau \sim 1/N$, (i.e., $\tau \sim h$) and not at $\tau \sim 1/N^2$, which is required by the contractivity condition. (The latter implies that for a fixed h the order of the error is defined only by τ.)

h	q_{opt}	error	error for $q = 1.5$	CPU ratio	q_{big}	CPU ratio
0.1	2	2.62(−6)	3.23(−6)	1.266	4.6	3.07
0.05	4	6.40(−7)	9.87(−7)	2.71	8.7	5.80
0.025	16	1.52(−7)	2.59(−7)	10.81	18	12.00
0.01	30	4.61(−9)	4.19(−8)	20.5	45	30.00
0.005	40	6.35(−9)	1.05(−8)	27.6	94	62.67
0.004	62.5	2.57(−9)	6.71(−9)	42.9	112.5	75.00

Table 2.4.7: Comparison of the accuracy and consumed CPU time.

q	4000	2000	400	200	100
CN-error	0.4765	0.4438	0.2397	7.86(−2)	2.30(−3)
BE-error	8.99(−4)	2.51(−4)	2.76(−5)	1.29(−5)	6.35(−6)
q	75	60	50	45	40
CN-error	8.18(−5)	1.48(−6)	3.68(−7)	3.69(−7)	3.71(−7)
BE-error	4.83(−6)	3.88(−6)	3.30(−6)	2.99(−6)	2.70(−6)
q	20	10	5	1.5	1
CN-error	3.73(−7)	3.74(−7)	3.74(−7)	3.74(−7)	3.74(−7)
BE-error	1.53(−6)	9.48(−7)	6.61(−7)	4.60(−7)	4.31(−7)

Table 2.4.8: The maximum norm error for $h = 0.005$ for the CN and BE methods for a non-smooth initial function.

2.4.4 Maximum norm contractivity for the modified Crank-Nicolson scheme

In order to guarantee the maximum norm contractivity, for any fixed space discretization parameter h we can select only such time discretization step-size Δt which is bounded from above by $1.5h^2$. As we have seen, this makes the Crank-Nicolson method less attractive. Our aim is to construct such a method which eliminates this requirement. We construct such a second order method for the one-dimensional heat equation which is contractive in the maximum norm, however, in fact, does not impose any condition for h and τ. (Clearly, due to the result of Spijker [129], such a method cannot be based solely on one rational approximation of the exponential function.)

Our approach follows that of Luskin and Rannacher, who introduced a second order stable approximation method with optimal convergence properties by combining the robust stability and approximation property of the backward Euler method with the second order accuracy of the Crank-Nicolson scheme (see [92], [112]). This result was generalized by Hansbo in [63] to Banach spaces. However, in these works, qualitative properties (such as the maximum norm contractivity) were not considered.

Because we will select the time discretization parameter τ for a fixed h, we will make a distinction in our notations. Let operator $\mathbf{A}$ be a generator of the semigroup $T(t)$ in a normed space $\mathbf{X}$. (We recall that the stability functions of the Crank-Nicolson and the backward Euler method for this operator read as

$$r_{CN}(\tau\mathbf{A}) = (I + \frac{\tau}{2}\mathbf{A})(I - \frac{\tau}{2}\mathbf{A})^{-1}, \quad r_{BE}(\tau\mathbf{A}) = (I - \tau\mathbf{A})^{-1} \tag{2.4.25}$$

respectively.)

Let us assume that the strongly continuous semigroup $T(t)$ is bounded, i.e., the relation

$$||T(t)|| \leq M \tag{2.4.26}$$

holds for some $M \geq 1$ and all $t \geq 0$. (For the heat equation (2.4.24) $M = 1$.) We say that an approximating operator family $\{V_n(\tau\mathbf{A})\}_{n=1}^{\infty}$ is *unconditionally bound preserving* (in the Banach space $\mathbf{X}$), if

$$||V_n(\tau\mathbf{A})|| \leq M \tag{2.4.27}$$

for all $\tau > 0$ and $n \in \mathbf{N}$ and for all M with the property (2.4.26). If (2.4.27) holds only for the values $\tau \leq \tau^\star$ with some $\tau^\star > 0$, we say that the approximating operator family $\{V_n(\tau A)\}_{n=1}^{\infty}$ is *conditionally bound preserving*. By the Hille-Yosida theorem, the backward Euler scheme

$$V_n(\tau \mathbf{A}) = r_{BE}^n(\tau \mathbf{A})$$

is unconditionally bound preserving. Whereas, as we have seen, the Crank-Nicolson scheme

$$V_n(\tau \mathbf{A}) = r_{CN}^n(\tau \mathbf{A})$$

is not bound preserving in an arbitrary Banach space.

In the next theorem we show how to construct second order unconditionally bound preserving schemes for exponentially decaying contraction semigroups.

Theorem 2.4.11 *Let* $\mathbf{A}$ *be a sectorial operator which generates the strongly continuous semigroup* $T(t)$*. Assume that* $||T(t)|| \leq e^{-\omega t}$ *for some* $\omega > 0$ *and all* $t \geq 0$*. Let* $r_2(\tau \mathbf{A})$ *be a conditionally bound preserving scheme for* $0 < \tau \leq \tau^\star$*. Then there exists a positive integer* n_0 *such that* $r_2^{n-n_0}(\tau \mathbf{A}) r_{BE}^{n_0}(\tau \mathbf{A})$ *is unconditionally bound preserving with the optimal second order error estimation.*

Proof. According to Proposition 1 in [50], the scheme $r_2^{n-n_0}(\tau \mathbf{A}) r_{BE}^{n_0}(\tau \mathbf{A})$ has the optimal second order error estimation. Since $r_2(z)$ is $\mathbf{A}(\theta)$-stable, it satisfies

$$||r_2^m(\tau \mathbf{A})|| \leq M_1 \tag{2.4.28}$$

for all $m = 1, 2, \ldots$ with some $M_1 > 0$ [140]. By the Hille-Yosida theorem

$$||r_{BE}^n(\tau \mathbf{A})|| \leq \frac{1}{(1 + \tau\omega)^n} \tag{2.4.29}$$

for all $n = 1, 2, \ldots$ and $\tau > 0$. Now, let n_0 be such that

$$||r_{BE}^{n_0}(\tau^\star \mathbf{A})|| \leq \frac{1}{M_1}. \tag{2.4.30}$$

For $\tau > \tau^\star$ by (2.4.29),(2.4.30) and (2.4.28) we have:

$$||r_2^{n-n_0}(\tau \mathbf{A}) r_{be}^{n_0}(\tau \mathbf{A})|| \leq ||r_2^{n-n_0}(\tau \mathbf{A})||\ ||r_{BE}^{n_0}(\tau \mathbf{A})|| \leq M_1 \frac{1}{M_1} = 1.$$

If $0 < \tau \leq \tau^\star$, then $||r_2^{n-n_0}(\tau \mathbf{A})|| \leq 1$ since $r_2(z)$ is conditionally bound preserving. Also, $||r_{BE}^{n_0}(\tau \mathbf{A})|| \leq \frac{1}{(1+\tau\omega)^{n_0}} \leq 1$ for all $\tau > 0$. Thus, $||r_2^{n-n_0}(\tau \mathbf{A}) r_{BE}^{n_0}(\tau \mathbf{A})|| \leq 1$. ■

In the following we construct such a second order method for the one-dimensional heat equation (2.4.24), for any $h > 0$, which is contractive in the maximum norm for any $\tau > 0$. In order to apply the above general results, we discretize first the space variable. We denote by $y_i(t)$, $(i = 0, 1, ..., N)$ the approximation of $u(ih, t)$, where $h := \frac{1}{N}$ and N is the dimension of the space discretization. Let the Banach space be $\mathbf{X} := (\mathbb{R}^{N+1}, ||\cdot||_\infty)$. Then the equation for the semidiscrete solution can be written as

$$\dot{y}(t) - \mathbf{Q}y(t) = 0, \quad t \geq 0, \quad y(0) = y^0, \tag{2.4.31}$$

where $\mathbf{Q} = (1/h^2)\mathbf{K}_0 = (1/h^2)\text{tridiag}[1, -2, 1]$ is sectorial and generates an analytic contraction semigroup on $\mathbf{X}$ with a growth bound ω less than zero.

Theorem 2.4.12 *The combined scheme $r_{CN}^{n-n_0}(\tau\mathbf{Q})r_{BE}^{n_0}(\tau\mathbf{Q})$ has second order accuracy. Moreover, for a suitable n_0, it is unconditionally bound preserving, i.e., contractive in the maximum norm for all $\tau > 0$ and $n \geq n_0$.*

PROOF. It is shown in [140] that $||r_{CN}^n(\tau\mathbf{Q})|| \leq M$ for all $n, N \in \mathbf{N}$ and $\tau > 0$. (In particular, for $q := \tau/h^2 \leq 3/2$ we have $||r_{CN}^n(\tau\mathbf{Q})|| \leq 1$ for all $n \in \mathbf{N}$ (see Theorem 2.4.8). Therefore, the statement follows from Theorem 2.4.11. ■

In the sequel we examine the question: how to define a suitable n_0 in Theorem 2.4.12 for some arbitrary fixed τ and h, i.e., for any q?. The backward Euler scheme satisfies

$$||r_{BE}(\tau\mathbf{Q})||_\infty = 1 - \frac{1}{\mathrm{ch}[\frac{N+1}{2}\mathrm{ch}(1+\frac{h^2}{2\tau})]} := g(\tau). \tag{2.4.32}$$

for all $\tau > 0$ (see [69]). According to Theorem 2.4.8, $||r_{CN}^n(\tau\mathbf{Q})||_\infty \leq 4.325$ for all $\tau > 0$. Therefore, if we fix the dimension of the space discretization N (or, equivalently, fix h) we seek n_0 such that the estimation $g(\tau^\star)^{-n_0} \leq 4.325$ is true. Then $g(\tau) \leq 4.325^{-\frac{1}{n_0}} =: \beta_1$ for all $\tau > \tau^\star$. Using the notation $\beta := (1-\beta_1)^{-1}$, the inequality $g(\tau) \leq \beta_1$ is equivalent to

$$\beta \geq \mathrm{ch}[\frac{N+1}{2}\mathrm{arch}(1+\frac{h^2}{2\tau})],$$

or

$$\frac{2\mathrm{arch}\beta}{N+1} \geq \mathrm{arch}(1+\frac{h^2}{2\tau}).$$

This yields

$$\tau \geq \frac{h^2}{2[\mathrm{ch}(2\frac{\mathrm{arch}\beta}{N+1}) - 1]}.$$

Finally, using the identity $\mathrm{ch}2\gamma - 1 = 2\mathrm{sh}^2\gamma$, we have

$$\tau \geq \frac{h^2}{4\mathrm{sh}^2\frac{\mathrm{arch}\beta}{N+1}}, \tag{2.4.33}$$

or, equivalently,

$$q \geq \frac{1}{4\mathrm{sh}^2\frac{\mathrm{arch}\beta}{N+1}}. \tag{2.4.34}$$

It is easy to see that for a fixed N the sequence

$$\frac{1}{4\mathrm{sh}^2\frac{\mathrm{arch}\beta}{N+1}} = \frac{1}{4\mathrm{sh}^2\frac{\mathrm{arch}[(1-(\frac{1}{4.325})^{\frac{1}{n_0}})^{-1}]}{N+1}}$$

tends to zero as n_0 tends to infinity. Therefore, there exists n_0 such that (2.4.34) holds. Using this formula n_0 can be determined. ■

Although (2.4.34) allows us to define $n_0 = n_0(\tau, N)$ such that the method is contractive in the maximum norm, if we choose the dimension of the space discretization N to be large, the value of n_0 becomes extremely big. Therefore, from the practical point of view it is reasonable to choose some smaller but fixed $n_0^\star$. In this case we select a suitable

	$n_0 = 1$	$n_0 = 2$	$n_0 = 3$	$n_0 = 4$	$n_0 = 5$	$n_0 = 10$	$n_0 = 50$
$N_0 = 1$	1.743	0.616	0.388	0.291	0.238	0.139	0.055
$N_0 = 50$	0.453	0.160	0.100	0.0076	0.0062	0.0036	0.0014
$N_0 = 10000$	0.436	0.154	0.0097	0.0073	0.0060	0.0035	0.0014

Table 2.4.9: The values of $\hat{\tau}$ uniform for all $N \geq N_0$.

τ at a fixed N and $n_0^\star$. In this case we cannot use an arbitrary $\tau > 0$, but only those $\tau > \tau_N^\star(N, n_0^\star)$, where $\tau_N^\star$ can be computed via (2.4.33). Then we have a choice. Either we use $\tau \leq 1.5h^2$ (the uniform contractivity condition for the Crank-Nicolson scheme) or $\tau > \tau_N^\star$.

For a fixed N and fixed n_0, using (2.4.33) we obtain a lower bound

$$\hat{\tau}(N, n_0) = \frac{1}{4N^2 \mathrm{sh}^2 \frac{\mathrm{arch}\beta}{N+1}} \tag{2.4.35}$$

for τ. This means that if we take any time step larger than $\hat{\tau}$, the combined method is contractive in the maximum norm. If we look at the relations (which can be checked directly)

$$\begin{aligned} \frac{1}{4\mathrm{arch}^2\beta}\left(\frac{N+1}{N}\right)^2 &> \frac{1}{4N^2\mathrm{sh}^2\frac{\mathrm{arch}\beta}{N+1}} \\ &> \frac{1}{4\mathrm{arch}^2\beta}\frac{1}{\left(\frac{N+1}{\mathrm{arch}\beta}\right)^2\mathrm{sh}^2\frac{\mathrm{arch}\beta}{N+1}}, \end{aligned} \tag{2.4.36}$$

then it is easy to see what is going to happen with this lower bound $\hat{\tau}(N, n_0)$ if we increase N by a fixed n_0; i.e.,

$$\lim_{N\to\infty} \hat{\tau}(N, n_0) = \lim_{N\to\infty} \frac{1}{4N^2 \mathrm{sh}^2 \frac{\mathrm{arch}\beta}{N+1}} = \frac{1}{4\mathrm{arch}^2\beta}. \tag{2.4.37}$$

We can also see that the sequence of the upper bounds in (2.4.36) decreases monotonically towards $1/(4\mathrm{arch}^2\beta)$. Therefore, we can give an upper bound, uniform in $N \geq N_0$, by taking some fixed value N_0 in $\hat{\tau}(N_0, n_0)$.

Table 2.4.9 shows the uniform lower bound for $\hat{\tau}$ for different values of N_0. The indicated discretizaton parameters (i.e., $h = 1/N$ and $\tau \geq \hat{\tau}$ in the table), by using n_0 times the backward Euler method and $n - n_0$ times the Crank-Nicolson method, we guarantee the maximum norm contractivity of the combined method. The method allows us to select much bigger τ for some fixed h. (E.g., according to the usual contractivity condition of the Crank-Nicolson scheme, in case $N = 10000$ for the maximum norm contractivity we must choose $\tau \leq 1.5 \times 10^{-8}$.) However, we can get low accuracy due to the large local error. Therefore, in what follows, we aim to get sharper uniform sufficient condition (e.g., smaller lower bound) for τ.

First we examine the monotonicity of the sequence of lower bounds in (2.4.35), i.e., introducing the notation $b = \mathrm{arch}\beta$, we want to define such an interval $[0, b)$ on which

$$\hat{a}_N = 4N^2 \mathrm{sh}^2 \frac{b}{N+1} \tag{2.4.38}$$

is a monotonically increasing sequence. Clearly, this is equivalent to the question of the monotonicity of the sequence

$$a_N = N\mathrm{sh}\frac{b}{N+1}. \tag{2.4.39}$$

Since

$$a_N = b\left[\frac{N+1}{b}\mathrm{sh}\frac{b}{N+1} - \frac{1}{b}\mathrm{sh}\frac{b}{N+1}\right], \tag{2.4.40}$$

we introduce a new function

$$\varphi(x) = \frac{1}{x}\mathrm{sh}x - \frac{1}{b}\mathrm{sh}x, \tag{2.4.41}$$

and we analyze its monotone decrease on $(0, b/2)$. After computing its derivative, we arrive at the condition

$$\mathrm{th}x > x - \frac{x^2}{b}. \tag{2.4.42}$$

The Taylor expansion of the function $\mathrm{th}x$ is

$$\mathrm{th}x = x - \frac{x^3}{3} + \frac{2x^5}{15} - \frac{17x^7}{315} + \cdots = \sum_{n=1}^{\infty} \frac{2^{2n}(2^{2n}-1)B_{2n}x^{2n-1}}{(2n)!}, \tag{2.4.43}$$

where B_n denotes the Bernoulli number, defined as

$$B_{2n} = 2(-1)^{n+1}\frac{\zeta(2n)\ (2n)!}{(2\pi)^{2n}} \tag{2.4.44}$$

and

$$\zeta(s) = \sum_{n=1}^{\infty}\frac{1}{n^s}; \quad s > 1 \tag{2.4.45}$$

is the Riemann zeta-function. Obviously, ζ is a monotonically decreasing function.
Due to the estimates

$$\frac{B_{2n+2}}{B_{2n}} = \frac{\zeta(2n+2)}{\zeta(2n)}\,\frac{(2n+2)!}{(2n)!}\,\frac{(2\pi)^{2n}}{(2\pi)^{2n+2}} \le \frac{(2n+1)(2n+2)}{4\pi^2} \tag{2.4.46}$$

and

$$\frac{2^{2n+2}-1}{2^{2n}-1} = 4 + \frac{3}{2^{2n}-1} \le 5, \tag{2.4.47}$$

the Taylor series for the function th is of Leibniz type whenever

$$x < \frac{\pi}{\sqrt{5}}. \tag{2.4.48}$$

Hence, for such values we have the estimate

$$\mathrm{th}x > x - \frac{x^3}{3}. \tag{2.4.49}$$

The relations (2.4.42) and (2.4.49) show that under the assumption

$$b < \frac{3\sqrt{5}}{\pi} \approx 2.13 \tag{2.4.50}$$

the function φ is monotonically decreasing on the interval $(0, b/2)$. This yields the requirement $\mathrm{arch}\beta \le 2.13$, i.e., $n_0 \le 5.48$.

Hence we get the following

	$n_0 = 1$	$n_0 = 2$	$n_0 = 3$	$n_0 = 4$	$n_0 = 5$
$N_0 = 1$	1.662	0.540	0.315	0.221	0.170
$N_0 = 50$	0.453	0.160	0.101	0.00759	0.0062
$N_0 = 10000$	0.436	0.154	0.0097	0.0073	0.0060

Table 2.4.10: Sharper lower bounds for the values of $\hat{\tau}$ uniform for all $N \geq N_0$.

n_0	0	1	2	3	5
$q = 5$	4.92(−5)	4.91(−5)	4.78(−5)	2.87(−5)	**4.51(−3)**
$q = 1$	1.01(−5)	1.49(−5)	1.91(−5)	2.36(−5)	3.34(−5)
$q = 0.1$	1.85E(−5)	1.86(−5)	1.86(−5)	1.86(−5)	1.87(−5)
n_0	10	25	50	100	250
$q = 5$	–	–	–	–	–
$q = 1$	**6.41(−5)**	–	–	–	–
$q = 0.1$	1.90(−5)	1.97(−5)	2.10(−5)	2.35(−5)	**3.18(−5)**

Table 2.4.11: Maximum norm errors for $h = 0.2$ for the damped method for a smooth initial function

Theorem 2.4.13 *For some fixed N_0 we select $\tau \geq \hat{\tau}(N_0, n_0)$, where $\hat{\tau}(N_0, n_0)$ is chosen according to* (2.4.35)*. If we execute $n_0 = 1, 2, 3, 4, 5$ steps by the backward Euler method, and $n - n_0$ steps by the Crank-Nicolson scheme, then the combined method is contractive in the maximum norm on each mesh (h, τ) where $h = 1/N$ and $N \geq N_0$.*

The different values of $\hat{\tau}(N_0, n_0)$ can be found in Table 2.4.10.

2.4.5 Numerical experiments with the modified Crank-Nicolson scheme

In the following we give numerical results for the damped method.

First, we analyze the behavior of the damped method on the problem with a smooth initial function, i.e., on the Example 2.4.9. Tables 2.4.11 - 2.4.13 show the numerical results for the damped method with different space discretization steps. Each table contains the maximum norm error for different numbers of smoothing steps n_0 and time discretization steps, (the values in the columns $n_0 = 0$ correspond to the Crank-Nicolson scheme and the errors with bold numbers are the results of the backward Euler method). We observe that with the increase of n_0 the damped method loses its accuracy and the "almost best" choice is $n_0 = 2$. Table 2.4.14 shows the results for this fixed choice with the small space discretization step size $h = 0.002$.

For a non-smooth initial function, i.e., for the Example 2.4.10, the behavior of the damped method for $n_0 = 1, 2, 3$ damping steps is given in Table 2.4.15.

Finally, we compare the damped method, the Crank-Nicolson scheme, and the backward Euler method on a mesh where the maximum norm is preserved for the damped method, that is, the mesh is chosen according to the condition (2.4.34). Clearly, on such a mesh the BE method is also maximum norm contractive, while the CN method is usually not. Tables 2.4.16-2.4.18 contain the results for the smooth initial function (Example 2.4.9), while Tables 2.4.19-2.4.21 contain the results for the non-smooth problem (Example 2.4.10). Especially remarkable is the advantage of the damped method on the non-smooth problem.

n_0	0	1	2	3	5
$q = 125$	9.57(−6)	5.01(−6)	4.79(−8)	5.65(−6)	1.87(−5)
$q = 50$	1.48(−6)	5.70(−7)	3.53(−7)	1.29(−6)	3.22(−6)
$q = 5$	1.52(−7)	1.62(−7)	1.72(−7)	1.82(−7)	2.02(−7)
n_0	10	25	50	100	250
$q = 125$	6.60(−5)	**2.77(−4)**	–	–	–
$q = 50$	8.36(−6)	2.01(−5)	**7.11(−5)**	–	–
$q = 5$	2.52(−7)	3.52(−7)	6.54(−7)	2.71(−6)	**5.40(−6)**

Table 2.4.12: Maximum norm errors for $h = 0.02$ for the damped method for a smooth initial function.

n_0	0	1	2	3	5
$q = 50000$	5.17(−5)	5.17(−5)	5.01(−5)	3.22(−5)	**4.20(−3)**
$q = 5000$	1.64(−6)	7.35(−7)	1.85(−7)	1.12(−6)	3.05(−6)
$q = 125$	6.43(−10)	1.27(−9)	1.90(−9)	2.53(−9)	3.78(−9)
n_0	10	50	100	500	2000
$q = 50000$	–	–	–	–	–
$q = 5000$	8.17(−6)	**7.07(−5)**	–	–	–
$q = 125$	6.93(−9)	3.21(−8)	6.53(−8)	3.16(−7)	**1.27(−6)**

Table 2.4.13: Maximum norm errors for $h = 0.002$ for the damped method for smooth initial function.

q	50000	37500	25000	15000	10000
max. error	5.17(−5)	1.95(−5)	7.38(−6)	8.39(−7)	1.74(−7)
q	5000	4000	3000	2000	1500
max. error	1.85(−7)	1.60(−7)	9.72(−8)	4.69(−8)	2.79(−8)
q	1000	500	375	250	125
max. error	1.44(−8)	5.06(−9)	3.59(−9)	2.55(−9)	1.90(−9)

Table 2.4.14: Maximum norm errors for $h = 0.002$ and $n_0 = 2$ for the damped method for a smooth initial function.

q	4000	2000	400	200	100
$n_0 = 1$	3.10(−3)	1.40(−3)	2.76(−3)	1.31(−4)	6.77(−6)
$n_0 = 2$	1.42(−4)	2.43(−5)	1.20(−6)	5.24(−7)	3.79(−7)
$n_0 = 3$	2.10(−5)	5.76(−6)	6.54(−7)	4.48(−7)	3.93(−7)
q	75	50	40	10	5
$n_0 = 1$	5.55(−7)	3.72(−7)	3.72(−7)	3.74(−7)	3.74(−7)
$n_0 = 2$	3.79(−7)	3.76(−7)	3.75(−7)	3.75(−7)	3.74(−7)
$n_0 = 3$	3.87(−7)	3.79(−7)	3.77(−7)	3.75(−7)	3.75(−7)

Table 2.4.15: Maximum norm errors for $h = 0.005$ for the damped method for a non-smooth initial function.

q	10.58	3.72	2.32	1.722	1.395
n_0	1	2	3	4	5
DM	0.064	1.01(−5)	2.97(−5)	3.32(−5)	4.58(−5)
CN	0.3216	2.88(−5)	1.32(−5)	1.72(−6)	4.17(−6)
BE	0.037	0.0019	8.54(−4)	4.49(−4)	3.89(−4)

Table 2.4.16: Maximum norm errors for $h = 0.2$ for the different methods on a maximum norm contractive mesh for a smooth initial function.

q	1066	380	240	86	35
n_0	1	2	3	10	50
DM	0.0685	1.79(−5)	2.38(−5)	2.70(−5)	2.99(−5)
CN	0.1262	2.63(−5)	4.29(−5)	4.70(−6)	6.77(−7)
BE	0.037	0.0016	0.0012	1.56(−4)	4.75(−5)

Table 2.4.17: Maximum norm errors for $h = 0.02$ for the different methods on a maximum norm contractive mesh for a smooth initial function.

q	106000	38000	24000	8620	3460
n_0	1	2	3	10	50
DM	0.0684	1.80(−5)	2.34(−5)	2.63(−5)	2.82(−5)
CN	0.1245	2.63(−5)	4.30(−5)	4.78(−6)	8.13(−7)
BE	0.037	0.0016	0.0012	1.53(−4)	4.56(−5)

Table 2.4.18: Maximum norm errors for $h = 0.002$ for the different methods on a maximum norm contractive mesh for a smooth initial function.

q	10.58	3.72	2.32	1.722	1.395
n_0	1	2	3	4	5
DM	0.057	4.75(−4)	2.03(−5)	2.03(−5)	2.83(−5)
CN	0.2894	0.0242	2.32(−4)	6.42(−6)	3.40(−6)
BE	0.0286	0.0015	6.37(−4)	3.37(−4)	2.89(−4)

Table 2.4.19: Maximum norm errors for $h = 0.2$ for the different methods on a maximum norm contractive mesh for a non-smooth initial function.

q	1066	380	240	86	35
n_0	1	2	3	10	50
DM	0.0667	3.66(−4)	2.54(−5)	2.43(−5)	2.70(−5)
CN	0.4970	0.4256	0.3905	0.2057	0.0220
BE	0.0332	0.0015	0.0011	1.40(−4)	4.28(−5)

Table 2.4.20: Maximum norm errors for $h = 0.02$ for the different methods on a maximum norm contractive mesh for a non-smooth initial function.

q	106000	38000	24000	8620	3460
n_0	1	2	3	10	50
DM	0.0667	3.63(−4)	2.53(−5)	2.39(−5)	2.56(−5)
CN	0.5210	0.4941	0.4927	0.4673	0.4192
BE	0.0336	0.0015	0.0011	1.38(−4)	4.14(−5)

Table 2.4.21: Maximum norm errors for $h = 0.002$ for the different methods on a maximum norm contractive mesh for a non-smooth initial function.

2.5 Summary

In the Sections 2.2- 2.4 we have analyzed various qualitative properties of the linear parabolic problems. We have considered three (in our opinion, the most important) qualitative properties, namely, the non-negativity preservation, the maximum principle, and the maximum norm contractivity. First we examined these qualitative properties for the linear continuous models. We pointed out the connection between these properties for the general parabolic case. Then we have defined the discrete analogues of these basic continuous properties and also established their connection. As a result, it was shown that the discrete non-negativity preservation, under some conditions implies all the other qualitative properties (Theorem 2.3.10). We considered the special, two-level discretizations in detail, and we have specified the above general conditions to this case (Theorems 2.3.16 and 2.3.17). The conditions were also formulated in terms of the mass and the stiffness matrices (Theorem 2.3.18), which are more natural during the practical applications. In Section 2.3.4 we have compared the different matrix maximum principles with our notions of maximum principle, pointing out the advantages of our approach. Then we have defined the conditions for two well-known and widely used approximations: for the finite difference and the linear finite element schemes. We have proven that these approximations automatically satisfy those conditions that guarantee that the non-negativity preservation implies all the other qualitative properties (Theorems 2.3.43 and 2.3.44). Hence, the whole investigation led to the analysis of the non-negativity preservation property. We gave the conditions in different space dimensions. For the heat equation in 1D, we gave the exact bounds for both kinds of discretizations (Theorems 2.3.57 and 2.3.57). In higher dimensions, we have formulated such sufficient conditions that, besides the the conditions for the ratio of the discretization step sizes, also includes some geometrical condition for the mesh (Theorems 2.3.76 and 2.3.78). In Section 2.4 we analyzed the stability constant of the Crank-Nicolson method, proving the existence of a lower bound and improving the known upper bounds. In Section 2.4.4 we suggested a new method, which is the combination of the backward Euler and the Crank-Nicolson method. We gave the algorithm of the method which guarantee both the second order of accuracy and the maximum norm contractivity, without any restriction to the choice of the mesh size (Theorem 2.4.4). The theoretical results are confirmed by numerical experiments.

Chapter 3

Analysis of operator splittings

Complex physical processes are frequently modelled by systems of linear or non-linear partial differential equations, which, as it was discussed in the previous chapter, implies the construction of a numerical model, too. Due to the complexity of these equations, typically, there is no numerical method which can provide a numerical solution accurate enough for such models, while taking reasonable integration time. Moreover, as we have also seen in the previous chapter, for a simpler problem we can formulate more easily those conditions which guarantee the preservation of the different qualitative properties in the mathematical (continuous/discrete) models, which makes the modelling process reliable.
Operator splitting means that the spatial differential operator appearing in the equations is split (decomposed) into a sum of different sub-operators having simpler forms, and the corresponding equations can be solved more easily. The sub-operators are usually chosen with regard to the different physical processes or geometric directions. Then instead of the original problem, a sequence of sub-models is solved, which gives rise to the splitting error.
Splitting techniques are commonly used when large-scale models, which appear in different fields of science and engineering, are treated numerically. In the treatment of large scientific and engineering problems splitting procedures are an excellent tool (and, very often, the only tool) by which huge computational tasks can be made tractable on the available computers by dividing them to a sequence of "smaller" and "simpler" tasks. This chapter is devoted to the investigation of this method.

3.1 History, motivation

Operator decomposition is perhaps the most widely used technique for solving multiscale, multiphysics problems. The general approach is to decompose a model into components involving simpler physics over a relatively limited range of scales, and then to seek the solution of the entire system by using numerical solutions for the individual components. According to our knowledge, the first simple operator splitting procedure for partial differential equations was proposed, as an example, in 1957 by Bagrinovskii and Godunov in [3]. This was probably the first attempt to apply a splitting procedure. Different splitting procedures have been developed and/or studied in many scientific papers; see, for example, Csomós et al. [25], Dynatron, [33], Lanser and Verwer, [86], Marchuk, [93], [96], Penenko and Obraztsov, [108], Strang, [135], Tikhonov and Samarski, [142], Yanenko, [150]. A detailed theoretical study and analysis of some splitting procedures can be found

in [94, 97] and [158].
Results related to the use of splitting procedures in the field of air pollution modelling can be found in [8], [95] and [101]. Splitting procedures for air pollution models are also discussed in [73]. Basic description of the models and the possibilities of using some splitting technique can be found in [157, 158].
Splitting techniques are, to our knowledge, used in all large-scale air pollution models, which are run operationally with many scenarios (emission scenarios, climatic scenarios, etc.). In order to simplify the task, the operator splitting procedure has been introduced ([135, 93]), which is widely used for solving advection–diffusion problems (see e.g., in [78, 99]), Hamilton – Jacobi equation (see e.g., in [75, 79]), and Navier – Stokes equation (see e.g., in [22]), including modelling of turbulence and interfaces (see e.g., [104]). More applications can also be found in [78].

Splitting schemes are also useful from the point of view of the preservation of qualitative properties (also called as geometric integration), which was in the focus of the previous chapter. The classical operator splittings preserve structural features of the flow as long as the split sub-problems also share these properties. Important examples include symplecticity, volume preservation, symmetries, etc. In this sense, the splitting schemes can be considered as geometric integrators, and as such, they show smaller error growth than standard integrators. It is not surprising then that a systematic search for splitting methods of higher order of accuracy has taken place during the last two decades and a large number of them exist in the literature (see [62, 100, 119] and references therein) which have been specifically designed for different families of problems.

In what follows, we pass to the mathematical motivation and description of the operator splittings. According to the previous general description, we consider the Cauchy problem in the Banach space $\mathbf{X}$ as follows

$$\begin{cases} \dfrac{dw(t)}{dt} = Aw(t) \equiv \displaystyle\sum_{i=1}^{d} A_i w(t), \quad t \in (0, t^\star] \\ \\ w(0) = w_0, \end{cases} \tag{3.1.1}$$

where $w : [0, t^\star] \to \mathbf{X}$ denotes the unknown function, $w_0 \in \mathbf{X}$ is a given element and $A_i : \mathbf{X} \to \mathbf{X}$ $(i = 1, 2, \dots d)$ are given operators. (We note that the boundary conditions, if they exist, are included into the domains of definition of the operators.)
The exact solution of problem (3.1.1) can be given directly when the operator A is generating a C_0-semigroup. Then the solution can be written as

$$w(t) = \exp(tA)w(0), \tag{3.1.2}$$

where $\exp(tA)$ $(t \geq 0)$ denotes the semigroup generated by A. In this work we always assume that the operators A_i and A are generators, that is, they generate semigroups, which will be denoted by $\exp(tA_i)$ and $\exp(tA)$, respectively. (We note that when A_i or A is a linear bounded operator, then the corresponding semigroup can be obtained by substitution of the operator tA_i or tA into the Taylor series of the scalar exponential function $\exp z$. For more details see e.g., [37].)
Since the representation of the solution in the form (3.1.2) is formal, typically we must apply some numerical method. In fact, this means that we approximate the exponential function (semigroup) by some rational function, i.e., we use formulas

$$\exp(z) \sim r(z). \tag{3.1.3}$$

Then the algorithm of the numerical method reads as

$$y^{n+1} = r(\tau A)y^n, \tag{3.1.4}$$

where $\tau > 0$ denotes the discretization parameter (step size) and y^n is the approximation at the time level $t = n\tau$.

EXAMPLE 3.1.1 *Let us define the approximation function $r(z)$ as the stability function of the θ-method, (cf.* (2.4.4) *in Section 2.4), i.e.,*

$$r(z) \equiv r_\theta(z) = \frac{1 + (1-\theta)z}{1 - \theta z}. \tag{3.1.5}$$

Then, for $d = 2$ we have

$$y^{n+1} = r_\theta(\tau(A_1 + A_2))y^n,$$

where

$$r_\theta(\tau(A_1 + A_2)) = (\ I - \theta\tau(A_1 + A_2))^{-1}(\ I + (1-\theta)\tau(A_1 + A_2))$$

and I denotes the identity operator.

The major imperfection of this approach is that we do not use the special structure of the operator A, namely, that it is the sum of simpler operators. Operator splitting is such a method which overcomes this problem.

Remark 3.1.2 *Let us notice that the scheme (3.1.4) yields a time-discretization method on the mesh $\omega_\tau = \{t_n = n\tau, n = 0, 1, \dots N; N\tau = t^\star\}$. Hence, when we apply it directly to a time-dependent PDE of parabolic type, then at each time level this results in a PDE of elliptic type. However, when (3.1.4) means a semi-discrete problem (obtained by the method of lines) and A_h denotes the discretized operator, then the problem*

$$y^{n+1} = r(\tau A_h)y^n \tag{3.1.6}$$

means a vector iteration process (however, it requires the solution of the system of linear algebraic equations for the implicit methods), which can directly be used as a computational algorithm.

When for the solution of the problem (3.1.1) we apply the approximation (3.1.3), then, in fact, we should approximate the function $\exp(\sum_1^d z_i)$ in a suitable way.

We have different possibilities.

- First we approximate the exponential function by some computationally simpler (typically rational) function $r(z)$, and then we replace z by the operator sum multiplied by τ. Example 3.1.1 shows that this way is not effective for our aims because it does not separate the operators A_i.

- We use the approximation $\exp(\sum_{i=1}^d z_i) \sim \Phi(\varphi_1(z_1), \varphi_2(z_2), \dots \varphi_d(z_d))$, where the one-variable functions φ_i correspond to some approximations of the exponential function, and $\Phi : \mathbb{R}^d \to \mathbb{R}$ is a given function. After this step, we replace $\varphi_i(z_i)$ by the operator $\varphi_i(\tau A_i)$, so, the approximation means the following:

$$\exp\left(\sum_{i=1}^d \tau A_i\right) \sim \Phi(\varphi_1(\tau A_1), \varphi_2(\tau A_2), \dots \varphi_d(\tau A_d)). \tag{3.1.7}$$

The idea of the basic operator splittings is that – at least, in the first stage of the approximation process, – we select φ_i as the exponential function, i.e., $\varphi_i(s) = \exp(s)$. Hence, the approximation reads as

$$\exp(\textstyle\sum_{i=1}^{d} tA_i) \sim \Phi(\exp(\tau A_1), \exp(\tau A_2), \dots \exp(\tau A_d)). \tag{3.1.8}$$

The operator splittings are defined by the specification of the function Φ. As we will see, the major part of those operator splittings that will be investigated in this work can be identified by some given function Φ.

Remark 3.1.3 *Sometimes we distinguish the cases when we apply the operator splitting to the original (usually rather complex) partial differential operators or to its semi-discretized form. The first is called differential splitting, while the second one is called algebraic splitting. Hence, for the differential splitting: the discretization ordering is time → space; the separate operators usually represent physical phenomena (like convection, diffusion, reaction etc.); and boundary conditions are needed for each subproblem. Our objective is decoupling the physical effects in the complex initial- boundary value problems. For the algebraic splitting: the discretization ordering is space → time; the separate operators represent discrete operators (usually sparse matrices of arbitrary origin); and boundary conditions are built into the operators beforehand. Our objective is segregated solution of the semi-discretized equations. We would like to emphasize that in both cases our subject is an (abstract) Cauchy problem, where the time derivative is still un-discretized. In the sequel we do not make a distinction between these approaches and we will use the unified notation "operator splitting". Hence, at this level, our approach is in contrast to the one where the decomposition is done for the already fully discretized schemes. This latter, (like Alternating Direction Implicit (ADI) method, Douglas-Rachford scheme, a lot of useful finite difference schemes in the Russian literature as predictor-corrector schemes, Iliin's generalization of the Douglas-Rachford scheme, locally one-dimensional (fractional step) schemes, and many others) has a vast literature and is widely examined in the different monographs. For the details we refer to [73, 133, 116, 151, ?]. Clearly, when we discretize the split sub-problems in the above split (abstract) Cauchy problems, we get again fully discretized schemes, which, in many cases, coincide with the second above listed group of the methods. This question will be investigated in Section 3.5.1 in more detail.*

The operator splitting can be considered as a special time-discretization method. Therefore, in analogue with the local approximation error, (see Definition 2.3.42 in Section 2.3.5), we can introduce the following notion.[1]

Definition 3.1.4 *For a given splitting process, the error after the first step, that is the expression*

$$Err_{\mathrm{spl}}(w_0, \tau) = Err_{\mathrm{spl}}(\tau) = w(\tau) - w_{\mathrm{spl}}^{(N)}(\tau) \tag{3.1.9}$$

is called local splitting error of the given splitting method[2]*.*

[1] This term is also called local truncation error, widely used in the theory of numerical methods for ODE's.

[2] The upper case index in the split solution $w_{\mathrm{spl}}^{(N)}$ refers to the number of the time sub-intervals in the splitting. Since, as we will see later, the splitting procedure is a time discretization process, for its convergence we analyze the sequence $\{w_{\mathrm{spl}}^{(N)}\}, N = 1, 2, \dots$.

We assume that the split sub-problems can be solved exactly. When the operators A_i in (3.1.1) are pairwise commuting, the operator splitting results in the exact solution, i.e., the local splitting error vanishes. However, this is not usually the case and $Err_{\rm spl}$ differs from zero. Then it is quite natural to require its convergence to zero when τ tends to zero. This motivates the introduction of the following

Definition 3.1.5 *We say that a given operator splitting is of order p when $Err_{\rm spl}(\tau) = \mathcal{O}(\tau^{p+1})$. The operator splitting method is called weakly consistent when $p > 0$.*[3]

Remark 3.1.6 *The above notion of weak consistency of a splitting method differs from the traditional notion of consistency of a numerical time discretization method used in the Lax-Richtmyer theory. (C.f. Definition 3.4.6 in the Section 3.4.3.) We note that the latter is a stronger one. However, they are in close relation: the weak consistency of a splitting method is a necessary condition of its consistency as a numerical method. Moreover, the order of the weak consistency influences the order of the convergence of the method, since it cannot be higher than p. Therefore a higher-order splitting, as a rule, results in faster convergence. (For more details see Section 3.4.3, [89] and [113].)*

3.2 Classical operator splittings: the sequential splitting and the Strang-Marchuk splitting

In this part we introduce and analyze the sequential splitting and the Strang-Marchuk splitting which are the most traditional and widely used operator splitting methods.

Definition 3.2.1 *The operator splitting corresponding to the choice*

$$\Phi(s_1, s_2, \ldots, s_d) = \prod_{i=1}^{d} s_{d+1-i} \tag{3.2.1}$$

is called sequential splitting. The operator splitting obtained by the choice

$$\Phi(s_1, s_2, \ldots s_d) = \left(\prod_{i=1}^{d-1} \frac{s_i}{2}\right)(s_d)\left(\prod_{i=1}^{d-1} \frac{s_{d-i}}{2}\right) \tag{3.2.2}$$

is called Strang-Marchuk splitting.

If we apply the above approximations on the mesh ω_τ, then the one step method (3.1.4) reads as

$$w_{\rm spl}^{(N)}((n+1)\tau) = r_{\rm spl}(\tau A) w_{\rm spl}^{(N)}(n\tau), \quad n = 0, 1, \ldots N, \tag{3.2.3}$$

where

$$r_{\rm spl}(\tau A) = r_{\rm seq}(\tau A) := \prod_{i=1}^{d} \exp(\tau A_{d+1-i}) \tag{3.2.4}$$

for the sequential splitting[4], and

$$r_{\rm spl}(\tau A) = r_{\rm SM}(\tau A) := \left(\prod_{i=1}^{d-1} \exp(\frac{1}{2}\tau A_i)\right) \exp(\tau A_d) \left(\prod_{i=1}^{d-1} \exp(\frac{1}{2}\tau A_{d-i})\right) \tag{3.2.5}$$

[3] When the operator splitting is exact, i.e., the local splitting error is zero, then we set $p = \infty$.

[4] It is also called Lie-Trotter or Yanenko splitting.

for the Strang-Marchuk splitting[5].

In the realization of the operator splitting algorithm (3.2.3) the central query is the computation of the exponential, namely, the computation of $w := \exp(\tau A_i)v$ with some given element v. Since clearly $w = w(\tau)$, where $w(t)$ is the solution of the Cauchy problem defined by the operator A_i and the initial value v on the time interval $[0, \tau]$, therefore the algorithm for the split solutions in the sequential splitting and Strang-Marchuk splitting are the following.

1. *Sequential splitting [3].* For each $n = 1, 2, \ldots, N$ we successively solve the Cauchy problems:

$$\begin{aligned} &\frac{dw_i^n}{dt}(t) = A_i w_i^n(t), \qquad (n-1)\tau < t \le n\tau, \\ &w_i^n((n-1)\tau) = w_{i-1}^n(n\tau), \end{aligned} \tag{3.2.6}$$

   where $i = 1, 2, \ldots, d$. The split solution is defined as

$$w_{\text{seq}}^{(N)}(n\tau) = w_d^n(n\tau). \tag{3.2.7}$$

   Here $w_0^n(n\tau) = w_{\text{seq}}^{(N)}((n-1)\tau)$, and $w_{\text{seq}}^{(N)}(0) = w(0)$ is known from the original continuous problem (3.1.1).

   That is, the algorithm is the following:

$$\underbrace{A_1 \to A_2 \to \cdots A_d}_{\text{step 1}} \Rightarrow \underbrace{A_1 \to A_2 \to \cdots A_d}_{\text{step 2}} \Rightarrow \cdots \Rightarrow \underbrace{A_1 \to A_2 \to \cdots A_d}_{\text{step N}}.$$

2. *Strang-Marchuk splitting [93, 135].* For each fixed $n = 1, 2, \ldots, N$ we solve the following Cauchy problems:

   For the values $i = 1, 2, \ldots, d-1$ we solve the problems

$$\begin{aligned} &\frac{dw_i^n}{dt}(t) = A_i w_i^n(t), \qquad (n-1)\tau < t \le (n-0.5)\tau, \\ &w_i^n((n-1)\tau) = w_{i-1}^n((n-0.5)\tau). \end{aligned} \tag{3.2.8}$$

   Then we define the solution of the problem

$$\begin{aligned} &\frac{dw_d^n}{dt}(t) = A_d w_d^n(t), \qquad (n-1)\tau < t \le n\tau, \\ &w_d^n((n-1)\tau) = w_{d-1}^n((n-0.5)\tau). \end{aligned} \tag{3.2.9}$$

   Finally, at a fixed n for the values $i = d+1, d+2, \ldots 2d-1$ we solve the problems

$$\begin{aligned} &\frac{dw_i^n}{dt}(t) = A_{2d-i} w_i^n(t), \qquad (n-0.5)\tau < t \le n\tau, \\ &w_i^n((n-0.5)\tau) = w_{i-1}^n(n\tau). \end{aligned} \tag{3.2.10}$$

[5]It is also called second order leapfrog, Störmer, Verlet splitting.

The split solution is

$$w_{\rm SM}^{(N)}(n\tau) = w_{2d-1}^{n}(n\tau). \tag{3.2.11}$$

Here $w_0^n((n-0.5)\tau) = w_{\rm SM}^N((n-1)\tau)$, and $w_{\rm SM}^{(N)}(0) = w(0)$ is known from (3.1.1). That is, the algorithm is the following:

$$\underbrace{\frac{1}{2}A_1 \to \frac{1}{2}A_2 \to \cdots \frac{1}{2}A_{d-1}}_{\text{step 1a}} \to \underbrace{A_d}_{\text{step 1b}} \to \underbrace{\frac{1}{2}A_{d-1} \to \frac{1}{2}A_{d-2} \to \cdots \frac{1}{2}A_1}_{\text{step 1c}} \Rightarrow$$

$$\cdots\cdots\cdots$$

$$\Rightarrow \underbrace{\frac{1}{2}A_1 \to \frac{1}{2}A_2 \to \cdots \frac{1}{2}A_{d-1}}_{\text{step Na}} \to \underbrace{A_d}_{\text{step Nb}} \to \underbrace{\frac{1}{2}A_{d-1} \to \frac{1}{2}A_{d-2} \to \cdots \frac{1}{2}A_1}_{\text{step Nc}}.$$

In the following we analyze the order of the sequential splitting and the Strang-Marchuk splitting for d linear bounded sub-operators. (For $d = 2$ and $d = 3$ similar results can be found in [73].)

Theorem 3.2.2 *For linear and bounded operators A_i $(i = 1, 2, \ldots, d)$ the sequential splitting has first, while the Strang-Marchuk splitting has second order of accuracy.*

Proof. The exact solution of (3.1.1) at $t = \tau$ is

$$w(\tau) = \exp(\tau A)w(0) = \sum_{n=0}^{\infty} \frac{1}{n!}\tau^n A^n w_0 = (I + \tau A + \frac{1}{2}\tau^2 A^2)w_0 + \mathcal{O}(\tau^3), \tag{3.2.12}$$

where, as before, $\mathcal{O}$ denotes the Landau symbol (see p.35). Since $A = \sum_{i=1}^{d} A_i$, we get

$$w(\tau) = \left(I + \tau \sum_{i=1}^{d} A_i + \frac{\tau^2}{2} \sum_{i,j=1}^{d} A_i A_j \right) w_0 + \mathcal{O}(\tau^3). \tag{3.2.13}$$

For the sequential splitting the split solution at $t = \tau$ is

$$w_{\rm seq}^{(N)}(\tau) = \prod_{i=1}^{d} \exp(\tau A_i) w_0 = \prod_{i=1}^{d} \left(I + \tau A_i + \frac{\tau^2}{2} A_i^2 \right) w_0 + \mathcal{O}(\tau^3). \tag{3.2.14}$$

Hence

$$w_{\rm seq}^{(N)}(\tau) = \left(I + \tau \sum_{i=1}^{d} A_i + \frac{\tau^2}{2} \sum_{i=1}^{d} A_i^2 + \tau^2 \sum_{\substack{i,j=1 \\ i<j}}^{d} A_i A_j \right) w_0 + \mathcal{O}(\tau^3). \tag{3.2.15}$$

Using the expressions (3.2.13) and (3.2.15), for the local splitting error we obtain

$$\begin{aligned} Err_{\rm seq}(\tau) &= \frac{\tau^2}{2} \left(\sum_{\substack{i,j=1 \\ i>j}}^{d} A_i A_j - \sum_{\substack{i,j=1 \\ i<j}}^{d} A_i A_j \right) w_0 + \mathcal{O}(\tau^3) = \\ &= \frac{\tau^2}{2} \sum_{\substack{i,j=1 \\ i>j}}^{d} (A_i A_j - A_j A_i)\, w_0 + \mathcal{O}(\tau^3). \end{aligned} \tag{3.2.16}$$

Hence, for arbitrary chosen operators A_i the right-hand side of (3.2.16) is $\mathcal{O}(\tau^2)$, which yields that the sequential splitting is of first order.

In order to prove the statement for the Strang-Marchuk splitting, we have to show that the operator $r_{\mathrm{SM}}(\tau A)$ in (3.2.5) approximates the operator $\exp(\tau A)$ in third order. Since

$$\begin{aligned} r_{\mathrm{SM}}(\tau A) = \prod_{i=1}^{d-1} \left(I + \frac{\tau}{2} A_i + \frac{\tau^2}{8} A_i^2 \right) \cdot \left(I + \tau A_d + \frac{\tau^2}{2} A_d^2 \right) \cdot \\ \cdot \prod_{i=1}^{d-1} \left(I + \frac{\tau}{2} A_{d-i} + \frac{\tau^2}{8} A_{d-i}^2 \right) + \mathcal{O}(\tau^3), \end{aligned} \tag{3.2.17}$$

we can write

$$r_{\mathrm{SM}}(\tau A) = B_1 B_2 B_3 + \mathcal{O}(\tau^3), \tag{3.2.18}$$

where

$$B_1 = \prod_{i=1}^{d-1} \left(I + \frac{\tau}{2} A_i + \frac{\tau^2}{8} A_i^2 \right) = I + \frac{\tau}{2} \sum_{i=1}^{d-1} A_i + \frac{\tau^2}{4} \sum_{\substack{i,j=1 \\ i<j}}^{d-1} A_i A_j + \frac{\tau^2}{8} \sum_{i=1}^{d-1} A_i^2 + \mathcal{O}(\tau^3); \tag{3.2.19}$$

$$B_2 = I + \tau A_d + \frac{\tau^2}{2} A_d^2; \tag{3.2.20}$$

and

$$B_3 = \prod_{i=1}^{d-1} \left(I + \frac{\tau}{2} A_{d-i} + \frac{\tau^2}{8} A_{d-i}^2 \right) = I + \frac{\tau}{2} \sum_{i=1}^{d-1} A_i + \frac{\tau^2}{4} \sum_{\substack{i,j=1 \\ i<j}}^{d-1} A_j A_i + \frac{\tau^2}{8} \sum_{i=1}^{d-1} A_i^2 + \mathcal{O}(\tau^3). \tag{3.2.21}$$

After some simple but tedious computation we get

$$B_1 B_2 B_3 = I + \tau \sum_{i=1}^{d} A_i + \frac{\tau^2}{2} \sum_{i=1}^{d} A_i^2 + \frac{\tau^2}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^{d} A_i A_j + \mathcal{O}(\tau^3), \tag{3.2.22}$$

which proves the statement. ■

In the following we give the conditions under which the order of the sequential splitting and the Strang-Marchuk splitting are higher.

The error formula (3.2.16) can be rewritten as

$$Err_{\mathrm{seq}}(\tau) = \frac{\tau^2}{2} \sum_{\substack{i,j=1 \\ i>j}}^{d} [A_i, A_j]\, w_0 + \mathcal{O}(\tau^3), \tag{3.2.23}$$

where

$$[A_i, A_j] = A_i A_j - A_j A_i \tag{3.2.24}$$

denotes the commutator of the operators A_i and A_j. Hence, we have

Theorem 3.2.3 *The sequential splitting has higher than first order accuracy if and only if the condition*

$$\sum_{\substack{i,j=1\\ i>j}}^{d} [A_i, A_j] = 0 \tag{3.2.25}$$

is satisfied.

Theorem 3.2.3 shows that the pairwise commutativity of the operators is a sufficient condition for (3.2.25). (This is quite natural because in this case the sequential splitting is exact.) Moreover, for $d = 2$ the commutativity is a necessary and sufficient condition. Therefore, when there are only two operators in the operator sum in (3.1.1), then there are only two cases: either the sequential splitting is exact or it has first order accuracy. However, it is not yet clear whether

- the pairwise commutativity is a necessary condition of the higher-order for $d > 2$;
- the local splitting error vanishes only under the pairwise commutativity condition.

We analyze these problems for $d = 3$ in a simple matrix case. We consider the matrix

$$A = \begin{bmatrix} 4 & 2 \\ 0 & 3 \end{bmatrix} \tag{3.2.26}$$

and let us split it into the sum $A_1 + A_2 + A_3$ with

$$A_1 = A_3 = \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} -2 & 0 \\ 0 & -1 \end{bmatrix}. \tag{3.2.27}$$

Then

$$e^{tA} = \begin{bmatrix} e^{4t} & 2e^{3t}(e^t - 1) \\ 0 & e^{3t} \end{bmatrix}, \tag{3.2.28}$$

$$e^{tA_1} = e^{tA_3} = \begin{bmatrix} e^{3t} & e^{2t}(e^t - 1) \\ 0 & e^{2t} \end{bmatrix}, \quad \text{and} \quad e^{tA_2} = \begin{bmatrix} e^{-2t} & 0 \\ 0 & e^{-t} \end{bmatrix}. \tag{3.2.29}$$

In this example A_1 and A_2 do not commute, since

$$[A_1, A_2] = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \tag{3.2.30}$$

but

$$e^{\tau A_1} e^{\tau A_2} e^{\tau A_3} = e^{\tau A}, \tag{3.2.31}$$

and so the sequential splitting is exact. Hence, the above example implies the following fact:

Theorem 3.2.4 *In the sequential splitting the commutation of each pair of sub-operators is not necessary for vanishing of the local splitting error or for achieving higher order when the number of sub-operators is greater than two*

As before, for the Strang-Marchuk splitting the commutativity of the sub-operators is sufficient for zero splitting error. However the necessity of the commutativity condition is unclear even for $d = 2$. Let us observe that the Strang-Marchuk splitting with two operators A_1 and A_2 is the same as the sequential splitting for three operators $0.5A_1$, A_2 and $0.5A_1$, respectively. Hence, by use of Theorem 3.2.4, we get

Theorem 3.2.5 *The commutation of the sub-operators in the Strang-Marchuk splitting is not a necessary condition for vanishing local splitting error.*

In the following, for $d = 2$ we give a condition under which the Strang-Marchuk splitting has third order of accuracy.
We have to analyze the $\mathcal{O}(\tau^3)$ term for the difference

$$\exp(\tau(A_1 + A_2)) - \exp\left(\frac{\tau}{2}A_1\right)\exp\left(\tau A_2\right)\exp\left(\frac{\tau}{2}A_1\right).$$

For the first exponential we have

$$\exp(\tau(A_1 + A_2)) = I + \tau\left(A_1 + A_2\right) + \frac{\tau^2}{2}\left(A_1 + A_2\right)^2 + \frac{\tau^3}{6}\left(A_1 + A_2\right)^3 + \mathcal{O}(\tau^4). \tag{3.2.32}$$

For the second expression we obtain

$$\begin{aligned}
&\exp\left(\frac{\tau}{2}A_1\right)\exp\left(\tau A_2\right)\exp\left(\frac{\tau}{2}A_1\right) = \left(I + \frac{\tau}{2}A_1 + \frac{\tau^2}{8}A_1^2 + \frac{\tau^3}{48}A_1^3\right)\cdot \\
&\cdot\left(I + \tau A_2 + \frac{\tau^2}{2}A_2^2 + \frac{\tau^3}{6}A_2^3\right)\cdot\left(I + \frac{\tau}{2}A_1 + \frac{\tau^2}{8}A_1^2 + \frac{\tau^3}{48}A_1^3\right) + \mathcal{O}(\tau^4) = \\
&= I + \tau(A_1 + A_2) + \frac{\tau^2}{2}(A_1^2 + A_2^2 + A_1A_2 + A_2A_1) + \\
&+\tau^3\left(\frac{1}{6}A_1^3 + \frac{1}{6}A_2^3 + \frac{1}{8}A_1^2A_2 + \frac{1}{8}A_2^2A_1 + \frac{1}{4}A_1A_2^2 + \frac{1}{4}A_2^2A_1 + \frac{1}{4}A_1A_2A_1\right) + \mathcal{O}(\tau^4).
\end{aligned} \tag{3.2.33}$$

Then, by use of the expressions (3.2.32) and (3.2.33), the local splitting error of the Strang-Marchuk splitting can be expressed by the commutators, namely we have

$$\begin{aligned}
Err_{\mathrm{SM}}(\tau) &= \frac{\tau^3}{12}\left(\frac{1}{2}A_1^2A_2 + A_2A_1^2 - A_1A_2^2 - A_2^2A_1 - A_1A_2A_1 + 2A_2A_1A_2\right)w_0 + \mathcal{O}(\tau^4) = \\
&= \left[\frac{1}{2}A_1 + A_2, [A_1, A_2]\right]w_0 + \mathcal{O}(\tau^4).
\end{aligned} \tag{3.2.34}$$

Hence, we have the following

Theorem 3.2.6 *The Strang-Marchuk splitting is of third order if and only if the operator* $0.5A_1 + A_2$ *commutes with the commutator* $[A_1, A_2]$*, i.e., when the condition*

$$[0.5A_1 + A_2, [A_1, A_2]] = 0 \tag{3.2.35}$$

holds.

Remark 3.2.7 *With some cumbersome computation, one can show that under the condition* (3.2.35) *the coefficient of not only* τ^3*, but also of* τ^4 *vanishes, i.e., the condition* (3.2.35) *guarantees the fourth-order accuracy [41].*

3.3 New operator splittings and their analysis

The traditional operator splittings are widely used, but they have some drawbacks. Namely, the sequential splitting is computationally simple, but it has low accuracy; the Strang-Marchuk splitting is more accurate but it requires more computational work, and its algorithm cannot be parallelized in a natural way (on the operator level). Moreover, for practical problems the condition of the higher order is rather unrealistic (c.f. [8, 31, 32]). This gives the motivation to define new operator splitting methods.

3.3.1 Weighted sequential splitting

The algorithm of the sequential splitting depends on the ordering of the operators. We can construct a further type of splitting techniques by symmetrizing the sequential splitting in the following manner. In each time step we apply the sequential splitting both in the order $A_1 \to A_2 \to \cdots \to A_d$ and $A_d \to A_{d-1} \to \cdots \to A_1$, and at the end of the time steps we combine the obtained solutions by taking a weighted average of the results. This means that the function Φ in (3.1.8) is defined as

$$\Phi(s_1, s_2, \dots s_d) = \theta \prod_{i=1}^{d} s_i + (1-\theta) \prod_{i=1}^{d} s_{d+1-i}, \tag{3.3.1}$$

where $\theta \in [0,1]$ is some fixed weighting parameter.
If we apply the above approximations on the mesh ω_τ, then (3.1.4) reads as

$$w_{\text{wss}}^{(N)}((n+1)\tau) = r_{\text{wss}}(\tau A) w_{\text{wss}}^{(N)}(n\tau), \quad n = 0, 1, \dots, N, \tag{3.3.2}$$

where

$$r_{\text{wss}}(\tau A) = \theta \prod_{i=1}^{d} \exp(\tau A_i) + (1-\theta) \prod_{i=1}^{d} \exp(\tau A_{d+1-i}). \tag{3.3.3}$$

Obviously, the values $\theta = 0$ and $\theta = 1$ are not interesting because these values result in the usual sequential splitting, however, as we shall see later, the value $\theta = 0.5$ is of special interest.

Definition 3.3.1 *The operator splitting method corresponding to the algorithm* (3.3.2) *is called weighted sequential splitting. The weighted sequential splitting with the choice* $\theta = 0.5$ *is called symmetrically weighted sequential splitting.*

Remark 3.3.2 *As an interesting historical fact, we note that the symmetrically weighted sequential splitting was already mentioned in the work [134], as a symmetrized version of the sequential splitting. However, this method was neither theoretically investigated, nor applied. (Perhaps due to the increased computational work on traditional computers.) The first theoretical investigation was done in [25], and later the method was successfully applied, among others, in air pollution modelling [16].*

The realization of the weighted sequential splitting is the following. For each fixed $n = 1, 2, \dots, N$ we solve the Cauchy problems:

$$\begin{aligned} &\frac{dv_i^n}{dt}(t) = A_i v_i^n(t), \qquad (n-1)\tau < t \le n\tau, \\ &v_i^n((n-1)\tau) = v_{i-1}^n(n\tau), \end{aligned} \tag{3.3.4}$$

and

$$\begin{aligned} &\frac{du_i^n}{dt}(t) = A_{d+1-i}u_i^n(t), \qquad (n-1)\tau < t \le n\tau, \\ &u_i^n((n-1)\tau) = u_{i-1}^n(n\tau), \end{aligned} \tag{3.3.5}$$

where $i = 1, 2, \ldots, d$. The split solution is defined as

$$w_{\text{wss}}^{(N)}(n\tau) = \theta u_d^n(n\tau) + (1-\theta)v_d^n(n\tau). \tag{3.3.6}$$

Here $v_0^n(n\tau) = u_0^n(n\tau) = w_{\text{wss}}^N((n-1)\tau)$, and $w_{\text{wss}}^{(N)}(0) = w(0)$ is known from (3.1.1). That is, the algorithm can be schematically given as

$$\left.\begin{aligned} A_1 \to A_2 \to \cdots A_d \\ A_d \to A_{d-1} \to \cdots A_1 \\ \text{(step 1)} \end{aligned}\right\} \Longrightarrow \cdots \Longrightarrow \left.\begin{aligned} A_1 \to A_2 \to \cdots A_d \\ A_d \to A_{d-1} \to \cdots A_1 \\ \text{(step N)} \end{aligned}\right\}. \tag{3.3.7}$$

In the following we determine the order of the weighted sequential splitting. Using (3.2.15), for the split solution we get

$$\begin{aligned} w_{\text{wss}}^{(N)}(\tau) = (1-\theta)\left(I + \tau\sum_{i=1}^{d} A_i + \frac{\tau^2}{2}\sum_{i=1}^{d} A_i^2 + \tau^2 \sum_{\substack{i,j=1 \\ i<j}}^{d} A_iA_j\right) w_0 + \\ +\theta\left(I + \tau\sum_{i=1}^{d} A_i + \frac{\tau^2}{2}\sum_{i=1}^{d} A_i^2 + \tau^2 \sum_{\substack{i,j=1 \\ i>j}}^{d} A_iA_j\right) w_0 + \mathcal{O}(\tau^3). \end{aligned} \tag{3.3.8}$$

Using the expressions (3.2.13) and (3.3.8), for the local splitting error of the weighted sequential splitting we obtain

$$\begin{aligned} Err_{\text{wss}}(\tau) = \tau^2\left((0.5-\theta)\sum_{\substack{i,j=1 \\ i<j}}^{d} A_iA_j + (\theta-0.5)\sum_{\substack{i,j=1 \\ i>j}}^{d} A_iA_j\right) w_0 + \mathcal{O}(\tau^3) = \\ = \tau^2(0.5-\theta)\sum_{\substack{i,j=1 \\ i\neq j}}^{d} \text{sign}(j-i)A_iA_jw_0 + \mathcal{O}(\tau^3). \end{aligned} \tag{3.3.9}$$

Hence we have

Theorem 3.3.3 *For arbitrarily chosen operators, the weighted sequential splitting has first order accuracy for any $\theta \neq 0.5$, and only the choice $\theta = 0.5$ (symmetrically weighted sequential splitting) has second order accuracy.*

Let us analyze the higher-order of the weighted sequential splitting. The local splitting error, according to (3.3.9), can be rewritten with the commutators as

$$Err_{\text{wss}}(\tau) = \tau^2(0.5-\theta)\sum_{\substack{i,j=1 \\ i<j}}^{d} [A_i, A_j]w_0 + \mathcal{O}(\tau^3). \tag{3.3.10}$$

Therefore, for $\theta \neq 0.5$ the method has second order accuracy if and only if the condition

$$\sum_{\substack{i,j=1 \\ i<j}}^{d} [A_i, A_j] = 0 \tag{3.3.11}$$

holds. As we have already seen for the sequential splitting, the pairwise commutativity is only a sufficient, but not a necessary condition for (3.3.11). For the vanishing of the local splitting error, the pairwise commutativity is not necessary, either. In order to show this, we consider the symmetrically weighted sequential splitting for two operators. Namely, we choose the matrices

$$A = \begin{bmatrix} 5 & 1 \\ 0 & 3 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3.3.12}$$

Then

$$\exp tA = \begin{bmatrix} \exp 5t & 0.5 \exp 5t - 0.5 \exp 3t \\ 0 & \exp 3t \end{bmatrix}, \tag{3.3.13}$$

$$\exp tA_1 = \begin{bmatrix} \exp 3t & \exp 3t - \exp 2t \\ 0 & \exp 2t \end{bmatrix}, \quad \text{and} \quad \exp tA_2 = \begin{bmatrix} \exp 2t & 0 \\ 0 & \exp t \end{bmatrix}. \tag{3.3.14}$$

It is easy to check that $[A_1, A_2] \neq 0$, but $r_{\text{swss}}(\tau A) = \exp \tau A$.

Theorem 3.3.4 *The commutation of the sub-operators in the symmetrically weighted sequential splitting is not a necessary condition for vanishing local splitting error, neither for the higher order.*

For two operators ($d = 2$) we can give a condition for the sub-operators under which the symmetrically weighted sequential splitting has third order accuracy. According to (3.3.3), for the symmetrically weighted sequential splitting we have

$$\begin{aligned} r_{\text{swss}}(\tau A) &= 0.5 \left(I + \tau A_1 + \frac{\tau^2}{2} A_1^2 + \frac{\tau^3}{6} A_1^3 \right) \left(I + \tau A_2 + \frac{\tau^2}{2} A_2^2 + \frac{\tau^3}{6} A_2^3 \right) + \\ &+ 0.5 \left(I + \tau A_2 + \frac{\tau^2}{2} A_2^2 + \frac{\tau^3}{6} A_2^3 \right) \left(I + \tau A_1 + \frac{\tau^2}{2} A_1^2 + \frac{\tau^3}{6} A_1^3 \right) + \mathcal{O}(\tau^4) = \\ &= \left(I + \tau(A_1 + A_2) + \frac{\tau^2}{2}(A_1 + A_2)^2 + \frac{\tau^3}{6}(A_1^3 + A_2^3) \right) + \\ &+ \frac{\tau^3}{4}(A_1^2 A_2 + A_1 A_2^2 + A_2^2 A_1 + A_2 A_1^2) + \mathcal{O}(\tau^4). \end{aligned} \tag{3.3.15}$$

Comparing (3.3.15) with the expression for $\exp(\tau(A_1 + A_2))$ in (3.2.32), for the local splitting error we get

$$\begin{aligned} Err_{\text{swss}}(\tau) &= \frac{\tau^3}{6} \left(-\frac{1}{2} A_1^2 A_2 - \frac{1}{2} A_1 A_2^2 - \frac{1}{2} A_2^2 A_1 - \frac{1}{2} A_2 A_1^2 + A_1 A_2 A_1 + A_2 A_1 A_2 \right) w_0 + \\ &+ \mathcal{O}(\tau^4) = -\frac{\tau^3}{12} [A_1 - A_2, [A_1, A_2]] w_0 + \mathcal{O}(\tau^4). \end{aligned} \tag{3.3.16}$$

Hence, we get

Theorem 3.3.5 *Assume that the difference of the operators commutes with the commutator of the operators, i.e.,*

$$[A_1 - A_2, [A_1, A_2]] = 0. \tag{3.3.17}$$

Then the symmetrically weighted sequential splitting has third order accuracy.

Remark 3.3.6 *Using some cumbersome computation, one can show that under the conditions* (3.3.17) *and*

$$[A_1, A_2]^2 = 0 \tag{3.3.18}$$

the symmetrically weighted sequential splitting has fourth order accuracy [25].

3.3.2 Additive splitting

The split solutions that we have considered before are weakly consistent at the mesh points of the fixed mesh ω_τ. When we solve the split sub-problems numerically (which is almost always necessary), for the step-size of the numerical method we can use smaller values than the splitting step-size. Our aim is to use this intermediate values as the approximation to the unknown function. (We discuss this question in Section 3.5 in more details.) Therefore, to investigate the consistency, we have to compare the split solution with the exact solution at any point of the split time-interval, i.e., on the whole $[0, \tau]$. First we should extend the split (discrete) solution to this interval. It is rather plausible to do it, e.g., for the sequential splitting, as follows:

$$w_{\text{seq}}^{(N)}(t) = \exp(tA_2)\exp(\tau A_1)w_0. \tag{3.3.19}$$

Hence, for the local splitting error at any time $t \in [0, \tau]$ we get

$$w(t) - w_{\text{seq}}^{(N)}(t) = (t - \tau)A_1 w_0 + \mathcal{O}(t^2). \tag{3.3.20}$$

This shows that $Err_{seq}(\tau) = \mathcal{O}(\tau)$, i.e., the method is not weakly consistent for arbitrary $t \to 0$ and a fixed τ, only in the case where $t = \tau \to 0$.

Remark 3.3.7 *For the Strang-Marchuk splitting and symmetrically weighted sequential splitting it is quite natural to define the extension of the split solution to the whole interval* $[0, \tau]$ *as*

$$w_{\text{SM}}^{(N)}(t) = \exp(0.5tA_1)\exp(\tau A_2)\exp(0.5\tau A_1)w_0. \tag{3.3.21}$$

and

$$w_{\text{swss}}^{(N)}(t) = 0.5\left(\exp(tA_1)\exp(\tau A_2) + \exp(tA_2)\exp(\tau A_1)\right)w_0. \tag{3.3.22}$$

An easy computation shows that they are weakly consistent only at $t = \tau$.

Hence, it is reasonable to introduce the following definition.

Definition 3.3.8 *A splitting method with a fixed step-size* τ *is called continuously weakly consistent when*

$$Err_{\text{spl}}(t) = w(t) - w_{\text{spl}}^{(N)}(t) = \mathcal{O}(t^{p+1}) \tag{3.3.23}$$

for all $t \in (0, \tau]$ *with some* $p > 0$.

The algorithm of the additive splitting is based on the following idea. Simultaneously (in a parallel way) we solve the Cauchy problems consisting only of one operator A_i and using the same initial value for each sub-problem, namely, the split solution at the previous time level. Then, by special averaging of the results (in order to ensure the weak consistency), we define the split solution at the next time level.

Definition 3.3.9 *The operator splitting corresponding to the choice*

$$\Phi(s_1, s_2, \ldots, s_d) = \sum_{i=1}^{d} s_i - (d-1) \tag{3.3.24}$$

is called additive splitting.

If we apply the above approximations on the mesh ω_τ, then the iteration (3.1.4) reads as

$$w_{\text{as}}^{(N)}((n+1)\tau) = r_{\text{as}}(\tau A) w_{\text{as}}^{(N)}(n\tau), \quad n = 0, 1, \ldots, N, \tag{3.3.25}$$

where

$$r_{\text{as}}(\tau A) := \sum_{i=1}^{d} \exp(\tau A_i) - (d-1)I. \tag{3.3.26}$$

In order to define the required exponentials in the algorithm, we solve the corresponding Cauchy problems, namely the realization is the following.

1. For each $n = 1, 2, \ldots, N$ we solve the Cauchy problems:

$$\begin{aligned} &\frac{dw_i^n}{dt}(t) = A_i w_i^n(t), \qquad (n-1)\tau < t \le n\tau, \\ &w_i^n((n-1)\tau) = w_{\text{as}}^{(N)}((n-1)\tau), \end{aligned} \tag{3.3.27}$$

   where $i = 1, 2, \ldots, d$.

2. The split solution is defined as

$$w_{\text{as}}^{(N)}(n\tau) = \sum_{i=1}^{d} w_i^n(n\tau) - (d-1) w_{\text{as}}^{(N)}((n-1)\tau). \tag{3.3.28}$$

   Here $w_{\text{as}}^{(N)}(0) = w(0)$ is known from the original continuous problem (3.1.1).

That is, the algorithm is as follows:

$$\left.\begin{matrix} A_1 \\ \vdots \\ A_d \\ (\text{step } 1) \end{matrix}\right\} \Longrightarrow \left.\begin{matrix} A_1 \\ \vdots \\ A_d \\ (\text{ step } 2) \end{matrix}\right\} \quad \ldots \quad \Longrightarrow \left.\begin{matrix} A_1 \\ \vdots \\ A_d \\ (\text{ step N}) \end{matrix}\right\}$$

In the following we compute the order of the additive splitting. For the split solution we get

$$w_{\text{as}}^{(N)}(\tau) = \sum_{i=1}^{d} \left(I + \tau A_i + \frac{\tau^2}{2} A_i^2 \right) w_0 + \mathcal{O}(\tau^3) - (d-1) w_0 =$$

$$= \left(I + \tau \sum_{i=1}^{d} A_i + \frac{\tau^2}{2} \sum_{i=1}^{d} A_i^2 \right) w_0 + \mathcal{O}(\tau^3).$$

Using the formula (3.2.13) for the solution of the un-split problem, for the local splitting error of the additive splitting we obtain

$$Err_{\text{as}}(\tau) = \frac{\tau^2}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^{d} A_i A_j w_0 + \mathcal{O}(\tau^3). \tag{3.3.29}$$

Let us notice that (3.3.29) also holds for an arbitrary $t \in (0, \tau]$, i.e.,

$$Err_{\text{as}}(t) = \frac{t^2}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^{d} A_i A_j w_0 + \mathcal{O}(t^3). \tag{3.3.30}$$

Hence we have

Theorem 3.3.10 *For arbitrary operators the additive splitting is continuously consistent, and it has first order accuracy.*

As opposed to the sequential splitting, the Strang-Marchuk splitting and the weighted sequential splitting, the order of the additive splitting is not influenced by the commutativity of the operators. However, if any pair A_i and A_j anticommute, i.e., if $A_i A_j + A_j A_i = 0$, then the local splitting error has second order. We note that in contrast to the previous splittings, where the pairwise commutativity resulted in the exactness of the splitting, for the additive splitting the anticommutativity of the operators does not imply higher than second-order accuracy.

To check the theoretically obtained order, we applied the additive splitting to two systems of simple ordinary differential equations with constant coefficients, where both the original problems and the sub-problems were solved exactly. In the first case the sub-matrices did not anticommute. Figure 3.3.1 shows the results on a log-log diagram. The points are located along a line with slope close to 2, which means that the method has first order.

In the second case we chose anticommuting matrices (so-called Pauli matrices, well-known in quantummechanics [102]). Here, as Figure 3.3.2 shows, second-order accuracy was achieved.

The subtraction in the second step of (3.3.27) - (3.3.28) causes significant theoretical difficulties in order to show the stability. Therefore we modify the method. We execute the separate splitting steps with the sub-operators multiplied by d, and then we compute the splitting approximation as an average of these results. Namely, the algorithm reads as follows:

1. For each $n = 1, 2, \ldots, N$ we successively solve the Cauchy problems:

$$\begin{aligned} &\frac{dw_i^n}{dt}(t) = dA_i w_i^n(t), \qquad (n-1)\tau < t \leq n\tau, \\ &w_i^n((n-1)\tau) = w_{\text{mas}}^{(N)}((n-1)\tau) \end{aligned} \tag{3.3.31}$$

   where $i = 1, 2, \ldots, d$.

2. The split solution is defined as

$$w_{\text{mas}}^{(N)}(n\tau) = \frac{1}{d} \sum_{i=1}^{d} w_i^n(n\tau). \tag{3.3.32}$$

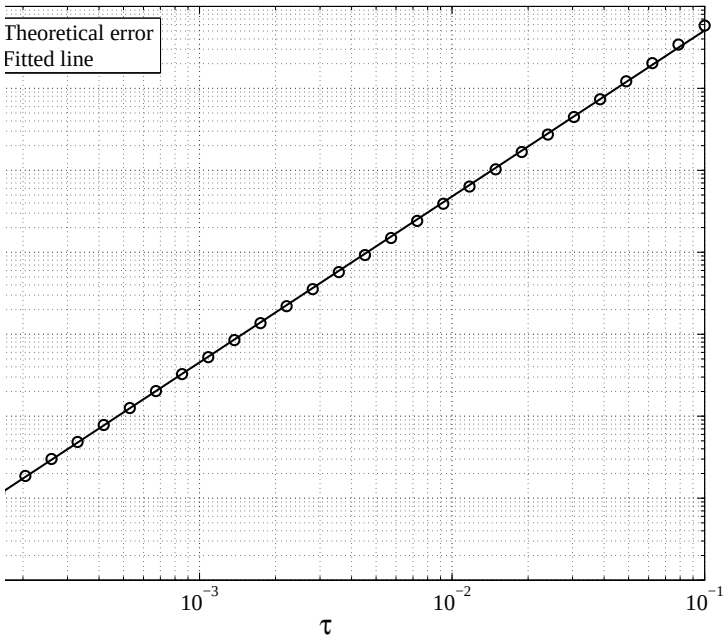

Figure 3.3.1: The local error of the additive splitting on a log-log diagram, when the sub-problems are solved exactly. The slope of the fitted line is 2.025.

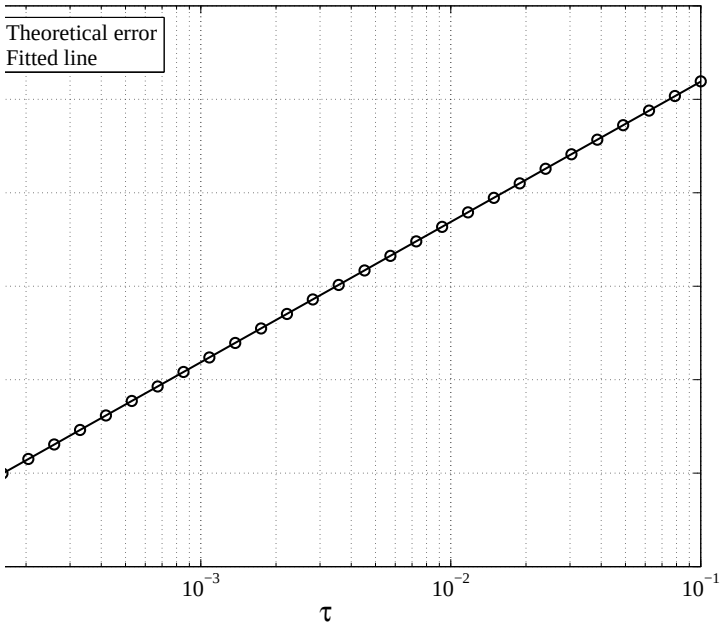

Figure 3.3.2: The local error of the additive splitting on a log-log diagram, when the sub-matrices anticommute, and the sub-problems are solved exactly. The slope of the fitted line is 3.002.

Here $w^{(N)}_{\text{mas}}(0) = w(0)$ is known from the original continuous problem (3.1.1).

We will refer to this method as *modified additive splitting* (mas).

Remark 3.3.11 *It is possible to replace the operators dA_i with A_i in (3.3.31). If such a replacement is performed, then the integration should be carried out on the interval*

$(n-1)\tau < t \leq (n+d-1)\tau$, and (3.3.32) should be replaced with $w_{\text{mas}}^{(N)}(n\tau) = \frac{1}{d}\sum_{i=1}^{d} w_i^n((n+d-1)\tau)$. If the operators are linear, then the two approaches sketched above are equivalent. However, for non-linear operators the results obtained by using these two approaches will in general be different.

The investigation of the local splitting error (for bounded linear operators) for the modified additive splitting (3.3.31)-(3.3.32) is similar as it was done in Theorem 3.3.10.

Theorem 3.3.12 *For bounded linear operators the modified additive splitting is a first order accurate, continuously consistent method.*

PROOF. The solution of the modified additive splitting at $t \in (0, \tau]$ can be given as

$$w_{\text{mas}}^{(N)}(t) = \frac{1}{d}\sum_{i=1}^{d} \exp(dA_i t) w_0. \tag{3.3.33}$$

Hence, we get

$$\begin{aligned} w_{\text{mas}}^{(N)}(t) &= \frac{1}{d}\sum_{i=1}^{d} \Big(I + dA_i t + \frac{1}{2}(dA_i t)^2\Big) w_0 + \mathcal{O}(t^3) = \\ &= \Big(I + \sum_{i=1}^{d} A_i t + \frac{d}{2}\sum_{i=1}^{d} A_i^2 t^2\Big) w_0 + \mathcal{O}(t^3). \end{aligned} \tag{3.3.34}$$

Comparing this expression with the similar Taylor expansion of the exact solution, we get the local splitting error

$$Err_{\text{mas}}(t) = \frac{t^2}{2}\left(\sum_{\substack{i,j=1 \\ i\neq j}}^{d} A_i A_j + (1-d)\sum_{i=1}^{d} A_i^2\right) w_0 + \mathcal{O}(t^3), \tag{3.3.35}$$

which proves the statement. ■

We note that for $d = 2$ (3.3.35) can be rewritten as

$$Err_{\text{mas}}(t) = -\frac{1}{2}\left((A_1 - A_2)^2 t^2\right) w_0 + \mathcal{O}(t^3). \tag{3.3.36}$$

3.3.3 Iterated splitting

As we have seen, the sequential splitting is not continuously consistent. The reason for this is the fact that on each sub-problem we passed through the time sub-interval by using one operator only. This observation gives the inspiration to modify the sequential splitting in such a way that each operator is involved in the sub-problems, but only one of them is applied to the unknown function, the other operators are applied to some known (previously defined) approximation. Then, we iterate consecutively on the same time interval $m \in \mathbb{N}$ times. The method, which will be called *iterated splitting*, is defined for two operators, (i.e., $d = 2$ in (3.1.1)) and was introduced in [39]. On some fixed split time interval $[(n-1)\tau, n\tau]$ the algorithm of the iterated splitting reads as follows:

$$\frac{dw_i^n(t)}{dt} = A_1 w_i^n(t) \ + \ A_2 w_{i-1}^n(t), \text{ with } \ w_i^n((n-1)\tau) = w_{\text{is}}^{(N)}((n-1)\tau) \tag{3.3.37}$$

$$\frac{dw_{i+1}^n}{dt} = A_1 w_i^n(t) \ + \ A_2 w_{i+1}^n(t), \text{ with } \ w_{i+1}^n((n-1)\tau) = w_{\text{is}}^{(N)}((n-1)\tau) \tag{3.3.38}$$

for $i = 1, 3, 5, \ldots, 2m-1$, where w_0^n is an arbitrarily chosen fixed starting function for the iteration and the index i refers to the number of the iteration on the fixed n-th time sub-interval. (We will discuss the suitable choice of the initial guess later.) Then the split solution at the mesh-points is defined as

$$w_{\text{is}}^{(N)}(n\tau) = w_{2m}^n(n\tau).$$

For $m = 1$ the method can be viewed as the continuous variant of the well-known ADI method (e.g., [94]).

Remark 3.3.13 *The iterated splitting can be written in a more compact way as follows:*

$$\frac{dw_i^n(t)}{dt} = \big[pA_1 + (1-p)A_2\big] w_i^n(t) + \big[(1-p)A_1 + pA_2\big] w_{i-1}^n(t), \quad t \in \big((n-1)\tau, n\tau\big],$$

$$w_i^n((n-1)\tau) = w_{\text{is}}^{(N)}((n-1)\tau), \tag{3.3.39}$$

where $i = 1, 2, \ldots, 2m$, $p = \mod(i, 2)$ *and the split solution is*

$$w_{\text{is}}^{(N)}(n\tau) = w_{2m}^n(n\tau). \tag{3.3.40}$$

This formulation has the advantage that it reveals the fact that the iterated splitting algorithm can be terminated not only after the second step (3.3.38).

We note that this method can be considered as an operator splitting method because we decompose the original problem into a sequence of two simpler sub-problems, in which the first sub-problem involves the first operator, while the second sub-problem the second operator.

Theorem 3.3.14 *Assume that on the time interval* $[0, \tau]$ *the starting function* $w_0^1(t)$ *for the iterated splitting satisfies the condition*

$$w_0^1(0) = w_0. \tag{3.3.41}$$

Then the iterated splitting is continuously consistent.

Proof. For $i = 1$ the exact solution of (3.3.37) is

$$w_1^1(t) = \exp(tA_1)w_0 + \int_0^t \exp((t-s)A_1)A_2 w_0^1(s)ds, \quad t \in [0, \tau]. \tag{3.3.42}$$

Using the obvious relation

$$w_0^1(s) = w_0^1(0) + \mathcal{O}(s) = w_0 + \mathcal{O}(s), \tag{3.3.43}$$

for the second term on the right-hand side of (3.3.42) we have

$$\int_0^t \exp((t-s)A_1)A_2 w_0^1(s)ds = \int_0^t \left[\sum_{j=0}^{\infty} \frac{1}{j!}(t-s)^j A_1^j\right] A_2(w_0+\mathcal{O}(s))ds = \tag{3.3.44}$$

$$= \textstyle\int_0^t A_2 w_0 ds + \mathcal{O}(t^2) = tA_2 w_0 + \mathcal{O}(t^2).$$

On the other hand,

$$\exp(tA_1)w_0 = (I + tA_1)w_0 + \mathcal{O}(t^2). \tag{3.3.45}$$

Hence, putting the relations (3.3.44) and (3.3.45) into (3.3.42), we get

$$w_1^1(t) = (I + tA_1 + tA_2)w_0 + \mathcal{O}(t^2), \tag{3.3.46}$$

which proves the continuous consistency. ■

Corollary 3.3.15 *Under the condition (3.3.41) the iterated splitting is consistent already after the first iteration. We can also observe that condition (3.3.41) is not only a sufficient, but also a necessary condition for the consistency in the first iterated step.*

Next we examine the order of the local splitting error when we apply the second step of the method, i.e., we solve (3.3.38) for $i = 1$, using $w_1^1(t)$ from (3.3.37).

Theorem 3.3.16 *On the time interval $[0, \tau]$ one complete step (3.3.37)–(3.3.38), under the condition (3.3.41) results in second order accuracy of the iterated splitting method.*

Proof. Clearly, the exact solution of the problem (3.3.38) can be written as

$$w_2^1(t) = \exp(tA_2)w_0 + \int_0^t \exp((t-s)A_2)A_1 w_1^1(s)ds. \tag{3.3.47}$$

Now, using the expression (3.3.46) for $w_1^1(s)$, we get

$$\exp((t-s)A_2)A_1 w_1^1(s) = [I - (t-s)A_2]\,(A_1 + A_1^2 s + A_1 A_2 s)w_0 + \mathcal{O}(s^2) = \tag{3.3.48}$$

$$= (A_1 + (t-s)A_2 A_1 + sA_1^2 + sA_1 A_2)w_0 + \mathcal{O}(s^2).$$

Hence, integrating on $[0, t]$, we get

$$\int_0^t \exp((t-s)A_2)A_1 w_1^1(s)ds = \left[A_1 t + \frac{1}{2}t^2(A_2 A_1 + A_1^2 + A_1 A_2)\right] w_0 + \mathcal{O}(t^3). \tag{3.3.49}$$

Using the Taylor series for $\exp(tA_2)$ and (3.3.49), we obtain the statement:

$$w_2^1(t) = (I + tA_2 + \frac{1}{2}t^2 A_2^2)w_0 + (A_1 t + \frac{1}{2}t^2 A_2 A_1 + \frac{1}{2}t^2 A_1^2 + \frac{1}{2}t^2 A_1 A_2)w_0 + \mathcal{O}(t^3) =$$

$$= \left[I + t(A_1 + A_2) + \frac{1}{2}t^2(A_1 + A_2)^2\right] w_0 + \mathcal{O}(t^3).$$

This completes the proof. ■

The choice of the initial iteration function has strong influence on the accuracy of the iterated splitting method. We pointed out that, when the initial guess function $w_0^1(t)$

satisfies the condition (3.3.41), then after the first complete step we obtain a higher (second) order accurate method.
It is quite natural to expect that we are able to increase the local order of the splitting with a more precisely chosen initial iteration function.

To show this conjecture, first we give the exact solution of problem (3.3.37) on the interval $[0,\tau]$, under the assumption that $w_{i-1}^1(t)$ is an analytical function on $[0,\tau]$. (For convenience, we omit the superscript, i.e., we will use the notation $w_i(t)$ instead of $w_i^1(t)$.)

Theorem 3.3.17 *The solution of problem (3.3.37) for an analytical function $w_{i-1}(t)$ has the form*

$$w_i(t) = \exp(A_1 t)w_0 + \sum_{s=0}^{\infty}\left(\sum_{n=0}^{\infty}\frac{t^{n+1+s}}{(n+1+s)!}A_1^n A_2 w_{i-1}^{(s)}(0)\right), \tag{3.3.50}$$

where $w_{i-1}^{(s)}(0)$ denotes the s-th derivative at the point $t=0$.

Proof. According to the assumption, $w_{i-1}(t)$ can be written as

$$w_{i-1}(t) = \sum_{s=0}^{\infty}\frac{t^s}{s!}w_{i-1}^{(s)}(0). \tag{3.3.51}$$

On the right-hand side of (3.3.50), the second term can be estimated in norm:

$$\sum_{s=0}^{\infty}\sum_{n=0}^{\infty}\left\|\frac{t^{n+1+s}}{(n+1+s)!}A_1^n A_2 w_{i-1}^{(s)}(0)\right\|$$

$$\le \sum_{s=0}^{\infty}\sum_{n=0}^{\infty}\frac{t^{n+1+s}}{(n+1+s)!}\|A_1\|^n\ \|A_2\ \|\ \ \|w_{i-1}^{(s)}(0)\| \tag{3.3.52}$$

$$\le \|A_2\|\sum_{s=0}^{\infty}\frac{t^s}{s!}\ \|w_{i-1}^{(s)}(0)\|\sum_{n=0}^{\infty}\frac{t^{n+1}}{(n+1)!}\|A_1\|^n < \infty.$$

(Here, we have used the inequality $(n+1+s)! \ge (n+1)!\ s!$ and the convergence of (3.3.51) in norm.)
Therefore, by the derivation of $w_i(t)$ in the sum (3.3.50) we can compute the derivatives elementwise. Hence,

$$\frac{dw_i(t)}{dt} = A_1\exp(A_1 t)w_0 + \sum_{s=0}^{\infty}\left(\sum_{n=0}^{\infty}\frac{t^{n+s}}{(n+s)!}A_1^n A_2 w_{i-1}^{(s)}(0)\right). \tag{3.3.53}$$

On the other hand, by using (3.3.53) and (3.3.51) we obtain

$$A_1 w_i(t) = A_1\exp(A_1 t)w_0 + \sum_{s=0}^{\infty}\left(\sum_{n=0}^{\infty}\frac{t^{n+1+s}}{(n+1+s)!}A_1^{n+1} A_2 w_{i-1}^{(s)}(0)\right)$$

$$= A_1\exp(A_1 t)w_0 + \sum_{s=0}^{\infty}\left(\sum_{n=0}^{\infty}\frac{t^{n+s}}{(n+s)!}A_1^n A_2 w_{i-1}^{(s)}(0) - \frac{t^s}{s!}A_2 w_{i-1}^{(s)}(0)\right) \tag{3.3.54}$$

$$= \frac{dw_i(t)}{dt} - \sum_{s=0}^{\infty}\frac{t^s}{s!}A_2 w_{i-1}^{(s)}(0) = \frac{dw_i(t)}{dt} - A_2 w_{i-1}(t),$$

which proves the statement. ■

Remark 3.3.18 *The solution (3.3.50) can be rewritten in a more compact form, namely*

$$w_i(t) = \exp(A_1 t)w_0 + \sum_{s=0}^{\infty}\sum_{k=s+1}^{\infty}\frac{t^k}{k!}A_1^{k-s-1}A_2 w_{i-1}^{(s)}(0). \tag{3.3.55}$$

Remark 3.3.19 *Since the solution (3.3.50) is an analytical function, therefore this theorem also shows that in the iteration process* (3.3.37) *(and similarly in* (3.3.38)*) the analyticity property is preserved. Therefore, it is enough to require that on each time-interval the starting function of the iteration is analytical.*

Now we will show that we can increase the order of the local splitting error by a suitable choice of the initial guess.

Theorem 3.3.20 *Assume that for the analytical starting iteration function $w_0(t)$ the condition*

$$w_0^{(s)}(0) = (A_1 + A_2)^s w(0); \quad s = 0, 1, \ldots, m \tag{3.3.56}$$

is satisfied. Then, after the first iteration, the order of the local splitting error is equal to $m + 2$.

Proof. We will show that for any $t \in [0, \tau]$ the relation

$$\exp(t(A_1 + A_2))w(0) - w_1(t) = \mathcal{O}(t^{m+2}) \tag{3.3.57}$$

holds. In order to do this, we take into account the assumption of the theorem and formula (3.3.55), and hence we have to prove the relation

$$\sum_{p=0}^{m+1}\frac{1}{p!}t^p(A_1 + A_2)^p = \sum_{p=0}^{m+1}\frac{1}{p!}t^p A_1^p + \sum_{s=0}^{m}\sum_{k=s+1}^{m+1}\frac{t^k}{k!}A_1^{k-s-1}A_2. \tag{3.3.58}$$

For the proof, we use the mathematical induction.
For $m = 0$ the formula (3.3.58) is obviously true.

Let us introduce the notation

$$S_m = \sum_{p=0}^{m+1}\frac{1}{p!}t^p A_1^p + \sum_{s=0}^{m}\sum_{k=s+1}^{m+1}\frac{t^k}{k!}A_1^{k-s-1}A_2. \tag{3.3.59}$$

Obviously, we have to show that

$$S_{m+1} = \sum_{p=0}^{m+2}\frac{1}{p!}t^p(A_1 + A_2)^p. \tag{3.3.60}$$

According to (3.3.60), we have the relation

$$S_{m+1} = \sum_{p=0}^{m+2}\frac{t^p}{p!}A_1^p + \sum_{s=0}^{m+1}\sum_{k=s+1}^{m+1}\frac{t^k}{k!}A_1^{k-s-1}A_2 = S_m + \frac{t^{m+2}}{(m+2)!}A_1^{m+2} +$$

$$+\sum_{s=0}^{m}\frac{t^{m+2}}{(m+2)!}A_1^{m+1-s}A_2(A_1 + A_2)^s + \frac{t^{m+2}}{(m+2)!}A_2(A_1 + A_2)^{m+1} = \tag{3.3.61}$$

$$= \sum_{p=0}^{m+1}\frac{t^p}{p!}(A_1 + A_2)^p + \frac{t^{m+2}}{(m+2)!}\left[A_1^{m+2} + \sum_{s=0}^{m+1}A_1^{m+1-s}A_2(A_1 + A_2)^s\right].$$

One can directly check the validity of the following identity:

$$\sum_{s=0}^{m+1} A_1^{m+1-s} A_2 (A_1+A_2)^s = \sum_{s=0}^{m+1} A_1^{m+1-s}(A_1+A_2-A_1)(A_1+A_2)^s =$$

$$= \sum_{s=0}^{m+1} A_1^{m+1-s}(A_1+A_2)^{s+1} - \sum_{s=0}^{m+1} A_1^{m+2-s}(A_1+A_2)^s = \tag{3.3.62}$$

$$= (A_1+A_2)^{m+2} - A_1^{m+2}.$$

Substituting (3.3.62) into (3.3.61), we get (3.3.60), which proves the theorem. ■

How can we guarantee the condition (3.3.56)? Let us choose such a starting function $w_0(t)$, which is continuously consistent in order m on the interval $[0,\tau]$, i.e.,

$$w(t) - w_0(t) = \mathcal{O}(t^{m+1}). \tag{3.3.63}$$

Hence, the relation

$$w^{(s)}(0) = w_0^{(s)}(0) \tag{3.3.64}$$

holds for $s = 0, 1, \ldots, m$. On the other hand, $w(t) = \exp((A_1+A_2)t)w(0)$, therefore $w^{(s)}(0) = (A_1+A_2)^s\ w(0)$, for $s = 0, 1, \ldots$.
Clearly, the same results can be formulated for the equation (3.3.38), too. Because the iterated splitting is continuously consistent, therefore each iterated step in the algorithm (3.3.37)-(3.3.38) the order of the local splitting error increases by one. So, we get

Theorem 3.3.21 *Assume that we execute the iteration (3.3.37)–(3.3.38) with a suitably chosen initial function on the time interval* $[0,\tau]$. *Then each iterative step increases the order of the local splitting error by one.*

As a main consequence, we can define the order of the iterated splitting when the starting iterated function $w_0(t)$ is obtained as a result of some other continuously consistent splitting method. Clearly, we have the following

Theorem 3.3.22 *Assume that the function* $w_0(t)$ *is a split solution of the Cauchy problem (3.1.1), obtained by using a continuously consistent operator splitting method of order* p. *Then, after* m *steps by the iterations (3.3.37) and (3.3.38), the local splitting error is* $\mathcal{O}(\tau^{m+p+1})$.

3.4 Further investigations of the operator splittings

In Sections 3.2 and 3.3 we introduced and analyzed the different operator splittings for linear bounded operators. In this part we give a more comprehensive analysis and extensions for those results.

3.4.1 Operator splittings for Cauchy problems with a source function

In this part we consider splitting methods for a linear system of differential equations $w'(t) = Aw(t) + f(t)$, $A \in \mathbb{R}^{n\times n}$ and $t \in \mathcal{I}$ (where $\mathcal{I} \subset \mathbb{R}$ is an interval), split into two

subproblems $w_1'(t) = A_1 w_1(t) + f_1(t)$ and $w_2'(t) = A_2 w_2(t) + f_2(t)$, $A = A_1 + A_2$, $f = f_1 + f_2$. This problem is motivated by many practical problems, e.g., the semi-discretization of the Maxwell equations leads to such a kind of problems. (See Appendix C.) We will analyze the second order methods. Namely, expressions for the leading term of the local error are derived for both the Strang-Marchuk splitting and the symmetrically weighted sequential splittings.

Using the power series representation of the matrix exponential, the following estimates clearly hold:

$$(\exp \tau A)\mathbf{v}(t) = \sum_{j=0}^{m} \frac{1}{j!} A^j \mathbf{v}(t) + \mathcal{O}(\tau^{m+1}), \qquad \forall t \in \mathcal{I},\ \forall m \in \mathbb{N}, \tag{3.4.1}$$

$$(\exp \tau A \cdot \exp \tau B)\,\mathbf{v}(t) = (I + \tau(A+B) + \frac{\tau^2}{2}(A^2 + B^2 + 2AB))\mathbf{v}(t) + \mathcal{O}(\tau^3), \tag{3.4.2}$$

$$\begin{aligned}(\exp \tau A \cdot \exp \tau B \cdot \exp \tau C)\,\mathbf{v}(t) &= (I + \tau(A+B+C) + \\ &+ \frac{\tau^2}{2}(A^2 + B^2 + C^2 + 2(AC + BC + AB)))\mathbf{v}(t) + \mathcal{O}(\tau^3),\end{aligned} \tag{3.4.3}$$

for all $t \in \mathcal{I}$. Let $\mathbf{v} : \mathcal{I} \to \mathbb{R}^n$ be a $p+1$ times continuously differentiable function ($\mathbf{v} \in C^{p+1}(\mathcal{I})$). Then the Taylor expansion of the function $\mathbf{v}$ can be defined as

$$\mathbf{v}(t+\tau) = \sum_{j=0}^{p} \frac{\tau^j}{j!} \mathbf{v}^{(j)}(t) + \frac{\tau^{p+1}}{(p+1)!} \mathbf{v}^{(p+1)}(t + \theta\tau), \qquad \forall \tau > 0\ (t + \tau \in \mathcal{I}), \tag{3.4.4}$$

where

$$\begin{aligned}\mathbf{v}^{(j)}(t) &= [v_i^{(j)}]_{i=1}^n \in \mathbb{R}^n, \qquad j = 1, \ldots, p, \\ \mathbf{v}^{(p+1)}(t + \theta\tau) &\equiv [v_i^{(p+1)}(t + \theta_i t)]_{i=1}^n \in \mathbb{R}^n, \qquad \theta_i \in (0,1), \quad i = 1, \ldots, n.\end{aligned}$$

Using the Landau notation of $\mathcal{O}(\tau^p)$, we can rewrite the Taylor series (3.4.4) as

$$\mathbf{v}(t+\tau) = \sum_{j=0}^{p} \frac{\tau^j}{j!} \mathbf{v}^{(j)}(t) + O(\tau^{p+1}). \tag{3.4.5}$$

For a function $\mathbf{v} : \mathcal{I} \to \mathbb{R}^n$ with integrable coordinate functions $v_i : \mathcal{I} \to \mathbb{R}$, the integral $\int_{\mathcal{I}} \mathbf{v}(t)dt$ is defined elementwise, i.e.,

$$\int_{\mathcal{I}} \mathbf{v}(t)dt \equiv \left[\int_{\mathcal{I}} v_i(t)dt\right]_{i=1}^n \in \mathbb{R}^n.$$

We will need the following result:

Lemma 3.4.1 *Assume that the interval $\mathcal{I}$ is of length $|\mathcal{I}| = \mathcal{O}(\tau^m)$, that is, there exists a constant $C_1 > 0$ such that for sufficiently small values $|\tau|$ the relation $|\mathcal{I}| \le C_1|\tau|^m$ holds. Assume also that $\tau > 0$ and $\tau \in \mathcal{I}$ and that $\mathbf{f}_\tau(t) = \mathcal{O}(\tau^p)$ and $\mathbf{g}_{\tau-s}(t) = \mathcal{O}(\tau - s)^p$. Then*

$$\int_{\mathcal{I}} \mathbf{f}_\tau(s)ds = \mathcal{O}(\tau^{p+m}), \text{ and } \int_{\mathcal{I}} \mathbf{g}_{\tau-s}(s)ds = \mathcal{O}(\tau^{m+mp}). \tag{3.4.6}$$

PROOF. There exist constants $C_2 > 0$ and $C_3 > 0$ such that

$$\|\mathbf{f}_\tau(t)\| \le C_2\tau^p, \qquad \|\mathbf{g}_{\tau-s}(t)\| \le C_3|\tau - s|^p,$$

uniformly for all $t \in \mathcal{I}$. Hence, the coordinate functions $f_{\tau,i}(t)$, $g_{\tau-s,i}(t)$, $i = 1, \ldots, n$, are also bounded as $|f_{\tau,i}(t)| \le C_2\tau^p$, $|g_{\tau-s,i}(t)| \le C_3|\tau - s|^p$. Thus we have

$$\left|\int_{\mathcal{I}} f_{\tau,i}(s)ds\right| \le \int_{\mathcal{I}} |f_{\tau,i}(s)|ds \le C_2\tau^p \int_{\mathcal{I}} ds \le C_1C_2\tau^{p+m},$$
$$\left|\int_{\mathcal{I}} g_{\tau-s,i}(s)ds\right| \le \int_{\mathcal{I}} |g_{\tau-s,i}(s)|ds \le C_3 \int_{\mathcal{I}} \underbrace{|\tau - s|^p}_{\le |\mathcal{I}|^p \le C_1\tau^{mp}} ds \le C_1C_3\tau^{mp} \int_{\mathcal{I}} ds \le C_1^2C_3\tau^{mp+m}.$$

■

Let $\mathbf{f} : [0, T] \to \mathbb{R}^n$ be a given vector function and $A \in \mathbb{R}^{n\times n}$. The solution of the initial value problem

$$\begin{cases} \mathbf{w}'(t) = A\mathbf{w}(t) + \mathbf{f}(t), & t \in [0, T], \\ \mathbf{w}(0) \text{ is given,} \end{cases} \tag{3.4.7}$$

reads

$$\mathbf{w}(t) = \exp(tA)\mathbf{w}(0) + \int_0^t \exp((t - s)A)\mathbf{f}(s)ds, \qquad t \in [0, T]. \tag{3.4.8}$$

In the following, we split (3.4.7) into the following two ODE systems:

$$\mathbf{w}_1' = A_1\mathbf{w}_1 + \mathbf{f}_1, \qquad \mathbf{w}_2' = A_2\mathbf{w}_2 + \mathbf{f}_2, \qquad \text{with} \quad A_1 + A_2 = A, \quad \mathbf{f}_1 + \mathbf{f}_2 = \mathbf{f}. \tag{3.4.9}$$

First, we consider the Strang-Marchuk splitting scheme. Assuming that the time integration is performed exactly for each split step, we can write the solution of this splitting scheme after one time step as

$$\begin{aligned} \mathbf{w}_{\mathrm{SM}}(\tau) = \exp(\frac{\tau}{2}A_1) \left\{\exp(\tau A_2)\mathbf{w}_1(\tau/2) + \int_0^\tau \exp((\tau - s)A_2)\mathbf{f}_2(s)ds\right\} + \\ + \int_{\tau/2}^\tau \exp((\tau - s)A_1)\mathbf{f}_1(s)ds, \end{aligned} \tag{3.4.10}$$

where $\mathbf{w}_1(\tau/2)$ is the solution of the first sub-step for the first sub-problem:

$$\mathbf{w}_1(\tau/2) = \exp(0.5\tau A_1)\mathbf{w}_1(0) + \int_0^{0.5\tau} \exp((0.5\tau - s)A_1)\mathbf{f}_1(s)ds,$$

with $\mathbf{w}_1(0) \equiv \mathbf{w}(0)$ being the initial data of the original problem (3.4.7). Substituting the

last expression for $\mathbf{w}_1(\tau/2)$ into (3.4.10), we obtain:

$$\mathbf{w}_{\mathrm{SM}}(\tau) = \exp(\frac{\tau}{2}A_1)\left\{\exp(\tau A_2)\left(\exp(\frac{\tau}{2}A_1)\mathbf{w}(0) + \int_0^{\tau/2}\exp((\tau/2-s)A_1)\mathbf{f}_1(s)ds\right) + \right.$$

$$\left. + \int_0^{\tau}\exp((\tau-s)A_2)\mathbf{f}_2(s)ds\right\} + \int_{\tau/2}^{\tau}\exp((\tau-s)A_1)\mathbf{f}_1(s)ds =$$

$$= \exp(\frac{\tau}{2}A_1)\exp(\tau A_2)\exp(\frac{\tau}{2}A_1)\mathbf{w}(0) +$$

$$+ \exp(\frac{\tau}{2}A_1)\exp(\tau A_2)\int_0^{\tau/2}\exp((\tau/2-s)A_1)\mathbf{f}_1(s)ds +$$

$$+ \exp(\frac{\tau}{2}A_1)\int_0^{\tau}\exp((\tau-s)A_2)\mathbf{f}_2(s)ds + \int_{\tau/2}^{\tau}\exp((\tau-s)A_1)\mathbf{f}_1(s)ds. \tag{3.4.11}$$

Theorem 3.4.2 *Assume that the functions* $\mathbf{f}_1$, $\mathbf{f}_2$ *are three times continuously differentiable vector functions:* $\mathbf{f}_i : [0,T] \to \mathbb{R}^n$, $\mathbf{f}_i \in C^3([0,T])$, $i = 1,2$. *Then the Strang-Marchuk splitting scheme* (3.4.11) *applied to the inhomogeneous ODE system* (3.4.7) *with splitting* (3.4.9) *has third order local error, i.e., the scheme has second order accuracy and for the local splitting error we have*

$$Err_{SM}(\tau) = \mathbf{w}(\tau) - \mathbf{w}_{\mathrm{SM}}(\tau) = \frac{\tau^3}{12}\left\{\left[\frac{1}{2}A_1 + A_2, [A_1,A_2]\right]\mathbf{w}(0) + \right.$$

$$+ \left(\frac{1}{2}A_2A_1 - A_1A_2 - A_2^2\right)\mathbf{f}_1\left(\frac{\tau}{2}\right) + \left(2A_2A_1 - A_1A_2 + \frac{1}{2}A_1^2\right)\mathbf{f}_2\left(\frac{\tau}{2}\right) - \tag{3.4.12}$$

$$\left. - A_1\mathbf{f}_2'\left(\frac{\tau}{2}\right) + \frac{1}{2}A_2\mathbf{f}_1'\left(\frac{\tau}{2}\right)\right\} + \mathcal{O}(\tau^4),$$

where, as before, $\mathbf{w}$ *is the exact solution of* (3.4.7) *defined by* (3.4.8) *and* $[A_1,A_2]$, *as before, denotes the commutator of* A_1 *and* A_2.

Proof. Comparing (3.4.11) and (3.4.8), using (3.2.34) we have

$$\left(\exp(\tau A) - \exp(\frac{\tau}{2}A_1)\exp(\tau A_2)\exp(\frac{\tau}{2}A_1)\right)\mathbf{w}(0) =$$

$$= \frac{\tau^3}{12}\left(\left[\frac{1}{2}A_1 + A_2, [A_1,A_2]\right]\right)\mathbf{w}(0) + \mathcal{O}(\tau^4). \tag{3.4.13}$$

The rest of the proof consists of estimating the differences in the terms containing $\mathbf{f}_1$ and $\mathbf{f}_2$ in (3.4.11) and (3.4.8). We first rewrite the terms of (3.4.11) containing $\mathbf{f}_1$ as

$$\exp(\frac{\tau}{2}A_1)\exp(\tau A_2)\int_0^{\tau/2}\exp((\tau/2-s)A_1)\mathbf{f}_1(s)ds + \int_{\tau/2}^{\tau}\exp((\tau-s)A_1)\mathbf{f}_1(s)ds$$

$$= \int_0^{\tau/2}\exp(\frac{\tau}{2}A_1)\exp(\tau A_2)\exp((\tau/2-s)A_1)\mathbf{f}_1(s)ds + \int_{\tau/2}^{\tau}\exp((\tau-s)A_1)\mathbf{f}_1(s)ds$$

and, using (3.4.6) and (3.4.1), we arrive at

$$= \int_0^\tau \mathbf{f}_1(s)ds + \int_0^\tau (\tau - s)A_1\mathbf{f}_1(s)ds + \int_0^\tau \frac{(\tau - s)^2}{2}A_1^2\mathbf{f}_1(s)ds \tag{3.4.14}$$

$$+ \int_0^{\tau/2} \left(\tau A_2 + \frac{\tau^2}{2}\left(A_2^2 + A_1A_2 + A_2A_1\right) - \tau s A_2A_1\right)\mathbf{f}_1(s)ds + \mathcal{O}(\tau^4). \tag{3.4.15}$$

We now introduce the notations

$$\mathbf{f}_{i,0} = \mathbf{f}_i(\tau/2) \in \mathbb{R}^n, \quad \mathbf{f}_{i,1} = \mathbf{f}_i'(\tau/2) \in \mathbb{R}^n, \quad \mathbf{f}_{i,2} = \mathbf{f}_i''(\tau/2) \in \mathbb{R}^n \qquad i = 1, 2,$$

and define the Taylor expansions of $\mathbf{f}_1$ and $\mathbf{f}_2$:

$$\mathbf{f}_i(s) = \mathbf{f}_{i,0} + (s - \tau/2)\mathbf{f}_{i,1} + \frac{(s - \tau/2)^2}{2}\mathbf{f}_{i,2} + \mathcal{O}(s - \tau/2)^3, \qquad i = 1, 2. \tag{3.4.16}$$

Replacing $\mathbf{f}_1$ in (3.4.15) by its Taylor expansion (3.4.16) and taking into account (3.4.6), we obtain

$$\int_0^\tau \mathbf{f}_1(s)ds = \int_0^\tau \left[\mathbf{f}_{1,0} + (s - \tau/2)\mathbf{f}_{1,1} + \frac{(s - \tau/2)^2}{2}\mathbf{f}_{1,2}\right]ds + \mathcal{O}(\tau^4)$$

$$= \int_0^\tau ds \cdot \mathbf{f}_{1,0} + \underbrace{\int_0^\tau (s - \tau/2)ds}_{=0} \cdot \mathbf{f}_{1,1} + \int_0^\tau \frac{(s - \tau/2)^2}{2}ds\mathbf{f}_{1,2} + \mathcal{O}(\tau^4) =$$

$$= \tau\mathbf{f}_{1,0} + \frac{\tau^3}{24}\mathbf{f}_{1,2} + \mathcal{O}(\tau^4).$$

Similarly, we have

$$\int_0^\tau (\tau - s)A_1\mathbf{f}_1(s)ds = \frac{\tau^2}{2}A_1\mathbf{f}_{1,0} - \frac{\tau^3}{12}A_1\mathbf{f}_{1,1} + \mathcal{O}(\tau^4),$$

$$\int_0^\tau \frac{(\tau - s)^2}{2}A_1^2\mathbf{f}_1(s)ds = \frac{\tau^3}{6}A_1^2\mathbf{f}_{1,0} + \mathcal{O}(\tau^4),$$

$$\int_0^{\tau/2} \tau A_2\mathbf{f}_1(s)ds = \frac{\tau^2}{2}A_2\mathbf{f}_{1,0} - \frac{\tau^3}{8}A_2\mathbf{f}_{1,1} + \mathcal{O}(\tau^4),$$

$$\int_0^{\tau/2} \frac{\tau^2}{2}(A_2^2 + A_1A_2 + A_2A_1)\mathbf{f}_1(s)ds = \frac{\tau^3}{4}(A_2^2 + A_1A_2 + A_2A_1)\mathbf{f}_{1,0},$$

and $$\int_0^{\tau/2} \tau s A_2A_1\mathbf{f}_1(s)ds = \frac{\tau^3}{6}A_2A_1\mathbf{f}_{1,0}.$$

Hence, we get

$$\exp\frac{\tau}{2}A_1 \cdot \exp\tau A_2 \int_0^{\tau/2} \exp(\tau/2 - s)A_1\mathbf{f}_1(s)ds + \int_{\tau/2}^\tau \exp(\tau - s)A_1\mathbf{f}_1(s)ds$$

$$= \tau\mathbf{f}_{1,0} + \frac{\tau^2}{2}(A_1 + A_2)\mathbf{f}_{1,0} \tag{3.4.17}$$

$$+ \frac{\tau^3}{12}\left(\frac{1}{2}\mathbf{f}_{1,2} - A_1\mathbf{f}_{1,1} + 2A_1^2\mathbf{f}_{1,0} + 3(A_2^2 + A_1A_2 + A_2A_1)\mathbf{f}_{1,0} - \frac{3}{2}A_2\mathbf{f}_{1,1} - \frac{3}{2}A_2A_1\mathbf{f}_{1,0}\right).$$

Next, we estimate the term with $\mathbf{f}_2$ in (3.4.11):

$$\exp\frac{\tau}{2}A_1\int_0^\tau \exp(\tau-s)A_2\mathbf{f}_2(s)ds =^{(3.4.2)} \int_0^\tau \mathbf{f}_2(s)ds+$$

$$+\int_0^\tau \left(\frac{\tau}{2}A_1+(\tau-s)A_2\right)\mathbf{f}_2(s)ds+ \tag{3.4.18}$$

$$+\int_0^\tau \left(\frac{\tau^2}{8}A_1^2+\frac{(\tau-s)^2}{2}A_2^2+\frac{\tau(\tau-s)}{2}A_1A_2\right)\mathbf{f}_2(s)ds+O(\tau^4).$$

Replacing here $\mathbf{f}_2(s)$ by its Taylor expansion (3.4.16), we have

$$\int_0^\tau \mathbf{f}_2(s)ds =^{(3.4.17)} \tau\mathbf{f}_{2,0}+\frac{\tau^3}{24}\mathbf{f}_{2,2}+O(\tau^4),$$

$$\int_0^\tau \left(\frac{\tau}{2}A_1+(\tau-s)A_2\right)\mathbf{f}_2(s)ds =^{(3.4.6)} \frac{\tau^2}{2}(A_1+A_2)\mathbf{f}_{2,0}-\frac{\tau^3}{12}A_2\mathbf{f}_{2,1}+O(\tau^4),$$

$$\int_0^\tau \left(\frac{\tau^2}{8}A_1^2+\frac{(\tau-s)^2}{2}A_2^2+\frac{\tau(\tau-s)}{2}A_1A_2\right)\mathbf{f}_2(s)ds$$

$$=^{(3.4.6)} \frac{\tau^3}{8}\left(A_1^2\mathbf{f}_{2,0}+\frac{3}{4}A_2^2\mathbf{f}_{2,0}+2A_1A_2\mathbf{f}_{2,0}\right)+O(\tau^4).$$

Thus, the $\mathbf{f}_2$-term in (3.4.11) can be estimated as

$$\exp\frac{\tau}{2}A_1\int_0^\tau \exp(\tau-s)A_2\mathbf{f}_2(s)ds=\tau\mathbf{f}_{2,0}+\frac{\tau^2}{2}(A_1+A_2)\mathbf{f}_{2,0} \tag{3.4.19}$$

$$\frac{\tau^3}{12}\left(\frac{1}{2}\mathbf{f}_{2,2}-A_2\mathbf{f}_{2,1}+\frac{3}{2}A_1^2\mathbf{f}_{2,0}+2A_2^2\mathbf{f}_{2,0}+3A_1A_2\mathbf{f}_{2,0}\right)+O(\tau^4).$$

This, together with (3.4.17), yields

$$\begin{aligned}\mathbf{w}_{\mathrm{SM}}(\tau)=\exp\frac{\tau}{2}A_1\exp\tau A_2\exp\frac{\tau}{2}A_1\mathbf{w}(0)+\tau(\mathbf{f}_{1,0}+\mathbf{f}_{2,0})+\frac{\tau^2}{2}(A_1+A_2)(\mathbf{f}_{1,0}+\mathbf{f}_{2,0})\\ +\tau^3\left(\frac{1}{24}\mathbf{f}_{1,2}-\left(\frac{1}{12}A_1+\frac{1}{8}A_2\right)\mathbf{f}_{1,1}+\frac{1}{24}(4A_1^2+6A_2^2+6A_1A_2+3A_2A_1)\mathbf{f}_{1,0}\right.\\ \left.+\frac{1}{24}\mathbf{f}_{2,2}-\frac{1}{12}A_2\mathbf{f}_{2,1}+\frac{1}{24}(3A_1^2+4A_2^2+6A_1A_2)\mathbf{f}_{2,0}\right)+O(\tau^4).\end{aligned} \tag{3.4.20}$$

Consider now the exact solution $\mathbf{w}(\tau)$ given by (3.4.8). Using again the Taylor series of $\mathbf{f}(s)$, we can rewrite the integral term in (3.4.8) as

$$\int_0^\tau \exp((\tau-s)A)\mathbf{f}(s)ds =^{(3.4.1),(3.4.6)} \int_0^\tau \left[I+(\tau-s)A+\frac{(\tau-s)^2}{2}A^2\right]\mathbf{f}(s)ds+O(\tau^4)$$

$$=\tau(\mathbf{f}_{1,0}+\mathbf{f}_{2,0})+\frac{\tau^2}{2}(A_1+A_2)(\mathbf{f}_{1,0}+\mathbf{f}_{2,0})+$$

$$+\tau^3\left(\frac{1}{6}A^2(\mathbf{f}_{1,0}+\mathbf{f}_{2,0})-\frac{1}{12}A(\mathbf{f}_{1,1}+\mathbf{f}_{2,1})+\frac{1}{24}(\mathbf{f}_{1,2}+\mathbf{f}_{2,2})\right)+O(\tau^4), \tag{3.4.21}$$

so that

$$\begin{aligned}\mathbf{w}(\tau) = \exp \tau A \mathbf{w}(0) + \tau(\mathbf{f}_{1,0} + \mathbf{f}_{2,0}) + \frac{\tau^2}{2}(A_1 + A_2)(\mathbf{f}_{1,0} + \mathbf{f}_{2,0}) \\ + \tau^3 \left(\frac{1}{6} A^2 (\mathbf{f}_{1,0} + \mathbf{f}_{2,0}) - \frac{1}{12} A(\mathbf{f}_{1,1} + \mathbf{f}_{2,1}) + \frac{1}{24}(\mathbf{f}_{1,2} + \mathbf{f}_{2,2}) \right) + O(\tau^4).\end{aligned} \tag{3.4.22}$$

Subtracting the last expression from (3.4.20) and taking into account (3.4.13), we arrive at the required statement given by (3.4.12). ■

For the symmetrically weighted sequential splitting a similar result holds, namely

Theorem 3.4.3 *Assume that the functions* $\mathbf{f}_1$, $\mathbf{f}_2$ *are three times continuously differentiable vector functions:* $\mathbf{f}_i : [0, T] \to \mathbb{R}^n$, $\mathbf{f}_i \in C^3([0, T])$, $i = 1, 2$. *Then the symmetrically weighted sequential splitting scheme, applied to the inhomogeneous ODE system* (3.4.7) *with splitting* (3.4.9) *has third order local error, i.e., the scheme has second order accuracy and for the local splitting error we have*

$$\begin{aligned}\mathbf{w}(\tau) - \mathbf{w}_{swss}(\tau) = -\frac{\tau^3}{12} [A_1 - A_2, [A_1, A_2]] \mathbf{w}(0) - \frac{\tau^3}{12} \big((A_2 A_1 - 2A_1 A_2 + A_2^2)\mathbf{f}_1(\tau/2) \\ + (A_1 A_2 - 2A_2 A_1 + A_1^2)\mathbf{f}_2(\tau/2) + A_2 \mathbf{f}_1'(\tau/2) + A_1 \mathbf{f}_2'(\tau/2) \big) + O(\tau^4).\end{aligned} \tag{3.4.23}$$

The proof of this statement is also very technical and we refer to [17].

3.4.2 Local error analysis

In Section 3.2 we gave the conditions under which the local splitting error vanishes for the sequential splitting. As it was proven, for two operators the commutativity was a necessary and sufficient condition. In other words, the operator norm of the commutator should equal zero to get zero local splitting error. In the following we analyze whether the dependence of the magnitude of the local splitting error is continuous on the norm of the commutator. (Is it true that to a small commutator norm corresponds a small local splitting error?)

When the order of the local splitting error is p, then this means the following: for sufficiently small values of τ the magnitude of $Err_{\text{spl}}(\tau)$ is defined by the leading term of the error, i.e., by the coefficient of τ^{p+1}. We observed that for the operator splittings under investigation this coefficient can be expressed by the commutators of the operators. This implies the direct dependence of the splitting error on the commutator norm. However, we should emphasize that this is true only for a sufficiently small τ. The following example shows that the magnitude of the norm of the commutator does not define the magnitude of the local splitting error in any case.

Example 3.4.4 *Assume that*

$$A_1 = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{3}{4} \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{3.4.24}$$

τ	$\frac{1}{2}\tau^2\|[A_2,A_1]e\|_\infty$	$\frac{1}{6}\tau^3\|R_Ae\|_\infty$
1	1.25(−1)	6.25(−2)
0.1	1.25(−3)	6.25(−5)
0.01	1.25(−5)	6.25(−8)
0.001	1.25(−7)	6.25(−11)

Table 3.4.1: The second and third terms in the splitting error for A.

$A = A_1+A_2$, $p \in \mathbb{R}$ $(p \neq 0)$ *arbitrary* , $B_1 = pA_1$, $B_2 = \dfrac{1}{p}A_2$, $B = B_1+B_2$, $v_0 = (1,1,1,1)$ *and* $\tau = 1$. *We consider the following problems:*

$$\left.\begin{array}{c} u'(t) = Au(t), \quad t \in (0,1] \\ u(0) = v_0 \end{array}\right\} \tag{3.4.25}$$

and

$$\left.\begin{array}{c} w'(t) = Bw(t), \quad t \in (0,1] \\ w(0) = v_0 \end{array}\right\}. \tag{3.4.26}$$

Then

$$\|[A_1,A_2]\|_\infty = \|[B_1,B_2]\|_\infty = \frac{1}{4},$$

that is the norm of the two different commutators are equal.
Let us select $p = 1000$, *and apply the sequential splitting for both problems (3.4.25) and (3.4.26). Then for the local splitting error we get*

$$\left\|Err^A_{\text{seq}}(1)\right\|_\infty = \left\|(e^{(A_1+A_2)} - e^{A_2}e^{A_1})v_0\right\|_\infty = 0.125$$

and

$$\left\|Err^B_{\text{seq}}(1)\right\|_\infty = \left\|(e^{(B_1+B_2)} - e^{B_2}e^{B_1})v_0\right\|_\infty = 20.8334.$$

One can see that in the second example we got a much (166.67 times) bigger local splitting error than in the first example. What is more, choosing bigger values of p, we can obtain arbitrarily big differences. This example suggests that the size of the splitting error is not completely determined by the commutator norm.

The reason is that in this example the value $\tau = 1$ is still quite big, and the contribution of the higher-order terms in the local splitting error is still important. When we decrease the value of τ, then the difference between the local splitting errors disappears. In order to analyze this phenomenon, we compute the coefficient at τ^3 for the error. The direct calculation shows that

$$Err_{\text{seq}}(\tau) = \frac{\tau^2}{2}[A_2,A_1]w_0 + \frac{\tau^3}{6}R_Aw_0 + \mathcal{O}(\tau^4), \tag{3.4.27}$$

where the operator R_A has the form

$$R_A = A_1A_2A_1 + A_2A_1A_2 + A_1^2A_2 + A_1A_2^2 - 2A_2^2A_1 - 2A_2A_1^2. \tag{3.4.28}$$

(The same expression is valid for B.) In our example, for the two different splittings, the second and the third order terms in the local splitting error for the operators A and B are shown in Table 3.4.1 and Table 3.4.2, respectively.
The total splitting errors for the different values of τ are shown in Table 3.4.3.

τ	$\frac{1}{2}\tau^2\|[B_2,B_1]e\|_\infty$	$\frac{1}{6}\tau^3\|R_Be\|_\infty$
1	1.25(−1)	20.833
0.1	1.25(−3)	2.08(−2)
0.01	1.25(−5)	2.08(−5)
0.001	1.25(−7)	2.08(−8)

Table 3.4.2: The second and third terms in the splitting error for B.

τ	splitting error for A	splitting error for B
1	1.25(−1)	20.833
0.1	1.25(−3)	2.08(−2)
0.01	1.25(−5)	2.08(−5)
0.001	1.25(−7)	1.25(−7)

Table 3.4.3: The total splitting errors for A and B.

We can see that the local splitting error for the operator A is determined by the second order term for each τ; however, for the splitting of B it does only for the last value ($\tau = 0.001$), for the bigger values the third order term defines the error.

The above example raises the question of a suitable choice of the splitting time discretization parameter τ: when it is too big, we lose the order; otherwise, when it is too small, we have to execute too many computations in order to get an approximation at some fixed time level. One can get a good hint by using the magnitude of $\|R_A\|$. However, formula (3.4.28) shows that the expressions for the coefficients at τ^k become more and more complicated with increasing k. To avoid this problem, i.e., to write operator R_A in a more compact form, we can use the Baker-Campbell-Hausdorff (BCH) formula [143], which is very useful for the sequential splitting with two operators. The idea is to write the product of two matrix exponentials as a matrix exponential of one matrix.[6] We seek such $C_n = C_n(A_2, A_1), n = 0, 1, \ldots$ for which the relation

$$e^{\tau A_2} e^{\tau A_1} = e^{\sum_{n=0}^{\infty} \tau^n C_n} \tag{3.4.31}$$

[6]The *Baker-Campbell-Hausdorff formula* is the solution to $Z = \log(e^X e^Y)$ for non-commuting X and Y. It was first noted in print by Campbell [21], elaborated by Poincaré [109] and Baker [6], and systematized by Hausdorff [66]. The formula below was introduced by Dynkin. For an arbitrary Lie algebra the formula is cumbersome and in practice it is not perspicuous. However, for a matrix Lie group we obtain a simpler formula:

$$Z = \sum_{n>0} \frac{(-1)^{n-1}}{n} \sum_{\substack{r_i+s_i>0 \\ 1\le i\le n}} \frac{X^{r_1}Y^{s_1}\cdots X^{r_n}Y^{s_n}}{r_1!s_1!\cdots r_n!s_n!}. \tag{3.4.29}$$

The Zassenhaus formula [155] is

$$e^{t(X+Y)} = e^{tX}\ e^{tY}\ e^{-\frac{t^2}{2}[X,Y]}\ e^{\frac{t^3}{6}(2[Y,[X,Y]]+[X,[X,Y]])}\ e^{t^4\ldots}\ldots \tag{3.4.30}$$

which shows the necessity and sufficiency of commutatitivity of the operators for the relation $e^{tX}\ e^{tY} = e^{tY}\ e^{tX}$. For some more details we refer to [60].

holds. It is not difficult to show that, for the first coefficients, we have

$$C_0 = C_0(A_2, A_1) = 0, \quad C_1 = C_1(A_2, A_1) = A_1 + A_2,$$
$$C_2 = C_2(A_2, A_1) = \tfrac{1}{2}[A_2, A_1], \quad C_3 = C_3(A_2, A_1) = \tfrac{1}{6}[A_2 - A_1, \tfrac{1}{2}[A_2, A_1]]. \tag{3.4.32}$$

For the further values n ($n = 4, 5 \ldots$) for the matrices $C_n = C_n(A_2, A_1)$ there is a recursion formula, see (3.4.29) in the footnote. We remark that from this formula follows also the fact that the sequential splitting error for two operators vanishes if and only if the operators commute. According to the definition of the exponential of operators, and using the formulae (3.4.32) in (3.4.31), we obtain

$$e^{\tau(A_1+A_2)} = I + \tau(A_1 + A_2) + \frac{1}{2!}\tau^2(A_1 + A_2)^2 + \frac{1}{3!}\tau^3(A_1 + A_2)^3 + \mathcal{O}(\tau^4), \tag{3.4.33}$$

and

$$e^{\tau A_2}e^{\tau A_1} = e^{\tau(A_1+A_2)+\sum_{n=2}^{\infty}\tau^n C_n} = I + \tau(A_1 + A_2) + \sum_{n=2}^{\infty}\tau^n C_n +$$
$$+\frac{1}{2!}[\tau(A_1 + A_2) + \sum_{n=2}^{\infty}\tau^n C_n]^2 + \frac{1}{3!}[\tau(A_1 + A_2) + \sum_{n=2}^{\infty}\tau^n C_n]^3 + \mathcal{O}(\tau^4). \tag{3.4.34}$$

It is easy to see that expressions (3.4.33) and (3.4.34) are equal up to the first order in τ. Moreover, the coefficient of the second order term in the difference (3.4.34)-(3.4.33) reads

$$\frac{1}{2!}\,[A_2, A_1]\,. \tag{3.4.35}$$

The resulting difference operator at τ^3 can be written as

$$R_A = C_3 + \frac{1}{2!}(A_1 + A_2)C_2 + \frac{1}{2!}C_2(A_1 + A_2).$$

Applying the corresponding expressions for C_2 and C_3, we get

$$R_A = \frac{1}{12}\,[A_2 - A_1, [A_2, A_1]] + \frac{1}{4}\,(A_2 + A_1)\,[A_2, A_1] + \frac{1}{4}[A_2, A_1]\,(A_2 + A_1)\,. \tag{3.4.36}$$

Using the trivial upper bound of the norm for the commutator, we get

$$\|R_A\| \le (\frac{1}{6}\,\|A_2 - A_1\|_\infty + \frac{1}{2}\,\|A\|_\infty)\,\|[A_2, A_1]\|_\infty\,. \tag{3.4.37}$$

From the above formula one can draw the following conclusions. In any non-trivial case, the estimating expression vanishes only if the operators commute. However, if the norm of the original operator or the norm of the difference of the split operators are big, then the obtained upper bound can also be big. Therefore the error may be (but not necessarily is) significant, even if the commutator norm is relatively small. This is in accordance with the results of our numerical experiment, since

$$\|A\|_\infty = 1, \quad \|B\|_\infty = 500.0005, \quad \|A_2 - A_1\|_\infty = \frac{1}{2}, \quad \|B_2 - B_1\|_\infty = 749.99925,$$

so $\|B_2 - B_1\|_\infty >> \|A_2 - A_1\|_\infty$ and $\|B\|_\infty >> \|A\|_\infty$.

The dominance of the second order term, by using the estimation (3.4.37) means the condition

$$\tau \leq \frac{3}{\|A_2 - A_1\|_\infty + 3\,\|A\|_\infty}. \tag{3.4.38}$$

In our example, for the splitting of operator B, this gives the bound $\tau \leq 1.33 \cdot 10^{-3}$. However, we can give another estimation, since, R_A can be rewritten as

$$R_A = \frac{1}{6}\big\{\,(2A_2 + A_1)\,[A_2, A_1] + [A_2, A_1]\,(2A_1 + A_2)\,\big\}. \tag{3.4.39}$$

Hence, the condition (3.4.38) can be modified as

$$\tau \leq \frac{3}{2\big\{\,\|2A_2 + A_1\|_\infty + \|2A_1 + A_2\|_\infty\,\big\}}. \tag{3.4.40}$$

In our example, for the splitting of operator B, this gives the bound $\tau \leq 1.0 \cdot 10^{-3}$, which is sharper (c.f. numerical results in Table 3.4.3).

3.4.3 Consistency and convergence of the operator splitting discretization methods

In this section we consider again the abstract Cauchy problem (3.1.1), but in a more general setting: the boundedness of the linear operator A will not be assumed.

We consider the Cauchy problem

$$\begin{cases} u'(t) = A_0 u(t) \quad t \in (0, t^\star], \\ u(0) = u_0, \end{cases} \tag{3.4.41}$$

in a Banach space $\mathbf{X}$, where $A_0 : \mathbf{X} \to \mathbf{X}$ is a closed, densely defined linear operator with the domain of definition $D(A_0)$. Assume that A_0 generates a C_0-semigroup $\{S_0(t)\}_{t\in[0,t^\star]}$. Then, according to the so-called growth estimation condition, there exist constants $\omega_0 \in \mathbb{R}$ and $M_0 \geq 1$ such that

$$\|S_0(t)\| \leq M_0 e^{\omega_0 t}, \quad t \in [0, t^\star]. \tag{3.4.42}$$

Moreover, for any $u_0 \in D(A_0)$, (3.4.41) has the unique classical solution (see e.g., [37])

$$u(t) = S_0(t)u_0, \quad t \in [0, t^\star]. \tag{3.4.43}$$

Assume that

$$A_0 = A_1 + A_2, \tag{3.4.44}$$

where A_1 and A_2 are generators of such C_0-semigroups $\{S_1(t)\}_{t\geq 0}$ and $\{S_2(t)\}_{t\geq 0}$, which can be approximated more easily than $\{S_0(t)\}_{t\geq 0}$ (Details about the approximation of semigoups can be found in [5].) Furthermore, let

$$\begin{aligned} D_k = D(A_1^k) \bigcap D(A_2^k) \bigcap D(A_0^k), \quad k = 1, 2, 3 \ \text{ dense in } X \\ \text{and} \quad A_{i\,|_{D_k}}^k, i = 0, 1, 2, \ k = 1, 2, 3 \text{ closed operators.} \end{aligned} \tag{3.4.45}$$

We will use the notation $D = \bigcap_{k=1}^3 D_k$.

Remark 3.4.5 *If we assume that the operators A_0, A_1 and A_2 are bounded, then the above conditions are automatically satisfied. If the operators are unbounded, but the assumption $D(A_1^k) = D(A_2^k) = D(A_0^k)$, $k = 1, 2, 3$, holds and the resolvent sets $\rho(A_i)$, $i = 0, 1, 2$ are not empty, as it is assumed for $k = 1, 2$ in [11], then (3.4.45) are automatically satisfied. (See [67] and also [37], Appendix B, B.14).*

As before, we divide the time interval $(0, t^\star]$ of the problem into N sub-intervals of equal length $\tau = t_{n+1} - t_n$, defining the mesh ω_τ. Then, on each sub-interval $(t_n, t_{n+1}]$, $n = 0, 1, \ldots, N-1$ the approximate solution $\nu_{\rm spl}^{n+1}$ of $u(t_{n+1})$ is computed as

$$\nu_{\rm spl}^{n+1} = S_{\rm spl}(\tau)\nu_{\rm spl}^{n}, \tag{3.4.46}$$

where $S_{\rm spl}(\tau)$ is one of the operator splitting methods, introduced before. For the most classical operator splittings they are clearly defined as

1. sequential splitting: $S_{\rm seq}(\tau) = S_2(\tau)S_1(\tau)$,
2. symmetrically weighted sequential splitting: $S_{\rm swss}(\tau) = \frac{1}{2}(S_1(\tau)S_2(\tau) + S_2(\tau)S_1(\tau))$,
3. Strang-Marchuk splitting: $S_{\rm SM}(\tau) = S_1(\tau/2)S_2(\tau)S_1(\tau/2)$.

Since the operator splitting is a time-discretization method, therefore it is quite reasonable to investigate its convergence as the discretization parameter τ tends to zero. In order to do this, we will use the basic concept of such problems, namely, the Lax-Richtmyer theory, which leads the question of convergence to the investigation of the consistency and stability. Therefore, we start with recalling three definitions from [11].

Definition 3.4.6 *Let $T_h : \mathbf{X} \times [0, t^\star - \tau] \to \mathbf{X}$ be defined as*

$$T_\tau(u_0, t) = S_0(\tau)u(t) - S_{\rm spl}(\tau)u(t). \tag{3.4.47}$$

For each u_0 and t, $T_\tau(u_0, t)$ is called the local truncation error of the corresponding splitting method.

Hence, the meaning of the local truncation error is the following: if we start from the exact solution $u(t)$ at any fixed point $t \in [0, t^\star]$, then $T_\tau(u_0, t)$ shows the difference between the exact and split solutions at time $t + \tau$, i.e.,

$$T_\tau(u_0, t) = Err_{\rm spl}(u(t), \tau). \tag{3.4.48}$$

(C.f. notation in (3.1.9).) Therefore, at $t = 0$ the local truncation error is the local splitting error, i.e., $T_\tau(u_0, 0) = Err_{\rm spl}(\tau)$.

Definition 3.4.7 *The splitting method is called consistent on $[0, t^\star]$ if*

$$\lim_{\tau \to 0} \sup_{0 \le t_n \le t^\star - \tau} \frac{\|T_\tau(u_0, t_n)\|}{\tau} = 0 \tag{3.4.49}$$

whenever $u_0 \in \mathcal{B}$, $\mathcal{B}$ being some dense subspace of $\mathbf{X}$.

Definition 3.4.8 *If in the consistency relation (3.4.49) we have*

$$\sup_{0 \le t_n \le T - t^\star} \frac{\|T_\tau(u_0, t_n)\|}{\tau} = \mathcal{O}(\tau^p), \quad p > 0, \tag{3.4.50}$$

then the method is said to be (consistent) of order p.

In the following we analyze the local truncation error, which from (3.4.47) can be rewritten as

$$T_\tau(u_0,t) = (S_0(\tau) - S_{\rm spl}(\tau))\, u(t). \tag{3.4.51}$$

First we assume that the operators $A_0, A_1, A_2 : \mathbf{X} \to \mathbf{X}$ are bounded and they are defined on the entire space. Then we have estimations for the local splitting error, and on the base of (3.4.51), we can write:

$$\|T_\tau(u_0,t)\| = \|Err_{\rm spl}(u(t),\tau)\| \le E(\tau)\|u(t)\|, \tag{3.4.52}$$

with $E(\tau) = \mathcal{O}(\tau^{p+1})$ (p is the order of the given splitting). Using (3.4.43) and (3.4.42), we have

$$\|T_\tau(u_0,t)\| \le E(\tau) M_0 e^{|\omega_0|t^\star}\|u_0\| = const \cdot E(\tau). \tag{3.4.53}$$

This proves

Theorem 3.4.9 *For bounded operators all the considered operator splitting methods are consistent and their order of consistency equals to the order of the operator splitting.*

The analysis of the consistency for unbounded operators is much more complicated. For the sequential splitting it is proven in [11] that it is consistent in first order. In the following we give the results for the higher-order operator splittings, namely, for the Strang-Marchuk splitting and the symmetrically weighted sequential splitting.

The following formula will play a basic role in our investigations.

Theorem 3.4.10 *For any C_0-semigroup $\{S(t)\}_{t\ge 0}$ of bounded linear operators with corresponding infinitesimal generator A, we have the Taylor series expansion*

$$S(t)u_0 = \sum_{j=0}^{n-1} \frac{t^j}{j!} A^j u_0 + \frac{1}{(n-1)!}\int_0^t (t-s)^{n-1} S(s)A^n u_0 \mathrm{d}s \text{ for all } u_0 \in D(A^n), \tag{3.4.54}$$

see [67], Section 11.8. Particularly, for $n = 3, 2$ and 1 we get the relations

$$S(\tau)u_0 = u_0 + \tau A u_0 + \frac{\tau^2}{2}A^2 u_0 + \frac{1}{2}\int_0^\tau (\tau - s)^2 S(s)A^3 u_0 \,\mathrm{d}s, \tag{3.4.55}$$

$$S(\tau)u_0 = u_0 + \tau A u_0 + \int_0^\tau (\tau - s) S(s)A^2 u_0 \,\mathrm{d}s \tag{3.4.56}$$

and

$$S(\tau)u_0 = u_0 + \int_0^\tau S(s)A u_0 \,\mathrm{d}s, \tag{3.4.57}$$

respectively. The following lemmas will also be helpful (see [154], Chapter II.6, Theorem 2).

Lemma 3.4.11 *Let A and B be closed linear operators from $D(A) \subset \mathbf{X}$ and $D(B) \subset \mathbf{X}$, respectively, into $\mathbf{X}$. If $D(A) \subset D(B)$, then there exists a constant $\hat{C}$ such that*

$$\|Bu_0\| \le \hat{C}(\|Au_0\| + \|u_0\|) \quad \text{for all} \quad u_0 \in D(A). \tag{3.4.58}$$

This implies that there exists a universal constant $\hat{C}$ by which for $u_0 \in D_k,\ k = 1, 2, 3$

$$\|A_i^k u_0\| \le \hat{C}(\|A_j^k u_0\| + \|u_0\|) \quad i, j = 0, 1, 2, \tag{3.4.59}$$

where D_k are according to (3.4.45).

Lemma 3.4.12 *Let A be an infinitesimal generator of a C_0-semigroup $\{S(t)\}_{t\geq 0}$. Let $t^\star > 0$ and $n \in \mathbb{N}$ arbitrary. If $u_0 \in D(A^n)$, then $u(t) = S(t)u_0 \in D(A^n)$ for $0 \leq t \leq t^\star$, and we have*

$$\sup_{[0,t^\star]} \|A^k u(t)\| \leq C_k(t^\star), \quad k = 0, 1, \ldots, n, \tag{3.4.60}$$

where $C_k(t^\star)$ are constants independent of τ.

Proof. Let $z(t) = A^{n-1}u(t) = A^{n-1}S(t)u_0 = S(t)A^{n-1}u_0$. Clearly, $u_0 \in D(A^n)$ implies $A^{n-1}u_0 \in D(A)$. It is known from the theory of C_0-semigroups (see [37], Chapter II, Lemma 1.3) that then $S(t)A^{n-1}u_0 \in D(A)$, i.e., $A^{n-1}u(t) \in D(A)$. Consequently, $u(t) \in D(A^n)$. Moreover,

$$\sup_{[0,t^\star]} \|A^k u(t)\| = \sup_{[0,t^\star]} \|A^k S(t)u_0\| = \sup_{[0,t^\star]} \|S(t)A^k u_0\| \leq Me^{|\omega|t^\star}\|A^k u_0\| \tag{3.4.61}$$

for $k = 0, 1, \ldots, n$. ■

Now we will consider the symmetrically weighted sequential splitting. Our aim is to show its second-order consistency for generators of C_0-semigroups. By using (3.4.54) for $n = 3$, for $u_0 \in D$ we have

$$\begin{aligned} S_2(\tau)S_1(\tau)u_0 &= S_1(\tau)u_0 + \tau A_2 S_1(\tau)u_0 + \frac{\tau^2}{2}A_2^2 S_1(\tau)u_0 + \\ &\quad \frac{1}{2}\int_0^\tau (\tau - s)^2 S_2(s)A_2^3 S_1(\tau)u_0 \, \mathrm{d}s \end{aligned} \tag{3.4.62}$$

and similarly,

$$\begin{aligned} S_1(\tau)S_2(\tau)u_0 &= S_2(\tau)u_0 + \tau A_1 S_2(\tau)u_0 + \frac{\tau^2}{2}A_1^2 S_2(\tau)u_0 + \\ &\quad \frac{1}{2}\int_0^\tau (\tau - s)^2 S_1(s)A_1^3 S_2(\tau)u_0 \, \mathrm{d}s. \end{aligned} \tag{3.4.63}$$

Applying (3.4.55), (3.4.56) and (3.4.57) for the semigroups $\{S_1(t)\}_{t\geq 0}$ and $\{S_2(t)\}_{t\geq 0}$ and substituting the corresponding expressions into the first, second and third terms on the right-hand side of (3.4.62), we get

$$\begin{aligned} &\frac{1}{2}[S_2(\tau)S_1(\tau)u_0 + S_1(\tau)S_2(\tau)u_0] = \\ &\quad u_0 + \tau(A_1 + A_2)u_0 + \frac{\tau^2}{2}(A_1 + A_2)^2 u_0 \\ &\quad + \frac{1}{4}\int_0^\tau (\tau - s)^2 S_1(s)A_1^3 u_0 \, \mathrm{d}s + \frac{1}{4}\int_0^\tau (\tau - s)^2 S_2(s)A_2^3 u_0 \, \mathrm{d}s \\ &\quad + \frac{1}{2}\tau A_2 \int_0^\tau (\tau - s) S_1(s)A_1^2 u_0 \, \mathrm{d}s + \frac{1}{2}\tau A_1 \int_0^\tau (\tau - s) S_2(s)A_2^2 u_0 \, \mathrm{d}s \\ &\quad + \frac{1}{4}\tau^2 A_2^2 \int_0^\tau S_1(s)A_1 u_0 \, \mathrm{d}s + \frac{1}{4}\tau^2 A_1^2 \int_0^\tau S_2(s)A_2 u_0 \, \mathrm{d}s \\ &\quad + \frac{1}{4}\int_0^\tau (\tau - s)^2 S_2(s)A_2^3 S_1(\tau)u_0 \, \mathrm{d}s + \frac{1}{4}\int_0^\tau (\tau - s)^2 S_1(s)A_1^3 S_2(\tau)u_0 \, \mathrm{d}s. \end{aligned} \tag{3.4.64}$$

On the other hand, we have

$$S_0(\tau)u_0 = u_0 + \tau A_0 u_0 + \frac{\tau^2}{2}A_0^2 u_0 + \frac{1}{2}\int_0^\tau (\tau - s)^2 S_0(s)A_0^3 u_0 \, \mathrm{d}s, \tag{3.4.65}$$

so the difference is

$$\frac{1}{2}[S_2(\tau)S_1(\tau)u_0 + S_1(\tau)S_2(\tau)u_0] - S_0(\tau)u_0 = \tag{3.4.66}$$

$$+\frac{1}{4}\int_0^\tau (\tau-s)^2 S_1(s)A_1^3u_0\,\mathrm{d}s + \frac{1}{4}\int_0^\tau (\tau-s)^2 S_2(s)A_2^3u_0\,\mathrm{d}s \tag{3.4.67}$$

$$+\frac{1}{2}\tau A_2\int_0^\tau (\tau-s)S_1(s)A_1^2u_0\,\mathrm{d}s + \frac{1}{2}\tau A_1\int_0^\tau (\tau-s)S_2(s)A_2^2u_0\,\mathrm{d}s \tag{3.4.68}$$

$$+\frac{1}{4}\tau^2A_2^2\int_0^\tau S_1(s)A_1u_0\,\mathrm{d}s + \frac{1}{4}\tau^2A_1^2\int_0^\tau S_2(s)A_2u_0\,\mathrm{d}s \tag{3.4.69}$$

$$+\frac{1}{4}\int_0^\tau (\tau-s)^2 S_2(s)A_2^3S_1(\tau)u_0\,\mathrm{d}s + \frac{1}{4}\int_0^\tau (\tau-s)^2 S_1(s)A_1^3S_2(\tau)u_0\,\mathrm{d}s \tag{3.4.70}$$

$$-\frac{1}{2}\int_0^\tau (\tau-s)^2 S_0(s)A_0^3u_0\,\mathrm{d}s. \tag{3.4.71}$$

Lemma 3.4.13 *Let A_0, A_1 and A_2 be infinitesimal generators of the C_0-semigroups $\{S_0(t)\}_{t\geq 0}$, $\{S_1(t)\}_{t\geq 0}$ and $\{S_2(t)\}_{t\geq 0}$, respectively. Assume that (3.4.44) and (3.4.45) are satisfied, and let $t^\star > 0$. Then for all $u_0 \in D$ the relation*

$$\left\|\frac{1}{2}[S_2(\tau)S_1(\tau)u_0 + S_1(\tau)S_2(\tau)u_0] - S_0(\tau)u_0\right\| \leq \tau^3 C(t^\star)(\|A_0^3u_0\|+\|A_0^2u_0\|+\|A_0u_0\|+\|u_0\|) \tag{3.4.72}$$

holds for $\tau \in [0, t^\star]$, where $C(t^\star)$ is a constant independent of τ.

Proof. We estimate the terms on the right-hand side of (3.4.66)–(3.4.71). We will often exploit the fact that the semigroups under consideration are C_0-semigroups, and so the growth estimation condition (3.4.42) is valid for each of them:

$$\|S_i(t)\| \leq M_i e^{\omega_i t}, \quad \forall t \in [0, t^\star], \quad i = 0, 1, 2, \tag{3.4.73}$$

where $M_i \geq 1, \omega_i \in \mathbb{R}$ are given constants. In the two terms under (3.4.67) and that under (3.4.71) we can make the following estimate:

$$\left\|\int_0^\tau (\tau-s)^2 S_i(s)A_i^3u_0\,\mathrm{d}s\right\| \leq M_i e^{|\omega_i|\tau}\|A_i^3u_0\|\frac{\tau^3}{3} \quad i = 0, 1, 2. \tag{3.4.74}$$

For the first term in (3.4.68) by using Lemma 3.4.11 we can write

$$\left\|\frac{1}{2}\tau A_2\int_0^\tau (\tau-s)S_1(s)A_1^2u_0\,\mathrm{d}s\right\| \leq \frac{\hat{C}}{2}\tau\left\|A_1\int_0^\tau (\tau-s)S_1(s)A_1^2u_0\,\mathrm{d}s\right\| \tag{3.4.75}$$

$$+\frac{\hat{C}}{2}\tau\left\|\int_0^\tau (\tau-s)S_1(s)A_1^2u_0\,\mathrm{d}s\right\|. \tag{3.4.76}$$

Using (3.4.56) twice and the fact that all semigroups commute with their generator, we get

$$A_1\int_0^\tau (\tau-s)S_1(s)A_1^2u_0\,\mathrm{d}s = A_1(S_1(\tau)u_0 - \tau A_1u_0 - u_0) =$$

$$= S_1(\tau)A_1u_0 - \tau A_1^2u_0 - A_1u_0 = \int_0^\tau (\tau-s)S_1(s)A_1^3u_0\,\mathrm{d}s. \tag{3.4.77}$$

Hence, for term (3.4.75) we obtain the estimate

$$\frac{\hat{C}}{2}\tau\left\|A_1\int_0^\tau(\tau-s)S_1(s)A_1^2u_0\,\mathrm{d}s\right\|\le M_1e^{|\omega_1|\tau}\|A_1^3u_0\|\frac{\tau^3}{4}\hat{C}. \tag{3.4.78}$$

Term (3.4.76) can be estimated by

$$\frac{\hat{C}}{2}\tau\left\|\int_0^\tau(\tau-s)S_1(s)A_1^2u_0\,\mathrm{d}s\right\|\le M_1e^{|\omega_1|\tau}\|A_1^2u_0\|\frac{\tau^3}{4}\hat{C}. \tag{3.4.79}$$

So,

$$\left\|\frac{1}{2}\tau A_2\int_0^\tau(\tau-s)S_1(s)A_1^2u_0\,\mathrm{d}s\right\|\le M_1e^{|\omega_1|\tau}(\|A_1^3u_0\|+\|A_1^2u_0\|)\frac{\tau^3}{4}. \tag{3.4.80}$$

Similarly, for the second term in (3.4.68) the following relation is valid:

$$\left\|\frac{1}{2}\tau A_1\int_0^\tau(\tau-s)S_2(s)A_2^2u_0\,\mathrm{d}s\right\|\le M_2e^{|\omega_2|\tau}(\|A_2^3u_0\|+\|A_2^2u_0\|)\frac{\tau^3}{4}. \tag{3.4.81}$$

For the estimate of the first term of (3.4.69) on the base of Lemma 3.4.11 we can write

$$\begin{aligned}\left\|\frac{1}{4}\tau^2A_2^2\int_0^\tau S_1(s)A_1u_0\,\mathrm{d}s\right\| &= \frac{\hat{C}}{4}\tau^2\left\|A_1^2\int_0^\tau S_1(s)A_1u_0\,\mathrm{d}s\right\|+ &(3.4.82)\\ &+ \frac{\hat{C}}{4}\tau^2\left\|\int_0^\tau S_1(s)A_1u_0\,\mathrm{d}s\right\|, &(3.4.83)\end{aligned}$$

where for term (3.4.82) we have

$$\frac{\hat{C}}{4}\tau^2\left\|A_1^2\int_0^\tau S_1(s)A_1u_0\,\mathrm{d}s\right\|=\frac{\hat{C}}{4}\tau^2\left\|\int_0^\tau S_1(s)A_1^3u_0\,\mathrm{d}s\right\|\le\frac{\hat{C}}{4}\tau^3M_1e^{|\omega_1|\tau}\|A_1^3u_0\|, \tag{3.4.84}$$

and for term (3.4.83):

$$\frac{\hat{C}}{4}\tau^2\left\|\int_0^\tau S_1(s)A_1u_0\,\mathrm{d}s\right\|\le\frac{\hat{C}}{4}\tau^3M_1e^{|\omega_1|\tau}\|A_1u_0\|. \tag{3.4.85}$$

Consequently,

$$\left\|\frac{1}{4}\tau^2A_2^2\int_0^\tau S_1(s)A_1u_0\,\mathrm{d}s\right\|\le M_1e^{|\omega_1|\tau}\hat{C}(\|A_1^3u_0\|+\|A_1u_0\|)\frac{\tau^3}{4}. \tag{3.4.86}$$

In a similar way, the second term of (3.4.69) is estimated by

$$\left\|\frac{1}{4}\tau^2A_1^2\int_0^\tau S_2(s)A_2u_0\,\mathrm{d}s\right\|\le M_2e^{|\omega_2|\tau}\hat{C}(\|A_2^3u_0\|+\|A_2u_0\|)\frac{\tau^3}{4}. \tag{3.4.87}$$

For the first term of (3.4.70) one can write

$$\begin{aligned}&\left\|\frac{1}{4}\int_0^\tau(\tau-s)^2S_2(s)A_2^3S_1(\tau)u_0\,\mathrm{d}s\right\|\le M_2e^{|\omega_2|\tau}\|A_2^3S_1(\tau)u_0\|\frac{\tau^3}{12}\le &(3.4.88)\\ &\le M_2e^{|\omega_2|\tau}\hat{C}(\|A_1^3S_1(\tau)u_0\|+\|S_1(\tau)u_0\|)\frac{\tau^3}{12}\le M_1e^{|\omega_1|\tau}M_2e^{|\omega_2|\tau}\hat{C}(\|A_1^3u_0\|+\|u_0\|)\frac{\tau^3}{12}.\end{aligned}$$

Here we have used that

$$\|A_2^3 S_1(\tau)u_0\| \le \hat{C}(\|A_1^3 S_1(\tau)u_0\| + \|S_1(\tau)u_0\|). \tag{3.4.89}$$

Finally, in a similar manner, the second term of (3.4.70) is estimated by

$$\left\| \frac{1}{4}\int_0^\tau (\tau - s)^2 S_1(s) A_1^3 S_2(\tau)u_0 \, \mathrm{d}s \right\| \le M_1 e^{|\omega_1|\tau} M_2 e^{|\omega_2|\tau} \hat{C}(\|A_2^3 u_0\| + \|u_0\|)\frac{\tau^3}{12}. \tag{3.4.90}$$

■

To prove the second-order consistency of the symmetrically weighted sequential splitting, we need a uniform bound, proportional to τ^3 on

$$\left\| \frac{1}{2}\left[S_2(\tau)S_1(\tau)u(t) + S_1(\tau)S_2(\tau)u(t)\right] - S_0(\tau)u(t) \right\| \tag{3.4.91}$$

as t runs from 0 to $t^\star - \tau$, where $u(t) = S_0(t)u_0$ is the exact solution of the original problem (3.4.41).

Lemma 3.4.13, (3.4.59) and Lemma 3.4.12 imply the following

Theorem 3.4.14 *Let the conditions of Theorem 3.4.13 be satisfied. Then for any $u_0 \in D$ we have a uniform bound*

$$\left\| \frac{1}{2}[S_2(\tau)S_1(\tau)u(t) + S_1(\tau)S_2(\tau)u(t)] - S_0(\tau)u(t) \right\| \le \tau^3 C(t^\star), \tag{3.4.92}$$

where $C(t^\star)$ is a constant independent of τ. Hence, the symmetrically weighted sequential splitting has second order consistency for unbounded generators, too.

For the Strang-Marchuk splitting we can prove a similar result.

Theorem 3.4.15 *Let the conditions of Theorem 3.4.13 be satisfied. Then for any $u_0 \in D$ we have a uniform bound*

$$\|S_1(0.5\tau)S_2(\tau)S_1(0.5\tau)u(t) - S_0(\tau)u(t)\| \le \tau^3 C(t^\star), \tag{3.4.93}$$

where $C(t^\star)$ is a constant independent of τ. Hence, the Strang-Marchuk splitting has second order consistency for unbounded generators, too.

The proof of Theorem 3.4.15 is similar to the proof of Theorem 3.4.14 and can be found in [42].

According to the Lax-Richtmyer concept, to show the convergence of a consistent operator splitting method, we have to show the stability, which is defined as follows.

Definition 3.4.16 *A splitting method is called stable on $[0, t^\star]$, if the relation*

$$\| \left(S_{\mathrm{spl}}(\tau)\right)^n \| \le C_{spl}, \tag{3.4.94}$$

holds for all $n\tau \le t^\star$, where C_{spl} is a constant independent of τ.

To prove the convergence for the sequential splitting, the Strang-Marchuk splitting and the symmetrically weighted sequential splitting, we have to show the stability, i.e., the property (3.4.94). Let us assume that for the semigroups $S_1(t)$ and $S_2(t)$, generated by the operators A_1 and A_2, respectively, (3.4.73) holds with $M_1 = M_2 = 1$. Then, for the sequential splitting we have

$$\begin{aligned} \| \left(S_{\text{seq}}(\tau)\right)^n \| = \| \left(S_1(\tau) \cdot S_2(\tau)\right)^n \| \leq \|S_1(\tau)\|^n \cdot \|S_2(\tau)\|^n \leq e^{(\omega_1+\omega_2)n\tau} = \\ = e^{(\omega_1+\omega_2)t^\star} = const, \end{aligned} \tag{3.4.95}$$

which proves the stability of the sequential splitting.
To get the same result under such a growth condition for the Strang-Marchuk splitting is obvious.
For the weighted sequential splitting under the above assumption we have

$$\begin{aligned} \| \left(S_{\text{wss}}(\tau)\right)^n \| = \| \left(\theta \cdot S_1(\tau) \cdot S_2(\tau) + (1-\theta) \cdot S_2(\tau) \cdot S_1(\tau)\right)^n \| \leq \\ \leq \|\theta \cdot S_1(\tau) \cdot S_2(\tau) + (1-\theta) \cdot S_2(\tau) \cdot S_1(\tau)\|^n \leq \\ \leq \|S_1(\tau)\|^n \cdot \|S_2(\tau)\|^n \leq e^{(\omega_1+\omega_2)n\tau} = e^{(\omega_1+\omega_2)t^\star} = const, \end{aligned} \tag{3.4.96}$$

which proves the stability of both the weighted sequential splitting and the symmetrically weighted sequential splitting.
For the modified additive splitting we have:

$$\begin{aligned} \| \left(S_{\text{mas}}(\tau)\right)^n \| = \left\| \left\{ \frac{1}{2} \left(S_1(2\tau) + S_2(2\tau)\right) \right\}^n \right\| \leq \left\| \left\{ \frac{1}{2} \left(S_1(2\tau) + S_2(2\tau)\right) \right\} \right\|^n \leq \\ \left\{ \frac{1}{2} (\|S_1(2\tau)\| + \|S_2(2\tau)\|) \right\}^n \leq \left\{ \frac{1}{2} (e^{2\tau\omega_1} + e^{2\tau\omega_2}) \right\}^n \leq \left\{ e^{2\tau \max\{\omega_1,\omega_2\}} \right\}^n = \\ = e^{2t^\star \max\{\omega_1,\omega_2\}} = const, \end{aligned} \tag{3.4.97}$$

which proves the stability of the modified additive splitting.
When a semigroup $S(\tau)$ is generated by the bounded operator A then, due to its representation as $S(\tau) = \exp(\tau A)$ the estimation

$$\|S(\tau)\| \leq \exp(\tau \|A\|) \tag{3.4.98}$$

holds, i.e., for bounded generators the growth bound holds with $M = 1$. Hence, we can summarize our results as

Theorem 3.4.17 *Assume that in the Cauchy problem* (3.4.41) *the operators A_1 and A_2 are bounded. Then the sequential splitting, the Strang-Marchuk splitting, the weighted sequential splitting, the symmetrically weighted sequential splitting and the modified additive splitting are all convergent.*

When the growth estimation (3.4.73) holds with $M_i = 1$ and $\omega_i = 0$, then the semigroups are called contractive. Many important unbounded operators generate such semigroups, e.g., under some assumptions the diffusion operator and the advection operator. (For more details, see [37].) When A is a matrix, then, using the notion of the logarithmic norm, we can show that (3.4.73) holds with $M = 1$ and $\omega \leq 0$ for many important discretizations, i.e., the generated semigroup is contractive. Hence, we have

Theorem 3.4.18 *Assume that in the Cauchy problem* (3.4.41) *the operators* A_0, A_1 *and* A_2 *are generators of contractive semigroups. Then, under the required smoothness of the initial function, the sequential splitting, the Strang-Marchuk splitting, the weighted sequential splitting, the symmetrically weighted sequential splitting and the modified additive splitting are all convergent and have the same order as for the bounded generators.*

Based on the work [28], we note that the above splittings are convergent under rather more general assumptions, too. Let us assume that the following General Assumptions are satisfied:

(a) The operator $(A; D(A))$ generates the strongly continuous semigroup $T(t)_{t\geq 0}$ on the Banach space $\mathbf{X}$.

(b) The operator $(B; D(B))$ generates the strongly continuous semigroup $S(t)_{t\geq 0}$ on $\mathbf{X}$.

(c) The sum $A + B$, defined on $D(A + B) := \overline{D(A) \bigcap D(B)}$, generates the strongly continuous semigroup $U(t)_{t\geq 0}$ on $\mathbf{X}$.

Moreover, let us assume that the stability condition

$$\|[S(t/n)T(t/n)]^n\| \leq Me^{\omega t} \text{ for all } t \geq 0 \text{ and } n \in \mathbb{N} \tag{3.4.99}$$

holds with some constants $M \geq 1$ and $\omega \in \mathbb{R}$. Then, based on the Chernoff theorem (e.g., [37]), the following statement can be proven.

Theorem 3.4.19 *Under the General Assumptions the sequential splitting, the Strang-Marchuk splitting, the weighted sequential splitting, and the symmetrically weighted sequential splitting are convergent at a fixed time level* $t > 0$ *for any initial function in* $\mathbf{X}$*, if the stability condition* (3.4.99) *is satisfied.*

At the same time, we emphasize that for this case the convergence might be extremely slow.

3.4.4 Higher-order convergence of operator splittings

The use of higher-order time discretization methods is useful, because, in case of stability, typically they result in a higher-order of convergence. Since operator splitting is a time discretization method, this question has a great importance for splitting methods, too. There are several operator splitting methods of higher accuracy obtained directly from the comparison of exponential series. The general one-step splitting scheme (3.2.3) for two operators can be defined by the operator

$$r_{\text{spl}}(\tau A) = \sum_{i=1}^{s} \alpha_i \left(\prod_{j=1}^{r} \exp(\tau \beta_{ij} A_1) \cdot \exp(\tau \gamma_{ij} A_2) \right), \tag{3.4.100}$$

where $\sum_{i=1}^{s} \alpha_i = 1$ [73]. One could try to find suitable parameter choices that give higher-order procedures. But, as it is shown in [123], for order $p > 2$, some of the coefficients must be negative. This result was refined in [61], that under the non-negativity condition for α_i at least one of the parameters β_{ij} and γ_{ij} is negative, which, due to the problem of stability, makes the splitting less attractive.

A well-known fourth-order splitting method is the Yoshida-Suzuki method [153], [136]. Here, in the general one-step splitting scheme (3.2.3), the operator is defined as

$$r_{\mathrm{YS}}(\tau A) = r_{\mathrm{SM}}(\theta\tau A) \cdot r_{\mathrm{SM}}((1-2\theta)\tau A) \cdot r_{\mathrm{SM}}(\theta\tau A), \tag{3.4.101}$$

where r_{SM} is the operator of the Strang-Marchuk splitting, defined in (3.2.5), and $\theta = (2-\sqrt[3]{2})^{-1}$. Since $1-2\theta < 0$, therefore in the second subproblem a negative step size is taken, which usually makes the sub-problem ill-posed. (The time is reversed.) So, this approach seems to be of limited value.

Remark 3.4.20 *We emphasize that the presence of the negative time step does not mean that the numerical method is unstable in any case. Let us consider the following example [127]*

$$\frac{\partial u}{\partial t} = \frac{\partial (xu)}{\partial x} + \frac{\partial^2 u}{\partial x^2}. \tag{3.4.102}$$

(here the two operators do not commute, however, we can define the particular exact solution.) Denoting by $(A_1)_h$ the fourth order discretization of the operator $\partial_x x$ and by $(A_2)_h$ the fourth order discretization of the operator ∂_x^2, respectively, by the notation $S_h(\tau)$ for the sequential splitting, we can introduce splittings with different orders. (Then $(S_h(\tau))^T$ denotes the sequential splitting with the other ordering.) Namely,

- $S_h(\tau)$*: first order,*
- $S_h(\tau)(S_h(\tau))^T$*: second order,*
- $(S_h(\tau))^T \cdot S_h(\tau) \cdot S_h(\tau) \cdot S_h(\tau) \cdot (S_h(\tau))^T \cdot (-2S_h(\tau)) \cdot S_h(\tau) \cdot S_h(\tau) \cdot S_h(\tau)$*: third order.*

(There are very complicated expressions for the fourth and higher order splittings, too [126].) It is shown that these higher order schemes are stable with the indicated orders of accuracy. In [156], a semi-implicit finite difference operator splitting Padé method is applied for solving the higher-order non-linear Schrödinger equation, which describes the optical solution wave propagation in fibers. The method achieves fourth order of accuracy in space and has been proven to be stable by linear stability analysis.

It is known from the literature (e.g., [122]) that applying the same ODE solver by using two different step sizes and combining appropriately the obtained numerical solutions at each time step we can increase the convergence order of the method. Moreover, this technique allows us to estimate the absolute error of the underlying method. In this part we apply this procedure, widely known as Richardson extrapolation, to the operator splittings.

Consider again the Cauchy problem

$$\left.\begin{aligned} \frac{du(t)}{dt} &= Au(t), \quad t > 0; \\ u(0) &= u_0. \end{aligned}\right\} \tag{3.4.103}$$

Assume that we apply some convergent discretization method of order p to solving the problem (3.4.103). Let $y_\tau(t^\star)$ denote the numerical solution at a fixed time level $t^\star$ on a mesh with step size τ, denoted again by ω_τ. Then we have

$$u(t^\star) = y_\tau(t^\star) + \alpha(t^\star)\tau^p + \mathcal{O}(\tau^{p+1}). \tag{3.4.104}$$

Then on the meshes ω_{τ_1} and ω_{τ_2} with two different step sizes $\tau_1 < \tau$ and $\tau_2 < \tau$, the equalities

$$u(t^\star) = y_{\tau_1}(t^\star) + \alpha(t^\star)\tau_1^p + \mathcal{O}(\tau^{p+1}) \tag{3.4.105}$$

and

$$u(t^\star) = y_{\tau_2}(t^\star) + \alpha(t^\star)\tau_2^p + \mathcal{O}(\tau^{p+1}) \tag{3.4.106}$$

hold, respectively. We assume that $\omega := \omega_{\tau_1} \cap \omega_{\tau_2} \neq \emptyset$ and our aim is to get a mesh function on ω with higher accuracy $\mathcal{O}(\tau^{p+1})$. We define a mesh function $y_{\text{comb}}(t^\star)$ as follows:

$$y_{\text{comb}}(t^\star) = c_1 y_{\tau_1}(t^\star) + c_2 y_{\tau_2}(t^\star). \tag{3.4.107}$$

Let us substitute (3.4.105) and (3.4.106) into (3.4.107). Then we get

$$y_{\text{comb}}(t^\star) = (c_1 + c_2)u(t^\star) - (c_1\tau_1^p + c_2\tau_2^p)\alpha(t^\star) + \mathcal{O}(\tau^{p+1}). \tag{3.4.108}$$

From (3.4.108) one can see that a necessary condition for the combined method to be convergent is that the relation

$$c_1 + c_2 = 1 \tag{3.4.109}$$

holds. Moreover, we will only have a convergence order higher than p if

$$c_1\tau_1^p + c_2\tau_2^p = 0. \tag{3.4.110}$$

The solution of system (3.4.109)–(3.4.110) is $c_1 = -\tau_2^p/(\tau_1^p - \tau_2^p)$, $c_2 = 1 - c_1$. For example, if $\tau_2 = \tau_1/2$, then $\omega = \omega_{\tau_1}$ and, for $p = 1$ we have $c_1 = -1$ and $c_2 = 2$, and for $p = 2$ we have $c_1 = -1/3$ and $c_2 = 4/3$ (c.f. [73], p. 331).

The application of the same method by using two different time steps allows us to estimate the global error of the underlying method, see e.g., [111], p. 513. Formulas (3.4.105) and (3.4.106) allow us to determine the coefficient $\alpha(t^\star)$ approximately. Let us subtract (3.4.105) from (3.4.106). Then we get

$$0 = y_{\tau_2}(t^\star) - y_{\tau_1}(t^\star) + \alpha(t^\star)(\tau_2^p - \tau_1^p) + \mathcal{O}(\tau^{p+1}).$$

Expressing $\alpha(t^\star)$ gives

$$\alpha(t^\star) = \frac{y_{\tau_2}(t^\star) - y_{\tau_1}(t^\star)}{\tau_1^p - \tau_2^p} + \mathcal{O}\left(\frac{\tau^{p+1}}{\tau_1^p - \tau_2^p}\right). \tag{3.4.111}$$

The second term on the right-hand side is $\mathcal{O}(\tau)$, so the ratio $\hat{\alpha}(t^\star) := (y_{\tau_2}(t^\star) - y_{\tau_1}(t^\star))/(\tau_1^p - \tau_2^p)$ approximates $\alpha(t^\star)$ to the first order in τ. Then the absolute errors of the methods (3.4.105) and (3.4.106) can be approximated by the expressions $\hat{\alpha}(t^\star)\tau_1^p$ and $\hat{\alpha}(t^\star)\tau_2^p$, respectively, to the order $\mathcal{O}(\tau^{p+1})$ (a posteriori error estimates).

In what follows, we apply the Richardson extrapolation to the operator splittings.

For the case $p = 1$, we apply this approach to the sequential splitting, since it is a first-order time discretization method. By the choice $\tau_2 = \tau_1/2$, we put $c_1 = -1$ and $c_2 = 2$ and the Richardson-extrapolated sequential splitting for $w_{\text{Riseq}}^{(N)}(n\tau)$ reads as follows

$$w_{\text{Riseq}}^{(N)}((n+1)\tau) = \{-S_2(\tau)S_1(\tau) + 2(S_2(\tau/2)S_1(\tau/2))^2\} w_{\text{Riseq}}^{(N)}(n\tau), \tag{3.4.112}$$

for $n = 0, 1, \ldots, N$. According to our result, the method (3.4.112) has second order convergence.

For the case $p = 2$, clearly, we can extend our method if we use more numerical approximations. As a simple example, we consider the case where we have three numerical results on three different meshes with step-sizes τ_m $(m = 1, 2, 3)$. Due to the relation

$$u(t_n) = y_\tau(t_n) + \alpha_1(t_n)\tau^p + \alpha_2(t_n)\tau^{p+1} + \mathcal{O}(\tau^{p+2}), \tag{3.4.113}$$

on the mesh with step-size $\tau_m \leq \tau$ we have

$$u(t_n) = y_\tau(t_n) + \alpha_1(t_n)\tau_m^p + \alpha_2(t_n)\tau_m^{p+1} + \mathcal{O}(\tau^{p+2}). \tag{3.4.114}$$

Then on ω, which is the intersection of the above three meshes, we define a mesh-function y_τ as follows:

$$y_{\text{comb}}(t_n) = c_1 y_{\tau_1}(t_n) + c_2 y_{\tau_2}(t_n) + c_3 y_{\tau_2}(t_n). \tag{3.4.115}$$

By substitution into the above relations, for the unknown coefficients c_1, c_2 and c_3 we obtain the conditions

$$c_1 + c_2 + c_3 = 0, \tag{3.4.116}$$

$$c_1\tau_1^p + c_2\tau_2^p + c_3\tau_3^p = 0 \tag{3.4.117}$$

and

$$c_1\tau_1^{p+1} + c_2\tau_2^{p+1} + c_3\tau_3^{p+1} = 0. \tag{3.4.118}$$

The solution of system (3.4.116)–(3.4.118) yields the values of the coefficients in the approximation (3.4.115). For example, when $\tau_3 = \tau_2/2 = \tau_1/4$, which is the most typical choice, and $p = 1$, we get

$$c_1 = \frac{1}{3}, \quad c_2 = -2, \quad c_3 = \frac{8}{3}, \tag{3.4.119}$$

which results in third order convergence for the approximation in (3.4.115).

It is also possible to construct new methods by changing both the step size and the splitting method. This is the case if we apply the Richardson extrapolated sequential splitting in such a way that during the calculation with the halved time step we swap the sub-operators:

$$w_{\text{Rssos}}^{(N)}((n+1)\tau) = \{c_1 S_2(\tau)S_1(\tau) + c_2(S_1(\tau/2)S_2(\tau/2))^2\}w_{\text{Rssos}}^{(N)}(n\tau). \tag{3.4.120}$$

Note that (3.4.120) is not a special case of (3.4.107), where y_{τ_1} and y_{τ_2} belong to the same method, because changing the ordering of the sub-operators usually changes the splitting method. In this way, for the first method we have

$$u(t_n) = y_1(t_n) + \alpha(t_n)\tau_1^p + \mathcal{O}(\tau_1^{p+1}), \tag{3.4.121}$$

and for the second one, with time step τ_2

$$u(t_n) = y_2(t_n) + \beta(t_n)\tau_2^p + \mathcal{O}(\tau_1^{p+1}). \tag{3.4.122}$$

Then the combined method

$$y_{\text{comb}}(t_n) = (c_1 + c_2)u(t_n) - c_1\alpha(t_n)\tau_1^p - c_2\beta(t_n)\tau_2^p + \mathcal{O}(\tau_1^{p+1})$$

has a convergence order higher than p if

$$c_1 + c_2 = 1 \tag{3.4.123}$$

and

$$c_1\alpha(t_n)\tau_1^p + c_2\beta(t_n)\tau_2^p = 0. \tag{3.4.124}$$

Let us give expressions for $\alpha(t_n)$ and $\beta(t_n)$ for the sequential splitting in case of bounded operators. We have

$$y_1(t_n) = (e^{A_2\tau}e^{A_1\tau})^n u_0. \tag{3.4.125}$$

Computing the operator product under (3.4.125) by induction and comparing it with the exact solution

$$u(t_n) = e^{At_n}u_0 = (I + At_n + \frac{1}{2}A^2n^2\tau^2 + \mathcal{O}(\tau^3))u_0, \tag{3.4.126}$$

we obtain

$$u(t_n) - y_1(t_n) = \frac{n}{2}[A_1, A_2]\tau^2 u_0 + \mathcal{O}(\tau^3). \tag{3.4.127}$$

Hence

$$\alpha(t_n) = \frac{n}{2}[A_1, A_2]u_0. \tag{3.4.128}$$

The solution of the sequential splitting with a reverse operator sequence will be

$$y_2(t_n) = (e^{\frac{\tau}{2}A_1}e^{\frac{\tau}{2}A_2})^{2n}u_0, \tag{3.4.129}$$

from which

$$u(t_n) - y_2(t_n) = \frac{n}{4}[A_2, A_1]\tau^2 u_0 + \mathcal{O}(\tau^3),$$

i.e.,

$$\beta(t_n) = \frac{n}{4}[A_2, A_1]u_0. \tag{3.4.130}$$

Therefore the condition (3.4.124) has the form

$$c_1\frac{n}{2}[A_1, A_2]\tau^2 u_0 + c_2\frac{n}{4}[A_2, A_1]\tau^2 u_0 = 0,$$

which holds for all u_0 if and only if $c_1 = c_2/2$, which, together with condition (3.4.123), implies the coefficients

$$c_1 = \frac{1}{3}, \quad c_2 = \frac{2}{3}. \tag{3.4.131}$$

Hence, the iteration is

$$w_{\text{Rssos}}^{(N)}((n+1)\tau) = \left\{\frac{1}{3}S_2(\tau)S_1(\tau) + \frac{2}{3}(S_1(\tau/2)S_2(\tau/2))^2\right\} w_{\text{Rssos}}^{(N)}(n\tau) \tag{3.4.132}$$

and it is called *Richardson-extrapolated sequential splitting with operator swap (Rssos).*

Finally, some numerical experiments are presented in order to confirm our theoretical results. We will check the convergence order of the Richardson-extrapolated sequential splitting in matrix examples by exact solution of the sub-problems. (The case where we use numerical methods to the solution of the sub-problems will be discussed later.) We consider the Cauchy problem

$$\begin{cases} w'(t) = Aw(t), \quad t \in [0, 1] \\ \\ \qquad\quad w(0) = w_0 \end{cases} \tag{3.4.133}$$

with

$$A = \begin{bmatrix} -7 & 4 \\ -6 & -4 \end{bmatrix} \quad \text{and} \quad w_0 = (1, 1). \tag{3.4.134}$$

We decompose the matrix A as

$$A = A_1 + A_2 = \begin{bmatrix} -6 & 3 \\ -4 & 1 \end{bmatrix} + \begin{bmatrix} -1 & 1 \\ -2 & -5 \end{bmatrix} \tag{3.4.135}$$

and we apply second-order splitting methods: the Richardson-extrapolated sequential splitting, the symmetrically weighted sequential splitting and the Strang-Marchuk splitting with decreasing time steps τ. In the case of the Richardson-extrapolated sequential splitting each result was obtained by solving the problem by the sequential splitting both with τ and $\tau/2$ and combining the results by using the weight parameters -1 and 2, respectively. The obtained error norms at the end of the time interval are shown in Table 3.4.4.

τ	Riseq	Rssos	swss	SM
1	9.6506(-2)	9.2331(-2)	1.0301(-1)	7.8753(-2)
0.1	5.8213(-4)	6.0287(-4)	1.2699(-3)	3.3685(-4)
0.01	5.9761(-6)	6.0031(-6)	1.2037(-5)	3.3052(-6)
0.001	5.9823(-8)	5.9853(-8)	1.1974(-7)	3.3046(-8)

Table 3.4.4: Comparing the errors of the solutions obtained by the Richardson-extrapolated sequential splitting, without and with operator swap, the symmetrically weighted sequential splitting and the Strang-Marchuk splitting in example (3.4.133).

One can conclude that while all the methods have second order, the Richardson-extrapolated sequential splitting (with as well as without operator swap) gives a smaller error for each time step than the symmetrically weighted sequential splitting, and performs almost as well as the Strang-Marchuk splitting.

If we assume that the sequential splitting has already been applied for some time step τ, then to complete the Richardson-extrapolated sequential splitting (one sequential splitting with halved step size) practically takes equally as much time as the symmetrically weighted sequential splitting. However, both methods require more CPU time for the same time step than the Strang-Marchuk splitting. (Here we assumed that all computations are performed sequentially.) Since it is more correct to compare the accuracy of equally expensive methods, therefore some of the further comparisons will be restricted to the symmetrically weighted sequential splitting.

The Richardson-extrapolated sequential splitting (3.4.115) with (3.4.119) gives a third-order method. To check this theoretical result, we solved the problem (3.4.133) by use of this method. Table 3.4.5 convinces us about the expected third-order convergence.

In (3.4.111) we saw that the error of the Richardson-extrapolated sequential splitting can be estimated by using the solutions obtained by two different time steps τ_1 and τ_2. Applying this result to the sequential splitting, which is a first-order method, its global error at some time t_n can be estimated as $\hat{\alpha}(t_n)\tau$, where

$$\hat{\alpha}(t_n) = \frac{y_{\tau/2}(t_n) - y_\tau(t_n)}{\tau - \tau/2}, \tag{3.4.136}$$

and we used the time steps $\tau_1 = \tau$ and $\tau_2 = \tau/2$. The results are shown in Table 3.4.6.

τ	Example (3.4.133)
1	7.9792(-2)
0.1	4.4186(-6)
0.01	6.4643(-9)
0.001	7.5281(-12)

Table 3.4.5: Errors obtained when the Richardson-extrapolated sequential splitting is applied with three different time steps for example (3.4.133).

τ	Exact error norm	Estimated error norm
1	0.0215	0.1168
0.1	0.0012	0.0011
0.01	1.4966(-4)	1.4792(-4)
0.001	1.5281(-5)	1.5264(-5)

Table 3.4.6: Comparing the exact and estimated error norms of the solutions obtained by the Richardson-extrapolated sequential splitting in the example (3.4.133).

3.5 Numerical solution of the split sub-problems

In all our previous analysis, we assumed that after the application of the investigated operator splitting the split sub-problems were solved exactly. Clearly, this assumption is mostly unrealistic: we have to apply some numerical method to their solution.

In this approach two important questions can be formulated:

- The combination "operator splitting + numerical methods to the sub-problems" results in numerical methods suitable for the practical computations. How can these combined discretized methods be characterized? Which methods do they result in?
- These numerical methods are approximation in themselves therefore they have own approximation error. Clearly, this error is added to the splitting error and we should analyze the interaction of these error sources. Hence, for the total error analysis of the combined method, we arrive at a complex problem. Our task is to give the error analysis of these methods.

3.5.1 Combined discretization methods

Let us apply some numerical method to the solution of the split sub-problems with discretization parameter Δt. Clearly, the condition $\Delta t \leq \tau$ should be satisfied. Aiming at finishing the numerical solving process at the endpoint of the time intervals where the sub-problems are imposed, we select $\Delta t = \tau/K$, where K is some positive integer.

In this manner, for the different sub-problems different numerical methods can be chosen with different discretization parameters. Hence, assuming the use of the same numerical method at each splitting step, the total discretization operator depends on the choice of the splitting, the splitting step, the applied numerical methods and their step sizes. For instance, for the sequential splitting, by the choice of some numerical method NM1 with step size $\Delta t_1 = \tau/K_1$ for the first sub-problem, and a numerical method NM2 with the step size $\Delta t_2 = \tau/K_2$ for the second sub-problem, the total numerical discretization operator (which is, in fact, the stability function, considered in Section 2.4, see p.58) can

be defined as $r_{tot} = r_{tot}(\tau, NM1, K_1, NM2, K_2)$. (For simplicity, we will use the notation $r(\tau)$ when $NM1, K_1, NM2$ and K_2 are fixed.)
As an example, let us consider the sequential splitting applied to the Cauchy problem (3.4.41) with (3.4.44). We solve each problem with the explicit Euler (EE) method and we choose $\Delta t = \tau$ for both sub-problems. (I.e., $r_{tot} = r_{tot}(\tau, EE, 1, EE, 1)$). If we denote by y_1^n and y_2^n the approximations to $w_1^{(N)}(n\tau)$ and $w_2^{(N)}(n\tau)$, respectively, the numerical schemes are

$$\frac{y_1^{n+1} - y_1^n}{\tau} = A_1 y_1^n, \quad \frac{y_2^{n+1} - y_2^n}{\tau} = A_2 y_2^n \tag{3.5.1}$$

and $y_2^n = y_1^{n+1}$. Hence

$$y_2^{n+1} = (I + \tau A_2)(I + \tau A_1) y_1^n. \tag{3.5.2}$$

Consequently,

$$r(\tau(A_1 + A_2)) = (I + \tau A_2)(I + \tau A_1). \tag{3.5.3}$$

If we choose now $\Delta t = \tau/K$ for both sub-problems, i.e., $r_{tot} = r_{tot}(\tau, EE, K, EE, K)$, then

$$r(\tau(A_1 + A_2)) = (I + \frac{\tau}{K} A_2)^K (I + \frac{\tau}{K} A_1)^K. \tag{3.5.4}$$

Obviously, in order to prove the convergence of the combined numerical discretization, we can apply the Lax-Richtmyer theorem. In a simple case we illustrate this by the following statement.

Theorem 3.5.1 *Assume that the operators A_1, A_2 are bounded. Then the sequential splitting with the explicit Euler method by the choice $\Delta t = \tau$ is convergent for the well-posed Cauchy problem (3.4.41) with (3.4.44) at $t = t^\star$].*

Proof. To prove the consistency, we use the concept of Definition 3.4.7, valid for any numerical method. By the use of (3.5.3), we get

$$\left\| \frac{\exp(\tau(A_1 + A_2)) - r(\tau(A_1 + A_2))}{\tau} u(t) \right\| \leq \tau \cdot const, \tag{3.5.5}$$

which yields the consistency of the combined method.
As for the stability, we put $n\tau = t^\star$, and for (3.5.3) we get the relation

$$\begin{aligned} \|(r(\tau(A_1 + A_2)))^n\| &= \| \left((I + \tau A_2)(I + \tau A_1)\right)^n \| \leq \\ &\leq (1 + \tau\|A_2\|)^n (1 + \tau\|A_1\|)^n \leq \\ &\leq \exp(n\tau\|A_1\|) \exp(n\tau\|A_2\|) = \exp\left(t^\star(\|A_1\| + \|A_2\|)\right), \end{aligned} \tag{3.5.6}$$

which proves the stability. ■

Now we fix some special parameter choices in the combined schemes, which result in widely known and used numerical schemes.

1. *Crank-Nicolson method.* For the Cauchy problem (3.4.41), let us use the trivial splitting $A_0 = \frac{1}{2}A_0 + \frac{1}{2}A_0$, then the sequential splitting reads as follows:

$$\begin{aligned} \frac{dw_1^1(t)}{dt} &= \frac{1}{2} A_0 w_1^1(t), \quad 0 < t \leq \tau \\ w_1^1(0) &= w_0 \end{aligned} \tag{3.5.7}$$

$$\begin{aligned} \frac{dw_2^1(t)}{dt} &= \frac{1}{2}A_0 w_2^1(t), \quad 0 < t \le \tau \\ w_2^1(0) &= w_1^1(\tau) \end{aligned} \tag{3.5.8}$$

(For simplicity, we wrote out only the first step). Applying the explicit Euler method for the sub-problem (3.5.7) and the implicit Euler (IE) method for (3.5.8) with $\triangle t = \tau$, we obtain

$$\frac{y_1^1 - y_1^0}{\tau} = \frac{1}{2}A_0 y_1^0; \quad y_1^0 = w_0,$$

$$\frac{y_2^1 - y_2^0}{\tau} = \frac{1}{2}A_0 y_2^1; \quad y_2^0 = y_1^1.$$

This implies that for the above special decomposition the discretization operator for the sequential splitting is $r_{tot} = r_{tot}(\tau, EE, 1, IE, 1)$ and has the form

$$r(\tau(A_1 + A_2)) = (I - \frac{\tau}{2}A_0)^{-1}(I + \frac{\tau}{2}A_0), \tag{3.5.9}$$

i.e., we obtained the operator of the Crank-Nicolson method ("trapezoidal rule"). In this example, the discretization error of the method consists of only the numerical integration part because obviously the splitting error equals zero.

2. *Componentwise split Crank-Nicolson method.* Let us consider the Cauchy problem (3.4.41)-(3.4.44) and use the sequential splitting and the Crank-Nicolson method with $\tau = \Delta t$, i.e.,

$$\begin{aligned} \frac{y_1^{n+1} - y_1^n}{\tau} &= A_1 \frac{y_1^{n+1} + y_1^n}{2} \\ y_1^n &= y_2^{n-1} \end{aligned} \tag{3.5.10}$$

$$\begin{aligned} \frac{y_2^{n+1} - y_2^n}{\tau} &= A_2 \frac{y_2^{n+1} + y_2^n}{2} \\ y_2^n &= y_1^{n+1}, \end{aligned} \tag{3.5.11}$$

where $y_1^0 = w_0$ and $n = 1, 2, \ldots;\ n\tau \le t^\star$.
An easy computation shows that

$$r(\tau(A_1 + A_2)) = (I - \frac{\tau}{2}A_2)^{-1}(I + \frac{\tau}{2}A_2)(I - \frac{\tau}{2}A_1)^{-1}(I + \frac{\tau}{2}A_1). \tag{3.5.12}$$

Obviously, in the above algorithm the inverse operators always exist for sufficiently small values of τ. Using the Neumann series we have

$$(I - \frac{\tau}{2}A_i)^{-1} = I + \frac{\tau}{2}A_i + \frac{\tau^2}{4}A_i^2 + \mathcal{O}(\tau^3) \tag{3.5.13}$$

for $i = 1, 2$. Hence,

$$r(\tau(A_1 + A_2)) = I + \tau(A_1 + A_2) + \frac{\tau^2}{2}(A_1^2 + A_2^2 + 2A_2A_1) + \mathcal{O}(\tau^3),$$

which shows the consistency (in first order) of this combined method. (If the operators A_1 and A_2 commute, then the order is higher.)

This method requires the solution of only two systems of linear equations, with (hopefully) simple linear operators $I - 0.5\tau A_i$, for $i = 1, 2$.

3. *Second order Yanenko method.* The following method is known as the second-order Yanenko method [151]:

$$\begin{aligned} y_1^{n+1} &= y_1^n + \tau A_1 \frac{y_1^n + y_1^{n+1}}{2}, \\ y_2^{n+1} &= y_1^{n+1} + \tau A_2 \frac{y_1^{n+1} + y_2^{n+1}}{2}. \end{aligned} \tag{3.5.14}$$

   As an easy computation shows, this method corresponds to the sequential splitting by use of the middle-point numerical integration rule with $\tau = \Delta t$.

4. *Sequential alternating Marchuk scheme.* Let us denote the Yanenko scheme (3.5.14) as $y^{n+1} = \Phi_{A_1A_2}(y^n)$. In order to restore the symmetry, we interchange the order of A_1 and A_2 in each step. This leads to the modification

$$y^{n+1} = \Phi_{A_1A_2}(y^n);\ y^{n+2} = \Phi_{A_2A_1}(y^{n+1}),\ n = 0, 2, 4 \ldots . \tag{3.5.15}$$

   This method was defined by Marchuk [98] and it corresponds to the case of the Strang-Marchuk splitting with the middle-point numerical integration method and $\tau = \Delta t$.

5. *Parallel alternating scheme.* Let us consider the method defined as

$$y^{n+1} = \frac{1}{2}\Phi_{A_1A_2}(y^n) + \frac{1}{2}\Phi_{A_2A_1}(y^n) \tag{3.5.16}$$

   [137]. One can easily see that this method corresponds to the symmetrically weighted sequential splitting with the middle-point numerical integration method with $\tau = \Delta t$.

6. *Local one-dimensional schemes.* We consider the heat conduction equation in 3D. The Yanenko scheme has the form:

$$\begin{aligned} &\frac{y_1^{n+1} - y^n}{\tau} = \Lambda_{xx} y_1^{n+1}, \quad \frac{y_2^{n+1} - y_1^{n+1}}{\tau} = \Lambda_{yy} y_2^{n+1}, \\ &\frac{y_3^{n+1} - y_2^{n+1}}{\tau} = \Lambda_{zz} y_3^{n+1}, \quad y^{n+1} = y_3^{n+1} \end{aligned} \tag{3.5.17}$$

   for $n = 0, 1 \ldots$ with $y^0 = w_0$. In this scheme $\Lambda_{xx}, \Lambda_{yy}$ and Λ_{zz} denote the usual discretizations of the one-dimensional differential operators $\partial^2/\partial x^2$, $\partial^2/\partial y^2$, and $\partial^2/\partial z^2$, respectively. (For more details, see [98], [116].)

   Clearly, this scheme corresponds to the sequential splitting with implicit Euler method and $\tau = \Delta t$.

Obviously, many other methods can also be investigated in this framework. (E.g., when in the iterated splitting we use the usual finite difference approximations, and we execute only one iteration step with $\tau = \Delta t$, then we obtain the ADI method.)

3.5.2 Error analysis of the combined discretization methods

Now we sketch the second problem, namely, the error analysis of the combined method. The naive approach suggests two "principles":

a. If the operator splitting is exact (the splitting error vanishes), then a given numerical method with the same step-size gives the same result without and with splitting;

b. The "principle of the weakest chain link" is valid, i.e., when we apply an operator splitting method and a numerical method of given orders, then the order of the combined method is determined by the lower order.

We show that both conjectures are false.

First we examine the question a. by illustrating the interaction between the splitting procedure and the numerical method on a simple problem applying the *sequential splitting* and the *explicit Euler method*.

Let $(x(t), y(t))^T$ denote a function of type $\mathbb{R} \to \mathbb{R}^2$, and $0 \leq t \leq t^\star$. The time-evolution of the *harmonic oscillator* is described by the following equation (see, e.g., [84]):

$$\left.\begin{aligned} \dot{x}(t) &= y(t) \\ \dot{y}(t) &= -x(t), \end{aligned}\right\} \qquad t \in [0, t^\star] \tag{3.5.18}$$

with $x(0) = x_0 \in \mathbb{R}$ and $y(0) = y_0 \in \mathbb{R}$. The *exact solution* of system (3.5.18) reads

$$\left.\begin{aligned} x(t) &= y_0 \sin t + x_0 \cos t \\ y(t) &= y_0 \cos t - x_0 \sin t \end{aligned}\right\} \tag{3.5.19}$$

where $x(t)$ and $y(t)$ represent the amplitude and the velocity of the oscillator, respectively. The simplest example of an oscillating system is a mass connected to a rigid foundation with a spring.

Equation (3.5.18) can be written as:

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}. \tag{3.5.20}$$

We use the decomposition (splitting) of the matrix appearing in (3.5.20)[7] :

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} -1 & 1 \\ -1 & -1 \end{pmatrix}. \tag{3.5.21}$$

Since the first matrix is the identity matrix, clearly the commutator is zero, i.e., there is no splitting error. If we apply the sequential splitting, the *split solution* in the points $k\tau$ $(k = 0, 1, \ldots, m)$ with $\tau := \frac{t^\star}{m}$ $(m \in \mathbb{N})$ are

$$\left.\begin{aligned} x_{\mathrm{spl}}\left((k+1)\tau\right) &= y_{\mathrm{spl}}(k\tau) \sin\tau + x_{\mathrm{spl}}(k\tau) \cos\tau \\ y_{\mathrm{spl}}\left((k+1)\tau\right) &= y_{\mathrm{spl}}(k\tau) \cos\tau - x_{\mathrm{spl}}(k\tau) \sin\tau, \end{aligned}\right\} \tag{3.5.22}$$

for $k = 0, 1, \ldots, m$, where $x_{\mathrm{spl}}(0) = x_0$ and $y_{\mathrm{spl}}(0) = y_0$. From (3.5.19) and (3.5.22) one can easily check that

$$\left.\begin{aligned} x_{\mathrm{spl}}(k\tau) &= x(k\tau) \\ y_{\mathrm{spl}}(k\tau) &= y(k\tau), \end{aligned}\right\}$$

for $k = 0, 1, \ldots, m$, hence the splitting, as it is expected, does not cause any error in this case.

[7] The stability of the sequential splitting for this decomposition is proven in [90]. For the more general equation with function coefficients the stability is considered in the recent paper [14].

Let us apply the explicit Euler method with step size τ for (3.5.18). Then the *numerical solution* has the form:

$$\left.\begin{aligned} x^{k+1} &= x^k + \tau y^k \\ y^{k+1} &= y^k - \tau x^k, \end{aligned}\right\} \tag{3.5.23}$$

for $k = 0, 1, \ldots, m$, where $x^0 = x_0$ and $y^0 = y_0$.

If the explicit Euler method is applied to the split problems, the *numerical split solution* has the following form:

$$\left.\begin{aligned} x_{\mathrm{spl}}^{k+1} &= x_{\mathrm{spl}}^k + \tau y^k - \tau^2 \left(x_{\mathrm{spl}}^k - y_{\mathrm{spl}}^k\right) \\ y_{\mathrm{spl}}^{k+1} &= y_{\mathrm{spl}}^k - \tau x^k - \tau^2 \left(x_{\mathrm{spl}}^k + y_{\mathrm{spl}}^k\right), \end{aligned}\right\} \tag{3.5.24}$$

for $k = 0, 1, \ldots, m$, where $x_{\mathrm{spl}}^0 = x_0$ and $y_{\mathrm{spl}}^0 = y_0$.

Since the split and the exact solutions do not differ, we expect that if we use the same numerical method for both the split and the unsplit problems, then the numerical solution and the numerical split solution do not differ, either. However, by comparing (3.5.23) and (3.5.24) it can be seen that a certain error appears in the numerical split solution. Hence, we can see that there is a certain interaction between the splitting procedure and the numerical method.

In our numerical experiments we used $\tau = \frac{2\pi}{200}$ while integrating system (3.5.18) on the time interval $[0, 4\pi]$ with and without applying splitting (the period of the harmonic oscillator is 2π). In the left panel of *Figure 3.5.1* the effect of the numerical error can be seen: the numerical solution is spiraling outwards compared to the exact solution. In the right panel of *Figure 3.5.1* the effect of the interaction of the errors can be seen: the numerical split solution is spiraling inwards compared to the exact solution (which coincides with the split solution in this case). Hence, in this case the interaction between the numerical method and the splitting procedure causes the "turn" of the spiral.

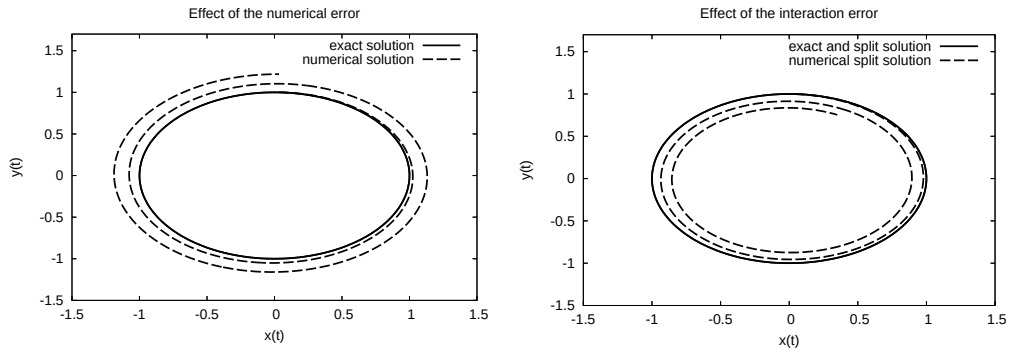

Figure 3.5.1: Effects of the numerical error (left panel) and interaction error (right panel).

To show that the conjecture b. is not true, it is enough to refer to the analysis of the Crank-Nicolson method, written by splitting+ numerical method in Section 3.5.1 on p.124: first order numerical methods were applied, however, the combined method had second order accuracy.

The above facts show how complex the error mechanism is in the combination "operator splitting + numerical methods to the sub-problems". Next, we give the way how to choose an appropriate numerical method and its discretization parameter for a given operator splitting.

Theorem 3.5.2 *Let us apply an operator splitting procedure of order $p < \infty$ together with a numerical method of order r, and set $\Delta t = \tau^s$ ($s \geq 1$, $\tau < 1$). Then the order of the local total error is: $\varrho = \min\{p, rs\}$.*[8]

This means that in order to preserve the accuracy, we can control it through the parameters p, r, and s for a given p. In practice, we apply a given splitting procedure, i.e., p and τ are fixed, and we want to preserve the order p to the local total error ($\varrho = p$). Hence, in this case we can estimate it, as well.

From Theorem 3.5.2 it follows that the interaction error causes a reduction in the order of $E_{\text{tot}}(\tau)$ unless $rs \geq p$.

Corollary 3.5.3 *From Theorem 3.5.2 it follows that $\varrho = p$ if $rs \geq p$.*

Therefore, we shall answer the following two questions for given values of p and τ.

1. How to choose Δt for a given numerical method of order r?
2. Fixing $\Delta t = \tau$, how to choose the order r of the numerical method?

From Theorem 3.5.2, the following answers can be stated.

Theorem 3.5.4 *When a given splitting procedure of order p is applied together with a given numerical method of order $r \leq p$, and the numerical step size is set as $\Delta t = \tau^s$, then the exponent s has to be chosen as $s = \frac{p}{r}$ in order to keep $\varrho = p$. For $r > p$ the identity $\varrho = p$ holds independently of the choice of Δt.*

Remark 3.5.5 *From an algorithmical point of view, it will be much easier to select the case $r > p$. For this case, clearly, the choice $\Delta t = \tau$ is optimal, because in this case the integration of the model needs the least computational work.*

Theorem 3.5.6 *When a given splitting procedure of order p is applied together with a certain numerical method of order r, and the numerical step size is set as $\Delta t = \tau^s$ ($s \geq 1$), then r has to be chosen as $r = \left[\frac{p}{s}\right] + 1 \in \mathbb{N}$ in order to keep $\varrho = p$.*

Remark 3.5.7 *Higher-order numerical methods could be chosen as well, but it would not lead to a higher-order total time discretization. It would only need more computational work.*

3.5.3 Richardson-extrapolated sequential splitting with numerical solution methods

Finally, we analyze the Richardson extrapolation when the sub-problems are solved numerically. Clearly, when the combined method is convergent, all results given in Section 3.4.4 are applicable, but now S_1 and S_2 are numerical solution operators.

In our experiments we compare the Richardson-extrapolated sequential splitting and the symmetrically weighted sequential splitting.

The results obtained by the explicit Euler method in the case of problem (3.4.133) are shown in Table 3.5.1. (Here and in the following tables the second numbers in the boxes are the ratios by which the errors decreased in comparison with the error corresponding

[8] For the proof of Theorem 3.5.2, a more detailed analysis and numerical examples we refer to [27] and [28].

to the previous step size.) In this case the Richardson-extrapolated sequential splitting shows second-order convergence, while the symmetrically weighted sequential splitting has only first order. This is understandable, since the sequential splitting applied together with a first-order numerical method has first order, and the application of this method by use of the Richardson extrapolation must have second order. However, the symmetrically weighted sequential splitting applied together with a first-order numerical method has only first order.

τ	Riseq	swss
1	2.7494e(+1)	1.2657(+1)
0.1	2.6149(-3) *(9.511(-5))*	5.4703(-3) *(4.322(-4))*
0.01	1.2927(-5) *(4.944(-3))*	7.2132(-4) *(1.319(-1))*
0.001	1.2322(-7) *(9.531(-3))*	7.3991(-5) *(1.026(-1))*

Table 3.5.1: Comparing the errors of the solutions obtained by the Richardson-extrapolated sequential splitting and the symmetrically weighted sequential splitting in example (3.4.133), when the sub-problems are solved by the explicit Euler method.

The results for the implicit Euler method are presented in Table 3.5.2. The conclusions are the same as for the explicit Euler method. We should stress: these results show that at the computer cost of a first order method we are able to get a second order method. This might be one of the biggest advantages of using the Richardson extrapolation.

If the sub-problems are solved by using a second order numerical method, then the order achieved by the symmetrically weighted sequential splitting will be two. Since the sequential splitting combined with any numerical method will only have first order, therefore the Richardson-version is also expected to have second order when combined with a second-order method. All this is confirmed by the results presented in Table 3.5.3.

τ	Riseq	swss
1	1.0896(-1)	1.0859(-1)
0.1	9.7726(-4) *(8.969(-3))*	9.2969(-3) *(8.561(-2))*
0.01	1.8353e-5 *(1.878(-2))*	7.6176(-4) *(8.194(-2))*
0.001	1.9900(-7) *(1.084(-2))*	7.4394(-5) *(9.766(-2))*

Table 3.5.2: Comparing the errors of the solutions obtained by the Richardson-extrapolated sequential splitting and the symmetrically weighted sequential splitting in example (3.4.133), when the sub-problems are solved by the implicit Euler method.

τ	Riseq	swss
1	2.9505(+1)	4.0007(+1)
0.1	4.4780(-4) *(1.533(-5))*	9.0087(-4) *(2.252(-5))*
0.01	2.4508(-6) *(5.473(-3))*	4.8727(-6) *(5.409(-3))*
0.001	2.3227(-8) *(9.478(-3))*	4.6428(-8) *(9.528(-3))*

Table 3.5.3: Comparing the errors of the solutions obtained by the Richardson-extrapolated sequential splitting and the symmetrically weighted sequential splitting in example (3.4.133), when the sub-problems are solved by the midpoint method.

3.5.4 Model for a stiff problem: reaction-diffusion equation

When the operator splitting theory is applied to the numerical solution of partial differential equations, it is usually assumed that in the first step we semi-discretize (in the space variable) the problem in order to get a Cauchy problem. Then we can use different splittings. Consequently, for this case, the operators that appear in the split tasks depend on the space discretization parameter. Typically they result in a stiff problem, which, due to the stability, has special features and requires the use of special numerical methods [30]. Therefore it is worth considering the different operator splitting and numerical methods for such problems separately.
Consider the diffusion-reaction equations with linear reaction [73]:

$$\left.\begin{aligned} \frac{\partial u}{\partial t} &= D_1 \frac{\partial^2 u}{\partial x^2} - k_1 u + k_2 v + s_1(x) \\ \frac{\partial v}{\partial t} &= D_2 \frac{\partial^2 v}{\partial x^2} + k_1 u - k_2 v + s_2(x), \end{aligned}\right\} \tag{3.5.25}$$

for the unknown concentration functions $u(x,t)$ and $v(x,t)$, where $0 < x < 1$ and $0 < t \leq T = \frac{1}{2}$, and the initial and boundary conditions are defined as follows:

$$\left\{\begin{array}{l} u(x,0) = 1 + \sin(0.5\pi x), \\ \\ v(x,0) = (k_1/k_2)u(x,0), \end{array}\right. \qquad \left\{\begin{array}{c} u(0,t) = 1, \\ v(0,t) = k_1/k_2, \\ \dfrac{\partial u}{\partial x}(1,t) = \dfrac{\partial v}{\partial x}(1,t) = 0. \end{array}\right. \tag{3.5.26}$$

We used the following parameter values:

- diffusion coefficients: $D_1 = 0.1; \quad D_2 = 0,$
- reaction rates: $k_1 = 1; \quad k_2 = 10^4,$
- source terms: $s_1(x) \equiv 1; \quad s_2(x) \equiv 0.$

After second-order finite-difference space discretization, a two-stage diagonally implicit Runge–Kutta (DIRK) [74] method was used for time discretization. Table 3.5.4 shows the coefficients of the method (with the notations of the Butcher tableau). For $\gamma = 1 - \frac{1}{2}\sqrt{2}$

γ	γ	0
$1-\gamma$	$1-2\gamma$	γ
	$1/2$	$1/2$

Table 3.5.4: Coefficients used in the two-stage DIRK method.

the method is L-stable and second-order.

The reference solution for the discretized problem was computed by the Matlab's ODE45 solver and is plotted in Figure 3.5.2.

In the different operator splittings the difference operator on the right-hand side of the semi-discrete problem was split into the sum $D + R$, where D contained the discretized diffusion and the inhomogeneous boundary conditions, and R the reaction and source terms. The spatial discretization of the diffusion terms and the big difference in the magnitude of the reaction rates give rise to stiffness, also indicated by the big operator norms $\|D\| = \mathcal{O}(10^3)$ and $\|R\| = \mathcal{O}(10^4)$.

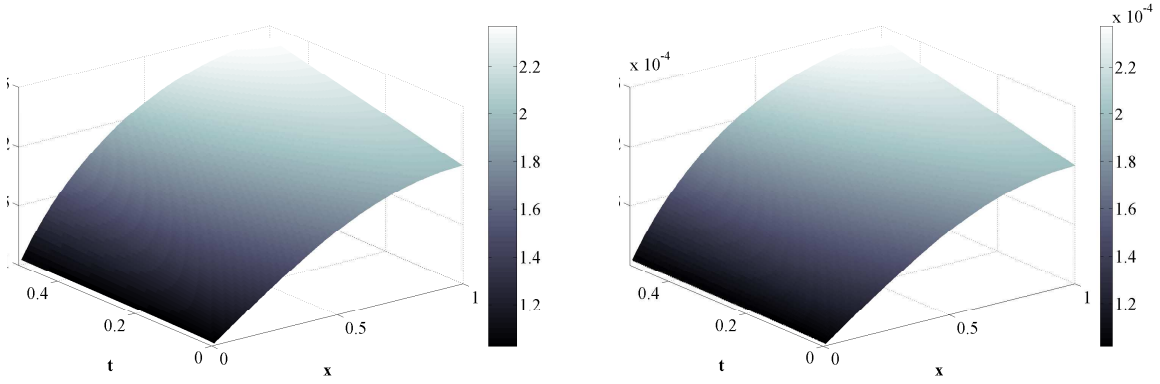

Figure 3.5.2: The reference solution of the reaction-diffusion problem.

First we compare the operator splittings of second order accuracy. We emphasize on the fact that here both sub-operators are stiff, while most studies are restricted to the case where one of the operators is stiff, the other is non-stiff. E.g., in [147] it is shown that in that case the ordering of the sub-operators in the sequential and Strang-Marchuk splitting can affect the accuracy considerably, and moreover, always the stiff operator should be put to the end. In our example, however, two stiff operators are present, so both sequences D-R and R-D should be tested for the Richardson-extrapolated sequential splitting and both D-R-D and R-D-R for the Strang-Marchuk splitting. (Obviously, the symmetrically weighted sequential splitting does not require ordering considerations.)

The sub-problems were solved by four different time integration methods: 1) the implicit Euler method, 2) the explicit Euler method, 3) the midpoint method and 4) the two-step DIRK method. During the experiments, the Richardson-extrapolated sequential splitting proved to be unstable for methods 2) and 3). Tables 3.5.5–3.5.6 show the maximum norms of the errors at the end of the time interval for methods 1) and 4). Note that the time steps were halved in these experiments, therefore a factor around 0.5 (in parentheses) corresponds to first-order convergence, while a factor around 0.25 to second-order convergence.

τ	Riseq. D-R	Riseq. R-D	swss	SM
1/10	7.07(-3)	5.86(-4)	4.80(-2)	1.26(-2)
1/20	3.67(-3) *(0.52)*	1.78(-4) *(0.30)*	2.40(-2) *(0.50)*	7.52(-3) *(0.60)*
1/40	1.86(-3) *(0.51)*	4.96(-5) *(0.28)*	1.20e(-2) *(0.50)*	4.28(-3) *(0.57)*
1/80	9.21(-4) *(0.50)*	1.31(-5) *(0.26)*	6.02(-3) *(0.50)*	2.35(-3) *(0.55)*
1/160	4.59(-4) *(0.50)*	3.22(-6) *(0.25)*	3.01e(-3) *(0.50)*	1.26(-3) *(0.54)*
1/320	2.22(-4) *(0.48)*	6.52(-7) *(0.20)*	1.50(-3) *(0.50)*	6.63(-4) *(0.53)*

Table 3.5.5: Comparing the errors of the solutions obtained by the Richardson-extrapolated sequential splittings, symmetrically weighted sequential splitting and Strang-Marchuk splitting (D-R-D) in the reaction-diffusion problem (3.5.25)–(3.5.26) for the implicit Euler method.

For the implicit Euler method the Richardson-extrapolated sequential splitting shows the expected second-order convergence in the sequence R-D. The errors obtained in the sequence D-R are one magnitude higher, moreover, here we only obtained first-order convergence (order reduction). The worst results were produced by the symmetrically weighted sequential splitting, which behaves as a first-order method, just like the Strang-

Marchuk splitting. (This is not surprising, because a second-order splitting method was combined with a first-order numerical method.) For the Strang-Marchuk splitting we only give the errors for the sequence D-R-D, which generally produced better results.

Table 3.5.6 illustrates that it is not worthwhile combining the Richardson-extrapolated sequential splitting with a second-order numerical method. The theoretically derived consistency order is still only two, furthermore, the obtained errors are bigger than for the first-order implicit Euler method. Note that the order of the symmetrically weighted sequential splitting and Strang-Marchuk splitting did not increase to two. This is the result of order reduction caused by stiffness.

τ	Riseq. D-R	Riseq. R-D	swss	SM
1/10	2.13(-2)	3.86(-3)	8.43(-2)	1.91(-2)
1/20	1.06(-2) *(0.50)*	1.86(-3) *(0.48)*	3.99(-2) *(0.47)*	9.48(-3) *(0.50)*
1/40	4.86(-3) *(0.46)*	8.95(-4) *(0.48)*	1.86(-2) *(0.47)*	4.42(-3) *(0.47)*
1/80	2.51(-3) *(0.52)*	3.47(-4) *(0.39)*	8.55(-3) *(0.46)*	2.25(-3) *(0.51)*
1/160	1.10(-3) *(0.44)*	9.50(-5) *(0.27)*	3.92(-3) *(0.46)*	1.01(-3) *(0.45)*
1/320	3.71(-4) *(0.34)*	2.04(-5) *(0.21)*	1.80(-3) *(0.46)*	3.76(-4) *(0.37)*
1/640	9.42(-5) *(0.25)*	8.54(-6) *(0.42)*	8.45(-4) *(0.47)*	1.15(-4) *(0.31)*

Table 3.5.6: Comparing the errors of the solutions obtained by the Richardson-extrapolated sequential splittings, symmetrically weighted sequential splitting and Strang-Marchuk splitting (D-R-D) in the reaction-diffusion problem (3.5.25)–(3.5.26) for the two-step DIRK method.

We have also tested the additive splitting and the iterated splitting on this example. (The iteration number in the iterative splitting was $m = 2$.)

The errors were computed in the maximum norm for different values of τ, see Table 3.5.7. The first column contains the values of the splitting time step. In the second column we can see the errors of the DIRK method without splitting; apparently we have a stable, second-order method. The third column gives the errors of the iterative splitting. One can see that even if an L-stable numerical method was used, the errors grow extremely fast. This instability can also be observed in Figure 3.5.3. The iterative splitting was also run with preconditioning. As the fourth column shows, this did not stabilize the method. The fifth column contains the errors of the additive splitting, which behaves as a stable, first-order method.

τ	DIRK	Iterative	Precond. iterative	Additive
$\frac{1}{10}$	2.3052(-4)	3.9243	3.5312	0.0843
$\frac{1}{20}$	5.6682(-5)	5745.9	5191	0.0398
$\frac{1}{40}$	1.4207(-5)	2.0694(+10)	1.8572(+10)	0.0185

Table 3.5.7: Global errors at $T = 1/2$ for different values of τ.

Remark 3.5.8 *We found that taking into account the intermediate values of the stages in the DIRK method stabilized the results. We applied preconditioning in several different ways, namely: i) by using the additive splitting on the whole interval, ii) by using the result of the additive splitting only at the endpoint of each splitting time interval, iii) by*

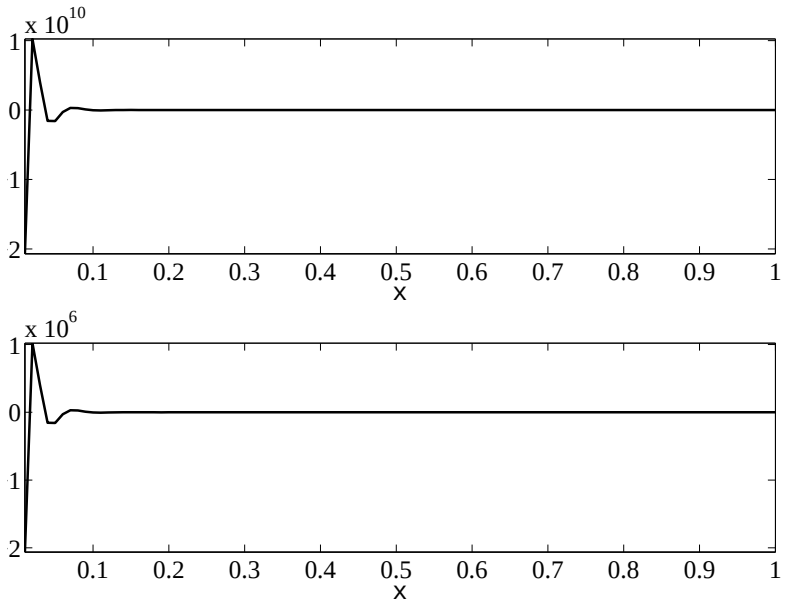

Figure 3.5.3: The solution of the iterative splitting for the two components at $T = 1/2$ ($\tau = 1/40$).

using the result of the sequential splitting at the endpoint of each splitting time interval. The best results, shown in Table 3.5.8, were obtained in case ii).

τ	Iterative	Precond. iterative
$\frac{1}{10}$	3.2904(-4)	2.9508(-4)
$\frac{1}{20}$	6.7570(-5)	6.2572(-5)
$\frac{1}{40}$	1.5476e(-5)	1.4767(-5)

Table 3.5.8: Global errors at $T = 1/2$ for different values of τ. The preconditioning was done by using the result of the Richardson-extrapolated sequential splitting only at the endpoint of each splitting time interval.

3.6 Air-pollution modelling - Danish Eulerian Model (DEM)

The operator splitting theory can be successfully applied to many real-life problems. The author of this dissertation has applied the operator splitting technique in air-pollution modelling and in the numerical solution of the Maxwell equations. Due to the limitation of the volume of this work, we consider only the first topic in more details in this section, and we will touch upon the Maxwell equation briefly in the Conclusion. Our investigation serves also for the analysis of some special features in the computer realization of the methods.

An air-pollution model containing N_s chemical species is normally described by a system of partial differential equations (see [157]):

$$\begin{aligned}\frac{\partial \mathbf{c}(\mathbf{x},t)}{\partial t} &= -\frac{\partial (u\mathbf{c})}{\partial x} - \frac{\partial (v\mathbf{c})}{\partial y} - \frac{\partial (w\mathbf{c})}{\partial z} + \\ &+\frac{\partial}{\partial x}\left(K_x \frac{\partial \mathbf{c}}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial \mathbf{c}}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z \frac{\partial \mathbf{c}}{\partial z}\right) + \\ &+E + Q(\mathbf{c}) - \kappa\ \mathbf{c},\end{aligned} \tag{3.6.1}$$

where $\mathbf{x} = (x, y, z)$, $\mathbf{c} = (c_1, c_2, \ldots, c_{N_s})^T$, $E = (E_1, E_2, \ldots, E_{N_s})^T$, $Q = (Q_1, Q_2, \ldots, Q_{N_s})^T$ and κ is a diagonal matrix the diagonal elements of which are $\kappa_{11}+\kappa_{21}, \kappa_{12}+\kappa_{22}, \ldots, \kappa_{1N_s}+\kappa_{2N_s}$. For $s = 1, 2, \ldots, N_s$, the different quantities that are involved in the mathematical model have the following physical meanings:

- the concentration of the s-th pollutant is denoted by c_s;
- u, v and w are the components of the wind velocities and $u = u(x, y, z, t), v = v(x, y, z, t)$ and $w = w(x, y, z, t)$;
- K_x, K_y and K_z are the diffusion coefficient functions and they are assumed to be non-negative;
- the emission sources in the space domain are described by the functions $E_s = E_s(x, y, z, t)$;
- κ_{1s} and κ_{2s} are the deposition coefficients (dry and wet, respectively) and $\kappa_{1s} \geq 0, \kappa_{2s} \geq 0$;
- the chemical reactions are described by the non-linear functions $Q_s(c_1, c_2, \ldots, c_{N_s})$. (We note that in some models Q also depends on x, y, z, t.)

The PDE system (3.6.1) is in fact describing the Danish Eulerian Model (DEM; see [157] and [158]), but all other large-scale air pollution models are described mathematically in a similar way. So far, the model has been mainly used with a chemical scheme containing 35 species (it may be necessary to involve more species in the future; experiments with chemical schemes containing 56 and 168 species have recently been carried out).

3.6.1 Examples of splitting procedures for air pollution models

During the application of the operator splitting, one of the most challenging problems is the following: how to split the original problem into sub-problems. In the air pollution model (3.6.1) we can do it in several ways. In the following we describe those decompositions which are used in the different program packages.

1. **DEM splitting**

   The first example for a splitting procedure for air pollution models is based on dividing the operator on the right-hand side of (3.6.1) into five simpler operators to obtain the following five simpler sub-models:

$$\frac{\partial \mathbf{c}^{[1]}(\mathbf{x},t)}{\partial t} = -\frac{\partial (u\mathbf{c}^{[1]})}{\partial x} - \frac{\partial (v\mathbf{c}^{[1]})}{\partial y}, \tag{3.6.2}$$

$$\frac{\partial \mathbf{c}^{[2]}(\mathbf{x},t)}{\partial t} = \frac{\partial}{\partial x}\left(K_x \frac{\partial \mathbf{c}^{[2]}}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial \mathbf{c}^{[2]}}{\partial y}\right), \tag{3.6.3}$$

$$\frac{\partial \mathbf{c}^{[3]}(\mathbf{x},t)}{\partial t} = E + Q(\mathbf{c}^{[3]}), \tag{3.6.4}$$

$$\frac{\partial \mathbf{c}^{[4]}(\mathbf{x},t)}{\partial t} = -\kappa\ \mathbf{c}^{[4]}, \tag{3.6.5}$$

$$\frac{\partial \mathbf{c}^{[5]}(\mathbf{x},t)}{\partial t} = -\frac{\partial (w\mathbf{c}^{[5]})}{\partial z} + \frac{\partial}{\partial z}\left(K_z \frac{\partial \mathbf{c}^{[5]}}{\partial z}\right). \tag{3.6.6}$$

Five physical and chemical processes (the horizontal advection, the horizontal diffusion, the chemistry, the deposition and the vertical exchange) are described with the systems (3.6.2)-(3.6.6).

2. **Physical splitting**

   The partition of operator A to a sum of sub-operators is not unique. An example that illustrates how the operator on the right-hand side of (3.6.1) can be split to the same number of operators (five), but in another way, is given below:

$$\frac{\partial \mathbf{c}^{[1]}(\mathbf{x},t)}{\partial t} = -\frac{\partial (u\mathbf{c}^{[1]})}{\partial x} - \frac{\partial (v\mathbf{c}^{[1]})}{\partial y} - \frac{\partial (w\mathbf{c}^{[1]})}{\partial z}, \tag{3.6.7}$$

$$\frac{\partial \mathbf{c}^{[2]}(\mathbf{x},t)}{\partial t} = \frac{\partial}{\partial x}\left(K_x \frac{\partial \mathbf{c}^{[2]}}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial \mathbf{c}^{[2]}}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z \frac{\partial \mathbf{c}^{[2]}}{\partial z}\right), \tag{3.6.8}$$

$$\frac{\partial \mathbf{c}^{[3]}(\mathbf{x},t)}{\partial t} = -\kappa\ \mathbf{c}^{[3]}, \tag{3.6.9}$$

$$\frac{\partial \mathbf{c}^{[4]}(\mathbf{x},t)}{\partial t} = E, \tag{3.6.10}$$

$$\frac{\partial \mathbf{c}^{[5]}(\mathbf{x},t)}{\partial t} = Q(\mathbf{c}^{[5]}). \tag{3.6.11}$$

   Let us notice that each operator on the right-hand sides describes different, geometrically independent physical processes (advection, diffusion, etc.). Therefore such a kind of operator splittings is called physical splitting.

3. **UNI-DEM splitting**

   The original air pollution model (3.6.1) can also be divided into three simpler submodels:

$$\frac{\partial \mathbf{c}^{[1]}(\mathbf{x},t)}{\partial t} = -\frac{\partial (w\mathbf{c}^{[1]})}{\partial z} + \frac{\partial}{\partial z}\left(K_z \frac{\partial \mathbf{c}^{[1]}}{\partial z}\right) \tag{3.6.12}$$

$$\frac{\partial \mathbf{c}^{[2]}(\mathbf{x},t)}{\partial t} = -\frac{\partial (u\mathbf{c}^{[2]})}{\partial x} - \frac{\partial (v\mathbf{c}^{[2]})}{\partial y} + \frac{\partial}{\partial x}\left(K_x \frac{\partial \mathbf{c}^{[2]}}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial \mathbf{c}^{[2]}}{\partial y}\right) \tag{3.6.13}$$

$$\frac{\partial \mathbf{c}^{[3]}(\mathbf{x},t)}{\partial t} = E + Q(\mathbf{c}^{[3]}) - \kappa\ \mathbf{c}^{[3]}. \tag{3.6.14}$$

The first of these sub-models, (3.6.12), describes the vertical exchange. The second sub-model, (3.6.13), describes the combined horizontal transport (the advection) and horizontal diffusion. The last sub-model, (3.6.14), describes the chemical reactions together with emission sources and deposition terms.

3.6.2 Some comments on the examples

The DEM (Danish Eulerian Model) splitting was used in a previous version of the software system called DEM. A pseudospectral method has been used in the discretization of the spatial derivatives in this version. This method is based on using a truncated expansion of the unknown function c in Fourier series. When the first-order derivatives are discretized, an expansion containing both sines and cosines is to be used. When the second-order derivatives are discretized, an expansion containing only cosines is to be used. Therefore it was worthwhile to split the advection and the diffusion processes. Moreover, very simple rules were used in the deposition process, which allowed to treat the sub-model arising in this part exactly. This is why the deposition was considered as a separate sub-model.

In the new versions of DEM finite elements are used instead of the pseudospectral method. Therefore, there is no need to split the advection and the diffusion processes (these could be treated together). More advanced rules for the deposition have been introduced. Therefore the deposition must be treated numerically and, thus, there is no need to split the chemistry and the deposition. A straightforward exploitation of these ideas led to the splitting procedure used in UNI-DEM; [1].

The sequential splitting procedure based on the formulae used in the physical splitting (3.6.7) - (3.6.11) was used by Dimov et al. [31] in the derivation of some important theoretical results for splitting procedures that are applicable in air pollution modelling.

Introduction of boundary conditions in the sub-models obtained by using splitting procedures is causing difficulties. This is especially true for the splitting procedures used in the DEM and physical splittings. The boundary conditions can be treated in a natural way when the splitting procedure used by UNI-DEM, (3.6.12) - (3.6.14), is used. The implementation of the boundary conditions is performed as follows:

- The boundary conditions on the top and the bottom of the space domain are treated in (3.6.12), where the computations are carried out along the vertical grid-lines.
- The lateral boundary conditions are handled in (3.6.13), where the computations are carried out in each of the horizontal grid-planes.

- The computations related to (3.6.14) are carried out by performing the chemical reactions at each grid-point. It is clear that the computations at any of the grid-points do not depend on the computations at the remaining grid-points. Therefore, no boundary conditions are needed when (3.6.14) is handled.

Finally, it should be mentioned that it is easy to use the splitting procedures defined by DEM and UNI-DEM also in the case where (3.6.1) is used as a two-dimensional model. If this is the case, then only the first four sub-models from DEM or the second and the third sub-models from UNI-DEM are to be handled. UNI-DEM can be used both as a two-dimensional (1-layer) model and as a three-dimensional model with 10 layers (see [1]).

3.6.3 Numerical results obtained by running UNI-DEM

To demonstrate the methods on a real application, we present several runs with UNI-DEM. The space domain covers the whole of Europe. Results obtained when a 2D version of the model is run on a 480×480 grid (which corresponds to 10 km $\times$ 10 km surface cells) are given in Table 3.6.1 (computing time obtained by using 8 processors) and Table 3.6.2 (comparison of observations taken in different European countries with results calculated with the sequential splitting procedure and with the Strang-Marchuk splitting procedure).

Process	seq. spl.	SM
Advection	18.96	40.17
Chemistry	25.94	26.65
Total	56.01	82.05

Table 3.6.1: Computing times, measured in hours, for the major sub-models (advection-diffusion, referred to as "advection" in the table, and chemistry-emission-deposition, referred to as "chemistry" in the table) and for the whole run when the sequential splitting and the Strang-Marchuk splitting techniques are used to treat UNI-DEM.

	Annual mean			Correlation	
Compound	Observations	Sequential	SM	Sequential	SM
NO_2	2.36	2.79	2.42	0.83	0.81
SO_2	1.51	1.61	1.55	0.83	0.83
SO_4	0.87	0.76	0.59	0.76	0.77
O_3	59.39	66.05	59.53	0.37	0.40

Table 3.6.2: Comparing the results calculated by the sequential splitting and the Strang-Marchuk splitting procedures with results obtained at observation stations located in different European countries.

The results shown in Table 3.6.1 indicate that the major increase of the computing time when the Strang-Marchuk splitting procedure is used is caused by the fact that the advection model is to be called twice per splitting time step (the computing time for the advection part is approximately doubled).

Potentially, the second order Strang-Marchuk splitting procedure should be more efficient than the first order sequential one. The results shown in Table 3.6.2 indicate that

the improvements from using the Strang-Marchuk splitting procedure are not very considerable. An explanation of this fact are perhaps the dominating errors from the input data and/or from the numerical methods. The use of more accurate input data and more accurate numerical methods in the future will probably require to reduce the errors caused by the splitting procedure and, thus, to use the more accurate Strang-Marchuk splitting procedure.

3.6.4 A simplified air pollution model of one air column

For testing the performance of the basic splitting methods we chose a simple one-column model with only vertical mixing, chemistry and emission operators.

The vertical mixing involves vertical diffusion and cumulus convection according to the TM3 global transport-chemistry model [141]. Cumulus convection represents vertical transport resulting from large-scale instabilities in the atmosphere. This process is of particular importance for short-lived gases, which would have no chance to reach the upper troposphere if only mean and eddy motions were considered. In the TM3 code the convection operator is defined as

$$V(z, c_i(z)) := \int_0^H [M(z,\zeta)c_i(\zeta) - M(\zeta, z)c_i(z)]d\zeta, \tag{3.6.15}$$

where $M(z,\zeta)$ gives the rate at which mass is transported from height ζ to height z, and H is the column height. The first term of the integrand expresses the gain at height z by transport from other heights, while the second term describes the loss at height z to other heights. In such a way the convection operator directly couples each vertical level to all others. This was done because the coupling along the vertical directions takes place on much shorter time scales than the time steps used for the numerical integration.

The chemical scheme is CBM-IV (Carbon Bond Mechanism IV), involving chemical reactions of 32 species. Emissions are set according to the CBM-IV urban scenario [117], which means that the emissions are high. The number of vertical layers is 19.

For our purpose, it is enough to consider the semi-discrete model, which has the form

$$\dot{\mathbf{y}} = \mathbf{V}\mathbf{y} + \mathbf{r}(\mathbf{y}) + \mathbf{E}, \quad \mathbf{y}(0) = \mathbf{y}_0 \tag{3.6.16}$$

where $\mathbf{V}$ is the vertical mixing matrix, $\mathbf{r}$ is the semi-discrete chemical operator, $\mathbf{E}$ is the emission and the vector $\mathbf{y}$ with 32 times 19 entries approximates the concentrations at the model layers. The way matrix $\mathbf{V}$ is computed and described in detail in [12].

For the numerical integration we have used the ROS3-AMF+ method, which can be described shortly as follows. Consider the autonomous ODE system

$$\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u}). \tag{3.6.17}$$

The main point in the Rosenbrock methods is the use of the Jacobian matrix of $\mathbf{F}$ instead of applying a Newton-type iteration process [30]. The third-order Rosenbrock method [87, 88] reads as

$$\begin{aligned} &\mathbf{u}^{n+1} = \mathbf{u}^n + \tfrac{5}{4}\mathbf{k_1} + \tfrac{3}{4}\mathbf{k_2} \\ &(\mathbf{I} - \gamma\Delta t\mathbf{J})\mathbf{k_1} = \Delta t\mathbf{F}(\mathbf{u}^n) \\ &(\mathbf{I} - \gamma\Delta t\mathbf{J})\mathbf{k_2} = \Delta t\mathbf{F}(\mathbf{u}^n + \tfrac{2}{3}\mathbf{k_1}) - \tfrac{4}{3}\mathbf{k_1}, \end{aligned} \tag{3.6.18}$$

where $\mathbf{J}$ denotes the Jacobian matrix $\mathbf{F}'(\mathbf{u}^n)$ and $\gamma = \frac{1}{2} + \frac{\sqrt{3}}{6}$. We remark that this specific γ yields A-stability, which is a desirable property if stiff problems are to be solved [88]. In our

case vector $\mathbf{u}$, approximating the concentration function, has mn_z entries, where m is the number of species and n_z is the number of vertical layers. Further, $\mathbf{F}(\mathbf{u}) = \mathbf{V}\mathbf{u}+\mathbf{r}(\mathbf{u})+\mathbf{E}$, where $\mathbf{V}$ is the vertical mixing matrix, $\mathbf{r}$ is the semi-discrete chemical operator, $\mathbf{E}$ is emission, and $\mathbf{J} = \mathbf{V} + \mathbf{R}$ with $\mathbf{R} = \frac{\partial \mathbf{r}}{\partial \mathbf{u}}(\mathbf{u^n})$. There exist modifications of the above scheme in which $\mathbf{J}$ is replaced by an approximate matrix. When standard AMF is used,

$$(\mathbf{I} - \gamma\Delta t\mathbf{J}) \approx (\mathbf{I} - \gamma\Delta t\mathbf{R})(\mathbf{I} - \gamma\Delta t\mathbf{V}). \tag{3.6.19}$$

Such types of decompositions are advantageous because they simplify the solution of the linear system (3.6.18) with respect to $\mathbf{k}_1$ and $\mathbf{k}_2$. The error of the above approximation is $(\gamma\Delta t)^2\mathbf{R}\mathbf{V}$, which may be large. Therefore, an improved version of this scheme was developed, which is called ROS3-AMF+. Here the approximation

$$(\mathbf{I} - \gamma\Delta t\mathbf{J}) \approx (\mathbf{L_V} - \gamma\Delta t\mathbf{R})\mathbf{U_V}, \tag{3.6.20}$$

is used, with the LU factors of $\mathbf{I}-\gamma\Delta t\mathbf{V} = \mathbf{L_V}\mathbf{U_V}$, diag$\mathbf{U_V} = \mathbf{I}$. This approximation still has an error of $\mathcal{O}(\Delta t^2)$, but it often can be shown to be bounded by $\gamma\Delta t\|\mathbf{R}\|$. Numerical experiments also show that this method is more accurate than standard AMF, while it requires the same computational costs [15].

In our experiments the model was run for a period of five days, starting from an initial concentration vector that had been obtained after a one-day integration of the model with a realistic initial concentration field as a starting vector. The vertical mixing matrix was updated in every six hours. The reference solution in our experiments was obtained by using a very small time step size. The sub-problems in the splitting schemes were the vertical mixing sub-problem and the chemistry sub-problem. Emission was treated together with the chemistry operator. The sub-problems were solved by the ROS3 method. The splitting time step was equal to the time integration step for all splitting schemes. Since the errors were largest in the surface layer, our observations are mostly based on this layer.

The first group of experiments was done with the sequential splitting. The order of the sub-operators in the sequential splitting can be chosen in two ways: we can begin the process either with the vertical mixing or the chemistry problem (V-R or R-V). Sportisse [130] suggests that whenever a stiff and a non-stiff operator are present, it is advisable to end the process always with the stiff one. This suggests that in our case the chemical operator should be put to the end.

The numerical results confirmed this suggestion. For time step $\tau = 15$ min splitting V-R performed significantly better than splitting R-V, the latter one producing unacceptable results, see a typical case in Figure 3.6.1. We can conclude that the chemistry operator must end the splitting process when the sequential splitting is used.

In the next group of experiments we compared the ROS3-AMF+, the symmetrically weighted sequential splitting and the Strang-Marchuk splitting. In our comparisons we used time step $\tau = 15$ min for all the methods. To get equal computational costs for all the methods compared, we modified the Strang-Marchuk splitting according to [12]: the middle operator was applied twice over half the time integration step. (This modification does not change the consistency order of the Strang-Marchuk splitting.)

Similarly to the sequential splitting, in the Strang-Marchuk splitting we can also choose two orders of the sub-operators: V-R-R-V or R-V-V-R, and the solution depends considerably on the order. Indications in the literature concerning which order should be taken

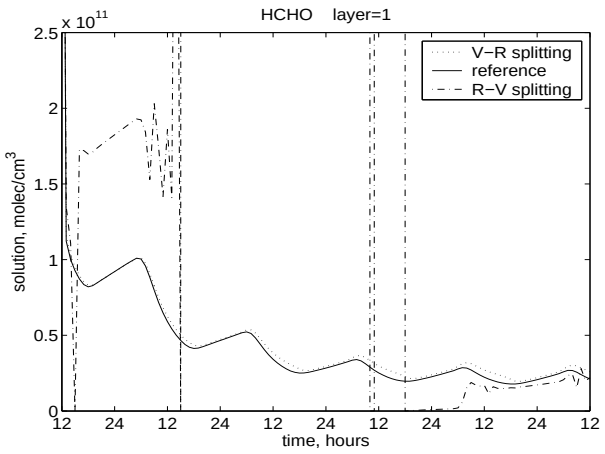

Figure 3.6.1: Solutions of sequential splittings V-R and R-V for trace gas HCHO on layer 1.

are ambiguous in this case: Sportisse [130] advocates ending the process with the stiff operator, while Verwer et al. [148] suggest the other way for the Strang-Marchuk splitting. Therefore, both Strang-Marchuk splitting, SM V-R-R-V and SM R-V-V-R were included into the experiments.

We can conclude that generally all the methods, ROS3-AMF+, symmetrically weighted sequential splitting, SM V-R-R-V and SM R-V-V-R give good results. The relative errors remain below 10% in most of the integration time and for most species. The most accurate method is unquestionably ROS3-AMF+ for all of the tracers. The fact that the method which is not based on splitting appeared to be the best one, conjectures the crucial role of the splitting error in the global one. Among the other three methods, which all are based on splitting, it is difficult to find a clear winner. The Strang-Marchuk splitting V-R-R-V method could be preferred to the symmetrically weighted sequential splitting and the other Strang-Marchuk splitting method. The quality of the symmetrically weighted sequential splitting solutions can be placed between those of the two Strang-Marchuk splitting solutions. A typical case is shown in Figure 3.6.2 for layer 1 and in Figure 3.6.3 for layer 5.

More precisely, SM V-R-R-V was better than SM R-V-V-R for 20 tracers and than the symmetrically weighted sequential splitting for 18 tracers. The symmetrically weighted sequential splitting was better than SM R-V-V-R for 21 tracers. It is interesting to examine also the number of those cases where the errors were significant:

- Comparing SM V-R-R-V versus the symmetrically weighted sequential splitting we see 10 tracers for which one of the schemes gave large errors (from which the symmetrically weighted sequential splitting is more accurate for 7 tracers).

- Comparing SM R-V-V-R versus symmetrically weighted sequential splitting we see 11 tracers for which one of the schemes gave large errors (from which symmetrically weighted sequential splitting is more accurate for 8 tracers).

We can state that for the most problematic stiff species the symmetrically weighted sequential splitting performs remarkably well. For three radicals, OH, HO_2 and NO_3, the symmetrically weighted sequential splitting gave much better results than any of the SM splittings. Figure 3.6.4 shows the results obtained for OH.

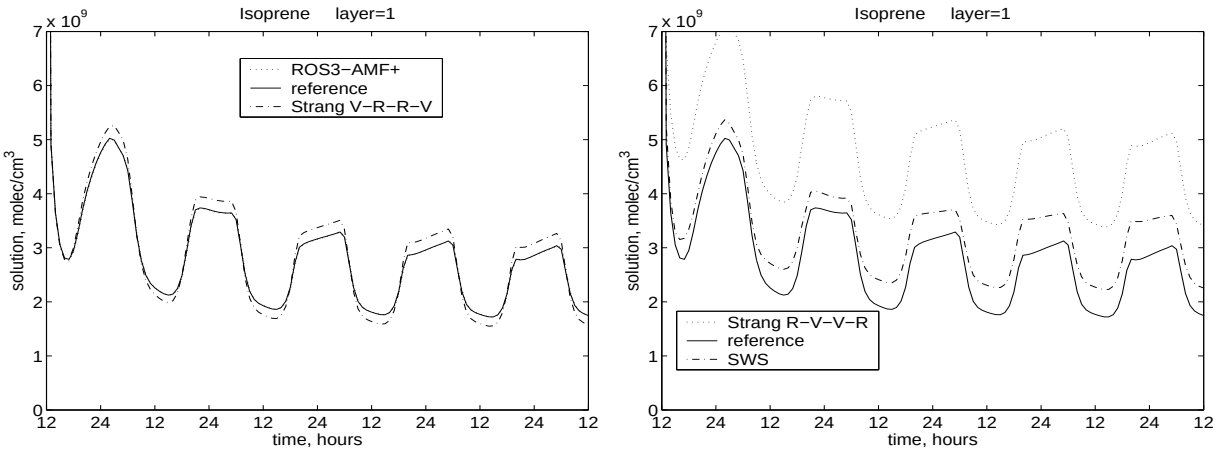

Figure 3.6.2: Solutions of ROS3-AMF+, SM V-R-R-V, SM R-V-V-R and symmetrically weighted sequential splitting for trace gas isoprene on layer 1. The dotted line in the left-hand side panel cannot be distinguished from the reference line, which demonstrates the remarkable accuracy of the ROS-AMF+ method for the chosen tracer.

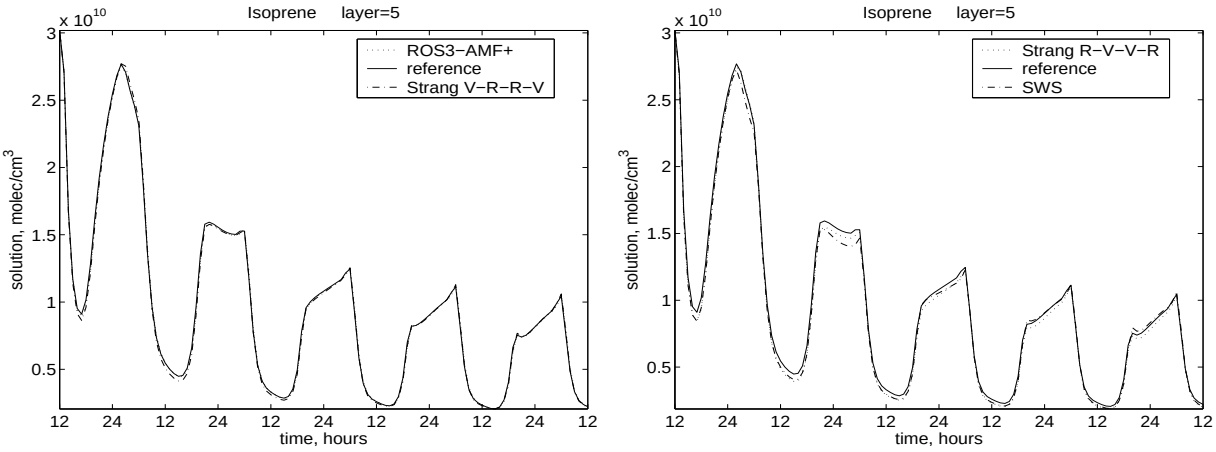

Figure 3.6.3: Solutions of ROS3-AMF+, Strang V-R-R-V, Strang R-V-V-R and SWS splitting for trace gas isoprene on layer 5.

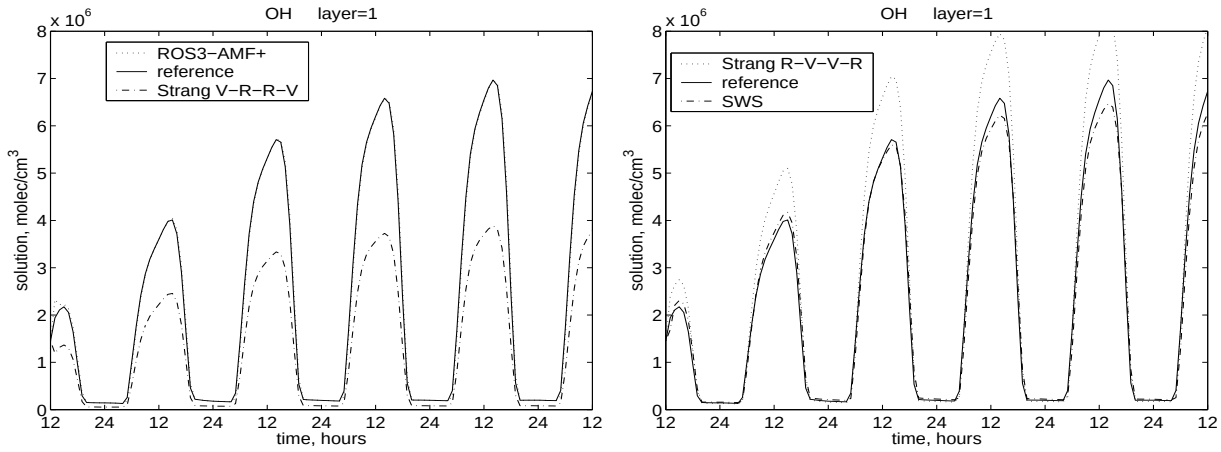

Figure 3.6.4: Solutions of ROS3-AMF+, SWS splitting, SM V-R-R-V and SM R-V-V-R for trace gas OH on layer 1. The dotted line in the left-hand panel can be hardly distinguished from the reference line.

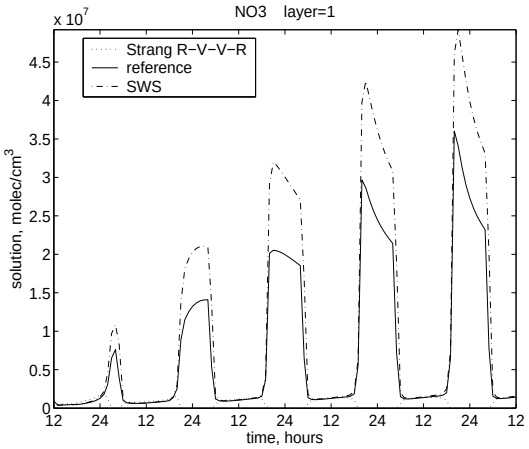

Figure 3.6.5: Solutions of SWS splitting and Strang R-V-V-R for trace gas NO_3 on layer 1.

In the experiments made with SM R-V-V-R we found two cases where the results were unacceptable: for N_2O_5 and NO_3, where the correct trend of the concentration changes was not reflected: there was no sign of the high peaks shown by the reference solution. Meanwhile, the symmetrically weighted sequential splitting was able to describe these peaks, see Figure 3.6.5. We can conclude that the symmetrically weighted sequential splitting is not only generally better than SM R-V-V-R, but, being free from some big errors produced by that method, is also more reliable. This feature should be appreciated all the more because, as we already mentioned, in many cases it is not possible to decide, which SM method would give better results.

Returning to the question of a proper ordering of the sub-operators in the Strang-Marchuk splitting, we note that in our case the choice proposed in [148], namely V-R-R-V, was better than the other one, advocated in [130].

3.7 Conclusion

Splitting methods are frequently used in practice to integrate differential equations numerically . They suggest a natural choice when the vector field associated with the differential equation can be split into a sum of two or more parts that are each simpler to integrate than the original problem.

We have introduced different operator splitting methods and investigated their nature. Section 3.2 treated the classical operator splittings, which are widely used in the different applications. We analyzed the local splitting error and we have shown that the commutativity of the operators is usually not necessary for the vanishing of the local splitting error (Theorems 3.2.4 and 3.2.5). We also gave the conditions under which the Strang-Marchuk splitting has higher order accuracy. In Section 3.3 we introduced new operator splittings and we examined their properties. The symmetrically weighted sequential splitting, which is based on the symmetrization of the sequential splitting, has higher (second) order accuracy, while the additive splitting is advantageous in the computer realization. The iterated splitting (which is, in some sense, the extension of the ADI method) differs from the other splitting methods in its high accuracy. We pointed out that the additive splitting and the iterated splitting have an extra qualitative property: both are continuously consistent approximation methods. In Section 3.4 we have done further analysis of the different operator splitting methods. We have shown the possibility using the second order operator splitting for the abstract Cauchy problems with inhomogeneous right-hand side (Theorems 3.4.2 and 3.4.3). We also examined the relation between the magnitude of the norm of the commutator and the local splitting error (Section 3.4.2). The consistency of the different operator splittings for unbounded operators is an important and less analyzed question. We have done it for the second order methods (Strang-Marchuk splitting and symmetrically weighted sequential splitting), proving the preservation of the order for this case, too (Theorems 3.4.14 and 3.4.15), and also the convergence of the different splittings for the contractive generators. In Section 3.4.4 we have shown the possibility of using the Richardson extrapolation method in order to increase the accuracy. Section 3.5 is devoted to the question of the choice of the suitable numerical integration method for the split sub-problems, in order to preserve (or, in several cases, to increase) the order of the operator splitting. In Section 3.6 we considered one of the most important applications of the operator splitting methods, namely, their application to the air pollution modelling. We have used the Danish Eulerian Model for the computer experiments. The numerical results confirmed well the theoretical results and the previous computer experiments on the model problems.

Finally, we finish this chapter with some generalizations.

- The rigorous theory that we have developed in this chapter for the operator splittings has serious limitations in their application for practical problems. Namely, we have assumed that the operators are linear and they are time-independent. However, as we have seen in Section 3.6, for some operators, arising from mathematical models of real-life problems, this is not the case. Although, due to practical need, the applicants are using the operator splitting also for these cases, however there is no well-based rigorous mathematical background behind them. In our opinion, the Magnus method is such an approach that makes possible to treat this problem. In the Appendix A. we sketch the main idea of this method.

- Although the non-linear problems are not the topic of this dissertation, due to the

practical need (see DEM model, chemical part) we touch this question in Appendix B.

- One of the benefits of using the operator splittings is their computer realization on parallel computers. The computational properties of three selected basic splitting procedures (sequential splitting, Strang-Marchuk splitting and weighted sequential splitting) are compared in the DEM computer program in the paper [26]. In this paper those conditions are formulated which can help the users in choosing the optimal splitting scheme.

- As it was mentioned, the operator splitting method can be applied to the numerical integration of the Maxwell equations, too. Namely, it was successfully applied to the 3D Maxwell equations, which describe the behaviour of time-dependent electromagnetic fields, in the absence of free charges and currents. Some results are given in Appendix C.

Bibliography

[1] Alexandrov, V. N., Owczarz, W., Thomsen, P. G., Zlatev, Z. (2004) *Parallel runs of a large air pollution model on a grid of Sun computers*. Math. Comp. Simul., 65, 557-577.

[2] Bachmann, P. (1894) *Analytische Zahlentheorie, Bd. 2: Die Analytische Zahlentheorie*. Leipzig, Teubner.

[3] Bagrinovskii, K. A., Godunov, S. K. (1957) *Difference schemes for multidimensional problems*. Dokl. Akad. Nauk USSR, 115, 431-433.

[4] Bakaev, N. Y. (1988) *On the resolvent bound of the second order difference operator with periodic conditions*. Prob. Sovr. Teor. Periodich. Dvizhen., 9, 73-84 (in Russian).

[5] Bakaev, N. Y. (2006) *Linear discrete parabolic problems*. North-Holland Mathematics Studies, 203, Elsevier.

[6] Baker, H. (1902) *Further applications of matrix notation to integration problems*. Proc. Lond. Math. Soc., 34, 347-360.

[7] Barash, D., Israeli, M. (2001) *Accurate operator splitting scheme for nonlinear diffusion filtering*. Scale-Space and Morphology in Computer Vision : Third International Conference, Lect. Notes in Comp. Sciences, 2106, 281-289.

[8] Bartholy, J., Faragó, I., Havasi, Á. (2001) *Splitting method and its application in air pollution modelling*. IDŐJÁRÁS, 105, 39-58.

[9] Berman, A., Plemmons, A. R. J. (1997) *Nonnegative matrices in the mathematical sciences*. Academic Press, New York.

[10] Berzins, M. (2001) *Modified mass matrices and positivity preservation for hyperbolic and parabolic PDEs*. Comm. Numer. Methods Engrg. 17, 9, 659-666.

[11] Bjørhus, M. (1998) *Operator splitting for abstract Cauchy problems*. IMA J. Numer. Anal., 18, 419-443.

[12] Berkvens, P. J. F., Botchev, M. A., Krol, M. C., Peters, W., Verwer. J. G. (2002) *Solving vertical transport and chemistry in air pollution models*. IMA Volumes in Mathematics and its Applications, 130, Atmospheric Modeling, eds: Chock, D. P., Carmichael, G. R., Springer, 1-20.

[13] Black, F., Scholes, M. (1973) *The pricing of options and corporate liabilities*. Journal of Political Economy, 81, 637-654.

[14] Blanes, S., Murua, A. (2007) *On the linear stability of splitting methods*. Found. Comput. Math., http://dx.doi.org/10.1007/s10208-007-9007-8

[15] Botchev, M., Verwer, J. G. (2003) *A new approximate matrix factorization for implicit time integration in air pollution modeling*. J. Comput. Appl. Math., 157, 309-327.

[16] Botchev, M., Faragó, I., Havasi, Á. (2004) *Testing weighted splitting schemes on a one-column transport-chemistry model*. Int. J. Environmental Pollution, 22, 3-16.

[17] Botchev, M., Faragó, I., Horváth, R. *Application of the operator splitting to the Maxwell equations including a source term*. Appl. Num. Math., (to appear), online at: http://eprints.eemcs.utwente.nl/9206/

[18] Borisov, V. S. (2003) *On discrete maximum principles for linear equation systems and monotonicity of difference schemes*. SIAM J. Matrix Anal. Appl. 24, 1110-1135.

[19] Borisov, V. S., Sorek, S. (2004) *On the monotonicity of difference schemes for computational physics*. SIAM J. Sci. Comput., 25, 1557-1584.

[20] Brandts, J., Korotov, S., Křížek, M. (2007) *Dissection of the path-simplex in R^n into n path-subsimplices*. Linear Algebra Appl., 421, 382-393.

[21] Campbell, J. (1898) *On a law of combination of operators*. Proc. Lond. Math. Soc., 28, 14-32.

[22] Christov, C. I., Marinova, R. S. (2001) *Implicit vectorial operator splitting for incompressible Navier – Stokes equations in primitive variables*. J. Comput. Technologies, 6, 92-119.

[23] Ciarlet, P. G. (1970) *Discrete maximum principle for finite-difference operators*. Aequationes Math., 4, 338-352.

[24] Ciarlet, P. G., Raviart, P. A. (1973) *Maximum principle and uniform convergence for the finite element method*. Comput. Methods Appl. Mech. Engrg., 2, 17-31.

[25] Csomós, P., Faragó, I., Havasi, Á. (2005) *Weighted sequential splitting and their analysis*. Comput. Math. Appl., 50, 1017-1031.

[26] Csomós, P., Dimov, I., Faragó, I., Havasi, Á., Ostromsky, Tz. (2007) *Computational complexity of weighted splitting scheme on parallel computers*. Int. J. Parallel Emergent Distrib. Syst., 22, 137-147.

[27] Csomós, P., Faragó, I. (2007) *Error analysis of the numerical solution obtained by applying operator splitting*. Math. Comput. Modelling (to appear).

[28] Csomós, P. (2007) *Theoretical and numerical analysis of operator splitting procedures*. PhD Thesis, Eötvös Loránd University, Budapest.

[29] Crank, J., Nicolson, P. (1947) *A practical method for numerical evaluation of solutions of partial differential equations of the heat conduction type*, Proc. Cambridge Philosophical Society, 43, 50–64.

[30] Dekker, K., Verwer, J. G. (1984) *Stability of Runge-Kutta methods for stiff nonlinear differential equations*. North Holland, Amsterdam.

[31] Dimov, I., Faragó, I., Havasi, Á., Zlatev, Z. (2001) *Commutativity of the operators in splitting methods for air pollution models*. Annales Univ. Sci. Budapest, 44, 129-150.

[32] Dimov, I., Faragó, I., Havasi, Á., Zlatev, Z. (2004) *Operator splitting and commutativity analysis in the Danish Eulerian Model*. Math. Comp. Simul., 67, 217-233.

[33] Dyakonov, E. G. (1962) *Difference schemes with splitting operators for multidimensional stationary problems*. Zh. Vychisl. Mat. i Mat. Fiz., 2, 57-79 (in Russian).

[34] Egorov, Yu. V., Subin, M. A. (1991) *Partial differential equations III*. Encyclopedia of Mathematical Sciences, 32, Springer Verlag.

[35] Elshebli, M. (2005) *Maximum principle and non-negativity preservation in linear parabolic problems*. Annales Univ. Sci. Budapest, 48, 99-108.

[36] Elshebli, M. (2007) *Discrete maximum principle for the finite element solution of linear nonstationary diffusion-reaction problems*. Appl. Math. Modelling, Published Online: July 6, 2007, doi:10.1016/j.apm.2007.03.014

[37] Engel, K.-J., Nagel, R. (2000) *One-parameter semigroups for linear evolution equations*. Graduate Texts in Mathematics, 194, Springer, New York.

[38] Faragó, I. (1996) *Nonnegativity of the difference schemes*. Pure Math. Appl., 6, 38-56.

[39] Faragó, I., Geiser, J. (2005) *Iterative operator- splitting methods for linear problems*. Weierstass Institute für Angewandte Analysis, 1043, 1-18. and Inter. J. Comp. Sci. Enging. (to appear).

[40] Faragó, I., Havasi, Á. (2002) *The mathematical background of operator splitting and the effect of non-commutativity*. In: Large Scale Scientific Computations III (Margenov, S., Yalamov P. and Wasniewski, J. eds.), Springer, Berlin, 264-271.

[41] Faragó, I., Havasi, Á. (2005) *On the convergence and local splitting error of different splitting schemes*. Progress in Computational Fluid Dynamics, 5, 495-504.

[42] Faragó, I., Havasi, Á. (2007) *Consistency analysis of operator splitting methods for C_0- semigroups*. Semigroup Forum, 74, 125-139.

[43] Faragó, I., Horváth, R. (2001) *On the non-negativity conservation of finite element solutions of parabolic problems. Proc. Conf. Finite Element Methods: Three-Dimensional Problems* (eds. P. Neittaanmäki, M. Křížek), GAKUTO Internat. Series Math. Sci. Appl., 15, Gakkotosho, Tokyo, 76-84.

[44] Faragó, I. Horváth, R., Korotov, S. (2004) *Discrete maximum principle for Galerkin finite element solutions to parabolic problems on rectangular meshes,*. in: Feistauer, M. et al ed; Numerical Mathematics and Advanced Applications, Springer Verlag, Berlin, 298-307.

[45] Faragó, I., Horváth, R., Schilders, W. (2005) *Investigation of numerical time integrations of the Maxwell equations using the staggered grid spatial discretization*. Int. J. Num. Modelling, 18, 149-169.

[46] Faragó, I., Horváth, R., Korotov, S. (2005) *Discrete maximum principle for linear parabolic problems solved on hybrid meshes*. Appl. Num. Math. 53 249-264.

[47] Faragó, I. Horváth, R. (2006) *Discrete maximum principle and adequate discretizations of linear parabolic problems*. SIAM Sci. Comput., 28, 2313-2336.

[48] Faragó, I. Horváth, R. (2007) *A review of reliable numerical models for three-dimensional linear parabolic problems*. Int. J. Numer. Meth. Engng., 70, 25-45.

[49] Faragó, I. Horváth, R. (2008) Continuous and discrete parabolic operators and their qualitative properties, IMA Numerical Analysis, (to appear)

[50] Faragó, I., Kovács, M. (2003) *On the maximum norm contractivity of second order damped single step methods*. Calcolo, 2, 91-108.

[51] Faragó, I., Palencia, C. (2002) *Sharpening the estimate of the stability bound in the maximum-norm of the Crank–Nicolson scheme for the one-dimensional heat equation*. Appl. Numer. Math., 42, 133-140.

[52] Faragó, I., Pfeil, T. (1995) *Preserving concavity in initial-boundary value problems of parabolic type and its numerical solution*. Period. Math. Hung., 30, 135-139.

[53] Faragó, I., Tarvainen, P. (1997) *Qualitative analysis of one-step algebraic models with tridiagonal Toeplitz matrices*, Period. Math. Hungar., 35, 177-192.

[54] Faragó, I., Tarvainen, P. (2001) *Qualitative analysis of matrix splitting method*. Comput. Math. Appl., 42, 1055-1067.

[55] Friedman, A. (1964) *Partial differential equations of parabolic type*. Prentice-Hall.

[56] Fiedler, M. (1964) *Special matrices and their applications in numerical mathematics*. Martnuus Nijhoff Publishers, Dordrecht.

[57] Fujii, H. (1973) *Some remarks on finite element analysis of time-dependent field problems*. Theory and practice in finite element structural analysis, Univ. Tokyo Press, Tokyo 91–106.

[58] Gantmacher, F. R., Krein, M. G. (1960) *Oscillationenmatrizen, Oscillatonskerne und kleine Schwingungen mechanisher Systeme*. Akademie Verlag, Berlin.

[59] Gilbarg, D., Trudinger, N. S. (1977) *Elliptic partial differential equations of second order*. Series of Comprehensive Studies in Mathematics, 224, Springer.

[60] Gilmore, R. (1974) *Baker-Campbell-Hausdorff formulas*. J. Math. Phys. 15, 2090-2092.

[61] Goldman, D., Kaper, T. J. (1996) *N-th order split operator schemes and non-revirsible systems*. SIAM J. Numer. Anal., 33, 349-367.

[62] Hairer, E., Lubich, C., Wanner, G. (2002) *Geometric numerical integration. structure-preserving algorithms for ordinary differential equations*. Springer Ser. Comput. Math., 31, Springer Verlag, Berlin.

[63] Hansbo, A. (1999) *Nonsmooth data error estimates for damped single step methods for parabolic equations in Banach spaces*. Calcolo, 36, 75–101.

[64] Hariton A. H. (1995) *Some qualitative properties of the numerical solution to the heat conduction equation*. Thesis for Cand. of Math. Science, Hungarian Academy of Sciences, Budapest.

[65] Harari, I., Hauke, G. (2007) *Semidiscrete formulations for transient transport at small time steps*, Int. J. Numer. Meth. Fluids, 54, 731-743.

[66] Hausdorff, F. (1906) Ber. Verh. Saechs. Akad. Wiss., Leipzig, 58, 19-48.

[67] Hille, E., Phillips, R. S. (1957) *Functional Analysis and Semi-groups*. Vol. XXXI of American Mathematical Society Colloquium Publications, revised edn. Providence, RI: American Mathematical Society.

[68] Horn, R. A. Johnson, C. (1986) *Matrix Analysis*. Cambrigde University Press.

[69] Horváth, R. (1999) *Maximum norm contractivity in the numerical solution of the one-dimensional heat equation*. Appl. Numer. Math., 31, 451-462.

[70] Horváth, R. (2000) *On the sign-stability of the numerical solution of the heat equation*. Pure Math. Appl., 11, 281-291.

[71] Horváth, R. (2001) *Some integral properties of the heat equation*. J. Comp. Math. Appl. 42, 1135-1141.

[72] Horváth, R. (2002) *On the monotonicity conservation in numerical solutions of the heat equation*. Appl. Numer. Math. 42, 189-199.

[73] Hunsdorfer, W., Verwer, J. G. (2003) *Numerical solution of time-dependent advection-diffusion-reaction equations*. Springer, Berlin.

[74] Iserles, A. (1996) *A first course in the numerical analysis of differential equations*. Cambridge University Press, New York.

[75] Jakobsen, E. R., Hvistendahl, K., Risebro, N. H. (2001) *On the convergence rate of operator splitting for Hamilton – Jacobi equations with source terms*. SIAM J. Numer. Anal., 39, 499–518.

[76] Karátson, J., Korotov, S. (2005) *Discrete maximum principles in finite element solutions of nonlinear problems with mixed boundary conditions*. Numer. Math., 99, 669-698.

[77] Korotov, S., Křížek, M., Neittaanmäki, P. (2001) *Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle*. Math. Comp., 70, 107-119.

[78] Karlsen, K., Lie, K.-A., Natvig, J. R., Nordhaug, H. F. Dahle, H. K. (2001) *Operator splitting methods for systems of convection–diffusion equations: nonlinear error mechanisms and correction strategies*. J. Comput. Phys., 173, 636–663.

[79] Karlsen, K., Risebro, N. H. (2002) *Unconditionally stable methods for Hamilton – Jacobi equations*. J. Comput. Phys., 180, 710–735.

[80] Kraaijevanger, J. (1992) *Maximum norm contractivity of discretization schemes for the heat equation*. Appl. Numer. Math., 9, 475-492.

[81] Křížek, M., Qun, L. (1995) *On diagonal dominance of stiffness matrices in* 3*D*.East-West J. Numer. Math., 3, 59-69.

[82] Křížek, M., Neittaanmäki, P. (1996) *Mathematical and numerical modelling in electrical engineering: Theory and applications*. Kluwer Academic Publishers, Dordrecht.

[83] Ladyzhenskaya, O. A., Solonnikov, V. A., Ural'tseva, N. N. (1967) *Linear and quasilinear equations of parabolic type*. Nauka, Moscow. (English translation: American Mathematical Society, Providence, Rhode Island, 1968.)

[84] Landau, L. D., Lifshitz, E. M. (1958) *Course on theoretical physics, 1: Mechanics*. Pergamon Press, Oxford.

[85] Landau, E. (1909) *Handbuch der Lehre von der Verteilung der Primzahlen*. Leipzig, Teubner. Landau, E. Handbuch der Lehre von (Reprinted by New York, Chelsea, 1953.)

[86] Lanser, D., Verwer, J. G. (1999). *Analysis of operators splitting in advection-diffusion-reaction problems in air pollution modelling*. J. Comput. Appl. Math., 111, 201-216.

[87] Lanser, D., Blom, J. G., Verwer, J. G. (2001) *Time integration of the shallow water equations in spherical geometry*. J. Comput. Phys., 1, 86-98.

[88] Lastdrager, B. Koren B., Verwer, J. G. (2001) *Solution of time-dependent advection-diffusion problems with the sparse-grid combination technique and a Rosenbrock solver*. Comput. Meth. Appl. Math., 1, 86-98.

[89] Lax, P., Richtmyer, R. (1956) *Survey of the stability of linear finite difference equations*. Comm. Pure Appl. Math., 9, 267–293.

[90] Lopez-Marcos, M. A., Sanz-Serna, J. M., Skeel, R. D. (1996) *An explicit symplectic integrators with maximal stability interval*. in: D. F. Griffiths and G. A. Watson, eds., Numerical Analysis, World Scietific, Singapore, 163-176.

[91] Lorenz, J. (1977) *Zur Inversmonotonie diskreter Probleme*. Numer. Math., 27, 227-238.

[92] Luskin, M., Rannacher, R., (1982) *On the smoothing property of the Crank-Nicolson scheme*. Appl. Anal., 14, 117-135.

[93] Marchuk, G. I. (1968) *Some application of splitting-up methods to the solution of mathematical physics problems*. Appl. Mat., 13, 103-132.

[94] Marchuk, G. I. (1971) *Methods of numerical mathematics*. Nauka, Moscow. (English transl. of 2nd rev. aug. ed., (Springer-Verlag), 1982.)

[95] Marchuk, G. I. (1982) *Mathematical modeling for the problem of the environment*. Nauka, Moscow (in Russian).

[96] Marchuk, G. I. (1986) *Mathematical modelling for the problem of the environment*. Studies in Mathematics and Applications, 16, North-Holland, Amsterdam.

[97] Marchuk, G. I. (1988) *Methods of splittings*. Nauka, Moscow (in Russian).

[98] Marchuk, G. I. (1990) *Splitting and alternating direction methods*. North Holland, Amsterdam.

[99] Marinova, R. S., Christov C. I., Marinov, T. T. (2003) *A fully coupled solver for incompressible Navier-Stokes equations using operator splitting*. Int. J. Comput. Fluid Dynamics, 17, 371-385.

[100] McLachlan, R. I., Quispel, R. G.(2002) *Splitting methods*. Acta Numerica, 11, 341-434.

[101] McRae, G. J., Goodin, W. R., Seinfeld, J. H. (1982) *Numerical solution of the atmospheric diffusion equations for chemically reacting flows*. J. Comput. Phys., 45, 1-42.

[102] Mendas, I., Milutinovic, P. (1990) *Anticommutator analogues of certain identities repeated commutators*. Phys. Lett. A, 23, 537-544.

[103] Merton, R. C. (1973) *Theory of rational option pricing*. Bell Journal of Economics and Management Science, 4, 141-183.

[104] Mimura, M., Nakaki, T., Tomoeda, K. (1984) *A numerical approach to interface curves for some nonlinear diffusion equations*. Japan. J. Appl. Math., 1, 93-139.

[105] Olver, F. W. J. (1979) *Asymptotics and special functions*. Academic Press, New York.

[106] Palencia, C. (1993) *A stability result for sectorial operators in Banach spaces*. SIAM J. Numer. Anal. 30, 1373-1384.

[107] Pfeil, T. (1993) *On the monotonicity in time of the solutions of linear second order homogeneous parabolic equations*. Ann. Univ. Sci. Budapest. Eötvös Sect. Math., 36, 139-146.

[108] Penenko V. V., Obraztsov, N. N. (1976) *A variational initialization model for the fields of meteorological elements*. Soviet Meteorol. Hydrol., 11, 1-11.

[109] Poincaré, H. (1899) Compt. Rend. Acad. Sci., Paris 128, 1065-1069.

[110] Protter, M., Weinberger, V. (1984) *Maximum principles in differential equations*. Springer-Verlag, New York.

[111] Quarteroni, A., Sacco, A., Saleri, R. F. (2000) *Numerical mathematics*. Springer, New York.

[112] Rannacher, R. (1984) *Finite element solution of diffusion problems with irregular data*. Numer. Math., 43, 309-327.

[113] Richtmyer, R., Morton, K. W. (1994) *Difference methods for initial-value problems*. Krieger Publishing, Malabar.

[114] Rózsa, P. (1976) *Linear algebra and its applications*. Müszaki Kiadó, Budapest (in Hungarian).

[115] Rump, S. M. (1999) *INTLAB-INTerval LABoratory*, in: T. Csendes (ed) Devolopments in Reliable Computing, Kluwer Acad. Publ., 77-104.

[116] Samarsky, A. A. (1977) *Theory of the difference schemes*, Nauka, Moscow (in Russian).

[117] Sandu, A., Verwer, J. G., Blom, J. G., Spee, E. J., Carmichael, G. R. (1997) *Benchmarking stiff ODE solvers for atmospheric chemistry problems II: Rosenbrock solvers*. Atmospheric Environment, 31, 3459-3472.

[118] Ruas Santos, V. (1982) *On the strong maximum principle for some piecewise linear finite element approximate problems of non-positive type*. J. Fac. Sci. Univ. Tokyo Sect. IA Math., 29, 473-491.

[119] Sanz-Serna, J. M., Calvo, M. P. (1994) *Numerical Hamiltonian problems*, Chapman Hall, London.

[120] Serdyukova, S. J. (1964) *The uniform stability with respect to initial data of a six-point symmetrical scheme for the heat conduction equation*. in: Numerical Methods for the Solution of Differential and Integral Equations and Quadrature Formulae, Nauka, Moscow, 212-216 (in Russian).

[121] Serdyukova, S. J. (1967) *The uniform stability of a sixpoint scheme of increased order of accuracy for the heat equation*. Zh. Vychisl. Mat. i Mat. Fiz., 7, 214-218 (in Russian).

[122] Shampine, L., Watts, H. (1976) *Global error estimation for ordinary differential equations*. ACM Trans. Math. Software, 2, 172-186.

[123] Sheng, Q. (1989) *Solving partial differential equations by exponential splittings*. IMA J. Numer. Anal., 9, 199-212.

[124] Simon, L., Baderko, E. (1983) *Linear partial differential equations of second order*. Tankönyvkiadó, Budapest (in Hungarian).

[125] Šolín, P., Vejhodský, T. (2007) *On a weak discrete maximum principle for hp-FEM*. J. Comp. Appl. Math., 209, 54-65.

[126] Sornborger, A., Stewart, E. (1999) *Higher-order methods for simulations on quantum computers*. Phys. Rev. A, 60, 1956-1965.

[127] Sornborger, A. (2007) *Higher-order operator splitting methods for deterministic parabolic equations*. Int. J. Comp. Math., 84, 887-893.

[128] Sperb, R. (1981) *maximum principles and their applications*, Academic Press, Inc. New York.

[129] Spijker, M. N. (1983) *Contractivity in the numerical solution of initial value problems*, Numer. Math., 42, 271-290.

[130] Sportisse, B., Djouad, R. (2002) *Some aspects of multi-timescales issues for the numerical modeling of atmospheric chemistry*. IMA Volumes in Mathematics and its Applications, 130, Atmospheric Modeling, eds.: Chock D. P. and Carmichael, G. R., Springer, 1-20.

[131] Stoyan, G. (1982) *On a maximum principle for matrices and on conservation of monotonicity with applications to discretization methods*. Z. Angew. Math. Mech., 62, 375-381.

[132] Stoyan, G. (1986) *On maximum principles for monotone matrices*. Lin. Alg. Appl., 78, 147-161.

[133] Stoyan, G., Takó, G. (1997) *Numerical methods 3.* TypoTex, Budapest (in Hungarian).

[134] Strang, G. (1963) *Accurate partial difference methods I: Linear Cauchy problems*. Arch. Rational Mech. Anal., 12, 392-402.

[135] Strang, G. (1968) *On the construction and comparison of differnce schemes*. SIAM J. Numer. Anal., 5, 506-517.

[136] Suzuki, M. (1990) *Fractal decomposition of exponential operatorss with applications to many-body theories and Monte-Carlo simulations*. Phys. Lett. A, 146, 319-323.

[137] Swayne, D. A. (1987) *Time dependent boundary and interior forcing in locally one-dimensional schemes*. SIAM J. Sci. Stat. Comput., 8, 755-767.

[138] Taflove, A. (2000) *Computational electrodynamics: The finite-difference time-domain method*. 2 ed., Artech House, Boston.

[139] Thomée, V. (1990) *Finite difference methods for linear parabolic equations*, Elsevier, North-Holland.

[140] Thomée, V. (1997) *Galerkin finite element methods for parabolic problems*. Springer, Berlin.

[141] The TM3 model homepage. http://www.phys.uu.nl/ peters/TM3/TM3S.html, Institute for Marine and Atmospheric Research Utrecht (IMAU), University of Utrecht.

[142] Tikhonov, A. N., Samarsky, A. A. (1977). *Equations of the mathematical physics*. Nauka, Moscow. (in Russian)

[143] Varadarajan, V. S. (1974) *Lie groups, Lie algebras and their representations*. Prentice-Hall Inc., Englewood Cliffs, New Jersey.

[144] Varga, R. (1966) *On discrete maximum principle*. J. SIAM Numer. Anal. 3, 355-359.

[145] Vejhodský, T. (2004) *On the non-negativity conservation in semidiscrete parabolic problems*. In: M. Křížekat all eds; Conjugate Gradient Algorithms and Finite Element Methods, Springer Verlag, Berlin, 282-295.

[146] Vejhodský, T., Šolín, P. (2007) *Discrete maximum principle for higher-order finite elements in 1D*. Math. Comp. 76, 1833-1846.

[147] Verwer, J. G. Sportisse, B. (1998) *A note on operator splitting in a stiff linear case*. MAS-R9830, CWI .

[148] Verwer, J. G., Hundsdorfer, W. H., Blom, J. G. (2002) *Numerical time integration for air pollution models*. Surveys on Mathematics for Industry, 10, 107-174.

[149] Wheeler, L. T. (1996) *Maximum principles in classical elasticity. Mathematical problems in elasticity*, in: Ser. Adv. Math. Appl. Sci., 38, 157-185.

[150] Yanenko, N. N. (1962) *On convergence of the splitting method for heat equation with variable coefficients*, Zh. Vychisl. Mat. i Mat. Fiz., 2, 933-937 (in Russian).

[151] Yanenko, N. N. (1971) *The method of fractional steps*. Springer, Berlin.

[152] Yee, K. S. (1966) *Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media*. IEEE Transactions on Antennas and Propagation, 14, 302-307.

[153] Yoshida, H. (1990) *Construction of higher order symplectic integrators*. Phys. Lett. A, 150, 262-268.

[154] Yosida, K. (1980) *Functional analysis*. Grundlehren der mathematischen Wissenschaften, 123, 6th ed., Springer, Berlin.

[155] Zassenhaus, H. (1940) *Uber Lie'schen Ringe mit Primzahlcharakteristik*. Abh. Math. Sem. Univ. Hamburg, 13, 1–100.

[156] Zhenli, Xu, Jingsong, H., Houde, H. (2006) *Semi-implicit operator splitting Padé method for higher-order nonlinear Schrodinger equations*. Appl. Math. Comput., 179, 596-605.

[157] Zlatev, Z. (1995) *Computer treatment of large air pollution models*. Kluwer Academic Publishers, Dordrecht-Boston-London.

[158] Zlatev, Z., Dimov, I. (2006) *Computational and numerical challenges in environmental modelling*. Studies in Computational Mathematics, 13, Elsevier.

Appendices

Appendix A

The Magnus method

We consider an example of the Cauchy problem

$$\begin{cases} \dfrac{\mathrm{d}y}{\mathrm{d}t} = A(t)y \quad t \in (0, t^\star] \\ \\ y(0) = y_0, \end{cases} \tag{A.1}$$

where $y_0 \in \mathbb{R}^d$, and $A(t) \in \mathbb{R}^{d\times d}$ is a time-dependent matrix function. If $d = 1$, then the solution is given as

$$y(t) = \exp\left(\int_0^t A(s)\mathrm{d}s\right) y_0.$$

If $d > 1$, then this formula does not hold any more, however we will see that the solution can still be written formally in the form

$$y(t) = \exp(\Omega(t))y_0.$$

Our aim is to define $\Omega(t)$ in this formula.

It is known that the exact solution of (A.1) is $y(t) = Y(t)y_0$, where $Y(t)$ is the fundamental solution, i.e., it is the matrix-function satisfying

$$Y'(t) = A(t)Y(t), \quad t \in (0, t^\star), \text{ with } Y(0) = I.$$

Therefore, it is enough to restrict ourselves to the expression of the fundamental solution. Integrating (A.1) we get

$$Y(t) - I = \int_0^t A(s)Y(s)\mathrm{d}s. \tag{A.2}$$

The right-hand side can be considered as an operator K applied to Y at time t:

$$K(Y)(t) := \int_0^t A(s)Y(s)\mathrm{d}s.$$

Hence $Y = KY + I$, which implies the relation $(I-K)Y = I$. This yields $Y = (I-K)^{-1}I$, which can be written as the sum of the Neumann series $\sum_{n=0}^{\infty} K^n(I)$. Therefore

$$Y(t) = \sum_{n=0}^{\infty} K^n(I)(t) = I + \int_0^t A(s)\mathrm{d}s + \int_0^t \int_0^{s_1} A(s_2)A(s_1)\mathrm{d}s_2\mathrm{d}s_1 + \dots$$

We formally need the logarithm of $Y(t)$, since by the choice $\Omega(t) := \log Y(t)$ we obtain the unknown $Y(t) = \exp(\Omega(t))$. We use the matrix equivalent of the known scalar equality

$$\log(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \dots$$

with $x := K(I)(t) + K^2(I)(t) + \dots$. Hence,

$$\begin{aligned}
\log Y(t) &= (K(I)(t) + K^2(I)(t) + \dots) - \frac{1}{2}(K(I)(t) + K^2(I)(t) + \dots)^2 + \dots \\
&= K(I)(t) + K^2(I)(t) - \frac{1}{2}(K(I)(t))^2 + \dots \\
&= \int_0^t A(s)\mathrm{d}s + \int_0^t \int_0^{s_1} A(s_2)A(s_1)\mathrm{d}s_2\mathrm{d}s_1 - \frac{1}{2}\left(\int_0^t A(s)\mathrm{d}s\right)^2 + \dots
\end{aligned}$$

With the following simple manipulation

$$\left(\int_0^t A(t)\ \mathrm{d}s\right)^2 = \int_0^t \int_0^t A(s_1)A(s_2)\ \mathrm{d}s_2\ \mathrm{d}s_1 = \int_0^t \int_0^{s_1} A(s_1)A(s_2)\ \mathrm{d}s_2\ \mathrm{d}s_1 +$$

$$\int_0^t \int_{s_1}^t A(s_1)A(s_2)\ \mathrm{d}s_2\ \mathrm{d}s_1 = \int_0^t \int_0^{s_1} A(s_1)A(s_2)\ \mathrm{d}s_2\ \mathrm{d}s_1 + \int_0^t \int_0^{s_1} A(s_2)A(s_1)\ \mathrm{d}s_2\ \mathrm{d}s_1$$

we get

$$\Omega(t) = \log Y(t) = \int_0^t A(s)\mathrm{d}s - \frac{1}{2}\int_0^t \int_0^{s_1} [A(s_2), A(s_1)]\ \mathrm{d}s_2\ \mathrm{d}s_1 \dots, \tag{A.3}$$

where $[\cdot,\cdot]$ denotes again the commutator. The series (A.3) is called the Magnus series. (We remark that the Magnus series can be used to derive the Baker-Campbell-Hausdorff formula for the product of two matrix exponentials, mentioned already in Section 3.4.2.)

The Magnus series can be used to construct a numerical method for solving problems of the form (A.1) in the following way.

1. First we truncate the series (A.3) as

$$\int_0^t A(s)\mathrm{d}s - \frac{1}{2}\int_0^t \int_0^{s_1} [A(s_2), A(s_1)]\ \mathrm{d}s_2\ \mathrm{d}s_1. \tag{A.4}$$

2. Further, we replace A(t) in (A.4) by interpolant at the Gauss-Legendre points $t_i = c_i t; i = 1; 2$, where $c_1 = \frac{1}{2} - \frac{\sqrt{3}}{6}$ and $c_2 = \frac{1}{2} + \frac{\sqrt{3}}{6}$, and evaluate the integrals. In this way we obtain the fourth-order approximation

$$y(t) = \exp\left(\frac{1}{2}t(A(t_1) + A(t_2)) - \frac{\sqrt{3}}{12}t^2[A(t_1), A(t_2)]\right) y_0 + \mathcal{O}(t^5). \tag{A.5}$$

The operator splitting methods can be successfully applied to the Magnus method, too.

Appendix B

Operator splittings for non-linear operators

In the general non-linear case the splitting error is related to the notion of L-commutativity, which can be defined in the following way. Let F and $G : \mathbf{X} \to \mathbf{X}$ be linear or non-linear differentiable mappings. (As before, $\mathbf{X}$ denotes a Banach space.) We define the operator $\{F, G\} : \mathbf{S} \to \mathbf{S}$ as follows:

$$\{F, G\}(s) := (F'(s) \circ G)(s) - (G'(s) \circ F)(s),$$

where the symbol $'$ stands for the derivative. The operators $F'(s)$ and $G'(s)$ are in $Lin(\mathbf{X})$.

Definition B.1 The operator $\{F, G\}$ is called the L-commutator of the mappings F and G.

Remark B.2 When F and G are linear operators, then the L-commutator is equal to the traditional commutator operator, denoted before as $[F, G]$.

Definition B.3 We say that the operators F and G L-commute if their L-commutator is zero, that is $\{F, G\} = 0$.

For this case the error analysis is similar as that for the linear operators ([8, 40, 73, 119]). E.g., for sequential splitting with two operators we obtain the following. When the operators do not L-commute, then the splitting error is $\mathcal{O}(\tau^2)$. Hence, for two non-linear operators the sequential splitting is exact if they are L-commuting, otherwise it has first order accuracy. (These results coincide with the results obtained for the linear operators.) Analogous results can be obtained for the other splitting methods.

Appendix C

Operators splittings for the Maxwell equations

The Maxwell equations can be written in the following form[1]:

$$-\nabla \times \mathbf{H} + \varepsilon \partial_t \mathbf{E} = \mathbf{0}, \tag{C.1}$$

$$\nabla \times \mathbf{E} + \mu \partial_t \mathbf{H} = \mathbf{0}, \tag{C.2}$$

$$\nabla(\varepsilon \mathbf{E}) = 0, \tag{C.3}$$

$$\nabla(\mu \mathbf{H}) = 0, \tag{C.4}$$

where $\mathbf{E}$ and $\mathbf{H}$ are the electric field strength and the magnetic field strength, respectively; ε is the electric permittivity and μ is the magnetic permeability. In [17] we have presented some numerical experiments with different (sequential, Strang-Marchuk, symmetrically weighted) splitting methods. In the experiments, both the finite difference and the finite element space discretizations are used. The numerical results demonstrate the following phenomena. For the Maxwell equations discretized in space with staggered central finite differences, the Strang-Marchuk splitting has been shown to be equally or more efficient (in terms of accuracy/ computational times) than the classical Yee method [152]. The symmetrically weighted sequential splitting, ideally suited for parallel implementation on two processors, outperforms the Yee method when implemented in parallel.

For the Maxwell equations discretized in space with edge vector finite elements, we have tested the Gautschi-Krylov scheme applied without splitting and with both the Strang-Marchuk splitting and the symmetrically weighted sequential splitting where the inhomogeneous source term was split. The time integration error of these two splitting schemes consists solely of the splitting error. The comparisons show that the Gautschi-Krylov scheme with splitting is slightly less accurate than without splitting. This can be explained by the fact that splitting off the source term, though it makes the scheme exact per split step, moves the error to the splitting level. This error can be significant since the time step size in the experiments is chosen large with respect to the characteristic time scale of the source term and the CFL restriction. The loss in accuracy due to splitting is, however, marginal. The computational costs of the unsplit Gautschi-Krylov scheme and the Gautschi-Krylov scheme with the Strang-Marchuk splitting are nearly the same. The costs of the Gautschi-Krylov scheme with the symmetrically weighted sequential splitting are twice as large. (Note again that the parallel symmetrically weighted sequential splitting has the same costs as the other schemes.) The connection between the traditional

[1] We use those notations that are common in the Maxwell theory.

numerical methods for the Maxwell equations and the operator splitting is investigated in [45].