Discrete Transparent Boundary Conditions for Systems of Evolution Equations

vorgelegt von

Dipl.-Math. techn. Andrea Zisowsky

aus Berlin

Von der Fakultät II, Mathematik und Naturwissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
- Dr. rer. nat. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr. Udo Simon
Berichter/Gutachter: Prof. Dr. Anton Arnold
Berichter/Gutachter: Prof. Dr. Andreas Unterrreiter

Tag der wissenschaftlichen Aussprache: 14.07.2003

Berlin 2003
D 83

Contents

Acknowledgement

I want to express my gratitude to my Ph.D. supervisor Prof. Dr. Anton Arnold. I thank Dr. Matthias Ehrhardt for his patience in answering my numerous questions and for motivating discussions.

Special thanks to Dr. Katinka Wolter for introducing me to the topic of fluid stochastic Petri nets and to Dipl.-Phys. Thomas Koprucki for fruitful discussions about $k{\cdot}p$-Schrödinger equations in quantum mechanics and for the idea to the numerical examples considered in the last part of this dissertation. I also thank Prof. Dr. Andreas Unterreiter for carefully reading the manuscript.

This work was financially supported by the German Research Council (DFG) under Grant No. MA 1662/1-2,3 and No. AR 277/3-1 and by the DFG Research Center "Mathematics for key technologies" (FZT 86) in Berlin.

Abstract

This dissertation is concerned with the derivation and implementation of *discrete transparent boundary conditions* for systems of evolution equations. *Transparent boundary conditions* (TBCs) are a special kind of artificial boundary conditions, that are constructed in such a way, that the solution on a bounded domain with TBCs is equal to the solution of the whole-space problem restricted to the bounded (computational) domain. The partial differential equations are discretised by finite differences (θ-scheme) and *discrete transparent boundary conditions* (DTBCs) are constructed for the discrete equation. Therefore, the DTBCs are well adapted to the numerical scheme. For scalar equations these DTBCs are well established. Compared to discretising the analytical TBC, in the scalar case it is known that these DTBCs have the advantage, not to destroy the stability properties of the underlying discrete scheme and to avoid any numerical reflections. In this dissertation we will deal with systems of partial differential equations (parabolic and Schrödinger type). For these systems the approach of DTBCs is completely new and involves additional problems not encountered in the scalar case. Since the numerical computation of these DTBCs is very costly, we give an approximation which greatly reduces the effort.

After a concise construction of the TBCs and DTBCs for the weakly coupled system of parabolic equations arising from the mathematical description of *fluid stochastic Petri nets*, we proceed to extend the results to a system of general parabolic equations. Finally we will consider DTBCs for a system of Schrödinger-type equations, which arise e.g. in the physics of layered semiconductor devices as the so called $k \cdot p$-Schrödinger equation of quantum mechanics.

For both kinds of systems we will give numerical examples, which show the very small error caused by the DTBC.

Zusammenfassung

Die vorliegende Doktorarbeit befasst sich mit der Herleitung und Implementierung von *diskreten transparenten Randbedingungen* für Systeme von Evolutionsgleichungen. *Transparente Randbedingungen* (TRBen) sind spezielle *künstliche Randbedingungen*, durch die die Lösung der Gleichung auf beschränktem und mit TRBen versehenem Gebiet mit der exakten Lösung des Ganzraumproblems (auf diesem beschränkten Gebiet) übereinstimmt. Die Differentialgleichungen werden durch ein Finite-Differenzen-Verfahren (θ-Schema) diskretisiert. Für die entstandene diskrete Gleichung werden diskrete transparente Randbedingungen (DTRBen) hergeleitet, wodurch die DTRBen besonders gut an das numerische Verfahren angepasst sind. Für skalare Gleichungen sind diese DTRBen bereits länger bekannt und man weiß, dass im Gegensatz zur ad-hoc Diskretisierung der analytischen TRBen Stabilitätsprobleme und künstliche Reflektionen am Rand vermieden werden können. Wir werden uns hier mit Systemen von Differentialgleichungen (parabolisch und Schrödinger Typ) befassen. Für diese Systeme ist der Ansatz der DTRBen gänzlich neu und wirft zusätzliche Probleme auf, die für skalare Gleichungen nicht auftreten.

Nachdem wir uns eingehend mit den TRBen und DTRBen für schwach gekoppelte Systeme von parabolischen Differentialgleichungen beschäftigt haben, die das Verhalten von fluiden stochastischen Petri-Netzen beschreiben, verallgemeinern wir unser Vorgehen auf ein beliebiges lineares parabolisches System. Im Anschluss daran betrachten wir ein System von Schrödinger-Gleichungen, wie es z.B. in der Halbleiterphysik als sogenannte $k \cdot p$-Schrödinger Gleichung der Quantenmechanik auftritt.

Die angeführten numerischen Beispiele zeigen, dass sowohl für parabolische als auch für Systeme von Schrödinger Gleichungen die DTRBen nur sehr kleine Fehler verursachen.

Introduction

Applied mathematics often describe physical problems by partial differential equations (PDEs), which frequently are posed on unbounded domains. But for numerical calculations a finite domain is necessary. Therefore, the computational domain can be restricted by introducing *artificial boundary conditions* or absorbing layers. Alternative approaches are boundary element methods (BEM) (cf. [**Kyt95**]) or infinite element methods (IEM) (see e.g. [**HBSS98**]).

In this dissertation we will be concerned with some special kind of artificial boundary conditions, the so called *transparent boundary conditions* (TBCs): if the initial data of the problem is supported on a finite domain Ω and the boundary conditions are constructed such, that the exact solution of the whole-space problem (restricted to Ω) is approximated, while the thus constructed initial boundary value problem (IBVP) is well-posed, then such BCs are called *absorbing boundary conditions*. Furthermore, if the approximated solution is equal to the exact solution, these BCs are called *transparent boundary conditions*.

The common way in numerics is to derive an analytic TBC, which is (often after an approximation) discretised to be used with an interior finite difference discretisation of the PDE. But this strategy proved not to be optimal: it can evoke stability problems and suffers from reduced accuracy (compared to the discretised whole-space problem) (cf. e.g. [**May89**]). At the boundary numerical reflections can be observed, particularly with a coarse grid. A *discrete approach* overcomes these problems by changing the order of the two steps of the usual strategy, i.e. first considering the discretisation of the PDE on the whole-space and then deriving the TBC for the difference scheme directly on a purely discrete level. The discrete approach completely avoids any numerical reflections at the boundary: no additional discretisation errors due to the boundary conditions occur. Moreover, the discrete TBC is already adapted to the inner scheme and therefore the numerical stability is often better–behaved than for a discretised differential TBC.

In the literature the discrete approach did not gain much attention for a long time. The first *discrete derivation of artificial boundary conditions* was presented in [**EM79**,

Section 5]. This discrete approach was also used in [**RC00**], [**Wag85**], [**Wil82**] for linear hyperbolic systems and in [**Hal82**] for the wave equation in one dimension, including error estimates for the reflected part. In [**Wag85**] a discrete (nonlocal) solution operator for general difference schemes (strictly hyperbolic systems, with constant coefficients in 1D) is constructed. Lill generalised in [**Lil92**] the approach of Engquist and Majda [**EM79**] to boundary conditions for a convection–diffusion equation and drops the standard assumption that the initial data is compactly supported inside the computational domain. However, the derived $\mathcal{Z}$–transformed boundary conditions were approximated in order to get local–in–time artificial boundary conditions after the inverse $\mathcal{Z}$–transformation. In [**Ehr01**] *discrete transparent boundary conditions* (DTBCs) for a Crank–Nicolson finite difference discretisation of general *Schrödinger–type pseudo–differential evolution equations* in 1D were constructed such that the overall scheme is unconditionally stable and as accurate as the discretised whole–space problem. The resulting DTBC is a generalisation of the DTBC for the Schrödinger equation in [**Arn98**]. The same strategy applies to the θ–scheme for scalar convection–diffusion equations [**Ehr97**] and was also used in [**JSSB93**] for the wave equation in the frequency domain.

For scalar equations research results are already advanced (cf. [**EA01**]) and DTBCs give outstanding results. Here, we are concerned with the construction and analysis of DTBCs for a *system of coupled partial differential equations* (parabolic or Schrödinger–type) in one spatial dimension. Such vector-valued parabolic equations arise for example at the analysis of second order fluid stochastic Petri nets [**Wol99**] to investigate performance and reliability of models for e.g. software systems [**WZ01**], linearised Navier-Stokes equations [**Hag94, Lil92**], mathematical biology [**WP98**]. Linear systems of Schrödinger–type are used in band structure calculations for layered semiconductor devices [**BKKR00**] and arise in so-called "*parabolic systems*" in electromagnetic wave propagation [**Lev00**].

To the author's knowledge the only work concerned with TBCs for systems of parabolic (or Schrödinger type) equations is by Hagstrom in [**Hag94**], which deals with a special 2×2 model problem. In general, the derivation and analysis of TBCs in the case of a coupled system is much more complicated than in the scalar case, i.e. in the scalar case it is possible to give an explicit analytic formula for the inverse $\mathcal{Z}$–transformed boundary equation. Due to the coupling this cannot be done in general for systems. Even for the simple 2×2 model problem Hagstrom uses early approximations to overcome this problem. Nevertheless, in

the application to Petri nets it is possible to prove some properties of the $\mathcal{Z}$–transformed boundary condition due to the special structure of the coupling term.

This dissertation is organised as follows. In Chap. 1 we will start with a brief introduction to Petri nets and stochastic processes and define *fluid stochastic Petri nets* (FSPNs), that can be described by a system of S parabolic equations. By means of these FSPNs we will explain two different approaches to derive DTBCs for a finite difference discretisation of a system of partial differential equations. Both approaches are based on the explicit solvability of the $\mathcal{Z}$–transformed difference equation. The first approach will use the standard power ansatz, the other reduces the problem to a system of first order difference equations. Both approaches give a different formulation of the boundary condition. In the first case only S unknowns (series of convolution coefficients) have to be computed, but the BC is constructed on S points at the boundary. The second approach needs to compute $S \times S$ unknowns, but the BC is located - as usually - at the boundary point and its next neighbour. We will be concerned with the relation between the two approaches. After an investigation of these convolution coefficients and their numerical computation we will give numerical examples for the constructed DTBCs. To conclude this chapter we show for which coefficients of a general parabolic system the TBC can be formulated. Discrete approximations of TBCs for parabolic systems can be found e.g. in [**Hag94**] and [**WP98**].

Since the second order differential equation arising from FSPNs was the origin for this work, this chapter is rather detailed. Here we will also examine the numerical error of the numerical inverse $\mathcal{Z}$-transformation.

In Chap. 2 we consider a system of Schrödinger-type equations in one space dimension. This arises e.g. in the physics of layered semiconductor devices as the so called $k \cdot p$-Schrödinger equation of quantum mechanics. For the considered system, which is described by a self-adjoint operator, we will construct DTBCs analogously to the parabolic case.

Due to historical reasons, the most detailed part of this work is concerned with TBCs for the system of parabolic equations, which describes a fluid stochastic Petri net. The generalisation to other systems of parabolic equations and the extension to systems of Schrödinger type equations, is held rather short.

Notation

Throughout this work we will use bold characters for vectors and matrices, where capitals are reserved for matrices. Scalars are always represented by ordinary (non-bold) letters.

We will deviate from this convention in one case: at the opening of chapter 2 we will use the notation of the physicists, which contains also small letters for matrices.

CHAPTER 1

Parabolic systems

In this chapter we will explain the concept of TBCs for systems of parabolic equations with the special structure, that describes a *fluid stochastic Petri net* (FSPN). These systems are weakly coupled, i.e. the coupling is restricted to the lowest order term of the differential equation. Stochastic Petri nets are a device to model and analyse the dynamical behaviour of complex technical systems. We will first give a description of stochastic Petri nets in general and derive the system of differential equations, that represents a FSPN. Then we will formulate and investigate the TBC, give numerical examples and at last give a generalisation to other parabolic systems.

1. Petri nets and stochastic processes

This section is designed for readers that have no previous knowledge of Petri nets and gives a brief introduction to Petri net theory. For a concise overview of stochastic Petri nets (SPN) refer to [**Ajm90**] and [**ABC$^+$95**].

We will explain the principle of a Petri net. Its dynamical behaviour can be described by a reduced reachability graph, which is isomorphic to a Markov process. Now, it is possible to give a differential equation, that describes the Markov process and thus the time depending behaviour of the Petri net. The basic structure of stochastic Petri nets was first used 1962 by C. A. Petri in [**Pet62**] and consists mainly of *places* and *transitions*, which frame the main nodes of a bipartite directed graph. Places are drawn as circles, transitions as rectangles and are connected by arrows. We call such an arrow *input arc* or *output arc* depending if it goes into or out of a transition. Whereas places usually model storage - e.g. computer memory or ware -, transitions represent its handling. Points in a circle indicate the contents of the place and are called *tokens*. The actual amount of tokens in every place is called *marking*. If each input place to a transition holds at least a certain number of tokens, which is defined by the multiplicity of the corresponding input arc, the transition is *enabled* and can *fire*. If an enabled transition fires, tokens can move along the arcs. Thus the system's dynamic is imitated.

Other arcs, called *inhibitor arcs*, are able to prevent the firing of transitions, if they go out of a suitably marked place.

In order to give capability, reliability or other quantitative results, these Petri nets have to be furnished with a time dependency. These *stochastic Petri nets* may consist of timed and immediate transitions. Immediate transitions fire prior to timed transitions. Among the set of timed transitions the priority is arranged by weights, that define the probability, that an enabled transition will fire. It is possible to consider deterministic, exponential or general firing times.

To illustrate stochastic Petri nets, we consider the M/M/1/K queueing system: namely a system with a *single-server* and a system storage (or queueing) *capacity* of K, where inter-arrival times of clients and service times are *distributed exponentially*. Fig. 1.1 shows an initial marking without clients in the system. The mark K in place P_3 indicates free capacity to serve K customers.

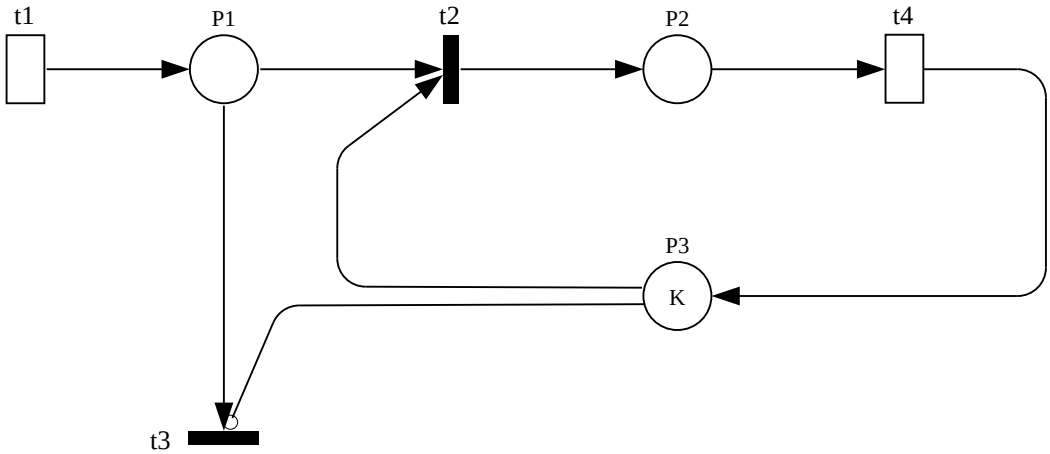


Figure 1.1: Petri net of a M/M/1/K queue

Transition t_1 models the arrival process. With exponentially distributed firing time it gives entrance to new clients. A new client coming to place P_1 is immediately transfered to the waiting place P_2, if a mark in P_3, i.e. free waiting space is available. If there is no mark in P_3 and all waiting space is occupied, the inhibitor arc is inactive and by the firing of transition t_3 the customer is lost to the system. A client in place P_2 waits for an exponentially distributed time until he is served by the firing of t_4. Thus the mark is put back into P_3 and a new waiting unit is free for another client.

A *reachability graph* is a directed graph, that pictures the succession of all markings, that can be reached ensuing from the initial marking. Each vertex of the graph is a reachable marking or a tuple ($\#P_1$,$\#P_2$,...), that gives the number of marks in every place. An edge indicates the transition into a new state and is provided with the firing transition's name. Fig. 1.2 shows the reachability graph of the M/M/1/K queue for K = 2.

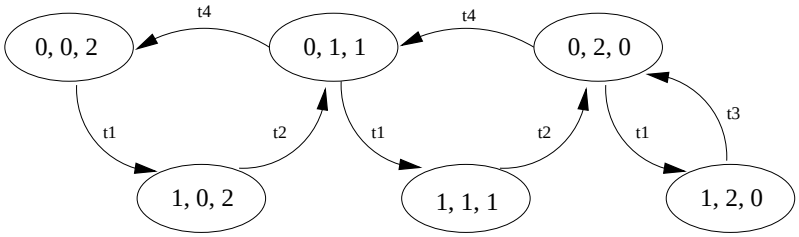


Figure 1.2: Reachability graph of M/M/1/2 queue

We distinguish between *vanishing* states, in which an immediate transition is enabled and thus no time is spend, and *tangible* states, in which only timed transitions are enabled and some time is spend. If we eliminate all vanishing states from the reachability graph, we obtain the *reduced reachability graph*, which is given in Fig. 1.3 for the case of the M/M/1/2 queue. The transition rate of a tangible state to another is given by the rate of the exponential distribution of the firing transition. Eliminating a vanishing state yields a new transition rate, that is the product of the rate of the exponential distribution with the probability that the immediate transition will fire.

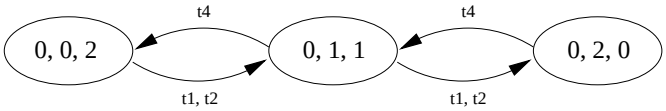


Figure 1.3: Reduced reachability graph of the M/M/1/2 queue

The reduced reachability graph describes graphically a stochastic process. The stochastic process underlying a Petri net is a Markov process. The tangible markings frame the states of a continuous-time Markov-chain with the state space $\{1, 2, \ldots, S\}$. We collect the state probabilities in a vector $\boldsymbol{\pi}(t) = (\pi_1(t), \ldots, \pi_S(t))^T \in \mathbb{R}^S$ with $\pi_i(t) = P[N(t) = i]$ if $N(t)$ is a random variable, which satisfies the Markov property. For $i, j \in \{1, 2, \ldots, S\}$ we define the transition rates from one state to another, i.e. the rates of the exponentially

distributed firing times of the timed transitions as

$$q_{ij} = \begin{cases} \text{rate from } i \text{ to } j & \text{if } i \neq j \\ -\sum\limits_{\substack{j=1 \\ j\neq i}}^{S} q_{ij} & \text{if } i = j \end{cases}. \tag{1.1.1}$$

The arising matrix $\mathbf{Q} = (q_{ij})$ is called *generator matrix* and has the typical property, that each of its row sums equals zero:

$$\sum_{j=1}^{S} q_{ij} = 0 \qquad \forall i = 1, \ldots, S. \tag{1.1.2}$$

If $\boldsymbol{\pi}_0$ is the initial marking, the Chapman-Kolmogorov equation [**Fel68**] yields a system of ordinary differential equations

$$\frac{d}{dt}\boldsymbol{\pi}(t)^T = \boldsymbol{\pi}(t)^T \cdot \mathbf{Q}, \qquad t > 0 \tag{1.1.3}$$

with the initial condition $\boldsymbol{\pi}(0) = \boldsymbol{\pi}_0$. The *transient* solution can be given in closed form by

$$\boldsymbol{\pi}(t)^T = \boldsymbol{\pi}_0^T \cdot e^{\mathbf{Q}t}. \tag{1.1.4}$$

For $t \to \infty$ (1.1.3) yields the steady state equation or *equilibrium distribution*. Together with the "normalisation property"

$$\mathbf{0} = \boldsymbol{\pi}^T \cdot \mathbf{Q}, \qquad \sum_{i=1}^{S} \pi_i = 1 \tag{1.1.5}$$

it has a unique solution, if the process obeys certain requirements [**CM65**]. Solving the system of equations yields a stationary solution, which is in many applications the aspect of the process of most interest [**CM65**].

2. Fluid Petri nets

Stochastic Petri nets are well suited for model-based performance and dependability evaluation of computer and communication systems. Due to the ever increasing complexity of the systems, the size of the state space explodes. In the last years fluid stochastic Petri nets have gained attention to approximate these extremely large state spaces or to model continuous quantities (cf. [**TK93**],[**HKNT98**],[**CNT97**],[**GSB99**],[**BGG+99**]). These extensions to SPNs have clearly divided discrete and continuous sub-models, that

can effect each other, but no probabilistic variation of the fluid flow is possible, because the fluid flow is modelled as a first order process.

Here we will give a formalism to describe *fluid stochastic Petri nets* (FSPN) of second order as defined in [**Wol99**], which we will simply call FSPNs. They have been used to analyse the reliability and performance in different models [**WZ01**],[**WZH02**] and are based on the previous described general stochastic Petri nets (GSPNs). An example will illustrate the new class of petri nets. We will describe parameters and give the underlying differential equations.

In addition to discrete places of SPNs, a FSPNs may contain places, that hold *fluid*. Thus, the state space of FSPNs consists of a discrete and a continuous part, that can affect each other.

Definition 1.1 (Second order fluid stochastic Petri net (FSPN)). *A FSPN is a 12-tuple*

$$FSPN = (\mathcal{P}_d, \mathcal{P}_c, \mathcal{T}, \mathcal{I}_d, \mathcal{I}_c, \mathcal{O}_d, \mathcal{O}_c, \mathcal{H}, \mathbf{m}_0, g, \lambda, r, w)$$

where

- $\mathcal{P}_d$ is a set of discrete places, that can hold a discrete number of *tokens*. A *marking* is defined by the number of tokens in each place $p \in \mathbb{N}^{|\mathcal{P}|}$. The number of tokens in a place is denoted as $\#(P)$.
- $\mathcal{P}_c = \{P_{1,c}, \ldots, P_{m,c}\}$ is a set of m *fluid places* that can hold a continuous amount of *fluid* rather than discrete tokens and that are graphically represented by two concentric circles. Its initial value is written as a real number in the middle of the circles. For the content of a fluid place the variables $x_i, i = 1, \ldots, m$ are used.

  All fluid places may have a restricted capacity, that is denoted by the interval $[x_i^{min}, x_i^{max}]$. Otherwise the fluid is defined on $[x_i^{min}, \infty)$. Often, $x_i^{min} = 0$ will be the lower boundary for it is the natural restriction. Therefore, if only one boundary is defined, it is assumed to be the upper bound and the lower bound is set to zero.

  The complete marking now is defined as composed of two parts, the discrete and the continuous state and is given by $\mathbf{m} = (s, \mathbf{x})$ where $s = (\#p_i, i \in \mathcal{P}_d)$ and $\mathbf{x} = (x_k, k \in \mathcal{P}_c)$. The initial marking is $\mathbf{m}_0 = (s_0, \mathbf{x}_0)$.

- $\mathcal{T} = \{t_1, \ldots, t_{|\mathcal{T}|}\} = \mathcal{T}_E \cup \mathcal{T}_G \cup \mathcal{T}_I$ is a set of transitions, that can be either immediate ($\mathcal{T}_I$) or timed with exponentially distributed firing times ($\mathcal{T}_E$) or with arbitrary firing time distribution ($\mathcal{T}_G$). The later transitions are drawn as grey rectangles.
- $\mathcal{I}_d \subseteq \mathcal{P}_d \times (\mathcal{T}_E \cup \mathcal{T}_I)$ is the set of all *discrete input arcs* from a discrete place to a transition.
- $\mathcal{I}_c \subseteq \mathcal{P}_c \times (\mathcal{T}_E \cup \mathcal{T}_G)$ is the set of all *fluid input arcs* leading from a fluid place to a timed transition. They are drawn like pipes. Fluid flows along the arc as long as the fluid place is not empty and the transition is enabled by all its discrete input places, or by a guard (see below), that can also depend on the continuous state.

  If the arc is not labelled with a flow rate function it has to be computed from the firing rates of the transition (see page 11). If the transition is only connected to fluid places and its firing has no effect on the discrete state of the model, the firing time can have any probability distribution. These are the transitions $t \in \mathcal{T}_G$.
- $\mathcal{O}_d \subseteq (\mathcal{T}_E \cup \mathcal{T}_I) \times \mathcal{P}_d$ is the set of all *discrete output arcs* from a transition to a discrete place.
- $\mathcal{O}_c \subseteq (\mathcal{T}_E \cup \mathcal{T}_G) \times \mathcal{P}_c$ is the set of all *fluid output arcs* analogous to the fluid input arcs.
- $\mathcal{H} \subseteq \mathcal{P}_d \times \mathcal{T}$ is the set of all *inhibitor arcs* from a discrete place to a transition.
- $\mathbf{m}_0 = (s_0, \mathbf{x}_0)$ is the *initial marking* consisting of a discrete part and a vector of initial fluid levels in all fluid places. The set of all markings that are reachable from the initial marking is called the *reachability set*

$$\mathcal{M} = \{\mathbf{m} | \mathbf{m}_0 \xrightarrow{*} \mathbf{m}\} = \mathcal{S} \times \mathcal{X} = \mathcal{S} \times (\bigtimes_{i=1}^{m} [x_i^{min}, x_i^{max}]) \subseteq \mathbb{N}^{|\mathcal{P}|} \times \mathbb{R}^m,$$

  where $\bigtimes$ means the cross product of the intervals.
- $g_\tau : \mathbb{N}^{|\mathcal{P}|} \times \mathcal{X} \to \mathbb{B}$ is the guard of transition t that can be a function of the discrete and the continuous state. A transition is enabled only, if the guard function evaluates to *true*. A guard is a short form for a sub-model, that not necessarily has to exist for every model.
- $\lambda_\tau : \mathbb{N}^{|\mathcal{P}|} \times \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is a function of both the continuous and the discrete marking. It is the rate of the exponential distribution of the firing time of transition t. Immediate transitions have rate infinity $\lambda_\tau(s) = \infty$ and their firing time is zero.

- $r_{t,p}$, $r_{p,t} : \mathbb{N}^{|\mathcal{P}|} \times [x_p^{min}, x_p^{max}] \rightarrow \mathbb{R}^2$, are the flow rate functions associated with the fluid input and output arcs that assign a cardinality in form of a normal distribution to each fluid arc in every marking. A transition is enabled only by the discrete marking and a guard, that can depend on the continuous state as well. Fluid flows along an arc as long as the transitions in its origin is enabled. To preserve the flow rates' independence they can only depend on the fluid level of the fluid place the arc is connected to.
- $w_\tau : \mathbb{N}^{|\mathcal{P}|} \rightarrow \mathbb{R}$ is the weight function of an immediate transition that is enabled in a vanishing marking. The firing probability of each immediate transition t is:

$$\frac{w_\tau(s)}{\sum\limits_{t \text{ enabled in } s} w_t(s)}.$$

■

With this definition we introduced the elements of a FSPN. We will illustrate them by the following example.

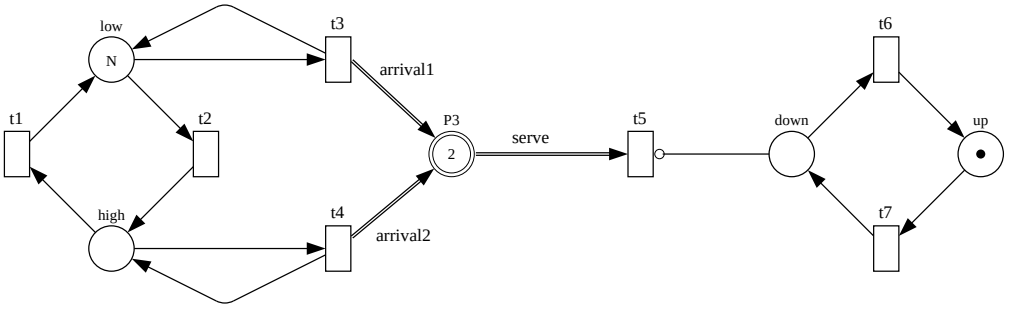


Figure 1.4: FSPN of a queueing system with failure and repair

Fig. 1.4 shows a queueing system, that represents the model of a node in a communication network. Its buffer is approximated by the fluid place P_3. Transitions t_3 and t_4 fire with different rates clients into the system. This imitates the existence of different peak times. If there is a token in place *down*, the firing (with exponentially distributed firing time) of transition t_5 is prevented as would be the case if the server fails. The parameters *arrival1*, *arrival2* and *serve* are defined by the rates of the transitions t_3, t_4 and t_5 respectively.

The state space of a FSPN consists of a discrete and a continuous part. Thus the vertex of the reduced reachability graph consists not only of discrete numbers of tokens, but contains real values for a fluid place. Thus, it is a tuple $(\#\mathcal{P}_{1,d}, \ldots, \#\mathcal{P}_{|\mathcal{P}|,d}, x_1, \ldots, x_m)$.

As an example we give in Fig. 1.5 the reduced reachability graph of the queueing system of Fig. 1.4. We choose $N = 2$ to obtain a concise graph. The expressions *up* and *down* give the partial markings $\#up = 1, \#down = 0$ and $\#up = 0, \#down = 1$ respectively. As well *high*, *high-low* and *low* signify the two places in the left part of the petri net, which currently share at least one of the two tokens, that were initially in place *low*.

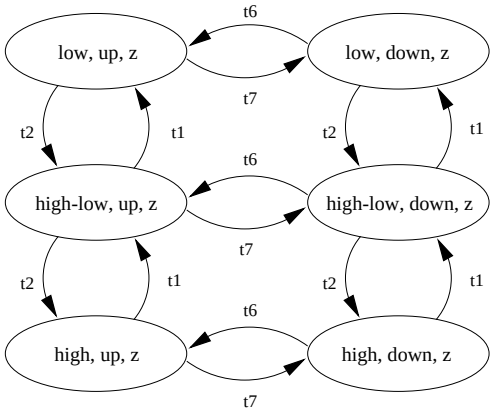


Figure 1.5: Reduced reachability graph of the queueing system of Fig. 1.4 for $N = 2$

Transition t_5 appears not in the reduced reachability graph, because it is not connected to a discrete place. Its rate influences the fluid flow parameters.

The fluid parameters have to be specified for each discrete state $i \in \mathcal{S}$ and for every fluid place. We first regard the continuous flow. The continuous flow for one fluid place $k \in \mathcal{P}_c$ in a discrete state $s \in \mathcal{S}$ is specified by the *fluid change rate function* $R : \mathcal{P}_c \times \mathcal{S} \to \mathbb{R}^2$, that is a normal distribution, defined through its first two moments, $R(k, s) = (\mu, \sigma^2)$.

There are two cases to be distinguished for the computation of the fluid change rate. Either the rates along the single fluid arcs have been specified, then

$$R(k,s) = (\mu_{k,s}, \sigma^2_{k,s}) = \left(\sum_{\substack{(t,k)\in\mathcal{O}_c \\ t\in\mathcal{T}}} \mu_r(x) - \sum_{\substack{(k,t)\in\mathcal{I}_c \\ t\in\mathcal{T}}} \mu_r(x), \sum_{\substack{(t,k)\in\mathcal{O}_c \\ (k,t)\in\mathcal{I}_c \\ t\in\mathcal{T}}} \sigma^2_r(x) \right) \tag{1.2.1}$$

where $(t, k) \in \mathcal{O}_c$ are all the fluid arcs leading from a transition, that is enabled in state s to the fluid place k and $(k, t) \in \mathcal{I}_c$ are the fluid arcs from a fluid place to a transition, respectively. μ_r and σ^2_r are *expectation* and *variance* of the rate specified for the arc.

If on the other hand no rate has been assigned to an arc in the FSPN, it is assumed, that the fluid place is meant to replace a discrete place for approximation reasons and

the firing rates of output and input transitions are used for the computations of the fluid change rate function.

Let $t_i \in \mathcal{T}_E \cup \mathcal{T}_G$ be the transitions with firing time distribution f_i, connected to the fluid place k by a fluid arc $(t_i, k) \in \mathcal{O}_c$ and let $t'_i \in \mathcal{T}_E \cup \mathcal{T}_G$ be the transitions with firing time distribution g_i, connected to the fluid place k by a fluid arc $(k, t_i) \in \mathcal{I}_c$ and let all be enabled in state $s \in \mathcal{S}$. The distributions f_i and g_i are assumed to have an expectation μ_{f_i} and μ_{g_i} and a variance $\sigma^2_{f_i}$ and $\sigma^2_{g_i}$. The *squared coefficient of variation* (scv) is defined as $c_{f_i} = \frac{\sigma^2_{f_i}}{(\mu_{f_i})^2}$, and c_{g_i} analogously. Then, $R(k, s)$ is computed as:

$$R(k, s) = (\mu_{k,s}, \sigma^2_{k,s}) \quad \text{with}$$

$$\mu_{k,s} = \sum_i \frac{1}{\mu_{f_i}} - \sum_i \frac{1}{\mu_{g_i}} \qquad \text{and} \qquad \sigma^2_{k,s} = \sum_i \frac{c_{f_i}}{\mu_{f_i}} + \sum_i \frac{c_{g_i}}{\mu_{g_i}}. \tag{1.2.2}$$

If f_i and g_i are exponential distributions with parameter λ_i and μ_i, respectively, the scv evaluates to $c = 1$ and the parameters $\mu_{k,s}$ and $\sigma^2_{k,s}$ reduce to

$$\mu_{k,s} = \sum_i \frac{1}{\lambda_i} - \sum_i \frac{1}{\mu_i} \qquad \text{and} \qquad \sigma^2_{k,s} = \sum_i \frac{1}{\lambda_i} + \sum_i \frac{1}{\mu_i}. \tag{1.2.3}$$

With (1.2.2) the fluid flow parameters for one discrete state are specified. For all states in the model and for each fluid place they are collected into diagonal matrices for the *expectation*

$$\mathbf{M}_k(x) = \operatorname{diag}\left(\mu_{k,1}(x) \ldots, \mu_{k,|\mathcal{S}|}(x)\right) \tag{1.2.4}$$

and for the *variance*

$$\mathbf{\Sigma}^2_k(x) = \operatorname{diag}\left(\sigma^2_{k,1}(x), \ldots, \sigma^2_{k,|\mathcal{S}|}(x)\right). \tag{1.2.5}$$

For Example 1.4 with $N = 2$ we get the following fluid parameters, if we enumerate the tangible states in Fig. 1.5 from left to right.

$$\boldsymbol{M} = \text{diag}\begin{pmatrix} arrival1 - serve \\ arrival1 \\ arrival1 + arrival2 - serve \\ arrival1 + arrival2 \\ arrival2 - serve \\ arrival2 \end{pmatrix} = \text{diag}\begin{pmatrix} -1.2 \\ 0.4 \\ 0.0 \\ 1.6 \\ -0.4 \\ 1.2 \end{pmatrix} \tag{1.2.6}$$

$$\boldsymbol{\Sigma}^2 = \text{diag}\begin{pmatrix} arrival1 + serve \\ arrival1 \\ arrival1 + arrival2 + serve \\ arrival1 + arrival2 \\ arrival2 + serve \\ arrival2 \end{pmatrix} = \text{diag}\begin{pmatrix} 2.0 \\ 0.4 \\ 3.2 \\ 1.6 \\ 2.8 \\ 1.2 \end{pmatrix} \quad , \tag{1.2.7}$$

if we choose $arrival1 = 0.4$, $arrival2 = 1.2$ and $serve = 1.6$.

To derive the differential equation, that describes the time dependent behaviour of a FSPN with one fluid place, we will introduce some concepts of the underlying stochastic processes. A fluid place can be deemed itself a queueing system, therefore we will use the corresponding terminology.

The important processes in a queueing system are the arrival and service processes. They are defined as follows

$$\begin{aligned} A(t) &= \text{Number of arrivals until time } t \\ D(t) &= \text{Number of departures until time } t \end{aligned} \tag{1.2.8}$$

Their difference gives the number of customers $N(t)$ in the system at time t.

$$N(t) = A(t) - D(t), \tag{1.2.9}$$

for $t = 0$ it is $N(0) = 0$. The random variable $N(t)$ is the length of the queue. For the time dependent change in the number of customers holds in the interval $[t, t + \triangle t]$

$$N(t + \triangle t) - N(t) = A(t + \triangle t) - A(t) - \{D(t + \triangle t) - D(t)\} \quad . \tag{1.2.10}$$

If the customers arrive at discrete arrival times $0 < t_{a1} < t_{a2} < \ldots$ and leave the system at the service times $0 < t_{d1} < t_{d2} < \ldots$, then the inter-arrival times $a_n = t_{an} - t_{an-1}$ and the

inter-service times $d_n = t_{dn} - t_{dn-1}$ are both independent and identically distributed with expectation μ_a, μ_d resp. and variance σ_a^2 and σ_d^2 respectively. It holds

$$P[(A(t+\triangle t) - A(t)) \geq n] = P[t_{an}^{\triangle t} \leq \triangle t] \quad , \tag{1.2.11}$$

where $t_{an}^{\triangle t}$ is the arrival time of the n-th customer in a time interval of length $\triangle t$. $t_{an}^{\triangle t}$ can be written as the sum of the inter-arrival times in $\triangle t$

$$t_{an}^{\triangle t} = a_1^{\triangle t} + a_2^{\triangle t} + \ldots + a_n^{\triangle t}. \tag{1.2.12}$$

Analogously holds for the service process

$$t_{dn}^{\triangle t} = d_1^{\triangle t} + d_2^{\triangle t} + \ldots + d_n^{\triangle t} \quad . \tag{1.2.13}$$

Then $t_{an}^{\triangle t}$ and $t_{dn}^{\triangle t}$ are the sum of n independent and identically distributed random variables and the central limit theorem holds. Thus, $t_{an}^{\triangle t}$ and $t_{dn}^{\triangle t}$ are normally distributed, if $\triangle t$ is sufficiently large (and therefore the number of arrivals and services in $[t, t+\triangle t]$ is sufficiently large). Thus, also $N(t+\triangle t)-N(t) = t_{an}^{\triangle t} - t_{dn}^{\triangle t}$ (cf. (1.2.11)) is as a difference of two normally distributed random variables again normally distributed with the first and second moment

$$E\left[N(t+\triangle t) - N(t)\right] = \left(\frac{1}{\mu_a} - \frac{1}{\mu_d}\right) \triangle t \tag{1.2.14}$$

and

$$Var\left[N(t+\triangle t) - N(t)\right] = \left(\frac{c_a}{\mu_a} - \frac{c_d}{\mu_d}\right) \triangle t \tag{1.2.15}$$

with

$$c_a = \frac{\sigma_a^2}{\mu_a^2} \qquad \text{and} \qquad c_d = \frac{\sigma_d^2}{\mu_d^2} \quad . \tag{1.2.16}$$

For the differential equation the infinitesimal mean and variance, which are derivatives with respect to the time of the conditional mean and variance (cf. [**Kle76**]) are required. They evaluate for all time homogeneous arrival and service processes to

$$E\left[\frac{d}{dt}N(t)\right] = \frac{1}{\mu_a} - \frac{1}{\mu_d} =: \mu \qquad \text{and} \qquad Var\left[\frac{d}{dt}N(t)\right] = \frac{c_a}{\mu_a} - \frac{c_d}{\mu_d} =: \sigma^2 \quad . \tag{1.2.17}$$

The approximate continuous process $Z(t)$ of $N(t)$ is then chosen such, that $dZ(t) = Z(t+\triangle t) - Z(t)$ is normally distributed with mean $\mu\triangle t$ and variance $\sigma^2\triangle t$.

We consider the *probability density function* (p.d.f.)

$$\pi_s(t,x) = \frac{\partial}{\partial x} P[S(t) = s, Z(t) \leq x] \quad , \tag{1.2.18}$$

which is collected for $s = 1, \ldots, S$ in the vector $\boldsymbol{\pi}(t,x) \in \mathbb{R}^S$. $\boldsymbol{\pi}(t,x)$ is a non-negative function, which satisfies the *Chapman-Kolmogorov equation*

$$\pi_s(t+\triangle t, x) = \int_\Omega \pi_s(t+\triangle t, x|t,\tilde{x})\pi_s(t,\tilde{x})\, d\tilde{x} \quad + (\boldsymbol{\pi}^T(t,x)\mathbf{Q})_s \triangle t \tag{1.2.19}$$

for $\Omega = \left[x^{\min}, x^{\max}\right]$ and where $\boldsymbol{\pi}(t+\triangle t, x \,\big|\, t, \tilde{x})$ is the conditional p.d.f.. The Chapman-Kolmogorov equation describes the change in the p.d.f. over a small time interval of length $\triangle t$ by considering all possible intermediate continuous states (first term) and summing up the discrete state transitions (second term). We will consider the integrands for each component. The function $\pi_s(t,\tilde{x})$ can be approximated by its Taylor expansion around x, where we regard terms up to order two (*second* order FSPNs):

$$\pi_s(t,\tilde{x}) = \pi_s(t,x) + (\tilde{x}-x)\frac{\partial}{\partial x}\pi_s(t,x) + \frac{(\tilde{x}-x)^2}{2}\frac{\partial^2}{\partial x^2}\pi_s(t,x) + \mathrm{O}((\tilde{x}-x)^3). \tag{1.2.20}$$

Since $\pi_s(t+\triangle t, x|t,\tilde{x})$ is a density function of a normally distributed random variable with the infinitesimal expectation $x - \triangle t\, \mu_s$ and the variance $\triangle t\, \sigma_s^2$, it holds

$$\pi_s(t+\triangle t, x|t,\tilde{x}) = \frac{1}{\sqrt{2\pi\triangle t\sigma_s^2}} \exp\left\{\frac{-(x-\tilde{x}-\triangle t\mu_s)^2}{2\triangle t\sigma_s^2}\right\}. \tag{1.2.21}$$

We simplify (1.2.19) with (1.2.20) and (1.2.21)

$$\begin{aligned}
&\int_\Omega \pi_s(t+\triangle t, x|t,\tilde{x})\pi_s(t,\tilde{x})\, d\tilde{x} \\
&\approx \frac{1}{\sqrt{2\pi\triangle t\sigma_s^2}} \int_\Omega e^{\frac{-(x-\tilde{x}-\triangle t\mu_s)^2}{2\triangle t\sigma_s^2}} \cdot \left[\pi_s(t,x) + (\tilde{x}-x)\frac{\partial}{\partial x}\pi_s(t,x) + \frac{(\tilde{x}-x)^2}{2}\frac{\partial^2}{\partial x^2}\pi_s(t,x)\right] d\tilde{x}.
\end{aligned}$$

Since we integrate a density function, $\frac{1}{\sqrt{2\pi\triangle t\sigma_s^2}} \int_\Omega e^{\frac{-(x-\tilde{x}-\triangle t\mu)^2}{2\triangle t\sigma^2}}\, d\tilde{x} = 1$ holds and yields

$$\begin{aligned}
&\int_\Omega \pi_s(t+\triangle t, x|t,\tilde{x})\pi_s(t,\tilde{x})\, d\tilde{x} \\
&\qquad \approx \pi_s(t,x) - x\frac{\partial}{\partial x}\pi_s(t,x) + \frac{\partial}{\partial x}\pi_s(t,x) \underbrace{\int_\Omega \frac{1}{\sqrt{2\pi\triangle t\sigma_s^2}} e^{\frac{-(x-\tilde{x}-\triangle t\mu_s)^2}{2\triangle t\sigma_s^2}} \cdot \tilde{x}\, d\tilde{x}}_{\text{Definition of expectation } = x-\triangle t\mu_s} \\
&\qquad\qquad + \frac{1}{2}\frac{\partial^2}{\partial x^2}\pi_s(t,x) \underbrace{\int_\Omega \frac{1}{\sqrt{2\pi\triangle t\sigma_s^2}} e^{\frac{-(x-\tilde{x}-\triangle t\mu_s)^2}{2\triangle t\sigma_s^2}} \cdot (\tilde{x}-x)^2\, d\tilde{x}}_{\text{Definition of the second moments}}.
\end{aligned}$$

Using the definition of expectation and second moment $\triangle t^2\mu_s^2 + \triangle t\sigma_s^2$ (not the central one, cf. [**New71**]), (1.2.19) reads

(1.2.22)
$$\boldsymbol{\pi}^T(t+\triangle t,x) = \boldsymbol{\pi}^T(t,x) - \triangle t\frac{\partial}{\partial x}\boldsymbol{\pi}^T(t,x)\boldsymbol{M} + \triangle t\frac{1}{2}\frac{\partial^2}{\partial x^2}\boldsymbol{\pi}^T(t,x)\left(\triangle t\boldsymbol{M}^2 + \boldsymbol{\Sigma}^2\right) + \boldsymbol{\pi}^T(t,x)\mathbf{Q}\triangle t.$$

Here, $\boldsymbol{M}$ and $\boldsymbol{\Sigma}^2$ are diagonal $S \times S$-matrices, with $\boldsymbol{M} = \operatorname{diag}(\mu_s)$ and $\boldsymbol{\Sigma}^2 = \operatorname{diag}(\sigma_s^2)$. We subtract $\boldsymbol{\pi}^T(t,x)$ and divide by $\triangle t$

(1.2.23)
$$\frac{\boldsymbol{\pi}^T(t+\triangle t,x) - \boldsymbol{\pi}^T(t,x)}{\triangle t} = -\frac{\partial}{\partial x}\boldsymbol{\pi}^T(t,x)\boldsymbol{M} + \frac{1}{2}\frac{\partial^2}{\partial x^2}\boldsymbol{\pi}^T(t,x)\left(\triangle t\boldsymbol{M}^2 + \boldsymbol{\Sigma}^2\right) + \boldsymbol{\pi}^T(t,x)\mathbf{Q}.$$

The limit $\triangle t \to 0$ yields the differential equation

$$\frac{\partial}{\partial t}\boldsymbol{\pi}^T(t,x) = -\frac{\partial}{\partial x}\boldsymbol{\pi}^T(t,x)\boldsymbol{M} + \frac{1}{2}\frac{\partial^2}{\partial x^2}\boldsymbol{\pi}^T(t,x)\boldsymbol{\Sigma}^2 + \boldsymbol{\pi}^T(t,x)\mathbf{Q} \qquad \forall x \in \mathring{\Omega} = \left(x^{\min}, x^{\max}\right), \quad t \geq 0$$

for $\boldsymbol{M}$ and $\boldsymbol{\Sigma}^2$ not depending on x.

According to [**CM65**] for the process with space depending $\boldsymbol{M} = \boldsymbol{M}(x)$ and $\boldsymbol{\Sigma}^2 = \boldsymbol{\Sigma}^2(x)$ results in

$$\frac{\partial}{\partial t}\boldsymbol{\pi}^T(t,x) = -\frac{\partial}{\partial x}\left\{\boldsymbol{\pi}^T(t,x)\boldsymbol{M}(x)\right\} + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left\{\boldsymbol{\pi}^T(t,x)\boldsymbol{\Sigma}^2(x)\right\} + \boldsymbol{\pi}^T(t,x)\mathbf{Q} \qquad \forall x \in \mathring{\Omega} = \left(x^{\min}, x^{\max}\right), \quad t \geq 0 \tag{1.2.24}$$

This partial differential equation is called *Kolmogorov forward equation*.

The boundary condition is determined by using the property of a mixed discrete-real random variable

$$\sum_{s=1}^{S}\int_{-\infty}^{\infty}\pi_s(t,x)\,dx = \sum_{s=1}^{S}\int_{x^{\min}}^{\infty}\pi_s(t,x)\,dx = 1. \tag{1.2.25}$$

The boundary $x^{\min}$ is called a reflecting barrier. (1.2.25) together with (1.2.24) yields the reflecting boundary condition and thus the initial boundary value problem reads

(1.2.26a)
$$\frac{\partial}{\partial t}\boldsymbol{\pi}(t,x) + \frac{\partial}{\partial x}\left(\boldsymbol{M}(x)\boldsymbol{\pi}(t,x)\right) = \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(\boldsymbol{\Sigma}^2(x)\boldsymbol{\pi}(t,x)\right) + \mathbf{Q}^T\boldsymbol{\pi}(t,x), \qquad x \in \Omega, \quad t \geq 0,$$

$$\frac{1}{2}\frac{\partial}{\partial x}\left(\boldsymbol{\Sigma}^2(x)\boldsymbol{\pi}(t,x)\right)\Big|_{x=x^{\min}} - \boldsymbol{M}(x)\boldsymbol{\pi}(t,x)\Big|_{x=x^{\min}} = \mathbf{0}, \tag{1.2.26b}$$

$$\boldsymbol{\pi}(0,x) = \delta(x-x_0)\boldsymbol{\pi}_0, \qquad x \in \Omega. \tag{1.2.26c}$$

Here δ denotes the Dirac-Delta distribution. From now on we assume, that $\boldsymbol{\Sigma}^2$ is a regular matrix.

Fluid places are bounded from below by x^{min}, which is often equal to zero. Thus the reflecting boundary condition, which prohibits any mass flow through the boundary, is postulated at the left boundary (cf. (1.2.26b)). If the fluid place is also bounded from above, the same kind of boundary condition is posed at the upper boundary. If e.g. the average fill level of a fluid place is looked for and no upper bound can be given beforehand, the initial boundary problem is posed on the half-space $x \geq x^{min}$. To calculate its solution numerically, TBCs have to be imposed. For the formulation of these TBCs it is necessary to know about some properties of the evolution equation (1.2.26) and in particular of the generator matrix $\boldsymbol{Q}$.

3. Properties of the generator matrix $\boldsymbol{Q}$

The generator matrix $\boldsymbol{Q}$ is the only coupling between the differential equations. Therefore, its structure and properties are very important for the behaviour and analysis of the solutions to the differential equation. These characteristics will also enter the derivation of the TBCs in Sec. 7.2. Hence, we will here investigate the generator matrix $\boldsymbol{Q}$ in detail. We summarise the essential properties:

- all off-diagonal entries are transition rates and thus nonnegative

$$q_{ij} \geq 0 \qquad \forall i, j = 1, \ldots, S \text{ with } i \neq j \tag{1.3.1}$$

- all diagonal entries are the negative sum of the off-diagonal entries in the row

$$q_{ii} = -\sum_{\substack{j=1 \\ j \neq i}}^{S} q_{ij} \qquad \forall i = 1, \ldots, S \tag{1.3.2}$$

and as a consequence of the two preceeding properties, we have

- the row sum is equal to zero (cf. (1.1.2))

$$\sum_{j=1}^{S} q_{ij} = 0 \qquad \forall i = 1, \ldots, S \tag{1.3.3}$$

In the following we will show, that $\lambda = 0$ is a simple eigenvalue of $\mathbf{Q}$, provided that $\mathbf{Q}$ is irreducible.

Definition 1.2 ([**HJ99a**]). *A matrix* $\mathbf{A}$ *of dimension* $n \times n$ *is said to be*

(a) reducible *if either*

- $n = 1$ *and* $\mathbf{A} = \mathbf{0}$*; or*
- $n \geq 2$*, there is a permutation matrix* $\mathcal{P}$ *and some integer* r *with* $1 \leq r \leq n-1$ *such that*

$$\mathcal{P}^{\mathbf{T}}\mathbf{A}\mathcal{P} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{pmatrix}$$

where $\mathbf{A}_{11}$ *is a* $r \times r$*-matrix and* $\mathbf{A}_{22}$ *is a* $(n-r)\times(n-r)$*-matrix.*

(b) irreducible *if it is not reducible.*

To study the spectrum of $\boldsymbol{Q}$, we define a matrix $\boldsymbol{P}$ by

$$\boldsymbol{P} := \boldsymbol{I} + \frac{1}{q_{max}}\boldsymbol{Q}, \qquad q_{max} = \max_{i=1,\ldots,S} |q_{ii}|. \tag{1.3.4}$$

The scaling factor q_{max} is chosen in a way such that all entries in $\boldsymbol{P}$ are nonnegative and the row sums equal one, i.e. $\boldsymbol{P}$ is a stochastic matrix. If $\boldsymbol{Q}$ is irreducible also $\boldsymbol{P}$ is irreducible and the following lemma holds.

LEMMA 1.1 ([**HJ99a**],Theorem 8.4.4). *If the stochastic matrix* $\boldsymbol{P}$ *is irreducible, then* $\rho(\boldsymbol{P})$ *is an algebraically (and hence geometrically) simple eigenvalue of* $\boldsymbol{P}$*.*

LEMMA 1.2. *The stochastic matrix* $\boldsymbol{P}$ *has the spectral radius* $\rho(\boldsymbol{P}) = 1$*.*

PROOF. Since the row sums of $\boldsymbol{P}$ constantly evaluate to one, the vector $\boldsymbol{x} = (1, \ldots, 1)^T$ is an eigenvector of $\boldsymbol{P}$ to the eigenvalue $||\boldsymbol{P}||_\infty = 1$ ($||\cdot||_\infty$ denotes the maximum row sum matrix norm):

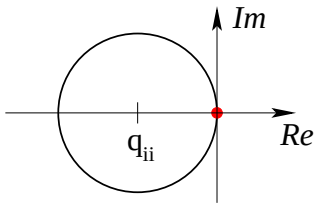
$$\boldsymbol{P}x = ||\boldsymbol{P}||_\infty x = 1 \cdot x.$$

Thus $\rho(\boldsymbol{P}) \geq ||\boldsymbol{P}||_\infty = 1$ holds. And since $\rho(\boldsymbol{P}) \leq ||\boldsymbol{P}||$ for any matrix norm induced by a vector norm, we conclude $\rho(\boldsymbol{P}) = ||\boldsymbol{P}||_\infty = 1$. □

Since the spectrum of $\boldsymbol{Q}$ is $\sigma(\boldsymbol{Q}) = q_{max}\,(\sigma(\boldsymbol{P}) - 1)$, the ensuing assertion follows directly from Lemma 1.1 and 1.2:

LEMMA 1.3. *If the generator matrix* $\boldsymbol{Q}$ *is irreducible, then* $\lambda = 0$ *is an algebraically (and hence geometrically) simple eigenvalue of* $\boldsymbol{Q}$*.*

REMARK 1.4. *$\lambda = 0$ is the eigenvalue of $\mathbf{Q}$ with the largest real part. All other eigenvalues have negative real part. This results from a consideration of Gerschgorin circles as in the right hand figure: In each row i of $\mathbf{Q}$ the diagonal entry is the negative sum of the off-diagonal entries (cf.* (1.3.2)*).*



4. Properties of the evolution equation

The evolution equation (1.2.26) is a system of linear second order partial differential equations of parabolic type in one spatial dimension (advection-diffusion equation, *Fokker-Planck equation*, cf. [**Ris84**]).

Here we will show two properties of equation (1.2.26). First we recall that the quantity

$$\sum_{s=1}^{S} \int_{x^{min}}^{\infty} \pi_s(t,x)dx \tag{1.4.1}$$

is constant in time (normalisation condition). Integrating (1.2.26) for $x^{min} < x < \infty$ yields

$$\frac{\partial}{\partial t} \int_{x^{min}}^{\infty} \boldsymbol{\pi}(t,x)\, dx = \left[\frac{1}{2}\frac{\partial}{\partial x}\boldsymbol{\Sigma}^2(x)\boldsymbol{\pi}(t,x) - \boldsymbol{M}(x)\boldsymbol{\pi}(t,x)\right]_{x^{min}}^{\infty} + \int_{x^{min}}^{\infty} \mathbf{Q}^T\boldsymbol{\pi}(t,x)\, dx. \tag{1.4.2}$$

The boundary term equals zero, due to the boundary condition (1.2.26b) and the decaying condition $\pi_s(t,\cdot) \in L^1(\mathbb{R}^+)$. The coupling term vanishes after summation w.r.t. the number of states s.

Second we want to show the positivity of the solution to (1.2.26), i.e. starting with nonnegative initial data, the solution will remain nonnegative. To do so we split (operator splitting) equation (1.2.26) in the following way [**CHMM78**]:

$$\begin{aligned}
&\boldsymbol{\pi}_t^1 = \frac{1}{2}\boldsymbol{\Sigma}^2\boldsymbol{\pi}_{xx}^1 - \mathbf{M}\boldsymbol{\pi}_x^1 =: \mathbf{G}\boldsymbol{\pi}^1, && \text{for } i\triangle t < t < (i+1)\triangle t, \quad x \geq x^{min} \\
&\boldsymbol{\pi}^1(i\triangle t, x) = \boldsymbol{\pi}^2(i\triangle t, x) \\
&\frac{1}{2}\frac{\partial}{\partial x}\left(\boldsymbol{\Sigma}^2(x^{min})\boldsymbol{\pi}^1(t,x^{min})\right) - \boldsymbol{M}(x^{min})\boldsymbol{\pi}^1(t,x^{min}) = \mathbf{0}
\end{aligned} \tag{1.4.3}$$

$$\begin{aligned}
&\boldsymbol{\pi}_t^2 = \mathbf{Q}^T\boldsymbol{\pi}^2, && \text{for } i\triangle t < t < (i+1)\triangle t, \quad x \geq x^{min} \\
&\boldsymbol{\pi}^2(0,x) = \delta(x-x_0)\boldsymbol{\pi}_0, && x \geq x^{min} \\
&\boldsymbol{\pi}^2(i\triangle t, x) = \boldsymbol{\pi}^1\left((i+1)\triangle t, x\right).
\end{aligned} \tag{1.4.4}$$

The solutions to (1.4.3) remain nonnegative, because the equation is nothing else but S scalar convection diffusion equations with Robin boundary data, for which the positivity preservations is well known ([**PW84**],Chap. 3, Sec. 2). Equation (1.4.4) has the solution $\boldsymbol{\pi}(t,x) = e^{\mathbf{Q}^T t}\boldsymbol{\pi}(0,x)$, which is nonnegative, since in

$$e^{\mathbf{Q}^T t} = e^{q_{max}t(\mathbf{P}^T - \mathbf{I})} = e^{q_{max}t\mathbf{P}^T}e^{-q_{max}t\mathbf{I}} \tag{1.4.5}$$

the first exponential is a nonnegative matrix for $t \geq 0$ and the second is diagonal.

$\mathbf{G}$ is the infinitesimal generator of a C_0 semigroup in $L^1(\mathbb{R}^+)$ [**Paz83**] and $\mathbf{Q}^T$ is a bounded linear operator. Then $\mathbf{G} + \mathbf{Q}^T$ is also the infinitesimal generator of a C_0 semigroup (cf. [**Paz83**],Theorem 3.1.1) and hence closed. If now $\mathbf{G}$ and $\mathbf{Q}^T$ are generators of *quasi-contractive* semigroups (i.e. $||\exp\{t\mathbf{G}\}||_{B(L^1(\mathbb{R}^+))} \leq \exp\{t\gamma\}$, $||\exp\{t\mathbf{Q}^T\}||_{B(L^1(\mathbb{R}^+))} \leq \exp\{t\eta\}$), then we know from [**CHMM78**], application 3.5, that the splitting converges for $\triangle t \to 0$, and also the solution of (1.2.26) is nonnegative. For the quasi-contractiveness we consider solutions $\boldsymbol{\pi}_1$ to (1.4.3). The L^1-norm of $\boldsymbol{\pi}_1$ is non-increasing $||\boldsymbol{\pi}_1(t,\cdot)||_{L^1(\mathbb{R}^+)} \leq ||\boldsymbol{\pi}_1(0,\cdot)||_{L^1(\mathbb{R}^+)}$ and thus

$$||\exp\{t\mathbf{G}\}||_{B(L^1(\mathbb{R}^+))} = \sup \frac{||\exp\{t\mathbf{G}\}\boldsymbol{\pi}_1(0,\cdot)||_{L^1(\mathbb{R}^+)}}{||\boldsymbol{\pi}_1(0,\cdot)||_{L^1(\mathbb{R}^+)}} = \sup \frac{||\boldsymbol{\pi}_1(t,\cdot)||_{L^1(\mathbb{R}^+)}}{||\boldsymbol{\pi}_1(0,\cdot)||_{L^1(\mathbb{R}^+)}} \leq 1, \tag{1.4.6}$$

i.e. γ can be chosen as zero.

For the quasi-contractiveness of $\mathbf{Q}^T$ we consider

$$\begin{aligned} ||\tilde{\boldsymbol{\pi}}(t)||_{L^1(\mathbb{R}^+)} &\leq ||\exp\{t\mathbf{Q}^T\}||_{B(L^1(\mathbb{R}^+))} \cdot ||\tilde{\boldsymbol{\pi}}_0||_{L^1(\mathbb{R}^+)} \\ &\leq \sum_{n=0}^{\infty} \frac{1}{n!}||\mathbf{Q}^T||^n_{B(L^1(\mathbb{R}^+))}t^n||\tilde{\boldsymbol{\pi}}_0||_{L^1(\mathbb{R}^+)} = e^{t||\mathbf{Q}^T||_{B(L^1(\mathbb{R}^+))}}||\tilde{\boldsymbol{\pi}}_0||_{L^1(\mathbb{R}^+)} \end{aligned} \tag{1.4.7}$$

and thus we have $\eta = ||\mathbf{Q}^T||_{B(L^1(\mathbb{R}^+))}$ and any solution of (1.2.26) with nonnegative initial data remains nonnegative.

After these preparations we can proceed to derive the transparent boundary condition at $x = x^{\max}$.

5. The transparent boundary condition

To derive the transparent boundary condition we consider the differential equation on the exterior domain. Performing a Laplace transformation gives a system of ordinary differential equations with constant coefficients, which can be solved explicitly. Therefore

we reduce the problem to a system of first order differential equations (*reduction of order method*). Its solution is given by calculating the Jordan form and the corresponding eigenvectors of the system matrix. We will prove, that the eigenvalues split in two groups: half have positive and half have negative real parts. The eigenvalues with a negative real part create a for $x \to \infty$ decaying solution, which can be used to formulate the *Dirichlet-to-Neumann map* on the boundary. We can give no explicit formulation of these TBCs, because their derivation involves the inverse Laplace transform of matrices of eigenvectors, that in general cannot be calculated explicitly.

For the derivation of the TBC we cut the original half-space problem at $x = x^{\max}$, where the TBC is supposed to be. This generates two subproblems: the interior and the exterior problem. They are coupled by the assumption, that $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_x$ are continuous at the boundary $x = x^{\max}$. The interior problem then reads

$$
\begin{aligned}
\frac{\partial}{\partial t}\boldsymbol{\pi} &= \frac{1}{2}\frac{\partial^2}{\partial x^2}(\boldsymbol{\Sigma}^2\boldsymbol{\pi}) - \frac{\partial}{\partial x}(\mathbf{M}\boldsymbol{\pi}) + \mathbf{Q}^T\boldsymbol{\pi}, \qquad x^{\min} < x < x^{\max}, \\
\boldsymbol{\pi}(0,x) &= \delta(x-x_0)\boldsymbol{\pi}_0, \\
\mathbf{0} &= \frac{1}{2}\frac{\partial}{\partial x}\big(\boldsymbol{\Sigma}^2(x)\boldsymbol{\pi}(t,x)\big)\Big|_{x=x^{\min}} - \mathbf{M}(x)\boldsymbol{\pi}(t,x)\Big|_{x=x^{\min}}, \\
\boldsymbol{\pi}_x(t,x^{\max}) &= (\mathbf{T}\boldsymbol{\pi})(t,x^{\max}),
\end{aligned}
\tag{1.5.1}
$$

where $\mathbf{T}$ denotes the Dirichlet-to-Neumann operator, that is determined by solving the exterior problem

$$
\begin{aligned}
\boldsymbol{\xi}_t &= \frac{1}{2}\boldsymbol{\Sigma}^2\boldsymbol{\xi}_{xx} - \mathbf{M}\boldsymbol{\xi}_x + \mathbf{Q}^T\boldsymbol{\xi}, \qquad x > x^{\max}, \\
\boldsymbol{\xi}(0,x) &= \mathbf{0}, \\
\boldsymbol{\xi}(t,x^{\max}) &= \boldsymbol{\eta}(t), \qquad t > 0, \qquad \boldsymbol{\eta}(0) = \mathbf{0}, \\
\boldsymbol{\xi}(t,\infty) &= \mathbf{0}, \\
(\mathbf{T}\boldsymbol{\eta})(t) &:= \boldsymbol{\xi}_x(t,x^{\max}).
\end{aligned}
\tag{1.5.2}
$$

The coefficient matrices of this exterior problem are all constant in x. Therefore, it can be solved explicitly by the Laplace-method. By this way, the Dirichlet-to-Neumann operator, that gives the TBC in (1.5.1), is determined .

5.1. Reduction of order method. We solve the Laplace-transformed matrix differential equation after reducing its order. We obtain a system of ordinary differential

equations of the form

$$\mathbf{w}_x = \mathbf{W}\mathbf{w} = \mathbf{P}_W \mathbf{J}_W \mathbf{P}_W^{-1}\mathbf{w}, \tag{1.5.3}$$

where $\mathbf{J}_W$ is the Jordan form of $\mathbf{W}$ and $\mathbf{P}_W^{-1}$ contains the corresponding (possibly generalised) eigenvectors. For the TBCs it is necessary to find a decaying solution to (1.5.3). We will show, that the $2S$ eigenvalues of $\mathbf{W}$ split into S eigenvalues with positive real part and S with negative real part. Since the solution to (1.5.3) is given by $\mathbf{w}(x) = e^{\mathbf{W}x} = e^{\mathbf{P}_W \mathbf{J}_W \mathbf{P}_W^{-1} x}$, the S negative eigenvalues yield S linearly independent solutions $\mathbf{w}_i(x)$, $i = 1, \ldots, S$, which decay for $x \to \infty$: $|\mathbf{w}_i(x)| \to 0$.

In the construction of the TBC we start with the differential equation on the exterior domain $x \geq x^{max}$ ("exterior problem")

$$\boldsymbol{\pi}_t = \frac{1}{2}\boldsymbol{\Sigma}^2 \boldsymbol{\pi}_{xx} - \mathbf{M}\boldsymbol{\pi}_x + \mathbf{Q}^T \boldsymbol{\pi}, \tag{1.5.4}$$

where $\mathbf{M}, \boldsymbol{\Sigma}^2$ and $\mathbf{Q}^T$ are assumed to be constant in x. Using the Laplace transformation (in t)

$$\hat{\boldsymbol{\pi}}(x,s) = \int_0^\infty e^{-st} \boldsymbol{\pi}(x,t)\, dt, \quad s = \alpha + i\xi, \quad \alpha > 0, \quad \xi \in \mathbb{R}, \tag{1.5.5}$$

of (1.5.4) yields a system of ordinary differential equations

$$\mathbf{M}\hat{\boldsymbol{\pi}}_x - \frac{1}{2}\boldsymbol{\Sigma}^2 \hat{\boldsymbol{\pi}}_{xx} - \left(\mathbf{Q}^T - s\mathbf{I}\right)\hat{\boldsymbol{\pi}} = \mathbf{0}, \qquad x \geq x^{max} \tag{1.5.6}$$

depending on the complex parameter $s \in \mathbb{C}$. We will discuss, under which conditions the exterior problem (1.5.6) has a unique solution at the end of this section in Lem. 1.8.

In order to derive the TBC we reformulate the exterior problem (1.5.6) as a first order system by introducing $\zeta_i = \hat{\pi}_i$ and $\eta_i = (\zeta_i)_x$:

$$\mathbf{A}\begin{pmatrix}\boldsymbol{\zeta}\\ \boldsymbol{\eta}\end{pmatrix}_x = \mathbf{B}\begin{pmatrix}\boldsymbol{\zeta}\\ \boldsymbol{\eta}\end{pmatrix}, \quad x > x^{max} \tag{1.5.7}$$

with

$$\mathbf{A} := \begin{pmatrix}\mathbf{M} & -\frac{1}{2}\boldsymbol{\Sigma}^2\\ -\frac{1}{2}\boldsymbol{\Sigma}^2 & \mathbf{0}\end{pmatrix}, \qquad \mathbf{B} := \begin{pmatrix}\mathbf{Q}^T - s\mathbf{I} & \mathbf{0}\\ \mathbf{0} & -\frac{1}{2}\boldsymbol{\Sigma}^2\end{pmatrix}. \tag{1.5.8}$$

If $\sigma_s^2 \neq 0$ for $s = 1, \ldots, S$, then $\mathbf{A}$ is regular and we can write

$$\begin{pmatrix}\boldsymbol{\zeta}\\ \boldsymbol{\eta}\end{pmatrix}_x = \mathbf{A}^{-1}\mathbf{B}\begin{pmatrix}\boldsymbol{\zeta}\\ \boldsymbol{\eta}\end{pmatrix} \quad \text{with} \quad \mathbf{A}^{-1} = \begin{pmatrix}\mathbf{0} & -2(\boldsymbol{\Sigma}^2)^{-1}\\ -2(\boldsymbol{\Sigma}^2)^{-1} & -4(\boldsymbol{\Sigma}^2)^{-1}\mathbf{M}(\boldsymbol{\Sigma}^2)^{-1}\end{pmatrix}. \tag{1.5.9}$$

In the following we will show, that the $2S$ eigenvalues of $\mathbf{A}^{-1}\mathbf{B}$ include exactly S eigenvalues with negative real parts, which correspond to a decaying solution.

THEOREM 1.5 (Splitting Theorem). *The eigenvalues of* $\mathbf{A}^{-1}\mathbf{B}$ *split in* S *eigenvalues with positive and* S *with negative real part, if* $Re(s) > \rho(\frac{\mathbf{Q}+\mathbf{Q}^T}{2})$.

PROOF. To prove this we first consider the eigenvalues of $\mathbf{A}$:

$$\begin{vmatrix} \mathbf{M} - \lambda\mathbf{I} & -\frac{1}{2}\mathbf{\Sigma}^2 \\ -\frac{1}{2}\mathbf{\Sigma}^2 & -\lambda\mathbf{I} \end{vmatrix} = 0, \tag{1.5.10}$$

from which we sort rows and columns to obtain the following block diagonal structure:

$$\begin{vmatrix} \mathbf{A}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \dots & \mathbf{0} \\ \mathbf{0} & & \ddots & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{A}_S \end{vmatrix} = 0 \qquad \text{with} \qquad \mathbf{A}_i = \begin{pmatrix} \mu_i - \lambda & -\frac{1}{2}\sigma_i^2 \\ -\frac{1}{2}\sigma_i^2 & -\lambda \end{pmatrix}. \tag{1.5.11}$$

This determinant can be calculated by the product of the determinants of the matrices on the diagonal

$$\det\mathbf{A} = \prod_{i=1}^{S} \det\left(\mathbf{A}_i\right) = 0. \tag{1.5.12}$$

Hence, the eigenvalues are

$$\lambda_i^+ = \frac{\mu_i}{2} + \frac{1}{2}\sqrt{\mu_i^2 + \sigma_i^4}, \qquad \lambda_i^- = \frac{\mu_i}{2} - \frac{1}{2}\sqrt{\mu_i^2 + \sigma_i^4}, \quad i = 1, \dots, S. \tag{1.5.13}$$

And thus, for each $i = 1, \dots, S$ the eigenvalues λ_i^+ and λ_i^- have different signs

$$\left.\begin{matrix} \lambda_i^+ > 0 \\ \lambda_i^- < 0 \end{matrix}\right\} \text{ for } \mu_i \geq 0 \qquad \text{and} \qquad \left.\begin{matrix} \lambda_i^+ < 0 \\ \lambda_i^- > 0 \end{matrix}\right\} \text{ for } \mu_i < 0, \quad i = 1, \dots, S. \tag{1.5.14}$$

We define the *inertia* of a matrix $\mathbf{M}$:

DEFINITION 1.3 ([**HJ99b**]). *Let* $\mathbf{M}$ *be a complex matrix. The* inertia *of* $\mathbf{M}$ *is the ordered triple*

$$i(\mathbf{M}) = \left(i_+\left(\mathbf{M}\right), i_-\left(\mathbf{M}\right), i_0\left(\mathbf{M}\right)\right), \tag{1.5.15}$$

where $i_+(\mathbf{M})$ *is the number of eigenvalues of* $\mathbf{M}$ *with positive real part,* $i_-(\mathbf{M})$ *is the number of eigenvalues of* $\mathbf{M}$ *with negative real part, and* $i_0(\mathbf{M})$ *is the number of eigenvalues of* $\mathbf{M}$ *with zero real part, all counting multiplicity.*

For the matrix $\mathbf{A}$ we know from (1.5.14) that $i(\mathbf{A}) = (S, S, 0) = i(\mathbf{A}^{-1})$. With the following theorem, we will deduce the inertia of $\mathbf{A}^{-1}\mathbf{B}$ from the inertia of $\mathbf{A}^{-1}$:

THEOREM 1.6 ([**HJ99b**]Theorem 2.4.15). *Let* $\mathbf{A}^{-1}$ *and* $\mathbf{B}$ *be quadratic matrices with* $\mathbf{A}^{-1}$ *Hermitian and* $\mathbf{B} + \mathbf{B}^*$ *positive definite. Then:* $i(\mathbf{A}^{-1}\mathbf{B}) = i(\mathbf{A}^{-1})$.

We can apply Thm. 1.6, since $\mathbf{A}^{-1}$ is symmetric and real and thus Hermitian with $i(\mathbf{A}) = (S, S, 0)$. Furthermore, $\mathbf{B} + \mathbf{B}^* = \begin{pmatrix} \mathbf{Q}^T + \mathbf{Q} - 2\mathrm{Re}(s)\mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{\Sigma}^2 \end{pmatrix}$ is negative definite for $Re(s)$ sufficiently large ($\mathrm{Re}(s) > 0$ for the Laplace-transformation). Then $-\mathbf{B} + (-\mathbf{B})^*$ is positive definite and we obtain for the inertia of $\mathbf{C} := \mathbf{A}^{-1}\mathbf{B}$:

$$i(\mathbf{C}) = i\left((-\mathbf{A}^{-1})(-\mathbf{B})\right) = i(-\mathbf{A}^{-1}) = i(\mathbf{A}^{-1}) = i(\mathbf{A}) = (S, S, 0), \tag{1.5.16}$$

which finishes the proof of Thm. 1.5. □

For the elements of $\mathbf{C}$ we have

$$\mathbf{C} = \mathbf{A}^{-1}\mathbf{B} = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ -2(\mathbf{\Sigma}^2)^{-1}(\mathbf{Q}^T - s\mathbf{I}) & 2(\mathbf{\Sigma}^2)^{-1}\mathbf{M} \end{pmatrix} \tag{1.5.17}$$

and (1.5.9) reads

$$\begin{pmatrix} \boldsymbol{\zeta} \\ \boldsymbol{\eta} \end{pmatrix}_x = \mathbf{C} \begin{pmatrix} \boldsymbol{\zeta} \\ \boldsymbol{\eta} \end{pmatrix} \text{ with } i(\mathbf{C}) = (S, S, 0). \tag{1.5.18}$$

Now we transform $\mathbf{C}$ into Jordan form $\mathbf{C} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$, where $\mathbf{P}$ holds the possibly generalised eigenvectors in columns and the eigenvalues of $\mathbf{C}$ are the diagonal elements of J. Since we know, that there are n eigenvalues with negative and n with positive real part, we can suppose the eigenvalues in $\mathbf{J}$, $\lambda_1, \ldots, \lambda_{2S}$, to be sorted with respect to increasing real parts. Accordingly we can split $\mathbf{J} = \left(\begin{smallmatrix} \mathbf{J}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{smallmatrix}\right)$, where $\mathbf{J}_1$ holds all Jordan blocks to eigenvalues with negative and $\mathbf{J}_2$ all eigenvalues with positive real part. Now the system looks as follows

$$\begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}_x = \begin{pmatrix} \mathbf{J}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{pmatrix} \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix} \quad \text{with} \quad \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix} := \mathbf{P}^{-1} \begin{pmatrix} \boldsymbol{\zeta} \\ \boldsymbol{\eta} \end{pmatrix} \tag{1.5.19}$$

and $\mathbf{f}$ decays for $x \to \infty$.

Let $\mathbf{P}^{-1} = \left(\begin{smallmatrix} \mathbf{P}_1 & \mathbf{P}_2 \\ \mathbf{P}_3 & \mathbf{P}_4 \end{smallmatrix}\right)$, then

$$\mathbf{P}^{-1} \begin{pmatrix} \boldsymbol{\zeta} \\ \boldsymbol{\eta} \end{pmatrix}_x = \begin{pmatrix} \mathbf{J}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{pmatrix} \begin{pmatrix} \mathbf{P}_1 & \mathbf{P}_2 \\ \mathbf{P}_3 & \mathbf{P}_4 \end{pmatrix} \begin{pmatrix} \boldsymbol{\zeta} \\ \boldsymbol{\eta} \end{pmatrix} = \begin{pmatrix} \mathbf{J}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{pmatrix} \begin{pmatrix} \mathbf{P}_1\boldsymbol{\zeta} + \mathbf{P}_2\boldsymbol{\eta} \\ \mathbf{P}_3\boldsymbol{\zeta} + \mathbf{P}_4\boldsymbol{\eta} \end{pmatrix}. \tag{1.5.20}$$

The lower row of this equation

$$\mathbf{P}_3\boldsymbol{\zeta} + \mathbf{P}_4\boldsymbol{\eta} = \mathbf{0} \tag{1.5.21}$$

yields the transparent boundary condition, since eigenvalues with positive real part involve increasing - and thus unphysical - solutions, which are thus eliminated. If $\mathbf{P}_4$ is regular the TBC can be formulated in Dirichlet-to-Neumann form

$$\boldsymbol{\pi}_x = -\mathbf{P}_4^{-1}\mathbf{P}_3\boldsymbol{\pi}. \tag{1.5.22}$$

REMARK 1.7. *The order of the eigenvalues and the eigenvectors in* $\mathbf{P}$ *is insignificant (within the two blocks of decaying and increasing solutions). To illustrate this we set* $\mathbf{X} = \mathbf{P}_3$ *and* $\mathbf{Y} = \mathbf{P}_4$*. Then for regular* $\mathbf{Y}$ *the TBC reads* $\boldsymbol{\zeta}_x = -\mathbf{Y}^{-1}\mathbf{X}\boldsymbol{\zeta}$*. We express* $\mathbf{Y}^{-1}$ *using Laplacian determinant expansion by minors. Then holds* $(\mathbf{Y}^{-1}\mathbf{X})_{i,j} = \sum_{l=1}^{S} \frac{(-1)^{i+j}}{\det(\mathbf{Y})}\det(\tilde{\mathbf{Y}}_{l,i})\mathbf{X}_{l,j}$*, where* $\tilde{\mathbf{Y}}_{l,i}$ *is a minor of* $\mathbf{Y}$*. Since the summation index* l *occurs as row index for* $\mathbf{X}$ *as well as for* $\mathbf{Y}$*, the order of rows in* $\mathbf{X}$ *and* $\mathbf{Y}$ *is of no importance.*

In this section we derived a transparent boundary condition using the solution of the exterior problem. It finally remains to discuss the existence and uniqueness of this solution:

LEMMA 1.8. (Existence and uniqueness of the solution to the Laplace-transformed exterior problem).

a) *If the solution of the boundary value problem* (1.5.6) *with the boundary data*

$$\hat{\boldsymbol{\pi}}(x = x^{max}) = \hat{\boldsymbol{\pi}}_{max}, \qquad \hat{\boldsymbol{\pi}}(x = \infty) = \mathbf{0} \tag{1.5.23}$$

exists, it is unique for $Re(s)$ *sufficiently large.*

b) *If the matrix* $\boldsymbol{\mathcal{P}}_1$ *is regular, the solution exists.*

PROOF. a) To prove uniqueness we assume, that there are two such solutions $\hat{\boldsymbol{\pi}}_1$ and $\hat{\boldsymbol{\pi}}_2$ of (1.5.6),(1.5.23) . Then their difference $\hat{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}}_1 - \hat{\boldsymbol{\pi}}_2$ is a solution of the problem with *homogeneous boundary data.* Multiplying (1.5.6) by $\bar{\hat{\boldsymbol{\pi}}}^T$ from the left and integrating from x^{max} to ∞ yields after integrating by parts

$$\frac{1}{2}\int_{x^{max}}^{\infty} \bar{\hat{\boldsymbol{\pi}}}_x^T\boldsymbol{\Sigma}^2\hat{\boldsymbol{\pi}}_x dx + \int_{x^{max}}^{\infty} \bar{\hat{\boldsymbol{\pi}}}^T\mathbf{M}\hat{\boldsymbol{\pi}}_x dx + \int_{x^{max}}^{\infty} \bar{\hat{\boldsymbol{\pi}}}^T(s\mathbf{I} - \mathbf{Q}^T)\hat{\boldsymbol{\pi}} dx = 0. \tag{1.5.24}$$

Taking real parts and denoting $\frac{1}{2}\mathbf{\Sigma}^2 \geq \sigma_0 > 0$ and $\lambda^0_{\mathbf{Q}} = 0$ to be the eigenvalue of $\mathbf{Q}$ with the largest real part gives

$$\begin{aligned} 0 &= \frac{1}{2}\int_{x^{max}}^{\infty} \bar{\hat{\boldsymbol{\pi}}}_x^T \mathbf{\Sigma}^2 \hat{\boldsymbol{\pi}}_x dx + \mathrm{Re}\int_{x^{max}}^{\infty} \bar{\hat{\boldsymbol{\pi}}}^T \mathbf{M}\hat{\boldsymbol{\pi}}_x dx + \int_{x^{max}}^{\infty} \bar{\hat{\boldsymbol{\pi}}}^T \left(\mathrm{Re}(s)\mathbf{I} - \frac{\mathbf{Q}^T + \mathbf{Q}}{2}\right)\hat{\boldsymbol{\pi}} dx \\ &\geq \sigma_0 \int_{x^{max}}^{\infty} |\hat{\boldsymbol{\pi}}_x|^2 dx - ||\mathbf{M}|| \int_{x^{max}}^{\infty} |\hat{\boldsymbol{\pi}}| \cdot |\hat{\boldsymbol{\pi}}_x| dx + (\mathrm{Re}(s) - \lambda^0_{\mathbf{Q}}) \int_{x^{max}}^{\infty} |\hat{\boldsymbol{\pi}}|^2 dx \\ &\geq \sigma_0 \int_{x^{max}}^{\infty} \left(|\hat{\boldsymbol{\pi}}_x| - \frac{||\mathbf{M}||}{2\sigma_0}|\hat{\boldsymbol{\pi}}|\right)^2 dx + \left(\mathrm{Re}(s) - \frac{||\mathbf{M}||^2}{4\sigma_0^2}\right) \int_{x^{max}}^{\infty} |\hat{\boldsymbol{\pi}}|^2 dx \geq 0, \end{aligned}$$

for all s with $\mathrm{Re}(s) > \frac{||\mathbf{M}||^2}{4\sigma_0^2}$. From this we conclude $\hat{\boldsymbol{\pi}} \equiv 0$ and thus the solution to (1.5.6),(1.5.23) is unique.

b) A general solution of (1.5.18) is given by

(1.5.25)
$$\begin{pmatrix} \boldsymbol{\zeta} \\ \boldsymbol{\eta} \end{pmatrix}(x) = e^{\mathbf{C}x}\tilde{\mathbf{c}} = \mathbf{P}e^{\mathbf{J}x}\mathbf{P}^{-1}\tilde{\mathbf{c}} = \mathbf{P}e^{\mathbf{J}x}\mathbf{c} = \begin{pmatrix} \boldsymbol{\mathcal{P}}_1 & \boldsymbol{\mathcal{P}}_2 \\ \boldsymbol{\mathcal{P}}_3 & \boldsymbol{\mathcal{P}}_4 \end{pmatrix} \begin{pmatrix} e^{\mathbf{J}_1 x} & \mathbf{0} \\ \mathbf{0} & e^{\mathbf{J}_2 x} \end{pmatrix} \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{pmatrix}, \quad x > x^{max}.$$

To fulfil the decaying condition $\hat{\boldsymbol{\pi}}(x) \to \mathbf{0} \quad (x \to \infty)$ the constant $\mathbf{c}_2$ is chosen as zero. Thus the general solution to (1.5.6) is given by $\hat{\boldsymbol{\pi}}(x) = \boldsymbol{\mathcal{P}}_1 e^{\mathbf{J}_1 x}\mathbf{c}_1$. The constant $\mathbf{c}_1$ can be determined by the boundary condition $\hat{\boldsymbol{\pi}}(x^{max}) = \hat{\boldsymbol{\pi}}_{max}$, if the matrix $\boldsymbol{\mathcal{P}}_1$ is regular . Thus, for regular $\boldsymbol{\mathcal{P}}_1$ the existence of a solution to the exterior problem is guaranteed. □

REMARK 1.9. *Standard methods to prove the existence do not work here. Adding an inhomogeneous term to* (1.5.6) *yields an homogeneous boundary condition. Then, by the variation of constant method we derive the integral representation of the solution*

$$(1.5.26)\quad \boldsymbol{\pi}(x) = \boldsymbol{\mathcal{P}}_1 e^{\mathbf{J}_1 x}\left(\mathbf{c}_1 + \int_{x^{max}}^{x} e^{\mathbf{J}_1 y}\mathbf{f}_1 e^{-y} dy\right) - \boldsymbol{\mathcal{P}}_2 e^{\mathbf{J}_2 x}\int_{x}^{\infty} e^{-\mathbf{J}_2 y}\mathbf{f}_2 e^{-y} dy, \quad x > x^{max}$$

with $\mathbf{f}_{1,2} = -\mathbf{P}_{2,4}\left(\mathbf{I} + 2(\mathbf{\Sigma}^2)^{-1}\left(\mathbf{Q}^T - s\mathbf{I} - \mathbf{M}\right)\right)\boldsymbol{\pi}_{max}$.

To apply the well-known technique of the Fredholm alternative, the integral operator in (1.5.26) *must be compact. But the solution operator given by* (1.5.26) *is of convolution type and thus it is not compact on* $X = \left\{f : \text{ continuous on } [0,\infty), \lim_{x\to\infty} f(x) \text{ exists}\right\}$, *since operators of convolution type are compact only on bounded domains* [**GLS90**] *but not on unbounded domains* [**Hac89**].

REMARK 1.10. *The regularity of* $\boldsymbol{\mathcal{P}}_1$ *is not clear in general, but was true for all considered examples.*

The TBC (1.5.21) is still in a Laplace-transformed formulation. The analytical inverse transform cannot be calculated for a general case as this is possible for a scalar parabolic equation [**Ehr01**]. In [**Hag94**] Hagstrom considered a simple 2×2 model problem. The parameters were chosen in such a way, that the characteristic polynomial factorises and the eigenvalues and eigenvectors can be given explicitly. Still these formulations include terms of the form $s^{1/4}$, which Hagstrom approximates before inverse Laplace-transforming to yield local boundary conditions. We will approximate as late as possible and inverse transform numerically.

6. Discretisation

In this section we briefly present a numerical method to solve the transient equation (1.2.26), which will be the basis for constructing the DTBCs in the following section. We use a finite difference discretisation (*θ-method*), which includes as special cases the fully implicit ($\theta = 1$), explicit ($\theta = 0$) and the Crank-Nicolson ($\theta = \frac{1}{2}$) scheme. The coupling term $\mathbf{Q}^T\boldsymbol{\pi}$ is discretised explicitly such that the system of equations, which has to be solved at each time step, can be decoupled to reduce the computational effort. Hence, one solves a separate equation for each state in which the coupling term appears as an inhomogeneity.

For a state s the equation to be discretised reads

(1.6.1)
$$\frac{\partial \pi_s(t,x)}{\partial t} + \frac{\partial}{\partial x}\Big(\pi_s(t,x)\mu_s(x)\Big) = \frac{1}{2}\frac{\partial^2}{\partial x^2}\Big(\pi_s(t,x)\sigma_s^2(x)\Big) + \sum_{l=1}^{S} q_{l,s}\pi_l, \qquad t \geq 0, x \in \Omega.$$

subject to the initial condition (1.2.26c) and the boundary conditions (1.2.26b).

The discretisation is carried out on an equidistant grid with step-size k in time direction and step-size h in space x-direction: $x_j = x^{\min} + jh$, $t_n = nk$ for $n = 0, \ldots, T$ and $j = 0, \ldots, J$. At the grid points the function $\pi_s(t,x)$ is approximated by the discrete function $u_{s,j}^n$ with $u_{s,j}^n \approx \pi_s(t_n, x_j)$ and $\mathbf{u}_j^n = (u_{s,j}^n)$ is a column vector. The coupling term is discretised explicitly by using the second order extrapolation from the two 'old' time levels n and $(n-1)$

$$\Big(\pi((n+\theta)k, x^{\min}+jh)\mathbf{Q}\Big)_s \approx (1+\theta)\left((u_{0,j}^n, \ldots, u_{S,j}^n)\mathbf{Q}\right)_s - \theta\left((u_{0,j}^{n-1}, \ldots, u_{S,j}^{n-1})\mathbf{Q}\right)_s. \tag{1.6.2}$$

In that way the coupling term is not involved in the implicit part and therefore the equations can be solved separately. We abbreviate the r.h.s. of (1.6.2) by $(\mathbf{Q}^T\mathbf{u})_{s,j}^{n+\theta}$. In the following the index s for the discrete state will be omitted and an arbitrary state will be used.

The second derivative is discretised with the standard central difference quotient. For the first order spatial derivatives an *upwind scheme* is used (cf. [**Str89**]). Upwind means that forward and backward differences are used, weighed with the *upwind parameter* ρ and $(1-\rho)$, respectively. The parameter ρ, $0 \le \rho \le 1$, depends on the sign of the convection parameter μ. This is motivated by the fact that if e.g. $\mu(x) > 0$, the mass moves to the right and the backward difference is more appropriate in describing this motion and thus yields a stable scheme.

With the abbreviation $u_{s,j}^{n+\theta} = (1-\theta)u_{s,j}^{n} + \theta u_{s,j}^{n+1}$ the discrete scheme for state s is

$$\begin{aligned}(1.6.3)\quad & \frac{u_{s,j}^{n+1} - u_{s,j}^{n}}{k} + (1-\rho_j)\frac{\mu_{s,j}u_{s,j}^{n+\theta} - \mu_{s,j-1}u_{s,j-1}^{n+\theta}}{h} + \rho_j \frac{\mu_{s,j+1}u_{s,j+1}^{n+\theta} - \mu_{s,j}u_{s,j}^{n+\theta}}{h} \\ & = \frac{1}{2}\frac{\sigma_{s,j-1}^2 u_{s,j-1}^{n+\theta} - 2\sigma_{s,j}^2 u_{s,j}^{n+\theta} + \sigma_{s,j+1}^2 u_{s,j+1}^{n+\theta}}{h^2} + \sum_{l=1}^{S}\left((1+\theta)\,u_{l,j}^{n} - \theta u_{l,j}^{n-1}\right) q_{s,l}.\end{aligned}$$

A basic property of the parabolic differential equation (1.2.26) is the maximum principle

$$(1.6.4)\qquad \sup_{x\in\Omega}\sum_{s=1}^{S}\pi_s(t,x) \le \sup_{x\in\Omega}\sum_{s=1}^{S}\pi_s(t',x), \qquad \text{if } t > t'.$$

The scheme (1.6.3) will have a similar property and satisfies a *discrete maximum principle*, if the following conditions hold [**Zis98**]:

$$(1.6.5)\qquad \begin{aligned} 0 \le\ \rho_j\ &< \ \frac{\sigma_j^2}{2h\,\mu_j} \qquad \text{for} \quad \mu_j > 0 \quad \text{and} \\ 1 \ge\ \rho_j\ &> 1 - \frac{\sigma_j^2}{2h|\mu_j|} \qquad \text{for} \quad \mu_j < 0. \end{aligned}$$

On the other hand the upwind scheme induces artificial diffusion. Therefore we choose ρ_j as close to $1/2$ as (1.6.5) allows, in order to minimise this artificial influence.

The transient solution of a second order FSPN is then numerically approximated by solving the system of linear equations

$$\begin{aligned}(1.6.6)\quad & u_{j-1}^{n+1}(-\theta r a_{j-1}) + u_j^{n+1}(1+\theta r(a_j+b_j)) + u_{j+1}^{n+1}(-\theta r b_{j+1}) \\ & = u_{j-1}^{n}(1-\theta)ra_{j-1} + u_j^n\left(1-(1-\theta)r(a_j+b_j)\right) + u_{j+1}^n(1-\theta)rb_{j+1} + (\mathbf{Q}^T\mathbf{u})_j^{n+\theta}, \\ & \qquad\qquad n = 0,\dots,T-1, \quad j = 1,\dots,N-1,\end{aligned}$$

with the abbreviations $a_j = h(1-\rho_j)\mu_j + \frac{1}{2}\sigma_j^2$, $b_j = -h\rho_j\mu_j + \frac{1}{2}\sigma_j^2$ and $r = \frac{k}{h^2}$.

7. Discrete boundary conditions

The considered BCs are derived on a discrete level. This can be important, if solutions of the discrete scheme should have the same qualitative properties as solutions of the continuous equation, such as satisfying a normalisation condition. In the next section we will derive *discrete reflecting boundary conditions*, that conserve the probability mass. Usually the analytic boundary condition is discretised in an ad-hoc way. Thus, an additional normalisation step is necessary. The use of our discrete reflecting BCs makes this normalisation step obsolete and avoids thereby not only extra errors and effort but is also a more adequate way to calculate the numerical solution.

For the TBCs it will be even more important to construct the BC on a completely discrete level, because the discretisation of the continuous BC can destroy the stability of the underlying difference scheme and induce artificial reflections.

7.1. The discrete reflecting boundary condition. The discrete reflecting BCs are a consistent discretisation of the BC (1.2.26b) which were derived by postulating that the *normalisation condition*

$$\sum_{s=1}^{S} \int_{\Omega} \pi_s(t,x)\,dx = 1 \tag{1.7.1}$$

should hold on the discrete level, i.e.

$$h \sum_{s=1}^{S} \sum_{j=0}^{J} u_{s,j}^n = 1, \qquad n = 0, \dots, T. \tag{1.7.2}$$

The summation of (1.6.6) over all interior points $j = 1, \dots, J-1$ and states $s = 1, \dots, S$ finally yields after applying (1.7.2) the reflecting boundary conditions at the left $x^{\min} = x_0$ and right boundary $x^{\max} = x_J$

$$u_0^{n+1}(1 + \theta r a_0) - u_1^{n+1}\theta r b_1 = u_0^n\,[1 - (1-\theta) r a_0] + u_1^n (1-\theta) r b_1, \tag{1.7.3a}$$

$$u_J^{n+1}(1 + \theta r b_J) - u_{J-1}^{n+1}\theta r a_{J-1} = u_J^n\,[1 - (1-\theta) r b_J] + u_{J-1}^n (1-\theta) r a_{J-1}, \tag{1.7.3b}$$

which are consistent with the analytic formulation, i.e. (1.7.3a) is consistent with (1.2.26b) with order $O(h, k^2)$.

For each time step the following system of linear equations has been obtained

$$\mathbf{L}\mathbf{u}^{n+1} = \mathbf{R}\mathbf{u}^n + (\mathbf{Q}^T\mathbf{u})^{n+\theta}, \tag{1.7.4}$$

with the matrices

$$\mathbf{L} = \begin{pmatrix} 1+r^+a_0 & -r^+b_1 & 0 & \cdots & 0 \\ -r^+a_0 & 1+r^+(a_1+b_1) & -r^+b_2 & & \\ \ddots & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & 0 & -r^+a_{J-2} & 1+r^+(a_{J-1}+b_{J-1}) & -r^+b_J \\ & & 0 & -r^+a_{J-1} & 1+r^+b_J \end{pmatrix} \tag{1.7.5}$$

and

$$\mathbf{R} = \begin{pmatrix} 1-r^-a_0 & r^-b_1 & 0 & \cdots & 0 \\ r^-a_0 & 1-r^-(a_1+b_1) & r^-b_2 & & \\ \ddots & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & 0 & r^-a_{J-2} & 1-r^-(a_{J-1}+b_{J-1}) & r^-b_J \\ & & 0 & r^-a_{J-1} & 1-r^-b_J \end{pmatrix} \tag{1.7.6}$$

and the abbreviations are

(1.7.7)

$$a_j = h(1-\rho_j)\mu_j + \frac{1}{2}\sigma_j^2, \qquad b_j = -h\rho_j\mu_j + \frac{1}{2}\sigma_j^2, \qquad r = \frac{k}{h^2}, \qquad r^+ = \theta r, \qquad r^- = (1-\theta)r.$$

This is the system of difference equations, that has to be solved, if the domain Ω is bounded. If Ω is unbounded for $x \to \infty$, then we have to introduce an artificial boundary at some point x_J. At this point we prescribe transparent boundary conditions.

7.2. The discrete transparent boundary condition. To derive the discrete transparent BC we consider the discrete difference equation on the exterior domain. Performing a $\mathcal{Z}$-transformation gives a system of ordinary difference equations, which for constant coefficients can be solved explicitly. An inverse $\mathcal{Z}$-transformation yields the discrete transparent boundary condition as a discrete convolution, which is non-local in time. We will present two approaches, which differ in the way to solve the system of ordinary difference equations: the first uses the usual power ansatz (*ansatz method*), the second reduces the problem to a system of first order difference equations (*reduction of order method*). Both methods yield DTBCs, that are quite different in appearance. While the second involves just the boundary point and its direct neighbour, the first is supported on the number

of points equal to the size of the system. We will illustrate later, that both boundary conditions are equivalent.

7.2.1. *The ansatz method.* In the *ansatz method* we solve the $\mathcal{Z}$-transformed difference equation using the standard power ansatz for difference equations with constant coefficients. Therefore, we consider the original discretisation scheme for some point $j \geq J$ outside of the computational domain, where all coefficients are constant

$$\begin{aligned} r\left(u_{s,j}^{n+1}-u_{s,j}^{n}\right) = \; & \frac{\theta}{2}\left[\sigma_s^2 u_{s,j+1}^{n+1} - 2\sigma_s^2 u_{s,j}^{n+1} + \sigma_s^2 u_{s,j-1}^{n+1}\right] + \frac{1-\theta}{2}\left[\sigma_s^2 u_{s,j+1}^{n} - 2\sigma_s^2 u_{s,j}^{n} + \sigma_s^2 u_{s,j-1}^{n}\right] \\ & -h\Big\{\theta\left[\rho_s\left(\mu_s u_{s,j+1}^{n+1} - \mu_s u_{s,j}^{n+1}\right) + (1-\rho_s)\left(\mu_s u_{s,j}^{n+1} - \mu_s u_{s,j-1}^{n+1}\right)\right] \\ & \qquad +(1-\theta)\left[\rho_s\left(\mu_s u_{s,j+1}^{n} - \mu_s u_{s,j}^{n}\right) + (1-\rho_s)\left(\mu_s u_{s,j}^{n} - \mu_s u_{s,j-1}^{n}\right)\right]\Big\} \\ & +h^2 \sum_{l=1}^{S} Q_{l,s}\left[(1+\theta)u_{l,j}^{n} - \theta u_{l,j}^{n-1}\right]. \end{aligned}$$

Using the $\mathcal{Z}$-transformation, which is defined as

$$\mathcal{Z}\left\{\boldsymbol{u}_j^n\right\} = \hat{\mathbf{u}}_j(z) := \sum_{n=0}^{\infty} \boldsymbol{u}_j^n \, z^{-n}, \quad z \in \mathbb{C}, \quad |z| > 1, \tag{1.7.8}$$

we obtain in matrix notation

$$\begin{aligned} r(z-1)\hat{\mathbf{u}}_j(z) \; = \; & \frac{1}{2}\boldsymbol{\Sigma}^2\left(\hat{\mathbf{u}}_{j+1}(z) - 2\hat{\mathbf{u}}_j(z) + \hat{\mathbf{u}}_{j-1}(z)\right)(\theta z + 1 - \theta) \\ & -h\left\{(\mathbf{I}-\mathbf{R})\mathbf{M}(\hat{\mathbf{u}}_j(z) - \hat{\mathbf{u}}_{j-1}(z)) + \mathbf{R}\mathbf{M}(\hat{\mathbf{u}}_{j+1}(z) - \hat{\mathbf{u}}_j(z))\right\}(\theta z + 1 - \theta) \\ & +h^2\left[1 + \theta - \theta\frac{1}{z}\right]\mathbf{Q}^T\hat{\mathbf{u}}_j(z). \end{aligned}$$

The shifted sequences additionally yield the terms $-z\mathbf{u}_j^0$ on the l.h.s. and an equivalent one on the r.h.s. for the first and second differences. We assume, that the function at the boundary $\mathbf{u}_j^0$ for $j \geq \bar{J}$ vanishes, where $\bar{J}$ is the innermost point involved in the construction of the DTBCs. Strategies to overcome this constraint were presented in [**EA01**]. Sorting the last equation according to $\hat{\mathbf{u}}_{j-1}(z)$, $\hat{\mathbf{u}}_j(z)$ and $\hat{\mathbf{u}}_{j+1}(z)$ this linear system reads

$$\mathbf{0} = \boldsymbol{\mathcal{M}}^{+}\hat{\mathbf{u}}_{j+1}(z) + \boldsymbol{\mathcal{M}}^{0}\hat{\mathbf{u}}_j(z) + \boldsymbol{\mathcal{M}}^{-}\hat{\mathbf{u}}_{j-1}(z) + h^2 m_Q \mathbf{Q}^T\hat{\mathbf{u}}_j(z), \tag{1.7.9}$$

with the abbreviations

$$\boldsymbol{\mathcal{M}}^{+} = \frac{1}{2}\boldsymbol{\Sigma}^2 - h\mathbf{R}\mathbf{M}, \qquad \boldsymbol{\mathcal{M}}^{0}(z) = -\boldsymbol{\Sigma}^2 + h(2\mathbf{R}-\mathbf{I})\mathbf{M} - r\frac{z-1}{\theta z + 1 - \theta}\mathbf{I},$$
$$\boldsymbol{\mathcal{M}}^{-} = \frac{1}{2}\boldsymbol{\Sigma}^2 + h(\mathbf{I}-\mathbf{R})\mathbf{M}, \qquad m_Q = \frac{1+\theta-\theta\frac{1}{z}}{\theta z + 1 - \theta}.$$

This system has solutions of the following form (power ansatz)

$$\hat{\mathbf{u}}_j(z) = \alpha^j(z)\mathbf{k}(z), \tag{1.7.10}$$

which decay for $j \to \infty$ if $\mid \alpha(z) \mid < 1$. For simplicity of notation we will frequently omit the z-dependency in the following. Using (1.7.10) in (1.7.9) yields

$$\alpha^{j-1}\left(\alpha^2\boldsymbol{\mathcal{M}}^{+} + \alpha\left(\boldsymbol{\mathcal{M}}^{0} + h^2 m_Q \mathbf{Q}^T\right) + \boldsymbol{\mathcal{M}}^{-}\right)\mathbf{k} = \mathbf{0}, \tag{1.7.11}$$

which has a non-trivial solution $\mathbf{k}$ if and only if the determinant of the system matrix equals zero. This condition leads to a polynomial in α of degree $2S$. In the case $S = 2$ there exists an explicit formula for the zeros of a polynomial with complex coefficients. For $S \geq 3$ a numerical algorithm (e.g. a MATLAB routine or the *Jenkins-Traub algorithm* [**JT72**]) is used.

The absolute value of α determines, if a solution of (1.7.11) increases or decays for $j \to \infty$. At the right boundary only decreasing solutions can form the boundary condition. The following theorem therefore investigates the absolute value of $\alpha = \alpha(z)$:

THEOREM 1.11 (Discrete Splitting Theorem). *Of the $2S$ zeros of (1.7.11) S have an absolute value larger than one, and the other S zeros have an absolute value smaller than one, if $|z|$ is sufficiently large.*

A proof will be given at the end of this section.

We compute the S zeros $\varkappa_1, \ldots, \varkappa_S$ with the smallest absolute values $|\varkappa_s| < 1, s = 1, \ldots, S$ numerically. Since not all zeros are necessarily simple, we assume that there are $\mathcal{N}$ different zeros $\{\alpha_1, \ldots, \alpha_{\mathcal{N}}\} \subset \{\varkappa_1, \ldots, \varkappa_S\}$ with the *algebraic multiplicities* $\nu_1, \ldots, \nu_{\mathcal{N}}$ and the *geometric multiplicities* $\gamma_1, \ldots, \gamma_{\mathcal{N}}$ with $S = \sum_{i=1}^{\mathcal{N}} \nu_i$. $\Gamma := \sum_{i=1}^{\mathcal{N}} \gamma_i$ is the number of Jordan blocks in $\mathbf{J}$ (see below). Furthermore we will consider the family $\{a_1, \ldots, a_\Gamma\}$, where the occurrence of each zero is its geometric multiplicity.

Then all decaying solutions of the exterior problem are given by

$$\hat{\mathbf{u}}_j = \mathbf{P}\mathbf{J}^j\mathbf{P}^{-1}\tilde{\mathbf{c}} =: \mathbf{P}\mathbf{J}^j\mathbf{c}, \tag{1.7.12}$$

where $\mathbf{P} \in \mathbb{C}^{S\times S}$ holds the eigenvectors and generalised eigenvectors of the system matrix (1.7.11) and $\mathbf{J}$ is the corresponding Jordan form with arbitrary constants $\mathbf{c} \in \mathbb{C}^S$ (cf. [**Ela96**]). Let J be the index of the right boundary point and $\hat{\mathbf{u}}_J$ the $\mathcal{Z}$-transform of the approximated function in each discrete state at the boundary. For the boundary condition a relation between $\mathbf{u}_J$ and its neighbours is needed. Consequently, we are looking for a representation of the kind

$$\hat{\mathbf{u}}_J = \sum_{m=1}^{S} \hat{\ell}_m \hat{\mathbf{u}}_{J-m}, \tag{1.7.13}$$

where the $\hat{\ell}_m = \hat{\ell}_m(z) \in \mathbb{C}$ for $m = 1, \ldots, S$ must be determined. Inserting (1.7.12) for $\hat{\mathbf{u}}_J$ and $\hat{\mathbf{u}}_{J-m}$ in (1.7.13) yields

$$\mathbf{P}\mathbf{J}^J\mathbf{c} = \sum_{m=1}^{S} \hat{\ell}_m \mathbf{P}\mathbf{J}^{J-m}\mathbf{c} \tag{1.7.14}$$

or for each Jordan block $\mathbf{J}_p$, $p = 1, \ldots, \Gamma$

$$\mathbf{J}_p^J\mathbf{c} = \sum_{m=1}^{S} \hat{\ell}_m \mathbf{J}_p^{J-m}\mathbf{c} \tag{1.7.15}$$

for arbitrary constants $\mathbf{c}$. The power of a Jordan block has the special upper triangular form

$$\mathbf{J}_p^n = \begin{pmatrix} a_p^n & \binom{n}{1} a_p^{n-1} & \binom{n}{2} a_p^{n-2} & \cdots & \binom{n}{\gamma_p-1} a_p^{n-\gamma_p+1} \\ & a_p^n & \binom{n}{1} a_p^{n-1} & \cdots & \binom{n}{\gamma_p-2} a_p^{n-\gamma_p+2} \\ & & \ddots & \ddots & \vdots \\ & & & a_p^n & \binom{n}{1} a_p^{n-1} \\ & & & & a_p^n \end{pmatrix}. \tag{1.7.16}$$

Therefore, backward substitution gives for each row of equation (1.7.15) one of the conditions

$$\binom{J}{k} a_p^{J-k} = \sum_{m=1}^{S} \hat{\ell}_m \binom{J-m}{k} a_p^{J-m-k}, \qquad \text{for } k = 0, \ldots, \gamma_p - 1, \tag{1.7.17}$$

because $\mathbf{c}$ is arbitrary; or equivalently

$$J^k a_p^S = \sum_{m=1}^{S} \hat{\ell}_m (J-m)^k a_p^{S-m}, \qquad \text{for } k = 0, \ldots, \gamma_p - 1, \quad p = 1, \ldots, \Gamma. \tag{1.7.18}$$

The equivalence of (1.7.17) and (1.7.18) can be seen easily by multiplying (1.7.17) with $k!\ \alpha_p^{S-J+k}$. This yields identical polynomials in J and $J-m$ on either side. Now successively inserting this equation for smaller k gives the equivalence for the highest powers.

Equation (1.7.18) gives S equations for the S unknown quantities $\hat{\ell}_m$. These equations are linearly independent if and only if the the geometric multiplicity of every eigenvalue is one. For any $i = 1, \ldots, \mathcal{N}$ with $\gamma_i > 1$ the associated Jordan blocks give equivalent conditions.

Therefore we define the following matrices and vectors for $i = 1, \ldots, \mathcal{N}$:

$$\left(\boldsymbol{\mathcal{A}}_i^{(S)}\right)_{p,q} := a_i^{S-q}\,(J-q)^{p-1}, \quad \boldsymbol{\mathcal{A}}_i^{(S)} \in M(\gamma_i \times S), \tag{1.7.19}$$

$$\left(\mathbf{b}_i^S\right)_p := J^{p-1} a_i^S, \qquad p = 1, \ldots, \gamma_i \tag{1.7.20}$$

and

$$\boldsymbol{\mathcal{A}}^{(S)} := \begin{pmatrix} \boldsymbol{\mathcal{A}}_1^{(S)} \\ \vdots \\ \boldsymbol{\mathcal{A}}_{\mathcal{N}}^{(S)} \end{pmatrix} \in M(S \times S), \qquad \mathbf{b}^S = \begin{pmatrix} \mathbf{b}_1^S \\ \vdots \\ \mathbf{b}_{\mathcal{N}}^S \end{pmatrix}, \qquad \hat{\boldsymbol{\ell}}^S := \begin{pmatrix} \hat{\ell}_1^{(S)} \\ \vdots \\ \hat{\ell}_S^{(S)} \end{pmatrix}. \tag{1.7.21}$$

The upper index S is introduced to allow a succeeding proof by induction over S. The equation to determine $\hat{\ell}_1^{(S)}, \ldots, \hat{\ell}_S^{(S)}$ now reads

$$\boldsymbol{\mathcal{A}}^{(S)} \hat{\boldsymbol{\ell}}^S = \mathbf{b}^S. \tag{1.7.22}$$

If all S zeros of $\boldsymbol{\mathcal{A}}^{(S)}$ are different, $\boldsymbol{\mathcal{A}}^{(S)}$ is the well-known *Vandermonde matrix* and the system has the special structure

$$\begin{pmatrix} \alpha_1^{S-1} & \alpha_1^{S-2} & \ldots & \alpha_1 & 1 \\ \alpha_2^{S-1} & \alpha_2^{S-2} & \ldots & \alpha_2 & 1 \\ \vdots & & & & \\ \alpha_S^{S-1} & \alpha_S^{S-2} & \ldots & \alpha_S & 1 \end{pmatrix} \begin{pmatrix} \hat{\ell}_1 \\ \hat{\ell}_2 \\ \vdots \\ \hat{\ell}_S \end{pmatrix} = \begin{pmatrix} \alpha_1^S \\ \alpha_2^S \\ \vdots \\ \alpha_S^S \end{pmatrix}. \tag{1.7.23}$$

The following lemma shows the regularity of the matrix $\boldsymbol{\mathcal{A}}^{(S)}$:

LEMMA 1.12. *The Matrix $\boldsymbol{\mathcal{A}}^{(S)}$ is regular if $\gamma_i = 1$ for all $i = 1, \ldots, \mathcal{N}$.*

PROOF. For the special case of a Vandermonde matrix (i.e. $\nu_i = 1$ for $i = 1, \ldots, S$), this is obvious, since

$$\det(\boldsymbol{\mathcal{A}}^{(S)}) = \prod_{\substack{i,j=1 \\ i>j}}^{S} (\alpha_i - \alpha_j) \tag{1.7.24}$$

and due to $\gamma_i = 1$ all zeros are simple. In the general case, Gauss elimination shows the regularity. E.g. Gauss elimination yields in each sub-matrix for $p \geq q$:$(\boldsymbol{\mathcal{A}}_i^{(S)})_{p,q} = \alpha_k^{S-q} \binom{q-1}{p-1}$. If we enlarge this sub-matrix by another sub-matrix for a single zero $\alpha_k, k \neq i$ the Gauss elimination of the matrix $\begin{pmatrix} \boldsymbol{\mathcal{A}}_i^{(S)} \\ \boldsymbol{\mathcal{A}}_k^{(S)} \end{pmatrix}$ gives in the last row $(\alpha_i - \alpha_k)^{\nu_i}$. □

If the solution of (1.7.22) is unique, it is given by the following theorem. If $\boldsymbol{\mathcal{A}}^{(S)}$ is not regular and the solution of (1.7.22) is not necessarily unique, we will nevertheless use the following formula:

THEOREM 1.13. (1.7.22) *is solved by*

$$\hat{\ell}_k^{(S)} = (-1)^{k+1} \sum_{1 \leq s_1 < \ldots < s_k \leq S} \varkappa_{s_1} \cdot \ldots \cdot \varkappa_{s_k}, \qquad k = 1, \ldots, S. \tag{1.7.25}$$

These $\hat{\ell}_k^{(S)}$ *obey the recursion formula*

$$\hat{\ell}_k^{(M)} = \hat{\ell}_k^{(M-1)} - \varkappa_M \hat{\ell}_{k-1}^{(M-1)}, \qquad k = 1, \ldots, S$$

$$\text{with } \hat{\ell}_0^{(M)} = -1, \quad \hat{\ell}_m^{(M)} = 0 \text{ if } \begin{cases} m < 0 \text{ or} \\ m > M \end{cases}, \quad M = 1, \ldots, S \tag{1.7.26}$$

and the γ_i*-level recursion*

$$\hat{\ell}_i^{(\gamma_1 + \ldots + \gamma_{\mathcal{N}})} = \sum_{k=0}^{\gamma_{\mathcal{N}}} \binom{\gamma_{\mathcal{N}}}{k} (-1)^k \varkappa_{\mathcal{N}}^k \hat{\ell}_{i-k}^{(\gamma_1 + \ldots + \gamma_{\mathcal{N}-1})}, \qquad i = 1, \ldots, S \tag{1.7.27}$$

The proof of Theorem 1.13 is deferred to the appendix.

The transformed DTBC reads

$$\hat{\mathbf{u}}_J = \sum_{m=1}^{S} \hat{\ell}_m \hat{\mathbf{u}}_{J-m}. \tag{1.7.28}$$

Finally an inverse $\mathcal{Z}$-transformation of (1.7.13) yields

$$\begin{aligned}\mathcal{Z}^{-1}\left\{\hat{\mathbf{u}}_J\right\} = \mathbf{u}_J^n &= \mathcal{Z}^{-1}\left\{\hat{\ell}_1\right\} * \left\{\mathbf{u}_{J-1}^n\right\} + \mathcal{Z}^{-1}\left\{\hat{\ell}_2\right\} * \left\{\mathbf{u}_{J-2}^n\right\} + \ldots + \mathcal{Z}^{-1}\left\{\hat{\ell}_S\right\} * \left\{\mathbf{u}_{J-S}^n\right\} \\ (1.7.29) \qquad &= \{\ell_1\} * \left\{\mathbf{u}_{J-1}^n\right\} + \{\ell_2\} * \left\{\mathbf{u}_{J-2}^n\right\} + \ldots + \{\ell_S\} * \left\{\mathbf{u}_{J-S}^n\right\} \\ &= \sum_{k=1}^{n} \ell_{1,n-k}\mathbf{u}_{J-1}^k + \sum_{k=1}^{n} \ell_{2,n-k}\mathbf{u}_{J-2}^k + \ldots + \sum_{k=1}^{n} \ell_{S,n-k}\mathbf{u}_{J-S}^k.\end{aligned}$$

This representation of $\mathbf{u}_J^n$ is implemented as a boundary condition. We have to compute the inverse $\mathcal{Z}$-transformations $\ell_1, \ldots, \ell_S$ of $\hat{\ell}_1(z), \ldots, \hat{\ell}_S(z)$ numerically. Therefore we use a MATLAB routine or the procedure *ENTCAF* ([**LS71**]), that calculates approximations to a set of normalised Taylor coefficients. Note that the $\mathcal{Z}$-transformation (1.7.8) can be regarded as a Taylor series in $\frac{1}{z}$.

PROOF OF THEOREM 1.11. Here, at the end of this section, we will return to the proof of Thm. 1.11. Therefore, we observe, that for $\mathbf{M} = \mathbf{0}$ the equation (1.7.11) remains invariant under the change of $j \to -j$. This implies, that the number of decaying and increasing solutions (and therefore the number of zeros with absolute value larger and smaller than one) is equal. We will show, that there exists no α with $|\alpha| = 1$. In that case, Thm. 1.11 holds for $\mathbf{M} = \mathbf{0}$, and a continuity argument shows the assertion for $\mathbf{M} \neq \mathbf{0}$: instead of $\mathbf{M}$ consider $\widetilde{\mathbf{M}} = \epsilon\mathbf{M}$. For ϵ increasing from zero to one the zeros of (1.7.11) are continuously depending on ϵ, and since $|\alpha| = 1$ is impossible, S zeros remain inside the unit circle, and S stay outside.

It remains to show, that $|\alpha| = 1$ is impossible for $|z|$ sufficiently large. Therefore, we multiply (1.7.11) with $\bar{\mathbf{k}}^T$ from the left

$$(1.7.30) \qquad \alpha^2\bar{\mathbf{k}}^T\boldsymbol{\mathcal{M}}^+\mathbf{k} + \bar{\mathbf{k}}^T\boldsymbol{\mathcal{M}}^-\mathbf{k} = -\alpha\bar{\mathbf{k}}^T\left(\boldsymbol{\mathcal{M}}^0 + h^2 m_Q \mathbf{Q}^T\right)\mathbf{k}.$$

We observe, that $\boldsymbol{\mathcal{M}}^0 = -(\boldsymbol{\mathcal{M}}^+ + \boldsymbol{\mathcal{M}}^-) + r\frac{z-1}{\theta z+1-\theta}\mathbf{I}$ and by taking absolute values we obtain

$$(1.7.31) \quad \left|\alpha^2\bar{\mathbf{k}}^T\boldsymbol{\mathcal{M}}^+\mathbf{k} + \bar{\mathbf{k}}^T\boldsymbol{\mathcal{M}}^-\mathbf{k}\right| = |\alpha|\left|\bar{\mathbf{k}}^T\left(\boldsymbol{\mathcal{M}}^+ + \boldsymbol{\mathcal{M}}^- + r\frac{z-1}{\theta z+1-\theta}\mathbf{I} - h^2 m_Q\mathbf{Q}^T\right)\mathbf{k}\right|.$$

We assume $|\alpha| = 1$ and use the triangle inequality:

$$(1.7.32) \qquad \bar{\mathbf{k}}^T(\boldsymbol{\mathcal{M}}^+ + \boldsymbol{\mathcal{M}}^-)\mathbf{k} \geq \left|\bar{\mathbf{k}}^T(\boldsymbol{\mathcal{M}}^+ + \boldsymbol{\mathcal{M}}^-)\mathbf{k} + \bar{\mathbf{k}}^T\left(r\frac{z-1}{\theta z+1-\theta}\mathbf{I} - h^2 m_Q\mathbf{Q}^T\right)\mathbf{k}\right|,$$

where $\boldsymbol{\mathcal{M}}^+$ and $\boldsymbol{\mathcal{M}}^-$ are diagonal matrices, which have only positive entries due to the maximum principle (cf. p. 29). Now, we introduce the abbreviation $\beta := \bar{\mathbf{k}}^T(\boldsymbol{\mathcal{M}}^+ + \boldsymbol{\mathcal{M}}^-)\mathbf{k} \in$

$\mathbb{R}^+$ and rewrite (1.7.32)

$$\beta \geq \left| \beta + \bar{\mathbf{k}}^T \left(r \frac{z-1}{\theta z + 1 - \theta} \mathbf{I} - h^2 m_Q \mathbf{Q}^T \right) \mathbf{k} \right|. \tag{1.7.33}$$

But the absolute value of the r.h.s. of (1.7.33) is strictly larger than β, if

$$\mathrm{Re} \left(r \frac{z-1}{\theta z + 1 - \theta} |\mathbf{k}|^2 - h^2 m_Q \bar{\mathbf{k}}^T \mathbf{Q}^T \mathbf{k} \right) > 0. \tag{1.7.34}$$

We now proceed investigating the two z-depending terms: the real part of the first term is

$$\mathrm{Re} \left(r \frac{z-1}{\theta z + 1 - \theta} \right) = \frac{h^2}{k} \frac{1}{\theta} \frac{(2 - \frac{1}{\theta}) \left[|z|^2 - \mathrm{Re}(z) \right] + \left(\frac{1}{\theta} - 1 \right) \left[|z|^2 - 1 \right]}{|z + \frac{1-\theta}{\theta}|^2}, \tag{1.7.35}$$

which is positive for $\frac{1}{2} \leq \theta \leq 1$ and $|z| > 1$. The factor $m_Q(z)$ in the second term is a heart shaped function for $|z| > 1$ and asymptotically behaves like

$$m_Q(z) = \frac{1 + \theta - \frac{\theta}{z}}{\theta z + 1 - \theta} = \frac{1+\theta}{\theta z} + \mathrm{O}(z^{-2}), \qquad z \to \infty. \tag{1.7.36}$$

Therefore the condition (1.7.34) holds for $|z|$ sufficiently large and (1.7.33) leads to a contradiction, i.e. we showed $|\alpha| \neq 1$ for $|z|$ sufficiently large. This finishes the proof of Thm. 1.11. □

7.2.2. *Reduction of order method.* This method to derive the discrete transparent boundary condition also arises when solving the $\mathcal{Z}$-transformed system of ordinary difference equations on the exterior domain. Here, the system is solved by reducing the difference equations to first order. It yields the DTBC in matrix form.

Again we start with the original discrete scheme for some point $j \geq J$ outside of the computational domain, where all coefficients are constant:

$$\begin{aligned} \frac{h^2}{k} \left(\mathbf{u}_j^{n+1} - \mathbf{u}_j^n \right) = \quad & \theta \left[\tfrac{1}{2} \boldsymbol{\Sigma}^2 \triangle^+ \triangle^- - h\mathbf{M} \left\{ \mathbf{R}\Delta^+ + (\mathbf{I} - \mathbf{R})\Delta^- \right\} \right] \mathbf{u}_j^{n+1} \\ & + (1-\theta) \left[\boldsymbol{\Sigma}^2 \Delta^+ \Delta^- - h\mathbf{M} \left\{ \mathbf{R}\Delta^+ + (\mathbf{I} - \mathbf{R})\Delta^- \right\} \right] \mathbf{u}_j^n \\ & + h^2 \mathbf{Q}^T \left[(1+\theta)\mathbf{u}_j^n - \theta \mathbf{u}_j^{n-1} \right]. \end{aligned} \tag{1.7.37}$$

We use again the $\mathcal{Z}$-transformation to get a system of ordinary difference equations

$$\begin{aligned} \frac{h^2}{k} (z-1) \hat{\mathbf{u}}_j \quad = \quad & (\theta z + 1 - \theta) \left[\frac{1}{2} \boldsymbol{\Sigma}^2 \Delta^+ \Delta^- \hat{\mathbf{u}}_j - h\mathbf{M} \left\{ \mathbf{R}\Delta^+ + (\mathbf{I} - \mathbf{R})\Delta^- \right\} \hat{\mathbf{u}}_j \right] \\ & + \left(1 + \theta - \frac{\theta}{z} \right) h^2 \mathbf{Q}^T \hat{\mathbf{u}}_j, \qquad j > J. \end{aligned} \tag{1.7.38}$$

LEMMA 1.14. *If the boundary value problem* (1.7.38) *with the boundary data*

$$\hat{\mathbf{u}}(x_j = x_J) = \hat{\mathbf{u}}_J, \qquad \hat{\mathbf{u}}(x_j = \infty) = \mathbf{0} \tag{1.7.39}$$

has a solution, it is unique for $|z|$ *sufficiently large.*

PROOF. For showing the uniqueness we assume, that there are two solutions $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_2$ of (1.7.38), (1.7.39). The difference $\hat{\mathbf{u}} = \hat{\mathbf{u}}_1 - \hat{\mathbf{u}}_2$ is then a solution of the problem with *homogeneous boundary data*. We multiply (1.7.38) with $\hat{\mathbf{u}}_j^H$ from the left and sum from J to infinity

$$\begin{aligned} -\frac{1}{2}\sum_{j=J}^{\infty}\hat{\mathbf{u}}_j^H\boldsymbol{\Sigma}^2\triangle^+\triangle^-\hat{\mathbf{u}}_j + h\sum_{j=J}^{\infty}\hat{\mathbf{u}}_j^H(\mathbf{I}-\mathbf{R})\mathbf{M}\triangle^-\hat{\mathbf{u}}_j + h\sum_{j=J}^{\infty}\hat{\mathbf{u}}_j^H\mathbf{R}\mathbf{M}\triangle^+\hat{\mathbf{u}}_j \\ + \sum_{j=J}^{\infty}\hat{\mathbf{u}}_j^H\left(\frac{h^2}{k}\frac{z-1}{\theta z+1-\theta}\mathbf{I} - h^2\frac{1+\theta-\frac{\theta}{z}}{\theta z+1-\theta}\mathbf{Q}^T\right)\hat{\mathbf{u}}_j = 0. \end{aligned} \tag{1.7.40}$$

We will abbreviate the first row with X and investigate it further. With the abbreviations from Sec. 7.2.1 it reads

$$X = \sum_{j=J}^{\infty}\hat{\mathbf{u}}_j^H\left(-\boldsymbol{\mathcal{M}}^+\hat{\mathbf{u}}_{j+1} + (\boldsymbol{\mathcal{M}}^+ + \boldsymbol{\mathcal{M}}^-)\hat{\mathbf{u}}_j - \boldsymbol{\mathcal{M}}^-\hat{\mathbf{u}}_{j-1}\right). \tag{1.7.41}$$

Mark that $\boldsymbol{\mathcal{M}}^+$ and $\boldsymbol{\mathcal{M}}^-$ are diagonal matrices with diagonal entries strictly larger than zero. This is due to the discrete maximum principle. We consider

$$\begin{aligned} \mathrm{Re}\left(\sum_{j=J}^{\infty}\hat{\mathbf{u}}_j^H\boldsymbol{\mathcal{M}}^-\hat{\mathbf{u}}_{j-1}\right) &= \mathrm{Re}\left(\sum_{J-1}^{\infty}\hat{\mathbf{u}}_{j+1}^H\boldsymbol{\mathcal{M}}^-\hat{\mathbf{u}}_j\right) \\ &= \mathrm{Re}\left(\sum_{j=J}^{\infty}(\overline{\hat{\mathbf{u}}_j^H\boldsymbol{\mathcal{M}}^-\hat{\mathbf{u}}_{j+1}})\right) = \mathrm{Re}\left(\sum_{j=J}^{\infty}\hat{\mathbf{u}}_j^H\boldsymbol{\mathcal{M}}^-\hat{\mathbf{u}}_{j+1}\right). \end{aligned} \tag{1.7.42}$$

Analogously for the left term we get $\mathrm{Re}\left(\sum_{j=J}^{\infty}\hat{\mathbf{u}}_j^H\boldsymbol{\mathcal{M}}^+\hat{\mathbf{u}}_{j+1}\right) = \mathrm{Re}\left(\sum_{j=J}^{\infty}\hat{\mathbf{u}}_j^H\boldsymbol{\mathcal{M}}^+\hat{\mathbf{u}}_{j-1}\right)$. In this way we can write $\mathrm{Re}(X) = -\mathrm{Re}\left(\sum_{j=J}^{\infty}\hat{\mathbf{u}}_j^H(\boldsymbol{\mathcal{M}}^+ + \boldsymbol{\mathcal{M}}^-)\triangle^+\hat{\mathbf{u}}_j\right)$ as well as $\mathrm{Re}(X) = \mathrm{Re}\left(\sum_{j=J}^{\infty}\hat{\mathbf{u}}_j^H(\boldsymbol{\mathcal{M}}^+ + \boldsymbol{\mathcal{M}}^-)\triangle^-\hat{\mathbf{u}}_j\right)$ and which together with summation by parts yields

$$\begin{aligned} \mathrm{Re}(X) &= -\frac{1}{2}\mathrm{Re}\left(\sum_{j=J}^{\infty}\hat{\mathbf{u}}_j^H(\boldsymbol{\mathcal{M}}^+ + \boldsymbol{\mathcal{M}}^-)\triangle^+\triangle^-\hat{\mathbf{u}}_j\right) \\ &= \frac{1}{2}\mathrm{Re}\left(\sum_{j=J}^{\infty}\triangle^-\hat{\mathbf{u}}_j^H(\boldsymbol{\mathcal{M}}^+ + \boldsymbol{\mathcal{M}}^-)\triangle^-\hat{\mathbf{u}}_j\right) \geq \frac{1}{2}\mathcal{M}_0^{\pm}\sum_{j=J}^{\infty}|\hat{\mathbf{u}}_j|^2, \end{aligned} \tag{1.7.43}$$

if $\mathcal{M}_0^{\pm}$ is the smallest off-diagonal entry of $\boldsymbol{\mathcal{M}}^+ + \boldsymbol{\mathcal{M}}^-$. Taking real parts in (1.7.40) yields

(1.7.44)
$$0 \geq \frac{1}{2}\mathcal{M}_0^{\pm} \sum_{j=J}^{\infty} |\hat{\mathbf{u}}_j|^2 - \sum_{j=J}^{\infty} \mathrm{Re}\left(\hat{\mathbf{u}}_j^H \frac{1+\theta-\frac{\theta}{z}}{\theta z+1-\theta} \mathbf{Q}^T \hat{\mathbf{u}}_j \right) + \mathrm{Re}\left(\frac{z-1}{\theta z+1-\theta} \right) \sum_{j=J}^{\infty} |\hat{\mathbf{u}}_j|^2 \geq 0,$$

if $|z|$ is sufficiently large. Because then $\frac{1+\theta-\frac{\theta}{z}}{\theta z+1-\theta}$ is small and the difference of the first two terms is positive. But $\mathrm{Re}\left(\frac{z-1}{\theta z+1-\theta}\right) > 0$ for all $|z| > 1$. Thus, we conclude $\hat{\mathbf{u}}_j = 0$ for $j > J$, which is a contradiction to the assumption. □

REMARK 1.15. *Analogously to the continuous problem the existence of a solution is guaranteed by the regularity of the $S \times S$ principal submatrix of the matrix of right eigenvectors (cf. Lem. 1.8 and Rem. 1.9), which holds for all considered examples.*

To solve the system (1.7.38) we reduce it to a system of first order by introducing $\hat{\mathbf{v}}_j = \Delta^- \hat{\mathbf{u}}_j$ and abbreviate $z_1 = \theta z + 1 - \theta$, $z_2 = 1 + \theta - \frac{\theta}{z}$, $m_Q = \frac{z_2}{z_1}$ and $t = \frac{z-1}{z_1}$:

(1.7.45)
$$\begin{pmatrix} h\mathbf{R}\mathbf{M} & -\frac{1}{2}\boldsymbol{\Sigma}^2 \\ -\mathbf{I} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \Delta^+ \hat{\mathbf{u}}_j \\ \Delta^+ \hat{\mathbf{v}}_j \end{pmatrix} = \begin{pmatrix} h^2 m_Q \mathbf{Q}^T - \frac{th^2}{k}\mathbf{I} & -(\mathbf{I}-\mathbf{R})\mathbf{M} \\ \mathbf{0} & -\mathbf{E} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{u}}_j \\ \hat{\mathbf{v}}_j \end{pmatrix},$$

where $\mathbf{E}$ is the *shift operator* $\mathbf{E}w_j = w_{j+1}$, which is eliminated by

(1.7.46)
$$\begin{pmatrix} h\mathbf{R}\mathbf{M} & -\frac{1}{2}\boldsymbol{\Sigma}^2 \\ -\mathbf{I} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{u}}_{j+1} - \hat{\mathbf{u}}_j \\ \hat{\mathbf{v}}_{j+1} - \hat{\mathbf{v}}_j \end{pmatrix} = \begin{pmatrix} h^2 m_Q \mathbf{Q}^T - \frac{th^2}{k} I & -h(\mathbf{I}-\mathbf{R})\mathbf{M} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{u}}_j \\ \hat{\mathbf{v}}_j \end{pmatrix}.$$

We define the matrices $\mathbf{A} = \begin{pmatrix} h\mathbf{R}\mathbf{M} & -\frac{1}{2}\boldsymbol{\Sigma}^2 \\ -\mathbf{I} & \mathbf{I} \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} h^2 m_Q \mathbf{Q}^T - \frac{th^2}{k}\mathbf{I} & -h(\mathbf{I}-\mathbf{R})\mathbf{M} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix}$ and have for regular $\mathbf{A}$

(1.7.47)
$$\begin{pmatrix} \Delta^+ \hat{\mathbf{u}}_j \\ \Delta^+ \hat{\mathbf{v}}_j \end{pmatrix} = \mathbf{A}^{-1}\mathbf{B} \begin{pmatrix} \hat{\mathbf{u}}_j \\ \hat{\mathbf{v}}_j \end{pmatrix}, \qquad j \geq J.$$

For the DTBC we have to reproduce a solution, that decays for $j \to \infty$. Therefore, we will now discuss, under which conditions a solution of the general equation

(1.7.48)
$$\Delta^+ \mathbf{w}_j = \mathbf{W}\mathbf{w}_j, \qquad j \geq 0$$

decays for $x \to \infty$. If $J_{\mathbf{W}}$ is the Jordan form of $\mathbf{W} + \mathbf{I}$ and $\mathbf{P}_{\mathbf{W}}$ holds the corresponding (possibly generalised) eigenvectors, we have

(1.7.49)
$$\mathbf{w}_{j+1} = (\mathbf{W}+\mathbf{I})\mathbf{w}_j = \mathbf{P}_{\mathbf{W}}\mathbf{J}_{\mathbf{W}}\mathbf{P}_{\mathbf{W}}^{-1}\mathbf{w}_j = \mathbf{P}_{\mathbf{W}}\mathbf{J}_{\mathbf{W}}^{j+1}\mathbf{P}_{\mathbf{W}}^{-1}\mathbf{w}_0.$$

And thus $\mathbf{P}_{\mathbf{W}}^{-1}\mathbf{w}_j$ decays in any matrix norm (see also [**QSS00**])

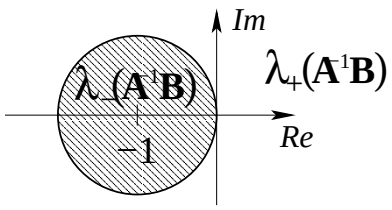
$$||\mathbf{P}_{\mathbf{W}}^{-1}\mathbf{w}_{j+1}|| \leq ||\mathbf{J}_{\mathbf{W}}|| \cdot ||\mathbf{P}_{\mathbf{W}}^{-1}\mathbf{w}_j||, \tag{1.7.50}$$

if $||\mathbf{J}_{\mathbf{W}}|| < 1$, where $||\cdot||$ denotes the associated matrix norm. Since $\mathbf{P}_{\mathbf{W}}^{-1}$ is regular, this implies that also $\mathbf{w}_j$ decays.

Equation (1.7.47) is equivalent to

$$\begin{pmatrix} \hat{\mathbf{u}}_{j+1} \\ \hat{\mathbf{v}}_{j+1} \end{pmatrix} = (\mathbf{A}^{-1}\mathbf{B} + \mathbf{I}) \begin{pmatrix} \hat{\mathbf{u}}_j \\ \hat{\mathbf{v}}_j \end{pmatrix} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1} \begin{pmatrix} \hat{\mathbf{u}}_j \\ \hat{\mathbf{v}}_j \end{pmatrix}, \qquad j \geq J, \tag{1.7.51}$$

when J is the Jordan form of $\mathbf{A}^{-1}\mathbf{B}+\mathbf{I}$ and $\mathbf{P}$ holds the corresponding (possibly generalised) eigenvectors. Thus, any solution of (1.7.51) involving eigenvalues $\lambda(\mathbf{A}^{-1}\mathbf{B})$ of $\mathbf{A}^{-1}\mathbf{B}$ with $|\lambda_-(\mathbf{A}^{-1}\mathbf{B}) + 1| < 1$ is decreasing for $j \to \infty$ in at least one vector norm (cf. Satz 6.9.2 [**SB90**]). The others with $|\lambda_+(\mathbf{A}^{-1}\mathbf{B}) + 1| > 1$ are increasing. This splitting of the eigenvalues was already shown in Thm. 1.11 for the ansatz method.



Thus, we may again split the Jordan form $\mathbf{J} = \left(\begin{smallmatrix} \mathbf{J}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{smallmatrix}\right)$, $\mathbf{J}_1$ holding the eigenvalues λ_- and $\mathbf{J}_2$ holding λ_+. With the matrix of eigenvectors $\mathbf{P}^{-1} = \left(\begin{smallmatrix} \mathbf{P}_1 & \mathbf{P}_2 \\ \mathbf{P}_3 & \mathbf{P}_4 \end{smallmatrix}\right)$ the equation

$$\begin{aligned} \mathbf{P}^{-1} \begin{pmatrix} \hat{\mathbf{u}}_{j+1} \\ \hat{\mathbf{v}}_{j+1} \end{pmatrix} &= \mathbf{P}^{-1}(\mathbf{A}^{-1}\mathbf{B} + \mathbf{I}) \begin{pmatrix} \hat{\mathbf{u}}_j \\ \hat{\mathbf{v}}_j \end{pmatrix} = \mathbf{P}^{-1}\mathbf{P} \begin{pmatrix} \mathbf{J}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{pmatrix} \begin{pmatrix} \mathbf{P}_1 & \mathbf{P}_2 \\ \mathbf{P}_3 & \mathbf{P}_4 \end{pmatrix} \begin{pmatrix} \hat{\mathbf{u}}_j \\ \hat{\mathbf{v}}_j \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{J}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{pmatrix} \begin{pmatrix} \mathbf{P}_1\hat{\mathbf{u}}_j + \mathbf{P}_2\hat{\mathbf{v}}_j \\ \mathbf{P}_3\hat{\mathbf{u}}_j + \mathbf{P}_4\hat{\mathbf{v}}_j \end{pmatrix}, \qquad j \geq J \end{aligned} \tag{1.7.52}$$

holds and the transformed discrete transparent boundary condition reads

$$\mathbf{P}_3\hat{\mathbf{u}}_J + \mathbf{P}_4\hat{\mathbf{v}}_J = \mathbf{0}. \tag{1.7.53}$$

If the matrix $\mathbf{P}_4$ is regular, then the boundary condition can be formulated in Dirichlet-to-Neumann form

$$\Delta^-\hat{\mathbf{u}}_J = \widehat{\mathbf{\Phi}}\hat{\mathbf{u}}_J, \tag{1.7.54}$$

where $\widehat{\mathbf{\Phi}} = -\mathbf{P}_4^{-1}\mathbf{P}_3$. The regularity of $\mathbf{P}_4$ is not clear, but it is true for every example we considered.

Thus, we have for each component s

$$\hat{u}_{s,J} - \hat{u}_{s,J-1} = \sum_{l=1}^{S} \widehat{\phi}_{s,l} \hat{u}_{l,J} \,, \tag{1.7.55}$$

or after an inverse $\mathcal{Z}$-transformation

$$\begin{aligned} u_{s,J}^{n+1} - u_{s,J-1}^{n+1} &= \sum_{l=1}^{S} \sum_{k=0}^{n+1} \phi_{s,l}^{n+1-k} u_{l,J}^{k} \\ &= \sum_{l=1}^{S} \left(\sum_{k=1}^{n} \phi_{s,l}^{n+1-k} u_{l,J}^{k} + \phi_{s,l}^{0} u_{l,J}^{n+1} \right) \\ u_{s,J}^{n+1} - u_{s,J-1}^{n+1} - \sum_{l=1}^{S} \phi_{s,l}^{0} u_{l,J}^{n+1} &= \sum_{l=1}^{S} \sum_{k=1}^{n} \phi_{s,l}^{n+1-k} u_{l,J}^{k}. \end{aligned} \tag{1.7.56}$$

Since we were able to confine the influence of the coupling term $\mathbf{Q}^T$ to time steps smaller or equal to n, the system of equations could be solved for each state s independently. The boundary condition in the form of (1.7.56) would destroy this advantage. Fortunately, we have $\phi_{s,l}^0 = 0$ for $l \neq s$. This can be seen, if the *initial value theorem* (cf. [**Doe67**])

$$\phi_{l,s}^{0} = \lim_{z \to \infty} \widehat{\phi}_{l,s}(z) \tag{1.7.57}$$

is regarded. $\widehat{\mathbf{\Phi}}$ is defined as the product of blocks of the matrix, that contains the eigenvectors of $\mathbf{A}^{-1}\mathbf{B} + \mathbf{I}$. If we change limit and eigenvalue calculation, we have

$$m_Q(z) = \frac{1 + \theta - \frac{\theta}{z}}{\theta z + 1 - \theta} = \frac{1+\theta}{\theta z + 1 - \theta} - \frac{\theta}{\theta z^2 + (1-\theta) z}, \qquad \lim_{z \to \infty} m_Q(z) = 0, \tag{1.7.58}$$

$$t(z) = \frac{z-1}{\theta z + 1 - \theta} = \frac{1}{\theta} \left(1 - \frac{1}{\theta z + 1 - \theta} \right), \qquad \lim_{z \to \infty} t(z) = \frac{1}{\theta} \tag{1.7.59}$$

and $\mathbf{A}^{-1}\mathbf{B} + \mathbf{I}$ is block-diagonal with a matrix of eigenvectors which is also block-diagonal. Of course, the product of two such blocks is diagonal.

Finally, we obtain for each state s the *discrete transparent boundary condition*

$$\left(1 - \phi_{s,s}^{0}\right) u_{s,J}^{n+1} - u_{s,J-1}^{n+1} = \sum_{l=1}^{S} \sum_{k=1}^{n} \phi_{s,l}^{n+1-k} u_{l,J}^{k}, \qquad s = 1, \ldots, S. \tag{1.7.60}$$

7.2.3. *The relation between the two discrete transparent boundary conditions.* In the preceeding sections we derived two different methods to formulate the DTBC. The construction and also the formulation differ significantly: The ansatz method calculates S scalar values and formulates the DTBC on S neighbouring grid points at the boundary. We have to presume, that the initial function vanishes at all boundary points involved. The reduction of order method needs to calculate a complete $S \times S$-matrix, but formulates the DTBC by a matrix equation for two boundary points. Only at these two boundary points, we assume the initial function to vanish.

In this section we will show, that the two different DTBCs (1.7.29) and (1.7.55) can be transformed into each other. Therefore we will show, that the DTBC of the *ansatz method* (1.7.29) is equal to

$$\left(\sum_{i=1}^{S} \hat{\ell}_i \mathbf{Y}_i\right) \hat{\mathbf{u}}_{J-1} = \left(\mathbf{I} - \sum_{i=2}^{S} \hat{\ell}_i \mathbf{X}_i\right) \hat{\mathbf{u}}_J \tag{1.7.61}$$

for appropriate definitions of the matrices $\mathbf{X}_i$ and $\mathbf{Y}_i$. Comparing this formulation with the DTBC of the *reduction of order method* (1.7.55) in matrix notation

$$\hat{\mathbf{u}}_{J-1} = (\mathbf{I} - \widehat{\mathbf{\Phi}})\hat{\mathbf{u}}_J, \tag{1.7.62}$$

yields a unique relation between the occurring matrices, if $\sum_{i=1}^{S} \hat{\ell}_i \mathbf{Y}_i$ is regular.

To show the equality of (1.7.29) and (1.7.61), we consider the $\mathcal{Z}$-transformed difference scheme in the form (1.7.9) or

$$\hat{\mathbf{u}}_{J-1} = \mathbf{B}^0 \hat{\mathbf{u}}_J + \mathbf{B}^+ \hat{\mathbf{u}}_{J+1} \tag{1.7.63}$$

with $\mathbf{B}^+ = -\left(\boldsymbol{\mathcal{M}}^-\right)^{-1} \boldsymbol{\mathcal{M}}^+$ and $\mathbf{B}^0 = -\left(\boldsymbol{\mathcal{M}}^-\right)^{-1} \left(\boldsymbol{\mathcal{M}}^0 + h^2 m_Q \mathbf{Q}^T\right)$, which exist, since all diagonal entries in $\boldsymbol{\mathcal{M}}^-$ are positive due to the discrete maximum principle and thus the diagonal matrix $\boldsymbol{\mathcal{M}}^-$ is regular. For simplicity we will only consider x-independent parameters $\mathbf{M}$, $\mathbf{\Sigma}^2$ and $\mathbf{Q}^T$ here. Otherwise the matrices $\boldsymbol{\mathcal{M}}^-, \boldsymbol{\mathcal{M}}^0$ and $\boldsymbol{\mathcal{M}}^+$ have a more complex structure, but still remain diagonal.

As a preparation we will now prove the following lemma:

LEMMA 1.16. *For the difference scheme* (1.7.63) *the following recursion holds:*

$$\hat{\mathbf{u}}_{J-m} = \mathbf{Y}_m \hat{\mathbf{u}}_{J-1} + \mathbf{X}_m \hat{\mathbf{u}}_J, \qquad \text{for } m \geq 2, \tag{1.7.64}$$

with the matrices

$$\mathbf{X}_1 = I, \qquad \mathbf{X}_2 = \mathbf{B}^+, \qquad \mathbf{X}_{m+1} = \mathbf{B}^0\mathbf{X}_m + \mathbf{B}^+\mathbf{X}_{m-1}, \qquad m \geq 2, \tag{1.7.65}$$

$$\mathbf{Y}_1 = I, \qquad \mathbf{Y}_2 = \mathbf{B}^0, \qquad \mathbf{Y}_{m+1} = \mathbf{B}^0\mathbf{Y}_m + \mathbf{B}^+\mathbf{Y}_{m-1}, \qquad m \geq 2. \tag{1.7.66}$$

PROOF. By induction:

For the induction basis $m = 2$ we have just the interior scheme $\hat{\mathbf{u}}_{J-2} = \mathbf{B}^0\hat{\mathbf{u}}_{J-1} + \mathbf{B}^+\hat{\mathbf{u}}_J$.

For the induction step $m \to m+1$ we consider the interior scheme and then use the induction hypothesis for m and $m-1$

$$\begin{aligned}
\hat{\mathbf{u}}_{J-(m+1)} &= \mathbf{B}^0\hat{\mathbf{u}}_{J-m} + \mathbf{B}^+\hat{\mathbf{u}}_{J-(m-1)} \\
&= \mathbf{B}^0(\mathbf{Y}_m\hat{\mathbf{u}}_{J-1} + \mathbf{X}_m\hat{\mathbf{u}}_J) + \mathbf{B}^+(\mathbf{Y}_{m-1}\hat{\mathbf{u}}_{J-1} + \mathbf{X}_{m-1}\hat{\mathbf{u}}_j) \\
&= (\mathbf{B}^0\mathbf{Y}_m + \mathbf{B}^+\mathbf{Y}_{m-1})\hat{\mathbf{u}}_{J-1} + (\mathbf{B}^0\mathbf{X}_m + \mathbf{B}^+\mathbf{X}_{m-1})\hat{\mathbf{u}}_J \\
&= \mathbf{Y}_{m+1}\hat{\mathbf{u}}_{J-1} + \mathbf{X}_{m+1}\hat{\mathbf{u}}_J.
\end{aligned}$$

□

After this preparation we are able to show the following lemma:

LEMMA 1.17. *The discrete TBC* (1.7.29) *of the ansatz method* $\hat{\mathbf{u}}_J = \sum_{i=1}^S \hat{\ell}_i\hat{\mathbf{u}}_{J-i}$ *is for all* $S \geq 1$ *an equivalent formulation of*

$$\hat{\mathbf{u}}_J = \sum_{i=1}^S \hat{\ell}_i\mathbf{Y}_i\hat{\mathbf{u}}_{J-1} + \sum_{i=2}^S \hat{\ell}_i\mathbf{X}_i\hat{\mathbf{u}}_J. \tag{1.7.67}$$

PROOF. By induction we will prove the equality of the r.h.s. and thus formulate the induction hypothesis

$$\sum_{i=1}^S \hat{\ell}_i\hat{\mathbf{u}}_{J-i} = \sum_{i=1}^S \hat{\ell}_i\mathbf{Y}_i\hat{\mathbf{u}}_{J-1} + \sum_{i=2}^S \hat{\ell}_i\mathbf{X}_i\hat{\mathbf{u}}_J. \tag{1.7.68}$$

For the induction basis we consider $S = 1$:

$$\hat{\ell}_1\hat{\mathbf{u}}_{J-1} = \hat{\ell}_1\mathbf{Y}_1\hat{\mathbf{u}}_{J-1}, \tag{1.7.69}$$

which is true, since $\mathbf{Y}_1 = \mathbf{I}$.

For the induction step $S \to S+1$ we use the induction hypothesis (IH) and Lem. 1.16:

$$\begin{aligned}
\sum_{i=1}^{S+1} \hat{\ell}_i \hat{\mathbf{u}}_{J-i} &= \sum_{i=1}^{S} \hat{\ell}_i \hat{\mathbf{u}}_{J-i} + \hat{\ell}_{S+1} \hat{\mathbf{u}}_{J-(S+1)} \\
&\overset{IH}{=} \sum_{i=1}^{S} \hat{\ell}_i \mathbf{Y}_i \hat{\mathbf{u}}_{J-1} + \sum_{i=2}^{S} \hat{\ell}_i \mathbf{X}_i \hat{\mathbf{u}}_J + \hat{\ell}_{S+1} \hat{\mathbf{u}}_{J-(S+1)} \\
&\overset{lem.1.16}{=} \sum_{i=1}^{S} \hat{\ell}_i \mathbf{Y}_i \hat{\mathbf{u}}_{J-1} + \sum_{i=2}^{S} \hat{\ell}_i \mathbf{X}_i \hat{\mathbf{u}}_J + \hat{\ell}_{S+1} (\mathbf{Y}_{S+1} \hat{\mathbf{u}}_{J-1} + \mathbf{X}_{S+1} \hat{\mathbf{u}}_J) \\
&= \sum_{i=1}^{S+1} \hat{\ell}_i \mathbf{Y}_i \hat{\mathbf{u}}_{J-1} + \sum_{i=2}^{S+1} \hat{\ell}_i \mathbf{X}_i \hat{\mathbf{u}}_J .
\end{aligned}$$

□

We showed, that the boundary condition of the ansatz method, that involves S boundary points is equivalent to a formulation based only on two boundary points. It is equal to the BC of the reduction of order method if the l.h.s. matrix of (1.7.61) is regular.

7.2.4. *Summed convolution coefficients.* For $S = 1$, i.e. a scalar problem, the DTBC (1.7.60) is equivalent to the DTBC for parabolic equations derived by Ehrhardt in [**Ehr01**]. He showed, that the coefficients $\phi_{s,l}$ are an oscillating series, and hence considered summed coefficients to avoid subtractive cancellation. An additional advantage of the summed coefficients is, that they are faster decreasing (in the scalar case like $\mathrm{O}(n^{-3/2})$) and the sum over k can be truncated to obtain local in time "simplified DTBCs". In Sec. 8 "Numerical examples" we will see, that also our coefficients for systems are oscillating. Therefore, we construct summed coefficients based on (1.7.56) by multiplying the boundary condition for two successive time steps by θ and $1-\theta$ respectively and adding the resulting equations. The DTBC with summed coefficients then results in

$$\theta(1-\psi_{s,s}^0) u_{s,J}^{n+1} - \theta u_{s,J-1}^{n+1} = \sum_{l=1}^{S} \sum_{k=0}^{n} \psi_{s,l}^{n+1-k} u_{l,J}^k + (1-\theta)(u_{s,J-1}^n - u_{s,J}^n) \tag{1.7.70}$$

with $\Psi_{s,l}^n := \theta\phi_{s,l}^n + (1-\theta)\phi_{s,l}^{n-1}$ for $n = 1, 2, \ldots$ and $\Psi_{s,l}^0 = \theta\phi_{s,l}^0$. The summed coefficients $\Psi_{s,l}^n$ are obtained by an inverse $\mathcal{Z}$-transformation:

$$\Psi_{s,l}^n := \mathcal{Z}^{-1}\Big\{ \frac{\theta z + 1 - \theta}{z} \, \widehat{\phi}_{s,l}(z) \Big\}. \tag{1.7.71}$$

These summed convolution coefficients yield very good numerical results as we will present in Chap. 8. Nevertheless the convolution with the "normal" as well as with the

summed coefficients yields nonlocal in time boundary conditions, which cause a linearly increasing numerical effort each time step, i.e. the total effort is a quadratic function of time. It depends on the application, if such high accuracy is necessary. If not, the high numerical effort makes the usage of the exact discrete TBCs rather unattractive. In the next section we will therefore present the derivation of approximated convolution coefficients, which allow for a fast evaluation of the discrete approximate convolution.

7.2.5. *Approximated convolution coefficients.* In order to reduce the numerical effort due to the boundary condition below that of the overall scheme, it is necessary to make some kind of approximation. We decided to approximate as late as possible and focused on the convolution coefficients. A simple approach is to cut off the convolution after a constant number of summands. But this method yields bad results. In this section we will use the approach of Arnold, Ehrhardt and Sofronov [**AES03**] to approximate the coefficients $\psi^n_{s,l}$ by the *sum of exponentials* ansatz. Afterwards we explain how these approximated convolution coefficients $\widetilde{\psi}^n_{s,l}$ enable us to fast evaluate the discrete convolution.

We use for each $s, \tau = 1, \dots, S$ the following ansatz, which uses a sum of exponentials

$$\psi^n_{s,\tau} \approx \widetilde{\psi}^n_{s,\tau} := \begin{cases} \psi^n_{s,\tau}, & n = 0, \dots, \nu - 1 \\ \sum\limits_{l=1}^{L_{s,\tau}} g_{s,\tau,l}\, h^{-n}_{s,\tau,l}, & n = \nu, \nu+1, \dots, \end{cases} \tag{1.7.72}$$

where $L_{s,\tau} \in \mathbb{N}^{S\times S}$ and $\nu \geq 0$ are fixed numbers. The approximation quality of this ansatz depends on $L_{s,\tau}$, ν and the sets $\{g_{s,\tau,l}\}$ and $\{h_{s,\tau,l}\}$ for all $s, \tau = 1, \dots, S$.

In the following we present a method to calculate these sets for given $L_{s,\tau}$ and ν. We consider the formal power series

$$f_{s,\tau}(x) := \psi^\nu_{s,\tau} + \psi^{\nu+1}_{s,\tau} x + \psi^{\nu+2}_{s,\tau} x^2 + \dots, \qquad \text{for } |x| \leq 1. \tag{1.7.73}$$

If there exists the Padé approximation of (1.7.73)

$$\tilde{f}_{s,\tau}(x) := \frac{n^{(L_{s,\tau}-1)}_{s,\tau}(x)}{d^{(L_{s,\tau})}_{s,\tau}(x)} \tag{1.7.74}$$

(where the numerator and the denominator are polynomials of degree $L_{s,\tau} - 1$ and $L_{s,\tau}$ respectively), then its Taylor series

$$\tilde{f}_{s,\tau}(x) = \widetilde{\psi}^\nu_{s,\tau} + \widetilde{\psi}^{\nu+1}_{s,\tau} x + \widetilde{\psi}^{\nu+2}_{s,\tau} x^2 + \dots \tag{1.7.75}$$

satisfies the conditions

$$\widetilde{\psi}^n_{s,\tau} = \psi^n_{s,\tau}, \qquad \text{for } n = \nu, \nu+1, \ldots, 2L_{s,\tau}+\nu-1 \tag{1.7.76}$$

according to the definition of the Padé approximation rule.

We now consider, how to compute the coefficient sets $\{g_{s,\tau,l}\}$ and $\{h_{s,\tau,l}\}$.

THEOREM 1.18 ([**AES03**], Theorem 3.1.). *Let $d^{L_{s,\tau}}_{s,\tau}$ have $L_{s,\tau}$ simple roots $h_{s,\tau,l}$ with $|h_{s,\tau,l}| > 1$, $l = 1, \ldots, L_{s,\tau}$. Then*

$$\widetilde{\psi}^n_{s,\tau} = \sum_{l=1}^{L_{s,\tau}} g_{s,\tau,l}\, h^{-n}_{s,\tau,l}, \qquad n = \nu, \nu+1, \ldots, \tag{1.7.77}$$

where

$$g_{s,\tau,l} := -\frac{n^{(L_{s,\tau}-1)}_{s,\tau}(h_{s,\tau,l})}{\big(d^{(L_{s,\tau})}_{s,\tau}\big)'(h_{s,\tau,l})}\, h^{\nu-1}_{s,\tau,l} \neq 0, \qquad l = 1, \ldots, L_{s,\tau}. \tag{1.7.78}$$

REMARK 1.19. *The asymptotic decay of the $\widetilde{\psi}^n_{s,\tau}$ is exponential. This is due to the sum of exponentials ansatz* (1.7.72) *and the assumption $|h_{s,\tau,l}| > 1$, $l = 1, \ldots, L_{s,\tau}$.*

Now we describe the fast evaluation of the discrete approximate convolution. The convolution

$$C^{(n+1)}_{s,\tau}(u) := \sum_{k=1}^{n+1-\nu} \widetilde{\psi}^{n+1-k}_{s,\tau}\, u^k_{\tau,k}, \tag{1.7.79}$$

with

$$\widetilde{\psi}^n_{s,\tau} := \sum_{l=1}^{L_{s,\tau}} g_{s,\tau,l}\, h^{-n}_{s,\tau,l}, \qquad n = \nu, \nu+1, \ldots \tag{1.7.80}$$

can be calculated efficiently by a simple recurrence formula:

THEOREM 1.20 ([**AES03**], Theorem 4.1.).

$$C^{(n+1)}_{s,\tau}(u) = \sum_{l=1}^{L_{s,\tau}} C^{(n+1)}_{s,\tau,l}(u) \tag{1.7.81}$$

with

$$\begin{aligned} C^{(n+1)}_{s,\tau,l}(u) &= h^{-1}_{s,\tau,l} C^{(n)}_{s,\tau,l} + g_{s,\tau,l} h^{-\nu}_{s,\tau,l} u^{n+1-\nu}_{\tau,J}, \qquad n = \nu, \nu+1, \ldots \\ C^{(\nu)}_{s,\tau,l}(u) &\equiv 0 \end{aligned} \tag{1.7.82}$$

Finally, we summarise the above method to evaluate approximate DTBCs:

Step 1: For each s,τ choose $L_{s,\tau}$ and ν and calculate the exact convolution coefficients $\psi^n_{s,\tau}$ for $n = 0,\dots,2L_{s,\tau}+\nu-1$.

Step 2: For each s,τ use the Padé approximation (1.7.74) for the series (1.7.75) with $\widetilde{\psi}^n_{s,\tau} = \psi^n_{s,\tau}$, for $n = \nu,\nu+1,\dots,2L_{s,\tau}+\nu-1$ to calculate the sets $\{g_{s,\tau,l}\}$ and $\{h_{s,\tau,l}\}$ for all $s,\tau = 1,\dots,S$ according to Theorem 1.18.

Step 3: Implement the recurrence formulas (1.7.81), (1.7.82) to calculate approximate convolutions.

7.3. Stability. To check the stability we have to assert that the l^1-norm of the numerical solution in the computational domain is bounded from above. Therefore, we define

$$||\mathbf{u}^n||_{l^1(0,d)} := \sum_{j=0}^{d-1}\sum_{s=1}^{S}|u^n_{s,j}| = \sum_{j=0}^{d-1}\sum_{s=1}^{S}u^n_{s,j}, \qquad d>1. \tag{1.7.83}$$

Thus, inserting the discrete equation (1.6.3) and remembering that the row sum of $\mathbf{Q}$ equals zero, the half-space problem satisfies

$$\begin{aligned}
\frac{h^2}{k}\triangle^+||\mathbf{u}^n||_{l^1(0,\infty)} &= \frac{h^2}{k}\sum_{s=1}^{S}(u^{n+1}_{s,0}-u^n_{s,0})\\
&\quad+\sum_{j=1}^{\infty}\sum_{s=1}^{S}\left(\frac{1}{2}\triangle^+\triangle^-(\sigma^2_{s,j}\mathbf{u}^{n+\theta}_{s,j}) - h\triangle^+(\mu_{s,j}\rho_{s,j}u^{n+\theta}_{s,j}) + h\triangle^-(\mu_{s,j}(1-\rho_{s,j})u^{n+\theta}_{s,j})\right)\\
&= \sum_{s=1}^{S}\left(\sum_{j=1}^{\infty}T^+_{s,j+1}u^{n+\theta}_{s,j+1} - \sum_{j=1}^{\infty}(T^+_{s,j}+T^+_{s,j})u^{n+\theta}_{s,j} + \sum_{j=1}^{\infty}T^-_{s,j-1}u^{n+\theta}_{s,j-1} + \frac{h^2}{k}(u^{n+1}_{s,0}-u^n_{s,0})\right)\\
&= \sum_{s=1}^{S}\left(-T^+_{s,1}u^{n+\theta}_{s,1} + T^-_{s,0}u^{n+\theta}_{s,0} + \frac{h^2}{k}(u^{n+1}_{s,0}-u^n_{s,0})\right) = 0
\end{aligned} \tag{1.7.84}$$

after an index transformation for the first and third sum over j. The abbreviations are $T^+_{s,j} = \frac{1}{2}\sigma^2_{s,j} - h\rho_{s,j}\mu_{s,j}$ and $T^-_{s,j} = \frac{1}{2}\sigma^2_{s,j} + h(1-\rho_{s,j})\mu_{s,j}$. The last equality is just the reflecting boundary condition (1.7.3a) at $j=0$.

Since with transparent boundary conditions at the right boundary x_J we just cut off a part of the solution of the half-space problem, the l^1-norm on the computational domain is bounded by the l^1-norm of the half-space problem:

$$||\mathbf{u}^n||_{l^1(0,J)} \le ||\mathbf{u}^n||_{l^1(0,\infty)} = ||\mathbf{u}^0||_{l^1(0,\infty)}. \tag{1.7.85}$$

But this is only valid for exact DTBCs. With numerically computed convolution coefficients a calculation analogous to (1.7.84) yields

$$\frac{h^2}{k}\triangle^+||\mathbf{u}^n||_{l^1(0,J)} = \sum_{s=1}^{S}\left(T^+_{s,J}u^+_{s,J}u^{n+\theta}_{s,J} - T^-_{s,J-1}u^{n+\theta}_{s,J-1}\right). \tag{1.7.86}$$

The aim is now, to show that (1.7.86) is non-positive. But using the DTBC does not yield any estimate, because our information about properties of the convolution matrix is to small. It remains the possibility to check the sign of (1.7.86) numerically. This we will do for the numerical example at the end of Sec. 8.

7.4. The numerical inverse $\mathcal{Z}$-transformation. The $\mathcal{Z}$-transformation (or in the analytical part the Laplace-transformation) is the mighty instrument, which enables us to solve occurring equations and to formulate this kind of transparent boundary conditions. In the implementation the numerical inverse $\mathcal{Z}$-transformation proved to be a more subtle problem than every other point including e.g. the calculation of eigenvectors. For that reason, we will investigate it further.

7.4.1. *Performing $\mathcal{Z}$-transformation with Fourier-transformation.* Many mathematical toolboxes contain ordinary transformations, including a (fast) Fourier-transformation as a standard routine. The less common $\mathcal{Z}$-transformation is rarely found. Here, we will present the easy coherence between both. The $\mathcal{Z}$-transform will be denoted by $\mathcal{Z}$, F is the discrete Fourier-transform. On the unit circle holds for the *finite $\mathcal{Z}$-transform* $\mathcal{Z}^N$ for $z = e^{i\varphi}$

$$\mathcal{Z}(f_j) \approx \mathcal{Z}^N(f_j) = \sum_{j=0}^{N} f_j z^{-j} = \sum_{j=0}^{N} f_j e^{-ij\varphi} = F(e^{i\varphi}). \tag{1.7.87}$$

On a circle with radius r holds

$$F(re^{i\varphi}) = \sum_{j=0}^{N} f_j r^{-j} e^{-ij\varphi} = \sum_{j=0}^{N} f_j r^{-j} z^{-j} = \mathcal{Z}^N(f_j r^{-j}) \approx \mathcal{Z}(f_j r^{-j}). \tag{1.7.88}$$

We observe, that applying the inverse $\mathcal{Z}$-transformation not on the unit circle necessitates a rescaling of the n-th convolution coefficient with r^n. For big circles this causes numerical problems.

7.4.2. *The error of the numerical inverse $\mathcal{Z}$-transformation.* In this section we will examine the numerical error caused by the inverse $\mathcal{Z}$-transformation, since it is the crucial point in the numerical implementation. For the transformation we have to choose a radius r

and a number N of points z_k to define the circle on which the transformation is performed. An intelligent choice of these parameters is essential to achieve good results.

The numerical error can be separated in ϵ_{approx} the error in the approximation on a finite number of sampling points and the roundoff error ϵ_{round}. ℓ_n^N denotes the approximation on a circle with N sampling points. A tilde on top of it indicates that the roundoff error is considered.

The $\mathcal{Z}$-transformation of $\{\ell_m\}$ at the sampling points z_k reads

$$\hat{\ell}_k := \hat{\ell}(z_k) = \sum_{m=0}^{\infty} \ell_m z_k^{-m}, \qquad \text{with } z_k = re^{-ik\frac{2\pi}{N}}. \tag{1.7.89}$$

If we assume, that $\hat{\ell}(z)$ is an analytic function for $|z| > R$, then the ℓ_n are just identical with the Laurent coefficients of $\hat{\ell}(z)$ given by

$$\ell_n = \frac{1}{2\pi i} \oint_{S_\rho} \hat{\ell}(z) z^{n-1} dz, \tag{1.7.90}$$

where S_ρ denotes the circle with radius $\rho > R$. If we substitute $z = \rho e^{i\varphi}$, we obtain

$$\ell_n = \frac{\rho^n}{2\pi} \int_0^{2\pi} \hat{\ell}\left(\rho e^{i\varphi}\right) e^{in\varphi} d\varphi. \tag{1.7.91}$$

Defining $M_{\hat{\ell}}^{\rho} = \max\limits_{0\le\varphi\le 2\pi} \left|\hat{\ell}\left(\rho e^{i\varphi}\right)\right|$ gives the estimate

$$|\ell_n| \le \rho^n M_{\hat{\ell}}^{\rho}. \tag{1.7.92}$$

The inverse $\mathcal{Z}$-transformation of $\hat{\ell}$ can be approximated on N discrete sampling points as follows

$$\ell_n^N = \frac{1}{N} r^n \sum_{k=0}^{N-1} \hat{\ell}_k \, e^{ink\frac{2\pi}{N}}, \quad n = 0, \dots N-1. \tag{1.7.93}$$

We insert (1.7.89) in (1.7.93), change the order of summation and use the orthogonality property

$$\begin{aligned}
\ell_n^N &= \frac{1}{N} r^n \sum_{m=0}^{\infty} \ell_m r^{-m} \sum_{k=0}^{N-1} e^{-imk\frac{2\pi}{N}} e^{ink\frac{2\pi}{N}} \\
&= \frac{1}{N} r^n \sum_{m=0}^{\infty} \ell_m r^{-m} \begin{cases} N & \text{, if } m = n + jN \quad , j \in \mathbb{N} \\ 0 & \text{, else} \end{cases} \\
&= r^n \sum_{k=0}^{\infty} \ell_{n+kN}\, r^{-(n+kN)}.
\end{aligned}$$

This gives

$$\ell_n^N - \ell_n = \sum_{k=1}^{\infty} \ell_{n+kN}\, r^{-kN}. \tag{1.7.94}$$

Now, we insert inequality (1.7.92) in (1.7.94) and sum the geometric series, which yields

$$|\ell_n^N - \ell_n| \leq \rho^n M_{\hat{\ell}}^{\rho} \sum_{k=1}^{\infty} \left(\frac{\rho}{r}\right)^{kN} = \rho^n M_{\hat{\ell}}^{\rho} \frac{\left(\frac{\rho}{r}\right)^N}{1 - \left(\frac{\rho}{r}\right)^N} \tag{1.7.95}$$

for $r > \rho > R$.

Similar estimates have been derived in the application of quadrature rules to numerical integration by Lubich, which involve Fourier transformation (cf. [**Lub88**], [**Hen79**]).

The other influential error is the roundoff error, that depends on the machine accuracy ϵ_m and the accuracy ϵ in the numerical computation of $\hat{\ell}_k$. For instance,we will use $\tilde{a} = a(1 + \epsilon_m)$ as the computer representation of an exact value a. The roundoff error of the inverse $\mathcal{Z}$-transformation is calculated from equation (1.7.93). The main part results from the N fold summation of $\hat{\ell}_k$ and the exponential function.

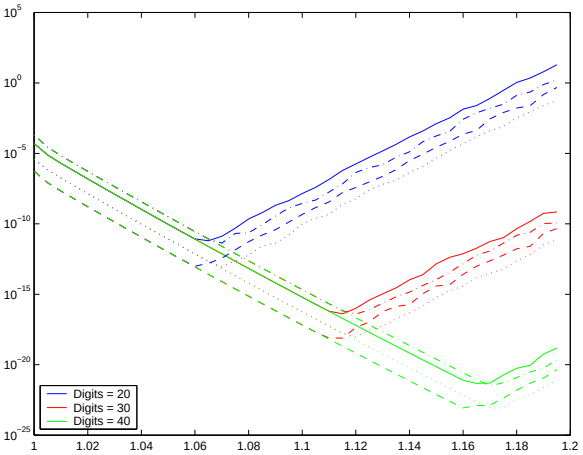
$$\left|\tilde{\ell}_n^N - \ell_n^N\right| \leq r^n \left(CN\epsilon_m + \epsilon\right) M_{\hat{\ell}_k}^r \tag{1.7.96}$$

Together with (1.7.95) the error is bounded by

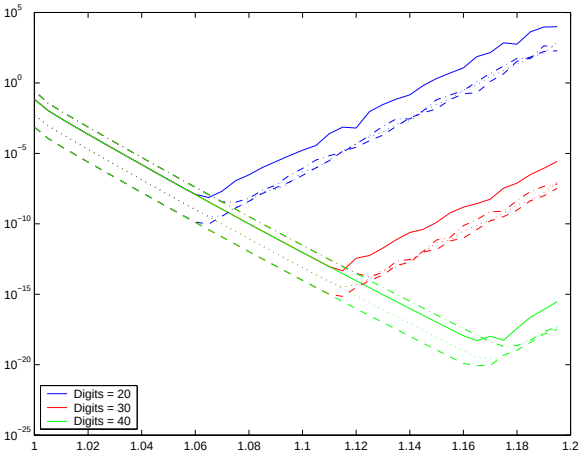
$$|\tilde{\ell}_n^N - \ell_n| \leq \rho^n M_{\hat{\ell}}^{\rho} \frac{\left(\frac{\rho}{r}\right)^N}{1 - \left(\frac{\rho}{r}\right)^N} + r^n \left((N+1)\epsilon_m + \epsilon\right) M_{\hat{\ell}_k}^r + O(\epsilon_m^2 + \epsilon\epsilon_m). \tag{1.7.97}$$

It is possible to show this behaviour of the error roughly in numerical examples. We calculated the series ℓ_n for an arbitrary problem with four states with different accuracy (20, 30 and 40 digits precision) and considered the solution obtained with 50 digits precision

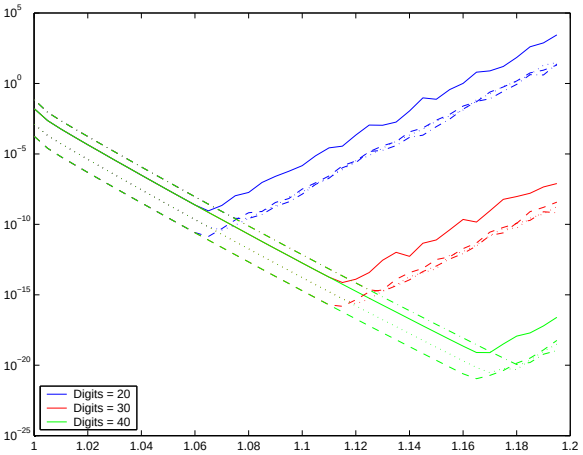
as a reference solution. We used $N = 256$ sampling points on the circle. The Euclidean norm of the error is shown in Fig. 1.6 for each of the 16 entries in the matrix $\boldsymbol{\ell}$. Each plot gives the four elements of one row. For each element the error decreases with growing radius, up to a r_{opt}, after which the roundoff error grows rapidly. Observe, that the y-axis of the plots are in logarithmic scale. The curves for 20, 30 and 40 digits coincide for small values of r up to the radius r_{opt}^{20}, r_{opt}^{30} respectively. This can be seen best in Fig. 1.7 which shows just the first diagonal element of $\boldsymbol{\ell}$.



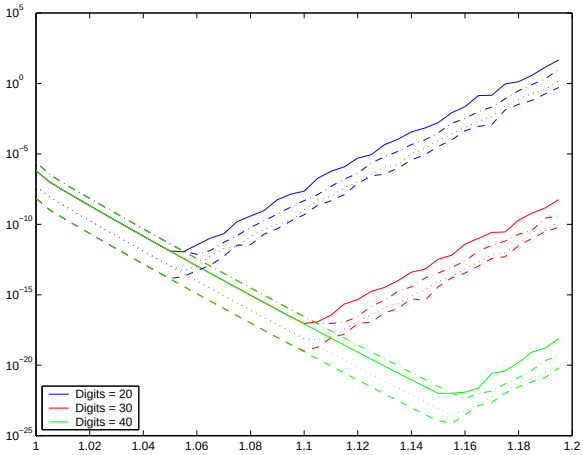
first row



second row



third row



fourth row

Figure 1.6: Error in the 4×4-matrix $\boldsymbol{\ell}$ compared to numerical solution calculated with 50 digits precision and $N = 256$ sampling points for the inverse $\mathcal{Z}$-transformation on radius $r \in [1.00001, 1.2]$

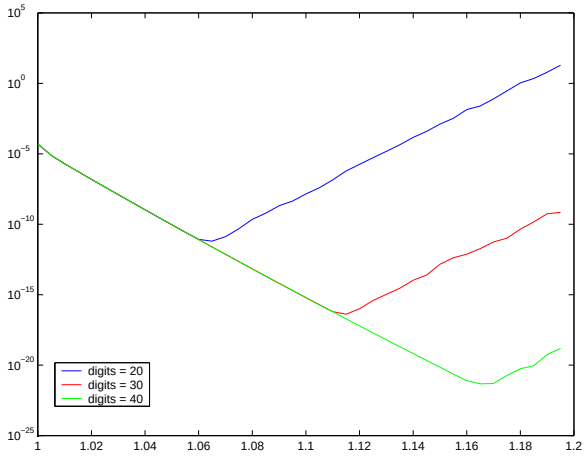


Figure 1.7: Error in one element of the matrix $\boldsymbol{\ell}$

Since the calculation for a system is rather expensive, it is desirable to predict a radius close to r_{opt}. From Fig. 1.6 we can learn, that even for different entries in our matrix, the optimal radius varies. But the dependency on the model parameters seems to be much less than the dependency on the working precision or the number of sampling points. In Fig. 1.6 we have $1.05 \leq r_{opt}^{20} \leq 1.075$ and thus $r = 1.06$ might be a good compromise for a working precision of 20 digits and $N = 256$ sampling points.

All subsequent plots are restricted to but one element of the matrix $\boldsymbol{\ell}$, since we could discern no perceptible qualitative difference between the elements.

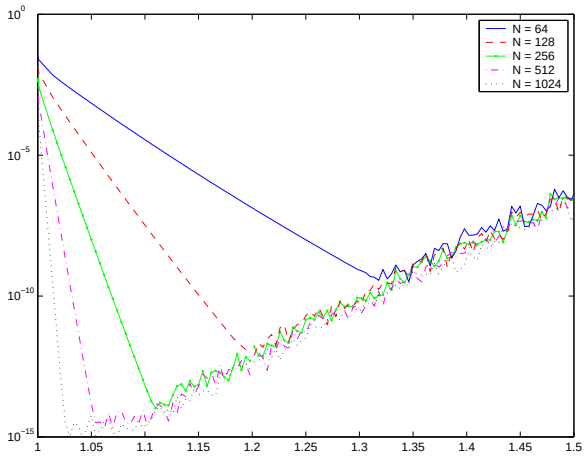


Figure 1.8: Error in one element of the matrix $\boldsymbol{\ell}$ calculated with 20 digits precision depending on the number N of sampling points for the inverse $\mathcal{Z}$-transformation

The preceeding figures showed the influence of the mantissa length on the accuracy of the calculation. Now, we want to show the dependence of the error on the number N of sampling points. Fig. 1.8 shows five error curves with 20 digits precision; one for N= 64, 128, 256, 512 and 1024 respectively. The error norm is summed up to 64. A higher number of sampling points yields a faster decreasing error, r_{opt} becomes smaller and of course the error at r_{opt} becomes less. An influence of N on the round off error is hardly discernable. Comparing the errors at the different N-depending r_{opt}, we notice, that the gain of taking the double number of points gets less with increasing N. Of course the error cannot become less than the precision in the calculation of $\hat{\ell}_n$. Fig. 1.8 shows that $N = 1024$ is absolutely sufficient. Just the need for a larger number of coefficients in time requires more sampling points.

In Fig. 1.9 we compare the error in the Euclidean norm with the error bound (1.7.97), i.e. with the separate bounds for the approximation error and roundoff error. We assumed $\epsilon = 10 \cdot \epsilon_m$ and calculated the maximum of $\hat{\ell}$ on all r for simplicity reasons.

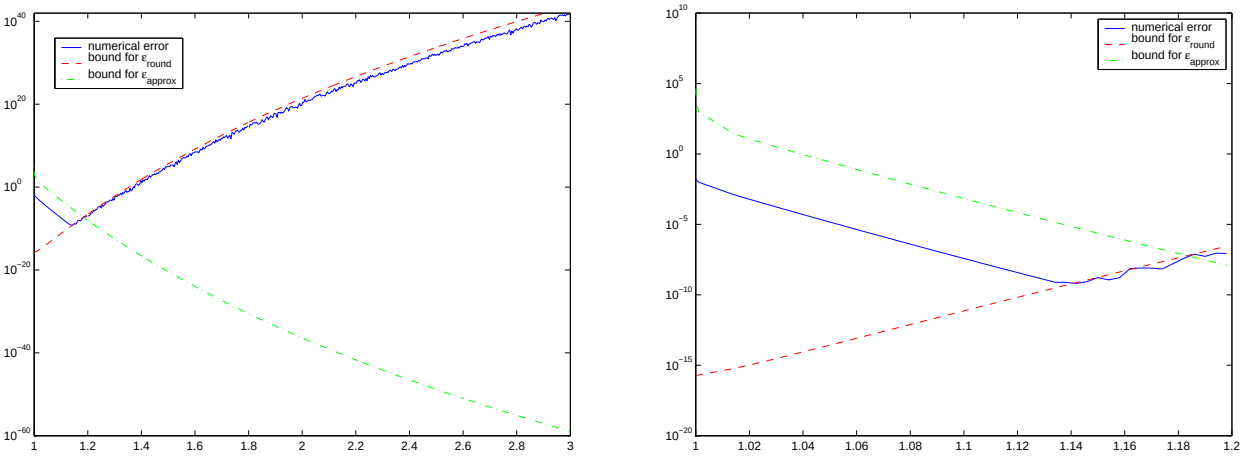


Figure 1.9: Error and error bounds

Estimate (1.7.97) and the plots reveal the reason, why it is so hard to find an appropriate radius r, on which we perform the transformation: the approximation error necessitates a larger radius; but for a bigger radius the roundoff error grows.

8. Numerical examples

In this section we will give some numerical examples to show the quality of the DTBCs. Before we reconsider the queueing system example of Fig. 1.4 in Chap. 2, we will start to give some confidence in the DTBC and especially in the numerical calculation of the convolution coefficients, whose accurate computation is the foundation for good numerical results.

We used the *reduction of order method* in all considered numerical examples for two reasons: first the numerical results for this method are better then those for the *ansatz method*. Second the convolution coefficients of the reduction of order method can be compared directly to those in the scalar case obtained by Ehrhardt [**Ehr01**]. This we will do in the first example.

8.1. Example 1 - the diagonalisable problem. In order to scrutinise the quality of the convolution coefficients, we give an example, in which the parabolic system (1.2.26a) can be transformed to a purely diagonal system

(1.8.1)
$$\frac{\partial}{\partial t}\underline{\boldsymbol{\pi}}(t,x)+\frac{\partial}{\partial x}\left(\underline{\boldsymbol{M}}(x)\underline{\boldsymbol{\pi}}(t,x)\right)=\frac{1}{2}\frac{\partial^2}{\partial x^2}\left(\underline{\boldsymbol{\Sigma}}^2(x)\underline{\boldsymbol{\pi}}(t,x)\right)+\underline{\mathbf{Q}}^T\underline{\boldsymbol{\pi}}(t,x),\qquad x\in\Omega,\quad t\geq 0,$$

where $\mathbf{Q^T}$ is diagonalisable with $\underline{\mathbf{Q}}^T=\mathbf{V}^T\mathbf{Q}^T(\mathbf{V}^T)^{-1}=\mathrm{diag}(\lambda_1,\ldots,\lambda_S)$, $\underline{\boldsymbol{M}}(x)=\boldsymbol{M}(x)=\underline{\mu}\,\mathbf{I}$, $\underline{\boldsymbol{\Sigma}}^2(x)=\boldsymbol{\Sigma}^2(x)=\underline{\sigma^2}\,\mathbf{I}$ and $\underline{\boldsymbol{\pi}}(t,x)=\mathbf{V}^T\boldsymbol{\pi}(t,x)$. Then the convolution coefficients are $\mathbf{\Phi^n}=(\mathbf{V^T})^{-1}\underline{\mathbf{\Phi}}^{\mathbf{n}}\mathbf{V^T}$. In this case we have for each component just a separate scalar equation and the convolution coefficients of the DTBC (1.7.56) can be computed explicitly by

$$\begin{aligned}
\underline{\phi}^0_{i,i} &= 1-\frac{1}{1+\tau_i^+}\left(H_i+\frac{\kappa_i}{K_i}+\frac{1+\sqrt{A_i}}{2\theta K_i}\right)\\
\underline{\phi}^n_{i,i} &= -\frac{(-1)^n}{2\theta(1-\theta)K_i}\left(\frac{1-\theta}{\theta}\right)^n\frac{1}{1+\tau_i^-}-\frac{1}{K_i\sqrt{A_i}}\,\frac{1}{1+\tau_i^-}\cdot\\
&\quad\cdot\left(\frac{A_i}{2\theta}\tilde{P}_n(\upsilon_i)+\frac{C_i}{2(1-\theta)}+\frac{1}{2\theta(1-\theta)^2}\sum_{k=0}^{n-1}\left(\frac{\theta-1}{\theta}\right)^{n-k}\tilde{P}_k(\upsilon_i)\right),\quad n\geq 1
\end{aligned}\tag{1.8.2}$$

(cf. [**Ehr01**]). The abbreviations are

$$K_i=\frac{1}{2}\underline{\sigma}^2_{i,i}\frac{k}{h^2},\qquad H_i=1-(1-2\rho_{i,i})\frac{\underline{\mu}_{i,i}}{\underline{\sigma}^2_{i,i}}h,$$
$$\kappa_i^+=1-\theta k\underline{\mathbf{Q}}_{i,i},\qquad \kappa_i^-=1+(1-\theta)k\underline{\mathbf{Q}}_{i,i},$$

$$\beta_i^+ = -\theta\underline{\mu}_{i,i}\frac{k}{h}, \qquad \beta_i^- = -(1-\theta)\underline{\mu}_{i,i}\frac{k}{h},$$
$$\tau_i^+ = -2\rho_{i,i}\frac{\underline{\mu}_{i,i}}{\underline{\sigma}_{i,i}^2}h, \qquad \tau_i^- = 2(1-\rho_{i,i})\frac{\underline{\mu}_{i,i}}{\underline{\sigma}_{i,i}^2}h,$$
$$A_i = (\kappa_i^+)^2 + 4\theta K_i H_i \kappa_i^+ + (\beta_i^+)^2, \qquad B_i = \kappa_i^+\kappa_i^- - 2K_iH_i[(1-\theta)\kappa_i^+ - \theta\kappa_i^-] - \beta_i^+\beta_i^-,$$
$$C_i = 1 - 4(1-\theta)K_iH_i + (\beta_i^-)^2,$$

and $\tilde{P}_n(\upsilon_i) := \left(\frac{\sqrt{C_i}}{\sqrt{A_i}}\right)^n P_n(\upsilon_i)$ are the "damped" Legendre polynomials with the argument $\upsilon_i = \frac{B_i}{\sqrt{A_i}\sqrt{C_i}}$.

Note that the above formulas just hold in the case of using a θ-discretisation also for the coupling term (instead of the described extrapolation, which leads to separated scalar difference equations), i.e. the factor m_Q in the reduction of order method in (1.7.45) evaluates to $m_Q \equiv 1$.

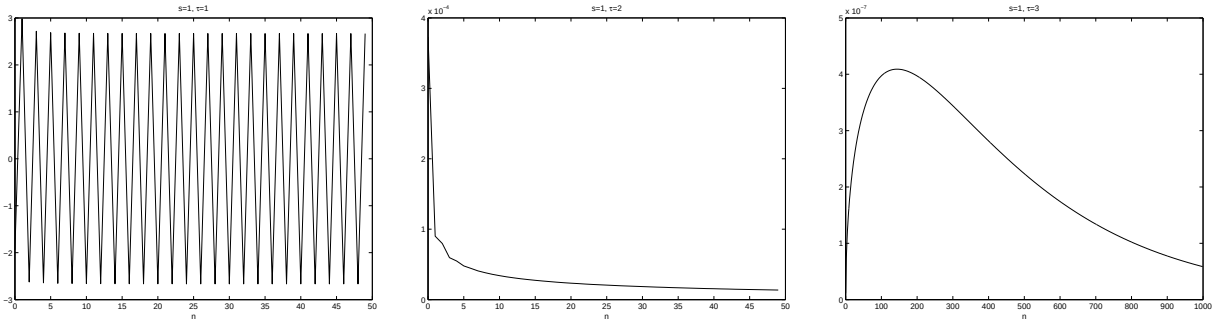


Figure 1.10: Example 1: Exact convolution coefficients $\underline{\phi}_{s,\tau}^n$ for $s = 1, \tau = 1, \ldots, 3$

We will compare the summed form $\underline{\psi}_{s,\tau}^n$ of these exact coefficients to the numerically calculated coefficients $\psi_{s,\tau}^n$ of the reduction of order method. Therefore we consider a simple 3×3 problem with $\mathbf{\Sigma} = 0.9 \cdot \mathbf{I}$, $\mathbf{M} = 3 \cdot \mathbf{I}$ and $\mathbf{Q} = \begin{pmatrix} -1 & 1 & 0 \\ 0.5 & -1 & 0.5 \\ 0 & 1 & -1 \end{pmatrix}$ in the interval $[0, 1]$ and choose the discretisation parameters as $k = 1/1000$, $h = 1/40$ and $\theta = 1/2$ (Crank-Nicolson scheme). The coefficients have been obtained by an inverse transformation on a circle of radius $rcirc = 1.002$ with 2^{13} sampling points. Fig. 1.10 shows the convolution coefficients $\phi_{1,1}^n, \ldots, \phi_{1,3}^n$. We observe the different behaviour of these coefficients: $\phi_{1,1}^n$ oscillates heavily and does not visibly decrease. The other two decrease; $\phi_{1,2}^n$ rapidly, $\phi_{1,3}^n$ after surmounting to a local maximum (observe the different numbers of plotted coefficients $n = 50$ and 1000 respectively as well as the different size). We can classify all convolution coefficients with one of these three types. To which class coefficients belong seems easy to decide:

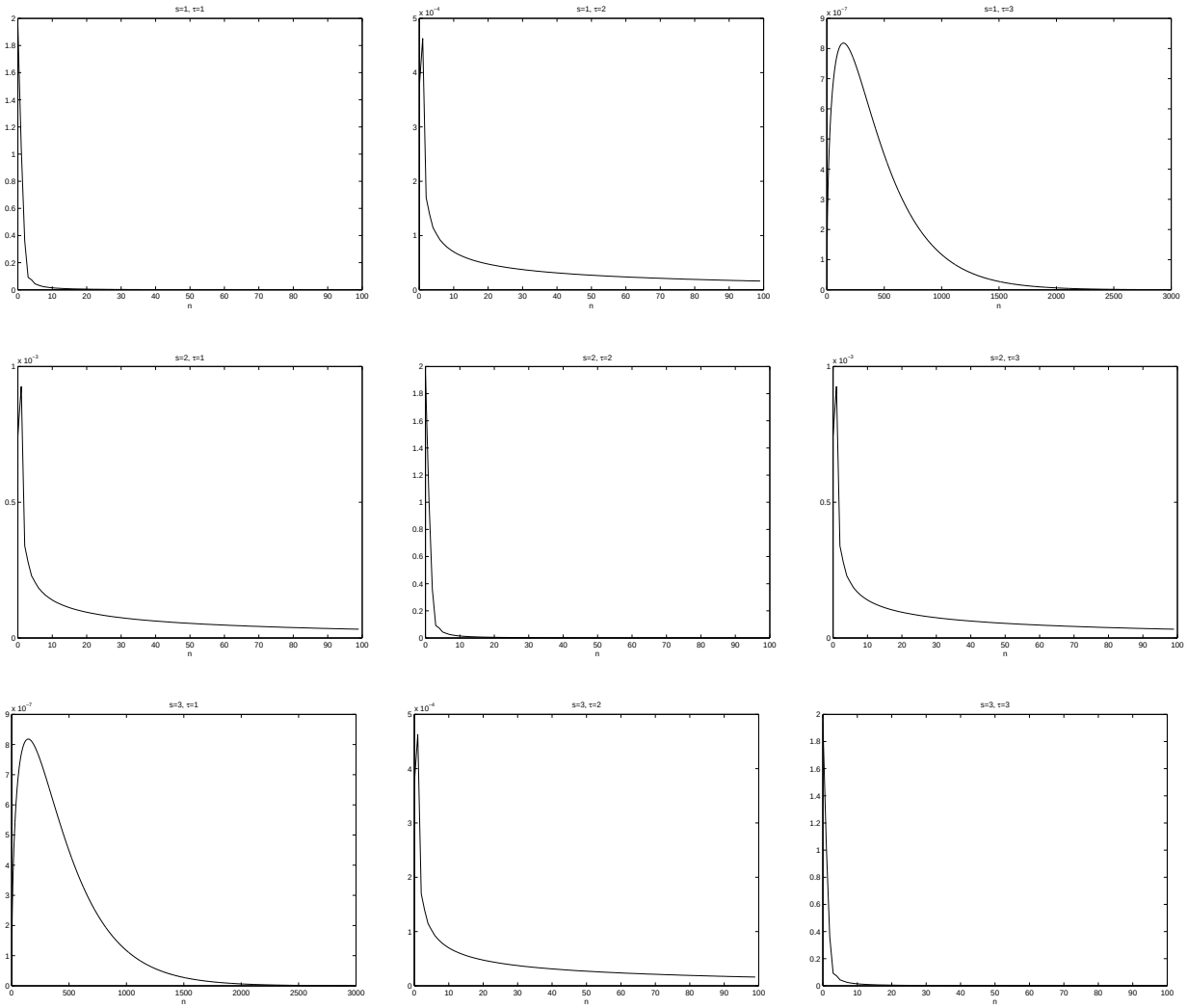


Figure 1.11: Example 1: Exact summed convolution coefficients $\underline{\psi}^n_{s,\tau}$ for $s, \tau = 1, \ldots, 3$

coefficients of diagonal entries belong to the first class. Non-diagonal entries $\phi^n_{s,\tau}$ belong to the second class, if the corresponding entry in the generator matrix is nonzero $q^T_{s,\tau} > 0$, otherwise to the third. We cannot prove this coherence, but the consideration of many different examples confirms this conjecture. Fig. 1.11 shows the the summed coefficients for each matrix entry of the problem. We observe, that also the diagonal summed coefficients decrease rapidly, which makes it extremely desirable to use them instead of the un-summed coefficients.

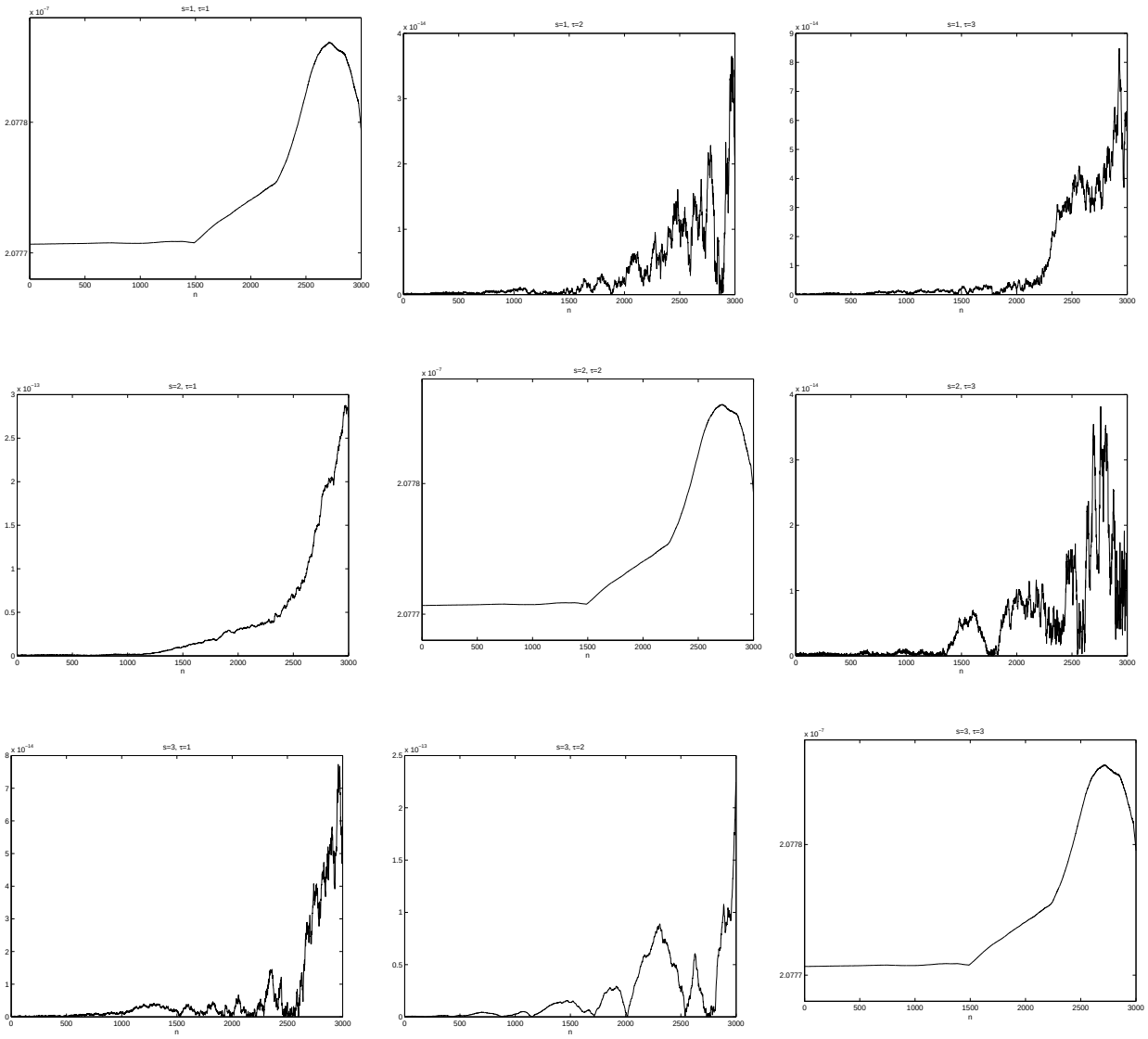


Figure 1.12: Example 1: Absolute error $|\phi^n_{s,\tau} - \underline{\phi}^n_{s,\tau}|$ for $s, \tau = 1, \ldots, 3$

Fig. 1.12 and 1.13 show the absolute error of the coefficients $\phi^n_{s,\tau}$ and the summed coefficients $\psi^n_{s,\tau}$ for $n = 0, \ldots, 3000$, respectively. We observe, that the error in the off-diagonal coefficients $\phi^n_{s,\tau}$ is close to the computational accuracy (MATLAB on our machine: $\epsilon_m = 10^{-16}$), whereas the error in the diagonal coefficients $\phi^n_{s,s}$ is approximately 10^{-7}. We also notice, that the error in the summed coefficients $\psi^n_{s,\tau}$ is nearly as small as ϵ_m, even for the diagonal coefficients. This improvement for the summed coefficients is due to the elimination of a singularity in the $\mathcal{Z}$-transform at $z = -1$ by the multiplication with the

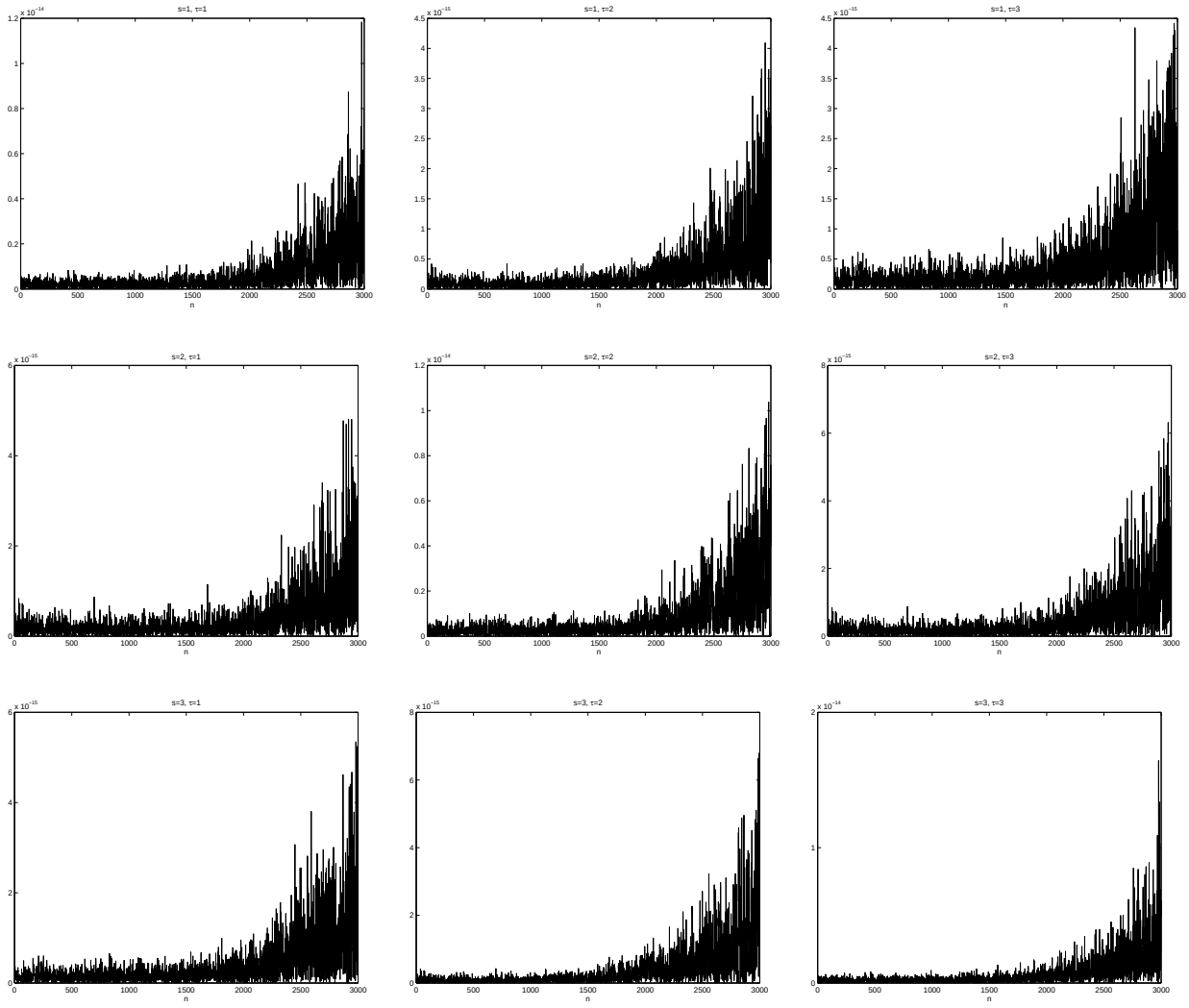


Figure 1.13: Example 1: Absolute error $|\psi^n_{s,\tau} - \underline{\psi}^n_{s,\tau}|$ for $s, \tau = 1, \ldots, 3$

factor $z + 1$ in (1.7.71). Fig. 1.14 illustrates this behaviour: It shows the $\mathcal{Z}$-transformed convolutions coefficients for $s = 1$ and $\tau = 1, \ldots, 3$ on the l.h.s. . For all three coefficients we can observe peaks in real and imaginary part of the $\mathcal{Z}$-transform. Choosing different amounts of sampling points verifies, that there is a singularity in $\hat{\phi}_{1,1}$ at $z = -1$, whereas the peak at $z = 1$ in $\hat{\phi}_{1,2}$ and $\hat{\phi}_{1,3}$ is none. By a multiplication with $\frac{z+1}{z}$ (cf. Sec. 7.2.4) we obtain the $\mathcal{Z}$-transformed summed coefficients. These are plotted in Fig. 1.14 on the r.h.s. for $\hat{\psi}_{1,1}, \hat{\psi}_{1,2}$ and $\hat{\psi}_{1,3}$. We observe, that the singularity is completely eliminated. Thus, the summed coefficients do not only avoid subtractive cancellation in the evaluation of the

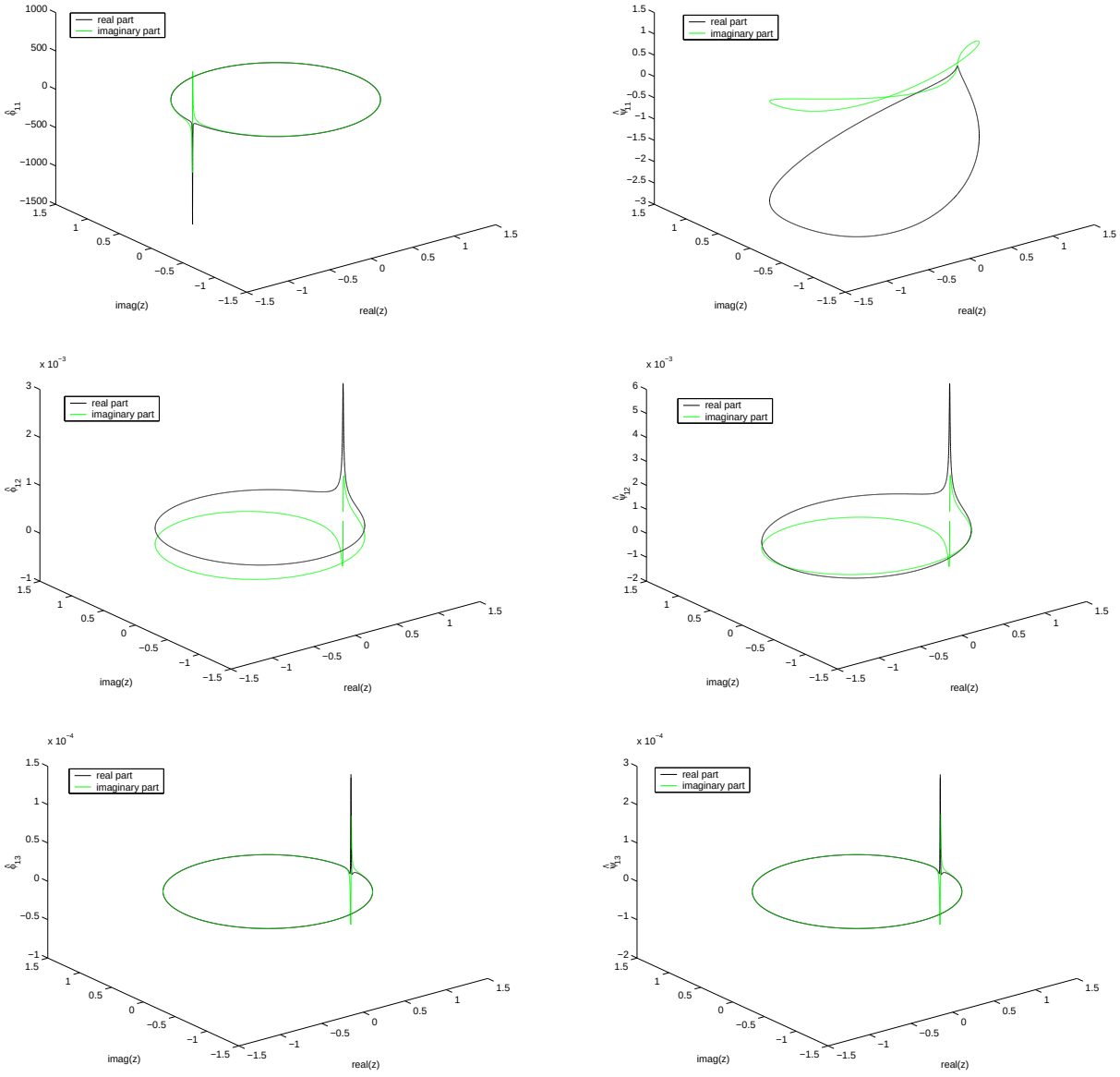


Figure 1.14: Example 1: $\mathcal{Z}$-transform of the coefficients $\hat{\phi}_{s,\tau}$ on the l.h.s. and of the summed coefficients $\hat{\psi}_{s,\tau}$ on the r.h.s. for $s = 1, \tau = 1, \ldots, 3$

convolution, but also is their $\mathcal{Z}$-transform better to inverse transform. All this caused us to use the summed coefficients in the further numerical calculations.

With the DTBC and these summed coefficients we calculated the density function $\boldsymbol{\pi}$. In order to check the quality of the boundary condition we computed a reference solution

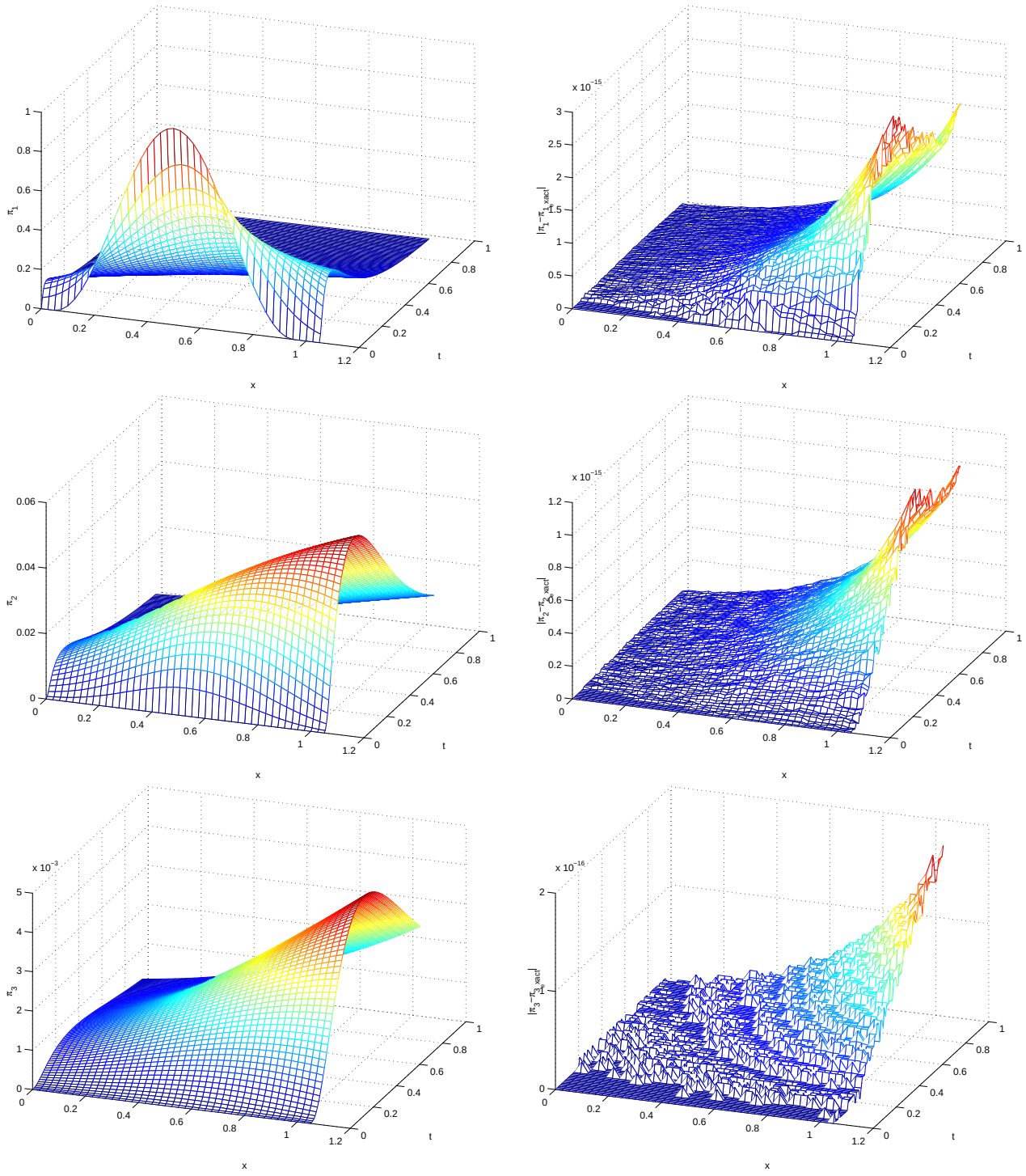


Figure 1.15: Example 1: The numerically calculated density function π_1, π_2 and π_3 on the l.h.s. and the error to a reference solution on the r.h.s.

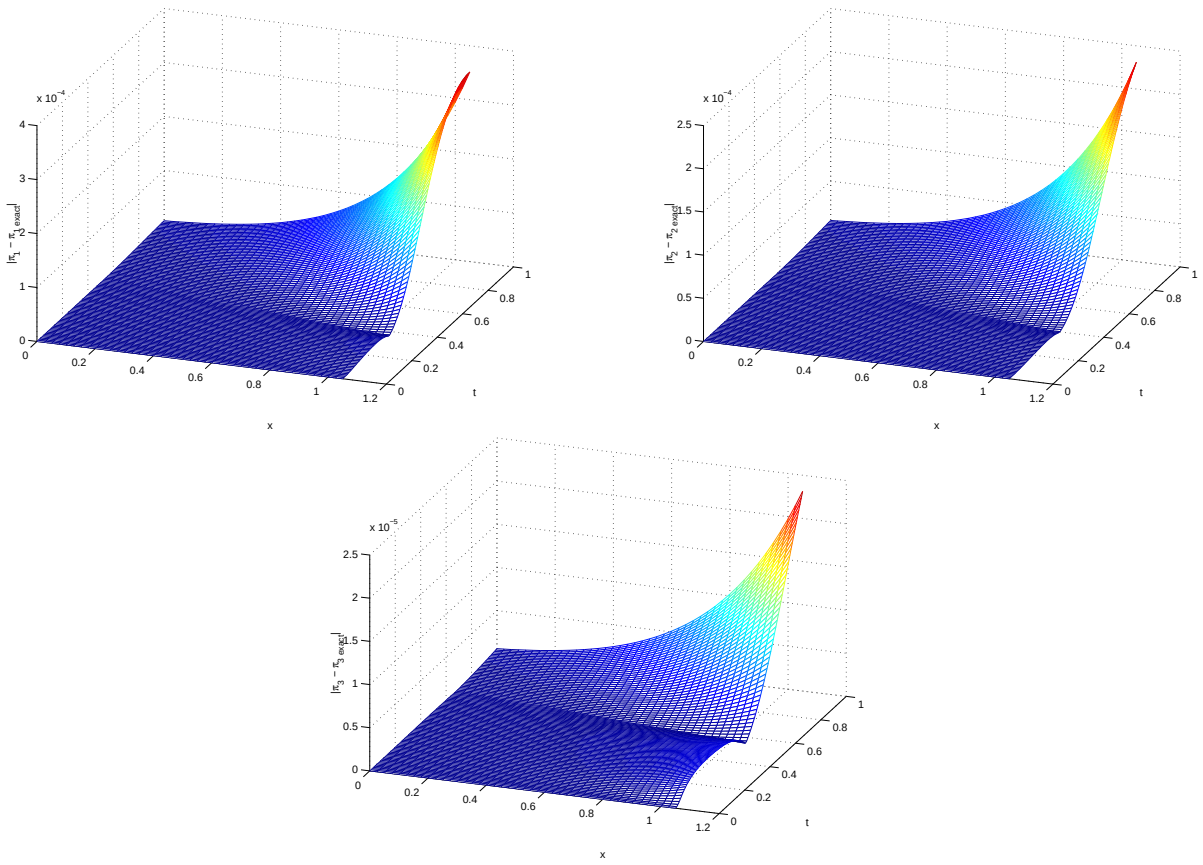


Figure 1.16: Example 1: Absolute error of the difference scheme with approximated coefficients for $s = 1, \ldots, 3$

$\boldsymbol{\pi}_{\text{exact}}$ on a much bigger interval $[0, 2.5]$. Fig. 1.15 shows the time evolution of the density function π_s for each state/component s on the l.h.s. and the error to the reference solution $\pi_{s_{\text{exact}}}$ on the r.h.s. . Not only the positive drift $\mu = 3$ can be seen in each state, but also the effect of diffusion $\sigma^2 = 0.9$ is obvious. Further we can observe, that the error is at first proportional to the amount of fluid at the transparent boundary: The error at the transparent boundary at first grows, then declines for a short time. Afterwards it keeps on growing. This grows is caused by the fact, that the convolution is increased in each time step by one summand.

The numerical computation of this example with 1000 time steps required 361,85 sec CPU time, thereof are 23,35 sec initialisation and calculation of the convolution coefficients. The main part of the remaining CPU time is due to the boundary condition, and this part increases with each step in time. To reduce the computational effort, we also used the

approximated coefficients (cf. Sec. 7.2.5)as an alternative. With this approximated boundary condition the algorithm required 28,5 sec CPU time for 1000 time steps. Not included is here the time for calculating the approximated coefficients. This required additional 240,2 sec for $L = 30$ and $\nu = 2$ including the computation of the first ν coefficients with 128 sampling points. The comparatively high effort is caused by the Padé approximation, which must be performed with high precision ($2L - 1$ digits mantissa length) to avoid a "nearly breakdown" by ill conditioned steps in the Lanczos algorithm (cf. [**BB97**]). If such problems still occur, the degree of the polynomial is successively reduced. Thus, the approximated DTBC shows its advantages for large-time calculations. The given CPU-times seem rather large. Nevertheless, there is no real alternative: even the solution on a bigger domain with some kind of more effective boundary conditions cannot yield good results, because due to the diffusion any boundary is reached after a few time steps and the boundary error is again "instantly" diffused on the whole domain. Fig. 1.16 shows the error of the numerical solution with approximated DTBCs to a reference solution. The error is with 10^{-4} and 10^{-5} respectively rather large compared to the error of the exact TBC (cf. Fig. 1.15). Still a direct comparison of the density functions gives no visible difference, i.e. the solution with the approximated TBC yields very good qualitative results.

8.2. Example 2 - the queueing system example. Next we will consider the queueing system example of Fig. 1.4 in Chap. 2. The generator matrix $\mathbf{Q}$ results from the the rates

$$\lambda_1 = 4.0, \qquad \lambda_2 = 5.0, \tag{1.8.3}$$

$$\lambda_6 = 1.0, \qquad \lambda_7 = 0.25 \tag{1.8.4}$$

of the transitions (t_1, t_2, t_6 and t_7), which have exponential distribution time and are part of the discrete petri net. Thus $\mathbf{Q}$ evaluates to

$$\mathbf{Q} = \begin{pmatrix} -5.25 & 0.25 & 5.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & -6.0 & 0.0 & 5.0 & 0.0 & 0.0 \\ 4.0 & 0.0 & -9.25 & 0.25 & 5.0 & 0.0 \\ 0.0 & 4.0 & 1.0 & -10.0 & 0.0 & 5.0 \\ 0.0 & 0.0 & 4.0 & 0.0 & -4.25 & 0.25 \\ 0.0 & 0.0 & 0.0 & 4.0 & 1.0 & -5.0 \end{pmatrix} . \tag{1.8.5}$$

The fluid parameters

$$\boldsymbol{M} = \operatorname{diag}(-1.2,\, 0.4,\, 0,\, 1.6,\, -0.4,\, 1.2)\,, \tag{1.8.6}$$

$$\boldsymbol{\Sigma}^2 = \operatorname{diag}(2.0,\, 0.4,\, 3.2,\, 1.6,\, 2.8,\, 1.2) \tag{1.8.7}$$

have already been given. At the beginning the system is in the state one shown in Fig. 1.4. Thus, the initial marking is

$$\boldsymbol{\pi}_0 = \Big(1,\ 0,\ 0,\ 0,\ 0,\ 0\Big)\,. \tag{1.8.8}$$

Again we used the Crank-Nicolson scheme ($\theta = 0.5$) with the step sizes $h = 0.025$ in space and $k = 0.001$ in time. We inverse transformed the convolution coefficients on a circle with radius $rcirc = 1.001$ with 2^{12} sampling points and used summed convolution coefficients.

Fig. 1.17 shows the xt-diagram of the density function $\boldsymbol{\pi}$ for all six states. We observe, that the mass in the states $s = 2, 4, 6$ moves to the right, what we expected due to $\mu_2, \mu_4, \mu_6 > 0$. The mass moving to the right is interpreted as an increasing number of waiting clients in the system, which grows since for $s = 2, 4, 6$ the server fails and the petri net is in the state "down". Due to the coupling, the mass in state $s = 3$ moves to the left.

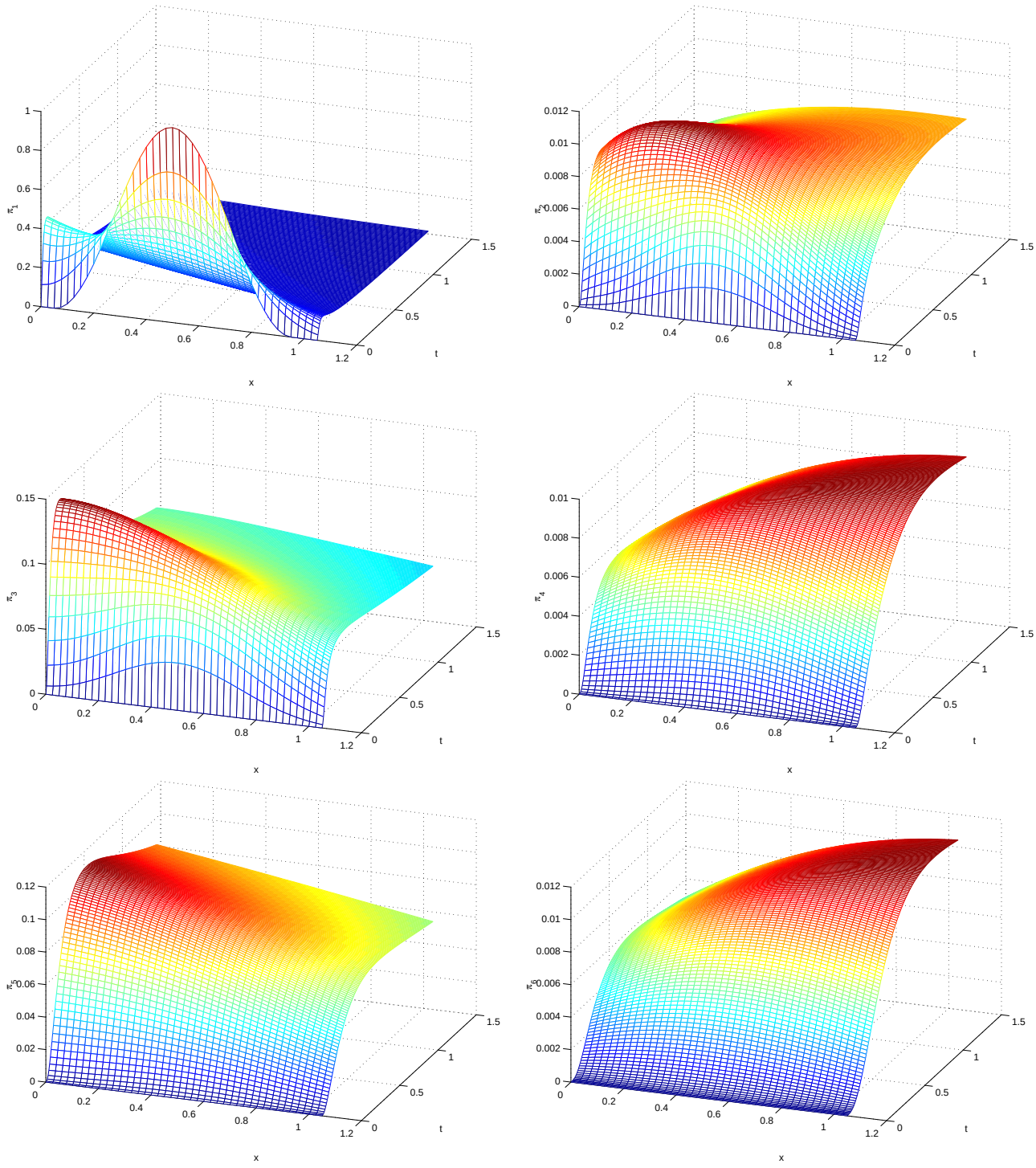


Figure 1.17: Example 2: The numerically calculated density $\pi_1, \ldots, \pi_6$

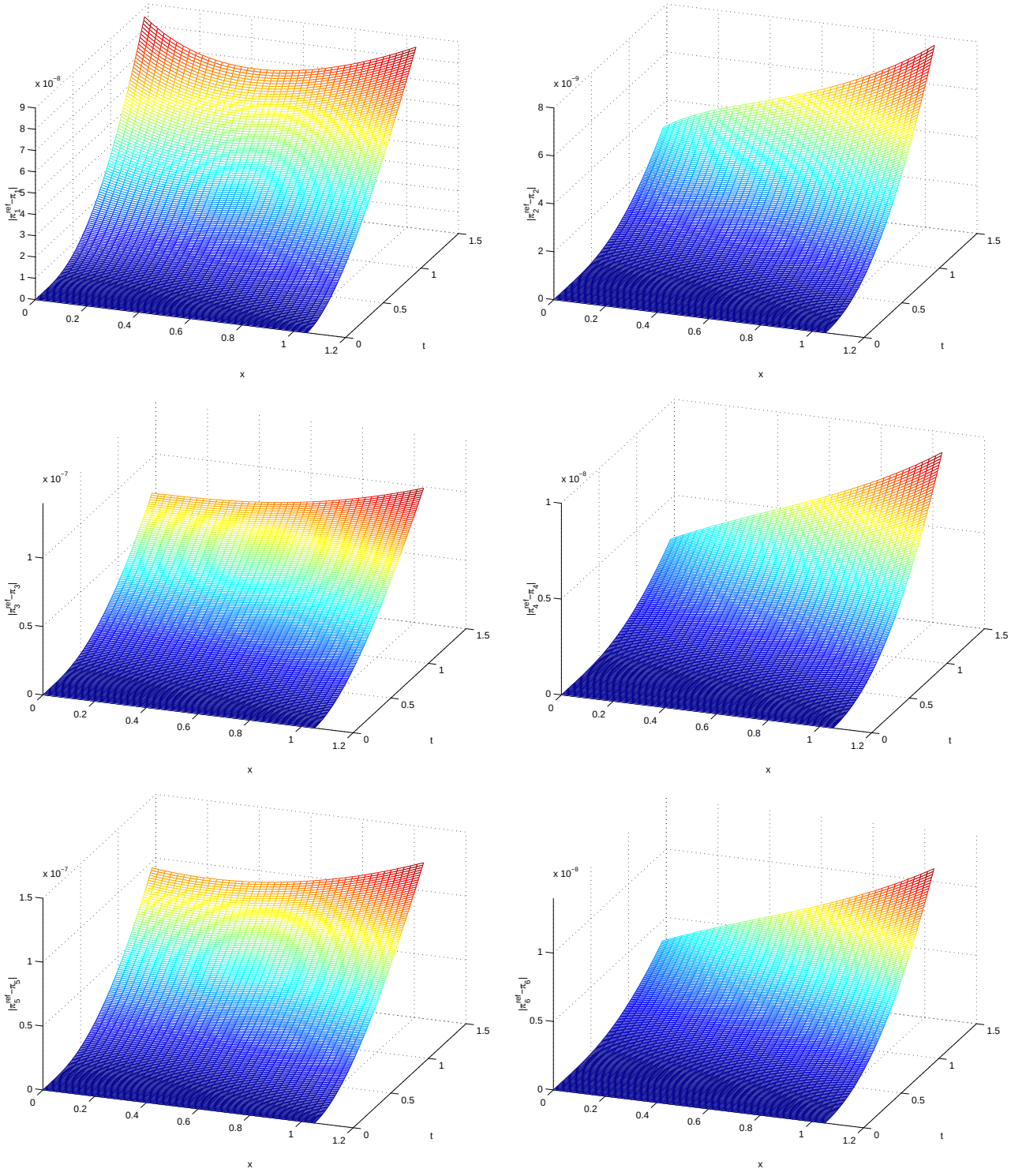


Figure 1.18: Example 2: Error $|\pi_s - \pi_{s_{\text{exact}}}|$of the numerical solution π_s compared to the reference solution $\pi_{s_{\text{exact}}}$ for $s = 1, \ldots, 6$

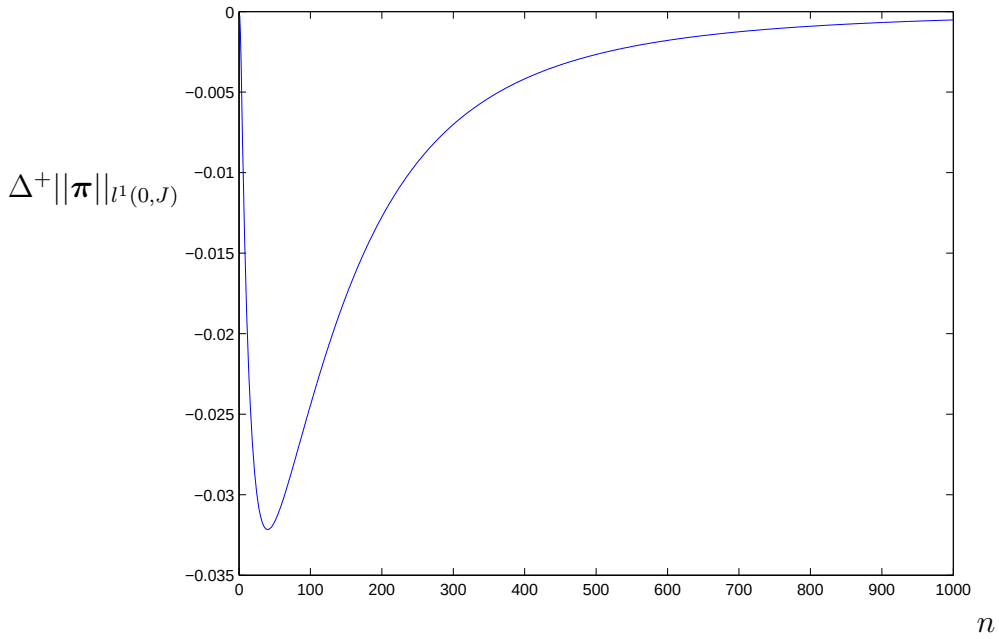


Figure 1.19: Example 2: The time dependent change in the l^1-norm of the solution: $\triangle_t^+||\mathbf{u}^n||_{l^1}$

μ_3 is zero, but the by the coupling in-flowing mass comes especially from state $s = 1$ and $s = 5$ (see above $q_{1,3} = 5, \quad q_{5,3} = 4$), which have negative μ.

Fig. 1.18 shows the error to the reference solution. It is with 10^{-7} to 10^{-9} after 1000 time steps very small, but growing in time. Probably the growth is not only due to the fact, that the convolution is increased in each time step by one summand. Also the accuracy of the convolution coefficients becomes worse with larger time.

Finally, we want to check numerically the stability of the θ-scheme with DTBCs for the current example. Therefore, we have to show, that (1.7.86) is non-positive for the whole computed time. Fig. 1.19 shows the time dependent change in the l^1-norm of the numerical solution. It is negative for each time step $n = 1, \ldots, 1000$. Thus, we used a stable scheme for this example.

9. The general system of parabolic equations

Here we will show, that it is possible to formulate the transparent boundary condition given in the preceeding sections, also for a more general system of parabolic equations. Whereas in the continuous case this is always possible, the discrete case necessitates additional restrictions. We consider

$$\mathbf{u}_t = \frac{\partial}{\partial x}(\mathbf{A}(x,t)\mathbf{u}_x) + \mathbf{M}(x,t)\mathbf{u}_x + \mathbf{V}(x,t)\mathbf{u}, \qquad x \in \mathbb{R}, t > 0, \tag{1.9.1}$$

where $\mathbf{A}$, $\mathbf{M}$ and $\mathbf{V}$ are real valued $n \times n$-matrices. We recall in which case a system is called parabolic:

Definition 1.4 ([**KL89**]). *The system* (1.9.1) *is called* parabolic *in* $0 \le t \le T$ *if there is a constant* $\delta > 0$ *such that for all* $x \in \mathbb{R}$, $0 \le t \le T$ *and for all eigenvalues* κ *of the matrix* $\mathbf{A}$ *holds*

$$\kappa \ge \delta > 0. \tag{1.9.2}$$

We again split the whole-space problem into three parts: the interior problem for $x_L \le x \le x_R$ and the left and right exterior problems. For the exterior problems

$$s\hat{\mathbf{u}} = \mathbf{A}_L\hat{\mathbf{u}}_{xx} + \mathbf{M}_L\hat{\mathbf{u}}_x + \mathbf{V}_L\hat{\mathbf{u}}, \qquad x < x_L, \tag{1.9.3a}$$

$$s\hat{\mathbf{u}} = \mathbf{A}_R\hat{\mathbf{u}}_{xx} + \mathbf{M}_R\hat{\mathbf{u}}_x + \mathbf{V}_R\hat{\mathbf{u}}, \qquad x > x_R, \tag{1.9.3b}$$

where the matrices $\mathbf{A}_{L,R}$, $\mathbf{M}_{L,R}$ and $\mathbf{V}_{L,R}$ are constant in x and t, the condition (1.9.2) reads:

$$\kappa > 0, \qquad \text{for all eigenvalues } \kappa \text{ of } \mathbf{A}_{L,R}. \tag{1.9.4}$$

Thus, we will restrict our considerations to *positive definite* matrices $\mathbf{A}_{L,R}$.

For the transparent boundary condition we again use the solution of the exterior problems. These will be solved using the Laplace transformation in time, yielding $2n$ solutions. We will show, that n solutions decay and n increase. The derivation of the left and right boundary condition just differs in the decision, which solutions increase and thus have to be extinguished by the DTBC. Therefore, we restrict the derivation to the right exterior problem. The Laplace transformed exterior problem reads

$$\begin{pmatrix} \mathbf{u} \\ \mathbf{u}_x \end{pmatrix}_x = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{A}^{-1}(s\mathbf{I} - \mathbf{V}) & -\mathbf{A}^{-1}\mathbf{M} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{u}_x \end{pmatrix} = \mathbf{C} \begin{pmatrix} \mathbf{u} \\ \mathbf{u}_x \end{pmatrix}, \qquad x \ge x_R, \tag{1.9.5}$$

where $\mathbf{A} = \mathbf{A}_R, \mathbf{M} = \mathbf{M}_R$ and $\mathbf{V} = \mathbf{V}_R$ are constant matrices. With the matrix $\mathbf{P}^{-1} = \left(\begin{smallmatrix} \mathbf{P}_1 & \mathbf{P}_2 \\ \mathbf{P}_3 & \mathbf{P}_4 \end{smallmatrix}\right)$ of left (possibly generalised) eigenvectors of $\mathbf{C}$ and $\mathbf{J} = \left(\begin{smallmatrix} \mathbf{J}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{smallmatrix}\right)$ the Jordan form of $\mathbf{C}$ split in $\mathbf{J}_1$ containing eigenvalues associated with decaying and $\mathbf{J}_2$ containing eigenvalues associated with increasing solutions, we have

$$\mathbf{P}\begin{pmatrix}\mathbf{u}\\ \mathbf{u}_x\end{pmatrix}_x = \begin{pmatrix}\mathbf{J}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{J}_2\end{pmatrix}\begin{pmatrix}\mathbf{P}_1\mathbf{u}+\mathbf{P}_2\mathbf{u}_x\\ \mathbf{P}_3\mathbf{u}+\mathbf{P}_4\mathbf{u}_x\end{pmatrix} \tag{1.9.6}$$

and claim, that no influence of for $x \to \infty$ / $x \to -\infty$ increasing solutions exists at the right/left boundary, which yields the TBCs

$$\mathbf{P}_3^R\mathbf{u} + \mathbf{P}_4^R\mathbf{u}_x = \mathbf{0}, \qquad x = x_R, \tag{1.9.7a}$$

$$\mathbf{P}_1^L\mathbf{u} + \mathbf{P}_2^L\mathbf{u}_x = \mathbf{0}, \qquad x = x_L. \tag{1.9.7b}$$

This splitting into eigenvalues associated to increasing and decreasing solutions gives then the right number of boundary conditions at each boundary, if it splits the eigenvalues into two equal groups. This will be asserted in the following lemma.

We remember from 5.1, that eigenvalues with a positive /negative real part, yield for $x \to \infty$ increasing/decreasing solutions.

THEOREM 1.21 (Splitting Theorem). *Of the $2n$ eigenvalues of $\mathbf{C}$ n have a positive and n a negative real part (and thus yield n for $x \to \infty$ increasing and n decaying solutions), if Re(s) is sufficiently large.*

PROOF. We will first show, that there is no purely imaginary eigenvalue λ of $\mathbf{C}$. Therefore, we use the ansatz

$$\hat{\mathbf{u}} = e^{\lambda x}\mathbf{u}_0 \tag{1.9.8}$$

in (1.9.3), which yields

$$\lambda^2\bar{\mathbf{u}}_0^T\mathbf{A}\mathbf{u}_0 + \lambda\bar{\mathbf{u}}_0^T\mathbf{M}\mathbf{u}_0 + \bar{\mathbf{u}}_0^T\mathbf{V}\mathbf{u}_0 - s|\mathbf{u}_0|^2 = 0. \tag{1.9.9}$$

Assume $\lambda = i\beta, \quad \beta \in \mathbb{R}$. We consider real parts. Since $\mathbf{A}$ is positive definite, it holds $\lambda^2\bar{\mathbf{u}}_0^T\mathbf{A}\mathbf{u}_0 < 0$ and thus, if the condition

$$\mathrm{Re}\left(\bar{\mathbf{u}}_0^T\mathbf{V}\mathbf{u}_0 + \lambda\bar{\mathbf{u}}_0^T\mathbf{M}\mathbf{u}_0 - s|\mathbf{u}_0|^2\right) < 0 \tag{1.9.10}$$

holds, (1.9.9) is a contradiction. But this condition is true, since

$$\frac{\mathbf{V}+\mathbf{V}^T}{2} - \beta \mathrm{Im}\left(\frac{\mathbf{M}-\mathbf{M}^T}{2}\right) - \mathrm{Re}(s) \tag{1.9.11}$$

is negative definite for $\mathrm{Re}(s)$ sufficiently large.

Now, instead of $\mathbf{M}$ consider $\epsilon\mathbf{M}$ in (1.9.3) with $\epsilon \in [0,1]$. For $\epsilon = 0$ equations (1.9.3) are invariant for $x \to -x$ and thus the number of increasing and decreasing solutions, i.e. the number of eigenvalues of $\mathbf{C}$ with positive and negative real parts must be the same. Now, for ϵ from zero to one, the eigenvalues of $\mathbf{C}$ are continuously depending on ϵ and there exists no purely imaginary eigenvalue for any $\epsilon \in [0,1]$. For $\epsilon = 1$, still n eigenvalues have positive and n have negative real part. □

Thus, the conditions (1.9.7) give for any parabolic system of the form (1.9.1) the right number of boundary conditions.

9.1. The discretisation of the general parabolic system. In this section we will shortly give an appropriate discretisation of the parabolic system. For the given discretisation, we then derive discrete transparent boundary conditions. To this end we solve the exterior problems using a $\mathcal{Z}$-transformation and assert, under which conditions this yields the right number of boundary conditions.

For a discretisation of (1.9.1) we use the θ-scheme with a central difference approximation for the first and second spatial derivative. An extrapolating discretisation of the lowest order term as for the Petri nets is no longer of advantage, because the coupling is not restricted to this term. Instead we use again the θ-scheme with the abbreviation $u_{s,j}^{n+\theta} = (1-\theta)u_{s,j}^n + \theta u_{s,j}^{n+1}$:

$$\frac{h^2}{k}(\mathbf{u}_j^{n+1} - \mathbf{u}_j^n) = \Delta^+(\mathbf{A}\Delta^-\mathbf{u}_j^{n+\theta}) + \frac{h}{2}\mathbf{M}(\Delta^+ + \Delta^-)\mathbf{u}_j^{n+\theta} + h^2\mathbf{V}\mathbf{u}_j^{n+\theta}. \tag{1.9.12}$$

For the derivation of the right DTBC we consider the $\mathcal{Z}$-transformed discrete equation for the right exterior problem

$$\frac{h^2}{k}\frac{z-1}{\theta z + 1 - \theta}\hat{\mathbf{u}}_j = \mathbf{A}\Delta^+\Delta^-\hat{\mathbf{u}}_j + \frac{h}{2}\mathbf{M}(\Delta^+ + \Delta^-)\hat{\mathbf{u}}_j + h^2\mathbf{V}\hat{\mathbf{u}}_j, \tag{1.9.13}$$

for $j \geq J$, where $\mathbf{A} = \mathbf{A}_R, \mathbf{M} = \mathbf{M}_R$ and $\mathbf{V} = \mathbf{V}_R$ are constant. We reduce the system of difference equations to first order

$$\begin{pmatrix} \frac{h}{2}\mathbf{M} & \mathbf{A} \\ -\mathbf{I} & \mathbf{I} \end{pmatrix}\begin{pmatrix} \Delta^+\hat{\mathbf{u}}_j \\ \Delta^+\Delta^-\hat{\mathbf{u}}_j \end{pmatrix} = \begin{pmatrix} \frac{h^2}{k}\frac{z-1}{\theta z+1-\theta}\,I - h^2\mathbf{V} & -\frac{h}{2}\mathbf{M} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix}\begin{pmatrix} \hat{\mathbf{u}}_j \\ \Delta^-\hat{\mathbf{u}}_j \end{pmatrix}, \tag{1.9.14}$$

or since $\mathbf{A}$ is regular equivalently

$$\begin{aligned}\begin{pmatrix}\Delta^+\hat{\mathbf{u}}_j\\ \Delta^+\Delta^-\hat{\mathbf{u}}_j\end{pmatrix} &= \begin{pmatrix}(\mathbf{T}^+)^{-1}\left[\frac{h^2}{k}\frac{z-1}{\theta z+1-\theta}\,I-h^2\mathbf{V}\right] & (\mathbf{T}^+)^{-1}\mathbf{T}^-\\ (\mathbf{T}^+)^{-1}\left[\frac{h^2}{k}\frac{z-1}{\theta z+1-\theta}\,I-h^2\mathbf{V}\right] & (\mathbf{T}^+)^{-1}\mathbf{T}^- - \mathbf{I}\end{pmatrix}\begin{pmatrix}\hat{\mathbf{u}}_j\\ \Delta^-\hat{\mathbf{u}}_j\end{pmatrix}\\ &= \widetilde{\mathbf{C}}\begin{pmatrix}\hat{\mathbf{u}}_j\\ \Delta^-\hat{\mathbf{u}}_j\end{pmatrix}, \qquad (1.9.15)\end{aligned}$$

with the abbreviations $\mathbf{T}^+ := \mathbf{A} + \frac{h}{2}\mathbf{M}$ and $\mathbf{T}^- := \mathbf{A} - \frac{h}{2}\mathbf{M}$. We claim, that $\mathbf{T}^+$ and $\mathbf{T}^-$ are positive definite matrices, which can be ensured by a sufficiently small space step size h.

We decompose the Jordan form $\mathbf{J}$ of $\tilde{\mathbf{C}}+\mathbf{I}$ in two blocks $\mathbf{J} = \left(\begin{smallmatrix}\mathbf{J}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{J}_2\end{smallmatrix}\right)$, where $\mathbf{J}_1$ holds the eigenvalues with an absolute value smaller than one, $\mathbf{J}_2$ those with an absolute value larger than one. Then (1.9.15) reads with the matrix $\mathbf{P}^{-1} = \left(\begin{smallmatrix}\mathbf{P}_1 & \mathbf{P}_2\\ \mathbf{P}_3 & \mathbf{P}_4\end{smallmatrix}\right)$ of left (possibly generalised) eigenvectors

$$\mathbf{P}\begin{pmatrix}\Delta^+\hat{\mathbf{u}}_j\\ \Delta^+\Delta^-\hat{\mathbf{u}}_j\end{pmatrix} = \begin{pmatrix}\mathbf{J}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{J}_2\end{pmatrix}\begin{pmatrix}\mathbf{P}_1 & \mathbf{P}_2\\ \mathbf{P}_3 & \mathbf{P}_4\end{pmatrix}\begin{pmatrix}\hat{\mathbf{u}}_j\\ \Delta^-\hat{\mathbf{u}}_j\end{pmatrix}. \qquad (1.9.16)$$

The eigenvalues in $\mathbf{J}_2$ yield for $j\to\infty$ increasing solutions. Therefore the DTBC reads

$$\mathbf{P}_3\hat{\mathbf{u}}_J + \mathbf{P}_4\Delta^-\hat{\mathbf{u}}_J = \mathbf{0}. \qquad (1.9.17)$$

Now we will justify the splitting of the eigenvalues:

THEOREM 1.22 (Discrete Splitting Theorem). *Of the $2n$ eigenvalues of $\widetilde{\mathbf{C}}+\mathbf{I}$ n have an absolute value strictly larger and n have an absolute value strictly smaller than one, if $\frac{\mathbf{V}+\mathbf{V}^T}{2}$ is negative definite, $\frac{1}{2}\le\theta\le 1$, $|z|>1$ and h sufficiently small.*

PROOF. The proof is analogous to that of Thm. 1.21. We will show, that no eigenvalue λ of $\widetilde{\mathbf{C}}+\mathbf{I}$ with an absolute value of one exists. As in the continuous case, equation (1.9.13) is invariant for $j\to -j$ for $\mathbf{M}=\mathbf{0}$ and a continuity argument proves the splitting.

To investigate the absolute value of the eigenvalues of $\widetilde{\mathbf{C}}+\mathbf{I}$ we insert the ansatz $\hat{\mathbf{u}}_j = \lambda^j\hat{\mathbf{u}}_0$ in (1.9.13)

$$\lambda^2\mathbf{T}^+\hat{\mathbf{u}}_0 + \mathbf{T}^-\hat{\mathbf{u}}_0 = \lambda\left(\mathbf{T}^+ + \mathbf{T}^- - h^2\mathbf{V} + \frac{h^2}{k}\frac{z-1}{\theta z+1-\theta}\mathbf{I}\right)\hat{\mathbf{u}}_0. \qquad (1.9.18)$$

We assume $|\lambda| = 1$, consider absolute values of (1.9.18) and use the triangle inequality after multiplication with $\bar{\hat{\mathbf{u}}}_0^T$ from the left

$$\bar{\hat{\mathbf{u}}}_0^T(\mathbf{T}^+ + \mathbf{T}^-)\hat{\mathbf{u}}_0 \geq \left| \bar{\hat{\mathbf{u}}}_0^T(\mathbf{T}^+ + \mathbf{T}^-)\hat{\mathbf{u}}_0 - h^2\bar{\hat{\mathbf{u}}}_0^T\mathbf{V}\hat{\mathbf{u}}_0 + \frac{h^2}{k}\frac{z-1}{\theta z + 1 - \theta}|\hat{\mathbf{u}}_0|^2 \right|, \tag{1.9.19}$$

where $\bar{\hat{\mathbf{u}}}_0^T(\mathbf{T}^+ + \mathbf{T}^-)\hat{\mathbf{u}}_0$ is a positive real value. But the absolute value on the r.h.s. is strictly larger than $\bar{\hat{\mathbf{u}}}_0^T(\mathbf{T}^+ + \mathbf{T}^-)\hat{\mathbf{u}}_0$, if

$$\mathrm{Re}\left(-h^2\bar{\hat{\mathbf{u}}}_0^T\mathbf{V}\hat{\mathbf{u}}_0 + \frac{h^2}{k}\frac{z-1}{\theta z + 1 - \theta}|\hat{\mathbf{u}}_0|^2 \right) > 0, \tag{1.9.20}$$

which is a contradiction. Thus for $\frac{\mathbf{V}+\mathbf{V}^T}{2}$ negative definite, $\frac{1}{2} \leq \theta \leq 1$ and $|z| > 1$ (cf. (1.7.35)) there exists no eigenvalue with absolute value one and the eigenvalues divide into two equal groups. □

REMARK 1.23. *We used the central difference to discretise the first spatial derivative, since this is possible for any matrix* $\mathbf{M}$. *If* M *is diagonalisable, it can be advantageous to use again an upwind discretisation. The upwind matrices* $\mathbf{R}$ *and* $\mathbf{I} - \mathbf{R}$ *are determined from* $\mathbf{M}_{diag}$ *the diagonalised* $\mathbf{M} = \mathbf{S}^{-1}\mathbf{M}_{diag}\mathbf{S}$. *This changes the matrices* $\mathbf{T}^+$ *and* $\mathbf{T}^-$ *into*

$$\mathbf{T}^+ = \mathbf{A} + h\mathbf{S}^{-1}\mathbf{M}_{\mathrm{diag}}\mathbf{R}\mathbf{S} \tag{1.9.21a}$$

and

$$\mathbf{T}^- = \mathbf{A} - h\mathbf{S}^{-1}\mathbf{M}_{\mathrm{diag}}(\mathbf{I} - \mathbf{R})\mathbf{S}, \tag{1.9.21b}$$

which still must be claimed to be positive definite.

In this discrete section, we could not formulate the DTBC without any restriction as in the continuous case. Nevertheless, for all systems with $\frac{\mathbf{V}+\mathbf{V}^T}{2}$ negative definite and by assuring that $\mathbf{T}^+$ and $\mathbf{T}^-$ are positive definite by choosing the step size h sufficiently small, we can derive the DTBC given in (1.9.17).

CHAPTER 2

Schrödinger–type systems

In the previous chapter we were dealing with parabolic systems. Here, we will be concerned with the derivation and analysis of DTBCs for systems of *Schrödinger-type equations* in one space dimension. These occur i.e. in the physics of *layered semiconductor devices* ([**Car96**],[**Kan82**]) as the so called $k\cdot p$*-Schrödinger equations*, which are a well established tool for *band structure calculations* [**Chu91**]. The $k\cdot p$ method in combination with an envelope function approximation ([**Bas88**],[**Sin93**],[**Chu95**,[**Car96**]) is frequently used to calculate the near band edge electronic band structure of *semiconductor heterostructures* ([**Sin93**],[**WZ96**],[**CC92**]), such as quantum wells. In the notation we follow Bandelow, Kaiser, Koprucki and Rehberg, who performed in [**BKKR00**] a rigorous analysis of spectral properties for the spatially one-dimensional $k\cdot p$-Schrödinger operators. The system then reads as follows

$$
\begin{aligned}
i\frac{\partial}{\partial t}\boldsymbol{\varphi} \;=\;& -\frac{\partial}{\partial x}(\mathbf{m}(x)\frac{\partial}{\partial x}\boldsymbol{\varphi}) + \mathbf{M}_0(x)\frac{\partial}{\partial x}\boldsymbol{\varphi} - \frac{\partial}{\partial x}(\mathbf{M}_0^H(x)\boldsymbol{\varphi}) \\
&+k_1\left(\mathbf{M}_1(x)\frac{\partial}{\partial x}\boldsymbol{\varphi} - \frac{\partial}{\partial x}(\mathbf{M}_1^H(x)\boldsymbol{\varphi})\right) + k_2\left(\mathbf{M}_2(x)\frac{\partial}{\partial x}\boldsymbol{\varphi} - \frac{\partial}{\partial x}(\mathbf{M}_2^H(x)\boldsymbol{\varphi})\right) \\
&+k_1\mathbf{U}_1(x)\boldsymbol{\varphi} + k_2\mathbf{U}_2(x)\boldsymbol{\varphi} + k_1^2\mathbf{U}_{11}(x)\boldsymbol{\varphi} + k_2^2\mathbf{U}_{22}(x)\boldsymbol{\varphi} + k_1k_2(\mathbf{U}_{12}(x)+\mathbf{U}_{21}(x))\boldsymbol{\varphi} \\
&+\mathbf{v}(x)\boldsymbol{\varphi} + \mathbf{e}(x)\boldsymbol{\varphi}, \qquad x\in\mathbb{R}, t>0, \quad k_1,k_2\in\mathbb{R},
\end{aligned} \tag{2.0.1}
$$

where $\boldsymbol{\varphi}(x,t)\in\mathbb{C}^d$ and $\mathbf{m}(x)$, $\mathbf{e}(x)$ are real diagonal $d\times d$-matrices. $\mathbf{U}_i(x)$, $\mathbf{U}_{ij}(x)$ and $\mathbf{v}(x)$ are Hermitian $d\times d$-matrices. The $d\times d$-matrices $\mathbf{M}_0(x)$, $\mathbf{M}_1(x)$ and $\mathbf{M}_2(x)$ are skew-Hermitian. This physical formulation is rather lengthy and we abbreviate

$$
\mathbf{M}_S(x) := \mathbf{M}_0(x) + k_1\mathbf{M}_1(x) + k_2\mathbf{M}_2(x), \tag{2.0.2a}
$$

(2.0.2b)

$$
\mathbf{V}(x) := k_1\mathbf{U}_1(x) + k_2\mathbf{U}_2(x) + k_1^2\mathbf{U}_{11}(x) + k_2^2\mathbf{U}_{22}(x) + k_1k_2(\mathbf{U}_{12}(x)+\mathbf{U}_{21}(x)) + \mathbf{v}(x) + \mathbf{e}(x).
$$

Then $\mathbf{M}_S(x)$ is skew-Hermitian, $\mathbf{V}(x)$ is Hermitian and (2.0.1) reads

$$
i\frac{\partial}{\partial t}\boldsymbol{\varphi} = -\frac{\partial}{\partial x}(\mathbf{m}(x)\frac{\partial}{\partial x}\boldsymbol{\varphi}) + \mathbf{M}_S(x)\frac{\partial}{\partial x}\boldsymbol{\varphi} - \frac{\partial}{\partial x}(\mathbf{M}_S^H(x)\boldsymbol{\varphi}) + \mathbf{V}(x)\boldsymbol{\varphi}, \quad x\in\mathbb{R}, t>0. \tag{2.0.3}
$$

1. Properties of the system of Schrödinger equations

An important property of the system (2.0.3) is the constancy in time of $||\boldsymbol{\varphi}||_{L^2}^2$ (conservation of mass). To verify this we multiply (2.0.3) with $\boldsymbol{\varphi}^H$ from the left and integrate by parts:

$$\begin{aligned}
\frac{\partial}{\partial t}||\boldsymbol{\varphi}||_{L^2}^2 &= \frac{\partial}{\partial t}\int_{\mathbb{R}} \boldsymbol{\varphi}^H\boldsymbol{\varphi}\,dx = \int_{\mathbb{R}} \boldsymbol{\varphi}^H\boldsymbol{\varphi}_t + \boldsymbol{\varphi}_t^H\boldsymbol{\varphi}\,dx = 2\,\mathrm{Im}\int_{\mathbb{R}} i\boldsymbol{\varphi}^H\boldsymbol{\varphi}_t \\
&= 2\,\mathrm{Im}\left(-\int_{\mathbb{R}} \boldsymbol{\varphi}^H\frac{\partial}{\partial x}(\mathbf{m}\frac{\partial}{\partial x}\boldsymbol{\varphi})\,dx + \int_{\mathbb{R}} \boldsymbol{\varphi}^H\mathbf{V}\boldsymbol{\varphi}\,dx\right. \\
&\quad \left. + \int_{\mathbb{R}} \boldsymbol{\varphi}^H\mathbf{M}_S\frac{\partial}{\partial x}\boldsymbol{\varphi}\,dx - \int_{\mathbb{R}} \boldsymbol{\varphi}^H\frac{\partial}{\partial x}(\mathbf{M}_S^H\boldsymbol{\varphi})dx\right) \\
&= 2\,\mathrm{Im}\left(\int_{\mathbb{R}} \boldsymbol{\varphi}_x^H\mathbf{m}\boldsymbol{\varphi}_x\,dx + \int_{\mathbb{R}} \boldsymbol{\varphi}^H\mathbf{V}\boldsymbol{\varphi}\,dx + \int_{\mathbb{R}} \underbrace{\boldsymbol{\varphi}^H\mathbf{M}_S\boldsymbol{\varphi}_x + \boldsymbol{\varphi}_x^H\mathbf{M}_S^H\boldsymbol{\varphi}}_{\in\mathbb{R}}\,dx\right) \\
&= 0.
\end{aligned}$$

The last equality can be seen by remembering that $\mathbf{V}$ and $\mathbf{m}$ are Hermitian and thus the imaginary part of the quadratic forms vanishes. The other term is of the form $y+y^H$ which is real.

2. Transparent boundary conditions

In this section we will derive transparent boundary conditions for the $k\cdot p$-Schrödinger equation (2.0.1) at the left $x = x_L$ and right $x = x_R$ boundary. In the scalar case (classical Schrödinger equation of quantum mechanics), the Laplace-transformed equation in the exterior domain can be solved explicitly. Afterwards the solution is inverse transformed, thus yielding the analytic TBC (cf. [**Arn98**]). Here (as for the parabolic system) the inverse Laplace transform can not be calculated explicitly for a system. Nevertheless, we will present the derivation of the Laplace transformed TBC.

For the derivation we proceed analogously to Chap. 1: we consider the Schrödinger equation in the left/right exterior domain. A Laplace transformation yields a system of ordinary differential equations, that can be reduced to first order. Then the solution of this system can be given in terms of its eigenvalues and eigenvectors. We will prove, that half of the eigenvalues have positive real parts and thus yield solutions increasing for $x \to \infty$; the other half has negative real parts, yielding decreasing solutions. Demanding that the part

of the increasing solutions in the right (and the decreasing solutions in the left) exterior domain vanishes, leads to the transparent boundary condition.

We consider equation (2.0.1) in the bounded domain $[x_L, x_R]$ together with TBCs at $x = x_L$ and $x = x_R$. The TBC at $x = x_R$ is constructed by considering (2.0.1) with constant coefficients for $x > x_R$, the so called *right exterior problem*

$$i\boldsymbol{\varphi}_t = -\mathbf{N}\boldsymbol{\varphi}_{xx} + i\mathbf{M}\boldsymbol{\varphi}_x + \mathbf{V}\boldsymbol{\varphi}, \qquad x > x_R,\ t > 0, \tag{2.2.1}$$

where $\mathbf{M} = \mathbf{M}^H$, $\mathbf{V} = \mathbf{V}^H$. $\mathbf{N}$ is diagonal, real and regular and given by

$$\mathbf{N} = \mathbf{m}, \tag{2.2.2a}$$

$$\mathbf{M} = -i\left(\mathbf{M}_0 - \mathbf{M}_0^H + k_1(\mathbf{M}_1 - \mathbf{M}_1^H) + k_2(\mathbf{M}_2 - \mathbf{M}_2^H)\right) \tag{2.2.2b}$$

$$= -i(\mathbf{M}_S - \mathbf{M}_S^H), \tag{2.2.2c}$$

$$\mathbf{V} = k_1\mathbf{U}_1 + k_2\mathbf{U}_2 + k_1^2\mathbf{U}_{11} + k_2^2\mathbf{U}_{22} + k_1k_2(\mathbf{U}_{12} + \mathbf{U}_{21}) + \mathbf{v} + \mathbf{e}. \tag{2.2.2d}$$

REMARK 2.1. *If $\mathbf{M}_S$ is skew-Hermitian, then $\mathbf{M}_S - \mathbf{M}_S^H = 2\mathbf{M}_S$ is also skew-Hermitian, thus $\mathbf{M} = -2i\mathbf{M}_S$ is Hermitian.*

We now use the Laplace-transformation given by

$$\hat{\boldsymbol{\varphi}}(x, s) = \int_0^\infty e^{-st}\boldsymbol{\varphi}(x, t)\, dt, \quad s = \alpha + i\xi, \quad \alpha > 0, \quad \xi \in \mathbb{R}, \tag{2.2.3}$$

on (2.2.1) and obtain the *transformed right exterior problem*

$$\mathbf{N}\hat{\boldsymbol{\varphi}}_{xx} - i\mathbf{M}\hat{\boldsymbol{\varphi}}_x = (\mathbf{V} - is\mathbf{I})\hat{\boldsymbol{\varphi}}, \qquad x > x_R. \tag{2.2.4}$$

LEMMA 2.2. *If the solution of the exterior problem* (2.2.4) *with the boundary data*

$$\hat{\boldsymbol{\varphi}}(x = x_R) = \hat{\boldsymbol{\varphi}}_R, \qquad \hat{\boldsymbol{\varphi}}(x = \infty) = \mathbf{0} \tag{2.2.5}$$

exists, it is unique for Re(s) sufficiently large.

PROOF. We assume, that there exist two such solutions of (2.2.4),(2.2.5) $\hat{\boldsymbol{\varphi}}_1$ and $\hat{\boldsymbol{\varphi}}_2$. Then the difference $\hat{\boldsymbol{\varphi}} = \hat{\boldsymbol{\varphi}}_1 - \hat{\boldsymbol{\varphi}}_2$ is a solution of (2.2.4) with homogeneous boundary data. Multiplying (2.2.4) with $\hat{\boldsymbol{\varphi}}^H$ from the left and integrating from x_R to ∞ yields after integrating by parts

$$-\int_{x_R}^\infty \hat{\boldsymbol{\varphi}}_x^H\mathbf{N}\hat{\boldsymbol{\varphi}}_x dx - i\int_{x_R}^\infty \hat{\boldsymbol{\varphi}}^H\mathbf{M}\hat{\boldsymbol{\varphi}}_x dx + \int_{x_R}^\infty is|\hat{\boldsymbol{\varphi}}|^2 dx - \int_{x_R}^\infty \hat{\boldsymbol{\varphi}}^H\mathbf{V}\hat{\boldsymbol{\varphi}} dx = 0. \tag{2.2.6}$$

Taking imaginary parts simplifies this, because the quadratic forms of the Hermitian matrices are purely real

$$\begin{aligned}
0 &= -\int_{x_R}^{\infty} \mathrm{Re}(\hat{\boldsymbol{\varphi}}^H \mathbf{M} \hat{\boldsymbol{\varphi}}_x) dx + \int_{x_R}^{\infty} \mathrm{Re}(s) |\hat{\boldsymbol{\varphi}}|^2 dx \\
&= -\int_{x_R}^{\infty} \sum_{k=1}^{d} \mathrm{Re}(\bar{\hat{\varphi}}_k m_{k,k} \hat{\varphi}_{kx}) dx - \int_{x_R}^{\infty} \sum_{k=1}^{d} \sum_{\substack{l=1 \\ l \neq k}}^{d} \mathrm{Re}(\bar{\hat{\varphi}}_k m_{k,l} \hat{\varphi}_{lx}) dx + \int_{x_R}^{\infty} \mathrm{Re}(s) |\hat{\boldsymbol{\varphi}}|^2 dx \\
&= -\frac{1}{2} \int_{x_R}^{\infty} \sum_{k=1}^{d} m_{k,k} \partial_x |\hat{\varphi}_k|^2 dx - \int_{x_R}^{\infty} \sum_{k=1}^{d} \sum_{l=k+1}^{d} \mathrm{Re}(\bar{\hat{\varphi}}_k m_{k,l} \hat{\varphi}_{lx} + \bar{\hat{\varphi}}_l \bar{m}_{l,k} \hat{\varphi}_{kx}) dx + \int_{x_R}^{\infty} \mathrm{Re}(s) |\hat{\boldsymbol{\varphi}}|^2 dx \\
&= \int_{x_R}^{\infty} \mathrm{Re}(s) |\hat{\boldsymbol{\varphi}}|^2 dx \geq 0 \qquad \text{for } \mathrm{Re}(s) > 0,
\end{aligned}$$

because $\partial_x |\hat{\varphi}_k|^2 = \bar{\hat{\varphi}}_{kx} \hat{\varphi}_k + \bar{\hat{\varphi}}_k \hat{\varphi}_{kx} = 2\mathrm{Re}(\bar{\hat{\varphi}}_k \hat{\varphi}_{kx})$ and with partial integration $\int m_{k,l} \bar{\hat{\varphi}}_k \hat{\varphi}_{lx} + \int \bar{m}_{l,k} \bar{\hat{\varphi}}_l \hat{\varphi}_{kx} = 2 \int \mathrm{Im}(m_{k,l} \bar{\hat{\varphi}}_k \hat{\varphi}_{lx})$, since $\mathbf{M}$ is Hermitian. From this we conclude $\hat{\boldsymbol{\varphi}} \equiv \mathbf{0}$, which is a contradiction to our assumption. □

REMARK 2.3. *As for the parabolic system the existence of a solution to the Laplace-transformed exterior problem is not clear (cf. Rem. 1.9). But in all considered examples the matrix of right eigenvectors has a regular submatrix, that gives a representation of the solution (cf. Lem. 1.8).*

To derive the transparent boundary condition we define $\boldsymbol{\nu} = \hat{\boldsymbol{\varphi}}$ and $\boldsymbol{\eta} = \hat{\boldsymbol{\varphi}}_x$ and thus reduce the order of the differential equation to obtain a system of first order differential equations

$$\underbrace{\begin{pmatrix} \mathbf{M} & i\mathbf{N} \\ -i\mathbf{N} & \mathbf{0} \end{pmatrix}}_{\mathbf{A}} \begin{pmatrix} \boldsymbol{\nu}_x \\ \boldsymbol{\eta}_x \end{pmatrix} = \underbrace{\begin{pmatrix} i\mathbf{V} + s\mathbf{I} & \mathbf{0} \\ \mathbf{0} & -i\mathbf{N} \end{pmatrix}}_{\mathbf{B}} \begin{pmatrix} \boldsymbol{\nu} \\ \boldsymbol{\eta} \end{pmatrix}. \tag{2.2.7}$$

We will show that the matrix $\mathbf{A}^{-1}\mathbf{B}$ is regular, because $\mathbf{A}^{-1}$ and $\mathbf{B}$ are regular. To this end we calculate the determinant of $\mathbf{A}$:

$$\det \begin{pmatrix} \mathbf{M} & i\mathbf{N} \\ -i\mathbf{N} & \mathbf{0} \end{pmatrix} = (-1)^n \det \begin{pmatrix} i\mathbf{N} & \mathbf{M} \\ \mathbf{0} & -i\mathbf{N} \end{pmatrix} = (-1)^n \det(\mathbf{N})^2 \neq 0, \tag{2.2.8}$$

since the determinant of a block-tridiagonal matrix is the product of the determinants of the block matrices on the diagonal. Here the determinant is obviously nonzero, since $\mathbf{N}$ is regular. Therefore the matrix $\mathbf{A}$ is regular with the inverse

$$\mathbf{A}^{-1} = \mathbf{N}^{-1} \begin{pmatrix} \mathbf{0} & i\mathbf{I} \\ -i\mathbf{I} & -\mathbf{M}\mathbf{N}^{-1} \end{pmatrix}. \tag{2.2.9}$$

The matrix $\mathbf{B}$ is a block diagonal matrix and thus its eigenvalues are the eigenvalues of the matrices on the diagonal. $\mathbf{V}$ is Hermitian, and thus it is diagonalisable and its eigenvalues $\mu_1, \ldots, \mu_n$ are real. Then $i\mathbf{V} + s\mathbf{I}$ is similar to $\mathrm{diag}(\alpha + (\mu_1 + \xi)i, \ldots, \alpha + (\mu_n + \xi)i)$ which is regular for $\mathrm{Re}(s) = \alpha > 0$, $s = \alpha + i\xi$, $\xi \in \mathbb{R}$. Therefore, $\mathbf{A}^{-1}\mathbf{B}$ as a product of regular matrices is regular for $\mathrm{Re}(s) > 0$.

In order to distinguish between increasing and decaying solutions of (2.2.4), we formulate the following lemma:

THEOREM 2.4 (Splitting Theorem). *For $Re(s) > 0$ the regular matrix $\mathbf{A}^{-1}\mathbf{B}$ has d eigenvalues with positive real part and d with negative real part.*

A proof of Thm. 2.4 will be given at the end of this section as a conclusion of Lem. 2.5 and Lem. 2.6.

We now transform $\mathbf{A}^{-1}\mathbf{B}$ into Jordan form with $\mathbf{A}^{-1}\mathbf{B} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$, where $\mathbf{P}$ contains the left eigenvectors in columns. We sort the Jordan blocks in $\mathbf{J}$ with respect to an increasing real part of the corresponding eigenvalue. Thus $\mathbf{J}$ can be written as $\mathbf{J} = \left(\begin{smallmatrix} \mathbf{J}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{smallmatrix}\right)$, where $\mathbf{J}_1$ holds all Jordan blocks to eigenvalues with negative real parts and $\mathbf{J}_2$ those with positive real parts. Due to Thm. 2.4 $\mathbf{J}_1$ and $\mathbf{J}_2$ are $d \times d$-matrices. With $\mathbf{P}^{-1} = \left(\begin{smallmatrix} \mathbf{P}_1 & \mathbf{P}_2 \\ \mathbf{P}_3 & \mathbf{P}_4 \end{smallmatrix}\right)$ equation (2.2.7) can be written as

$$\mathbf{P}^{-1} \begin{pmatrix} \boldsymbol{\nu}_x \\ \boldsymbol{\eta}_x \end{pmatrix} = \begin{pmatrix} \mathbf{J}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{pmatrix} \begin{pmatrix} \mathbf{P}_1 & \mathbf{P}_2 \\ \mathbf{P}_3 & \mathbf{P}_4 \end{pmatrix} \begin{pmatrix} \boldsymbol{\nu} \\ \boldsymbol{\eta} \end{pmatrix} = \begin{pmatrix} \mathbf{J}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{pmatrix} \begin{pmatrix} \mathbf{P}_1\boldsymbol{\nu} + \mathbf{P}_2\boldsymbol{\eta} \\ \mathbf{P}_3\boldsymbol{\nu} + \mathbf{P}_4\boldsymbol{\eta} \end{pmatrix} \tag{2.2.10}$$

Obviously, the upper equation yields parts of the solution, which decrease for $x \to \infty$ and increase for $x \to -\infty$. The opposite is true for the lower equation. We define the *left exterior problem* for $x < x_L$ analogous to the right exterior problem and denote the occurring matrices with "$\tilde{}$". Then, an analogous equation holds for the left exterior problem. Thus the *transformed transparent boundary conditions* for the left (a) and right (b) boundary is obtained by extinguishing the respectively increasing parts of the exterior

solutions:

$$\tilde{\mathbf{P}}_2\hat{\boldsymbol{\varphi}}_x(0,s) = -\tilde{\mathbf{P}}_1\hat{\boldsymbol{\varphi}}(0,s), \tag{2.2.11a}$$

$$\mathbf{P}_4\hat{\boldsymbol{\varphi}}_x(L,s) = -\mathbf{P}_3\hat{\boldsymbol{\varphi}}(L,s). \tag{2.2.11b}$$

If the matrices $\tilde{\mathbf{P}}_2$ and $\mathbf{P}_4$ are regular, then the Laplace-transformed TBC can be written in Dirichlet-to-Neumann form. It is not clear, if these matrices are regular, but for all considered examples this applies.

In the remainder of this section we will prove Thm. 2.4 proceeding as follows: Lem. 2.5 will show that under certain assumptions the inertia of $\mathbf{A}^{-1}\mathbf{B}$ can be ascribed to the inertia of $\mathbf{A}$. Thereafter we will show, that exactly d eigenvalues of $\mathbf{A}$ are positive and d are negative.

LEMMA 2.5 (Lemma 2 in [**CS63**]). *Let $\mathbf{F}, \mathbf{G}$ be $d \times d$-matrices with $\mathbf{G}$ Hermitian and regular, suppose $\mathbf{H} := \mathbf{GF} + \mathbf{F}^H\mathbf{G}$ is positive semi-definite and $i_0(\mathbf{F}) = 0$. Then $i(\mathbf{F}) = i(\mathbf{G})$.*

This semi-definite case of the *general inertia theorem* was proved by Carlson and Schneider by showing that for $\mathbf{G}$ Hermitian and regular and $\mathbf{H}$ positive semi-definite then $\mathrm{In}(\mathbf{F}) \leq \mathrm{In}(\mathbf{G})$ $\big($with Theorem 1 of [**OS62**] and where $\mathrm{In}(\mathbf{F}) \leq \mathrm{In}(\mathbf{G})$ means: $i_+(\mathbf{G}) \leq i_+(\mathbf{F})$ and $i_-(\mathbf{G}) \leq i_-(\mathbf{F})\big)$ and $\mathrm{In}(\mathbf{G}) \leq \mathrm{In}(\mathbf{F})$ if $i_0(\mathbf{F}) = 0$ and $\mathbf{H}$ Hermitian and positive semi-definite. A similar result was verified by Chen in [**Che73**]: the condition "$\mathbf{G}$ regular" can be weakened to "the rang of the $d \times d^2$ matrix $[\mathbf{H}\,\mathbf{FH}\,\ldots,\mathbf{F}^{d-1}\mathbf{H}]$ is maximal".

We will now check the assumptions of Lem. 2.5. Since $\mathbf{M}$ is Hermitian, $\mathbf{A} = \mathbf{A}^H$ as well. We already showed, that $\mathbf{G} := \mathbf{A}$ is regular. For $\mathbf{F} := \mathbf{A}^{-1}\mathbf{B}$ we have

$$\begin{aligned} \mathbf{H} &= \mathbf{GF} + \mathbf{F}^H\mathbf{G} = \mathbf{A}(\mathbf{A}^{-1}\mathbf{B}) + \mathbf{B}^H(\mathbf{A}^{-1})^H\mathbf{A}^H \\ &= \mathbf{B} + \mathbf{B}^H = \begin{pmatrix} 2\mathrm{Re}(s)\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \geq 0. \end{aligned} \tag{2.2.12}$$

It remains to show, that $i_0(\mathbf{A}^{-1}\mathbf{B}) = 0$. Therefore we prove the following lemma:

LEMMA 2.6. *For $Re(s) > 0$ the matrix $\mathbf{A}^{-1}\mathbf{B}$ has no purely imaginary eigenvalues.*

PROOF. We assume that $i\lambda$ with $\lambda \in \mathbb{R}$ is eigenvalue of $\mathbf{A}^{-1}\mathbf{B}$. In that case $\hat{\boldsymbol{\varphi}}(x) = \check{\boldsymbol{\varphi}}e^{i\lambda x}$ is a solution of (2.2.4) and yields

$$is\check{\boldsymbol{\varphi}} = (i\alpha - \xi)\check{\boldsymbol{\varphi}} = (\mathbf{N}\lambda^2 - \mathbf{M}\lambda + V)\check{\boldsymbol{\varphi}}. \tag{2.2.13}$$

This means, that $i\alpha - \xi$ is an eigenvalue of $\mathbf{N}\lambda^2 - \mathbf{M}\lambda + V$. But since $\mathbf{N}\lambda^2 - \mathbf{M}\lambda + V$ is - as a sum of Hermitian matrices - again Hermitian, all its eigenvalues must be real, and therefore $\alpha = 0$ which is a contradiction. □

CONCLUSION 2.7. *For any eigenvalue λ of $\mathbf{A}^{-1}\mathbf{B}$, $Re(\lambda) = 0$ implies $\lambda = 0$. Thus, since $\mathbf{A}^{-1}\mathbf{B}$ is regular, we have $i_0(\mathbf{A}^{-1}\mathbf{B}) = 0$.*

Finally, it remains to verify, that d eigenvalues of $\mathbf{A}$ have positive and d have negative real parts. Therefore we will use a continuity argument: we consider the matrix

$$\mathbf{A}(\epsilon) := \begin{pmatrix} \epsilon\mathbf{M} & i\mathbf{N} \\ -i\mathbf{N} & \mathbf{0} \end{pmatrix}, \qquad \epsilon \in [0,1]. \tag{2.2.14}$$

$\mathbf{A}(0)$ has d positive and d negative eigenvalues, which are given by

$$\lambda^0_{2k-1} = n_{k,k} \text{ and } \lambda^0_{2k} = -n_{k,k}, \qquad k = 1, \ldots, d. \tag{2.2.15}$$

Furthermore for all $\epsilon \in [0,1]$ the matrix $\mathbf{A}(\epsilon)$ has no zero eigenvalue (cf. 2.2.8). Then for ϵ from zero to one d eigenvalues of $\mathbf{A}(\epsilon)$ are positive and d are negative, since the eigenvalues are continuous in ϵ.

Thus, $i(\mathbf{A}) = (d, d, 0)$ holds and with Lem. 2.5 follows $i(\mathbf{A}^{-1}\mathbf{B}) = (d, d, 0)$, if $\mathrm{Re}(s) > 0$, which finishes the proof of Thm. 2.4.

3. Discretisation

As in Chap. 1 we do not discretise equation (2.2.11) (by a numerical inverse Laplace transformation), but derive discrete TBCs for a discretisation of (2.0.3). For the discretisation we choose a uniform grid with the step sizes h in space and k in time: $x_j = K + jh, t_n = nk$ with $j = 0, \ldots, J$, $n = 0, \ldots, N$. We discretise (2.0.3) using the Crank-Nicolson scheme in time and the central differences for the first and second spatial derivatives. The *discrete $k \cdot p$-Schrödinger equation* then reads

$$i\frac{h^2}{k}(\boldsymbol{\varphi}_j^{n+1} - \boldsymbol{\varphi}_j^n) = -\Delta^0_{\frac{h}{2}}(\mathbf{N}_j \Delta^0_{\frac{h}{2}} \boldsymbol{\varphi}_j^{n+\frac{1}{2}}) + \mathbf{M}_{Sj}\Delta^0 \boldsymbol{\varphi}_j^{n+\frac{1}{2}} - \Delta^0(\mathbf{M}^H_{Sj}\boldsymbol{\varphi}_j^{n+\frac{1}{2}}) + V_j \boldsymbol{\varphi}_j^{n+\frac{1}{2}} \tag{2.3.1}$$

for $j = 1, \ldots, J-1$ and $n = 0, \ldots, N$ with the difference operators

$$\Delta^0_{\frac{h}{2}} \boldsymbol{\varphi}_j^n = \boldsymbol{\varphi}^n_{j+\frac{1}{2}} - \boldsymbol{\varphi}^n_{j-\frac{1}{2}}, \tag{2.3.2a}$$

$$\Delta^0 \boldsymbol{\varphi}_j^n = \frac{1}{2}(\Delta^+ + \Delta^-)\boldsymbol{\varphi}_j^n = \boldsymbol{\varphi}^n_{j+1} - \boldsymbol{\varphi}^n_{j-1} \tag{2.3.2b}$$

and $\boldsymbol{\varphi}_j^{n+\frac{1}{2}} = \frac{\boldsymbol{\varphi}_j^{n+1}+\boldsymbol{\varphi}_j^n}{2}$.

3.1. Properties of the discrete equation. An appropriate discretisation scheme should carry over properties of the continuous equation to the difference equation. This is the case for the Crank-Nicolson scheme: it conserves the whole-space l^2-norm and thus it is unconditionally stable for the whole-space problem. To shows this, we first observe, that

(2.3.3)

$$\begin{aligned}\triangle_t^+ ||\boldsymbol{\varphi}^n||_{l^2}^2 &= \sum_{j=-\infty}^{\infty} \triangle_t^+ \left(\left(\boldsymbol{\varphi}_j^n\right)^H \boldsymbol{\varphi}_j^n \right) = \sum_{j=-\infty}^{\infty} \left(\left(\boldsymbol{\varphi}_j^{n+\frac{1}{2}}\right)^H \triangle_t^+ \boldsymbol{\varphi}_j^n + (\triangle_t^+ \boldsymbol{\varphi}_j^n)^H \boldsymbol{\varphi}_j^{n+\frac{1}{2}} \right) \\ &= 2\mathrm{Re} \left(\sum_{j=-\infty}^{\infty} \left(\boldsymbol{\varphi}_j^{n+\frac{1}{2}}\right)^H \triangle_t^+ \boldsymbol{\varphi}_j^n \right) = 2\mathrm{Im} \left(\sum_{j=-\infty}^{\infty} i \left(\boldsymbol{\varphi}_j^{n+\frac{1}{2}}\right)^H \triangle_t^+ \boldsymbol{\varphi}_j^n \right).\end{aligned}$$

Using this equality together with (2.3.1) and summation by parts yields

$$\begin{aligned}\frac{h^2}{k} \triangle_t^+ ||\boldsymbol{\varphi}^n||_{l^2}^2 &= 2\mathrm{Im} \left(-\sum_{j=-\infty}^{\infty} \left(\boldsymbol{\varphi}_j^{n+\frac{1}{2}}\right)^H \triangle_{\frac{h}{2}}^0 \left(\mathbf{N}_j \triangle_{\frac{h}{2}}^0 \boldsymbol{\varphi}_j^{n+\frac{1}{2}} \right) + \sum_{j=-\infty}^{\infty} \left(\boldsymbol{\varphi}_j^{n+\frac{1}{2}}\right)^H \mathbf{M}_{Sj} \triangle^0 \boldsymbol{\varphi}_j^{n+\frac{1}{2}} \right. \\ &\quad \left. - \sum_{j=-\infty}^{\infty} \left(\boldsymbol{\varphi}_j^{n+\frac{1}{2}}\right)^H \triangle^0 \left(\mathbf{M}_{Sj}^H \boldsymbol{\varphi}_j^{n+\frac{1}{2}} \right) + \sum_{j=-\infty}^{\infty} \left(\boldsymbol{\varphi}_j^{n+\frac{1}{2}}\right)^H \mathbf{V}_j \boldsymbol{\varphi}_j^{n+\frac{1}{2}} \right) \qquad (2.3.4) \\ &= \mathrm{Im} \left(\sum_{j=-\infty}^{\infty} \left(\triangle_{\frac{h}{2}}^0 \boldsymbol{\varphi}_j^{n+\frac{1}{2}} \right)^H \mathbf{N}_j \triangle_{\frac{h}{2}}^0 \boldsymbol{\varphi}_j^{n+\frac{1}{2}} \right. \\ &\quad \left. + \sum_{j=-\infty}^{\infty} \left(\boldsymbol{\varphi}_j^{n+\frac{1}{2}}\right)^H \mathbf{M}_{Sj} \triangle^0 \boldsymbol{\varphi}_j^{n+\frac{1}{2}} + \sum_{j=-\infty}^{\infty} \left(\left(\boldsymbol{\varphi}_j^{n+\frac{1}{2}}\right)^H \mathbf{M}_{Sj} \triangle^0 \boldsymbol{\varphi}_j^{n+\frac{1}{2}} \right)^H \right) \\ &= 0,\end{aligned}$$

because the matrices $\mathbf{N}_j$ and $\mathbf{V}_j$ are Hermitian. Thus, for the whole-space problem the discrete l^2-norm is constant in time.

4. Discrete transparent boundary conditions

For the case of a scalar Schrödinger equation Arnold [**Arn98**] derived a discrete transparent boundary condition. This DTBC is reflection-free compared to the discrete whole-space solution and conserves the stability properties of the whole-space Crank-Nicolson scheme. The DTBC has the form of a discrete convolution. The convolution coefficients

are a function of Legendre polynomials but can be obtained more easily by a three-term recurrence formula. Ehrhardt and Arnold showed in [**EA01**], that the imaginary parts of the convolution coefficients are not decaying and therefore introduced summed coefficients.

To derive the DTBC for (2.3.1) we solve the $\mathcal{Z}$-transformed system of ordinary difference equations in the exterior domain. Then all its solutions are determined by eigenvalues and eigenvectors, which can be distinguished into decaying and increasing solutions by the absolute value of the involved eigenvalue. We obtain the DTBC by claiming, that no influence of increasing solutions exists.

In the exterior space $j \geq J$ ($x_J = x_R$) the Crank-Nicolson scheme (2.3.1) simplifies to

$$i\frac{h^2}{k}(\boldsymbol{\varphi}_j^{n+1} - \boldsymbol{\varphi}_j^n) = -\mathbf{N}\Delta^+\Delta^-\boldsymbol{\varphi}_j^{n+1/2} + ih\mathbf{M}\frac{1}{2}(\Delta^+ + \Delta^-)\boldsymbol{\varphi}_j^{n+1/2} + h^2\mathbf{V}\boldsymbol{\varphi}_j^{n+1/2} \tag{2.4.1}$$

for $j \geq J$ and $n \geq 0$. The $\mathcal{Z}$-transformation given by

$$\mathcal{Z}\{\boldsymbol{\varphi}_j^n\} = \hat{\boldsymbol{\varphi}}_j(z) := \sum_{n=0}^{\infty} z^{-n}\boldsymbol{\varphi}_j^n, \qquad z \in \mathbb{C}, \qquad |z| > 1, \tag{2.4.2}$$

transforms (2.4.1) to

$$2i\frac{h^2}{k}\frac{z-1}{z+1}\hat{\boldsymbol{\varphi}}_j = -\mathbf{N}\Delta^+\Delta^-\hat{\boldsymbol{\varphi}}_j + ih\mathbf{M}\frac{1}{2}(\Delta^+ + \Delta^-)\hat{\boldsymbol{\varphi}}_j + h^2\mathbf{V}\hat{\boldsymbol{\varphi}}_j, \qquad j \geq J. \tag{2.4.3}$$

LEMMA 2.8. *If the solution of the $\mathcal{Z}$-transformed exterior problem* (2.4.3) *with the boundary data*

$$\hat{\boldsymbol{\varphi}}_{j=J} = \hat{\boldsymbol{\varphi}}_J, \qquad \hat{\boldsymbol{\varphi}}_\infty = 0 \tag{2.4.4}$$

exists, it is unique.

PROOF. We assume, that there exist two solutions of (2.4.3), (2.4.4) $\hat{\boldsymbol{\varphi}}_1$ and $\hat{\boldsymbol{\varphi}}_2$. The difference $\hat{\boldsymbol{\varphi}} = \hat{\boldsymbol{\varphi}}_1 - \hat{\boldsymbol{\varphi}}_2$ is then a solution of (2.4.3) with homogeneous boundary data. For this solution $\hat{\boldsymbol{\varphi}}$ we consider (2.4.3) multiplied by $\hat{\boldsymbol{\varphi}}_j^H$ from the left and take imaginary

parts:

$$
\begin{aligned}
(2.4.5) \qquad 0 &= \operatorname{Im}\left(2i\frac{h^2}{k}\frac{z-1}{z+1}\sum_{j=J}^{\infty}|\hat{\boldsymbol{\varphi}}_j|^2+\sum_{j=J}^{\infty}\hat{\boldsymbol{\varphi}}_j^H\mathbf{N}\triangle^+\triangle^-\hat{\boldsymbol{\varphi}}_j\right.\\
&\quad \left.-\frac{1}{2}ih\sum_{j=J}^{\infty}\hat{\boldsymbol{\varphi}}_j^H\mathbf{M}\left(\triangle^+-\triangle^-\right)\hat{\boldsymbol{\varphi}}_j-h^2\sum_{j=J}^{\infty}\hat{\boldsymbol{\varphi}}_j^H\mathbf{V}\hat{\boldsymbol{\varphi}}_j\right)\\
&= 2\frac{h^2}{k}\operatorname{Re}\left(\frac{z-1}{z+1}\right)\sum_{j=J}^{\infty}|\hat{\boldsymbol{\varphi}}_j|^2+\operatorname{Im}\left(-\sum_{j=J}^{\infty}\triangle^-\hat{\boldsymbol{\varphi}}_j^H\mathbf{N}\triangle^-\hat{\boldsymbol{\varphi}}_j\right.\\
&\quad \left.-\frac{1}{2}ih\sum_{j=J}^{\infty}\hat{\boldsymbol{\varphi}}_j^H\mathbf{M}\triangle^-\hat{\boldsymbol{\varphi}}_j+\frac{1}{2}ih\sum_{j=J}^{\infty}\left(\hat{\boldsymbol{\varphi}}_j^H\mathbf{M}\triangle^-\hat{\boldsymbol{\varphi}}_j\right)^H\right)\geq 0,
\end{aligned}
$$

if $|z|>1$ (cf. (1.7.35)) and even strictly larger than zero, if $\sum_{j=J}^{\infty}|\hat{\boldsymbol{\varphi}}_j|^2\neq 0$. But this is a contradiction. □

REMARK 2.9. *Analogously to the continuous problem the existence of a solution is guaranteed by the regularity of the $S\times S$ principal submatrix of the matrix of right eigenvectors (cf. Lem. 1.8 and Rem. 1.9), which holds for all considered examples.*

We proceed to solve the $\mathcal{Z}$-transformed exterior problem and define $\hat{\boldsymbol{\xi}}_j=\triangle^-\hat{\boldsymbol{\varphi}}_j$ and use the *reduction of order method*, i.e. we write (2.4.3) as a system of first order difference equations

$$
(2.4.6) \qquad \begin{pmatrix} i\frac{h}{2}\mathbf{M} & -\mathbf{N}\\ -\mathbf{I} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\triangle^+\hat{\boldsymbol{\varphi}}_j\\ \triangle^+\hat{\boldsymbol{\xi}}_j\end{pmatrix}=\begin{pmatrix} h^2 2\frac{z-1}{z+1}\frac{1}{k}i\mathbf{I}-h^2\mathbf{V} & -i\frac{h}{2}\mathbf{M}\\ \mathbf{0} & -\mathbf{I}\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{\varphi}}_j\\ \hat{\boldsymbol{\xi}}_j\end{pmatrix},
$$

i.e.

$$
(2.4.7) \quad \begin{pmatrix}\triangle^+\hat{\boldsymbol{\varphi}}_j\\ \triangle^+\hat{\boldsymbol{\xi}}_j\end{pmatrix}=\mathbf{A}^{-1}\mathbf{B}\begin{pmatrix}\hat{\boldsymbol{\varphi}}_j\\ \hat{\boldsymbol{\xi}}_j\end{pmatrix} \qquad \text{or} \qquad \begin{pmatrix}\hat{\boldsymbol{\varphi}}_{j+1}\\ \hat{\boldsymbol{\xi}}_{j+1}\end{pmatrix}=(\mathbf{A}^{-1}\mathbf{B}+\mathbf{I})\begin{pmatrix}\hat{\boldsymbol{\varphi}}_j\\ \hat{\boldsymbol{\xi}}_j\end{pmatrix}.
$$

The regularity of $\mathbf{A}$ follows from Lem. 2.13.

Solutions of (2.4.3),that are constructed with an eigenvalue λ of $\mathbf{A}^{-1}\mathbf{B}+\mathbf{I}$, are decaying for $x\to\infty$ if $|\lambda+1|<1$ and increasing if $|\lambda+1|>1$ (cf. Sec. 5.1). Analogously to Thm. 2.4 we prove a splitting property of $\mathbf{A}^{-1}\mathbf{B}+\mathbf{I}$:

THEOREM 2.10 (Discrete Splitting Theorem). *d of the $2d$ eigenvalues of $\mathbf{A}^{-1}\mathbf{B}+\mathbf{I}$ have an absolute value larger than unity and d have a smaller absolute value, if $|z|\neq 1$.*

A proof of Thm. 2.10 will be given succeeding to the DTBC at the end of this section.

If the eigenvalues $\lambda_1, \dots, \lambda_{2d}$ of $\mathbf{A}^{-1}\mathbf{B}$ split into two commensurate groups, then the solutions involving those with $|\lambda_k + 1| < 1$ for $k = 1, \dots, d$ decay for $j \to \infty$ and those with $|\lambda_k + 1| > 1$ for $k = d, \dots, 2d$ decay for $j \to -\infty$.

Thus, we may again split the Jordan form $\mathbf{J} = \left(\begin{smallmatrix} \mathbf{J}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{smallmatrix}\right)$ of $\mathbf{A}^{-1}\mathbf{B} + \mathbf{I}$, $\mathbf{J}_1$ containing the Jordan blocks corresponding to solutions decaying for $j \to \infty$ and $\mathbf{J}_2$ those which increase. With the matrix of left eigenvectors $\mathbf{P}^{-1} = \left(\begin{smallmatrix} \mathbf{P}_1 & \mathbf{P}_2 \\ \mathbf{P}_3 & \mathbf{P}_4 \end{smallmatrix}\right)$ the equation

$$\begin{aligned} \mathbf{P}^{-1} \begin{pmatrix} \hat{\boldsymbol{\varphi}}_{j+1} \\ \hat{\boldsymbol{\xi}}_{j+1} \end{pmatrix} &= \mathbf{P}^{-1}(\mathbf{A}^{-1}\mathbf{B} + \mathbf{I}) \begin{pmatrix} \hat{\boldsymbol{\varphi}}_j \\ \hat{\boldsymbol{\xi}}_j \end{pmatrix} = \mathbf{P}^{-1}\mathbf{P} \begin{pmatrix} \mathbf{J}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{pmatrix} \begin{pmatrix} \mathbf{P}_1 & \mathbf{P}_2 \\ \mathbf{P}_3 & \mathbf{P}_4 \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\varphi}}_j \\ \hat{\boldsymbol{\xi}}_j \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{J}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{pmatrix} \begin{pmatrix} \mathbf{P}_1\hat{\boldsymbol{\varphi}}_j + \mathbf{P}_2\hat{\boldsymbol{\xi}}_j \\ \mathbf{P}_3\hat{\boldsymbol{\varphi}}_j + \mathbf{P}_4\hat{\boldsymbol{\xi}}_j \end{pmatrix} \end{aligned} \tag{2.4.8}$$

holds and the *transformed discrete transparent boundary conditions* read

$$\tilde{\mathbf{P}}_1\hat{\boldsymbol{\varphi}}_1 + \tilde{\mathbf{P}}_2\hat{\boldsymbol{\xi}}_1 = 0, \tag{2.4.9a}$$

$$\mathbf{P}_3\hat{\boldsymbol{\varphi}}_J + \mathbf{P}_4\hat{\boldsymbol{\xi}}_J = 0 \tag{2.4.9b}$$

for the left (a) and right (a) boundary respectively.

REMARK 2.11. *In all considered examples the matrices* $\mathbf{P}_1, \dots, \mathbf{P}_4$ *and* $\tilde{\mathbf{P}}_1, \dots, \tilde{\mathbf{P}}_4$ *were regular, but this is not clear in general.*

For regular matrices $\mathbf{P}_4$ and $\tilde{\mathbf{P}}_2$ the $\mathcal{Z}$-transformed DTBC can be given in Dirichlet-to-Neumann form

$$\Delta^-\hat{\boldsymbol{\varphi}}_1 = \hat{\tilde{\mathbf{D}}}\hat{\boldsymbol{\varphi}}_1, \tag{2.4.10a}$$

$$\Delta^-\hat{\boldsymbol{\varphi}}_J = \hat{\mathbf{D}}\hat{\boldsymbol{\varphi}}_J, \tag{2.4.10b}$$

where $\hat{\mathbf{D}} = -\mathbf{P}_4^{-1}\mathbf{P}_3$ and $\hat{\tilde{\mathbf{D}}} = -\tilde{\mathbf{P}}_2^{-1}\tilde{\mathbf{P}}_1$. After an inverse $\mathcal{Z}$-transformation the *discrete transparent boundary conditions* read

$$\boldsymbol{\varphi}_1^{n+1} - \boldsymbol{\varphi}_0^{n+1} - \tilde{\mathbf{D}}^0\boldsymbol{\varphi}_1^{n+1} = \sum_{k=1}^{n} \tilde{\mathbf{D}}^{n+1-k}\boldsymbol{\varphi}_1^k, \tag{2.4.11a}$$

$$\boldsymbol{\varphi}_J^{n+1} - \boldsymbol{\varphi}_{J-1}^{n+1} - \mathbf{D}^0\boldsymbol{\varphi}_J^{n+1} = \sum_{k=1}^{n} \mathbf{D}^{n+1-k}\boldsymbol{\varphi}_J^k. \tag{2.4.11b}$$

REMARK 2.12. *Note, that in equation (2.4.10a) and (2.4.11a) the left boundary condition is given at $j = 1$. Of course, the boundary condition can also be formulated at $j = 0$ using $\hat{\boldsymbol{\xi}}_j = \Delta^+ \hat{\boldsymbol{\varphi}}_j$. This changes the lower row in $\mathbf{A}$ and $\mathbf{B}$ and thus the matrix $\tilde{\mathbf{D}}$ differs from $\mathbf{D}$. Posing the boundary condition at $j = 1$ as we do, has the advantage, that these matrices coincide, if the constant coefficients for $x < x_L$ and $x > x_R$ are equal, what occurs often in the application. In that case we can reduce the numerical effort to calculate the convolution coefficients by half.*

As for parabolic systems we will define summed coefficients (cf. Sect. 7.2.4 in Chap. 1). For a scalar Schrödinger equation Ehrhardt and Arnold showed in [**EA01**] that the imaginary parts of the coefficients were not decaying but oscillating. Therefore they introduced summed coefficients. These decay rapidly like $O(n^{-3/2})$. Since the scalar equation is as a special case included in our system, it suggests itself to use the summed coefficients, although we can give no asymptotic behaviour of the systems' coefficients, because they cannot be formulated explicitly. In Sec. 5 we will give some examples of the numerically calculated coefficients. The diagonal elements show the same properties as those for the scalar case. For the summed coefficients $\widehat{\tilde{\mathbf{S}}}_{s,l} = \frac{z+1}{z}\widehat{\tilde{\mathbf{D}}}_{s,l}$ and $\widehat{\mathbf{S}}_{s,l} = \frac{z+1}{z}\widehat{\mathbf{D}}_{s,l}$ the DTBCs read

$$\boldsymbol{\varphi}_1^{n+1} - \boldsymbol{\varphi}_0^{n+1} - \tilde{\mathbf{S}}^0 \boldsymbol{\varphi}_1^{n+1} = \sum_{k=1}^{n} \tilde{\mathbf{S}}^{n+1-k} \boldsymbol{\varphi}_1^k - \boldsymbol{\varphi}_1^n + \boldsymbol{\varphi}_0^n, \tag{2.4.12a}$$

$$\boldsymbol{\varphi}_J^{n+1} - \boldsymbol{\varphi}_{J-1}^{n+1} - \mathbf{S}^0 \boldsymbol{\varphi}_J^{n+1} = \sum_{k=1}^{n} \mathbf{S}^{n+1-k} \boldsymbol{\varphi}_J^k - \boldsymbol{\varphi}_J^n + \boldsymbol{\varphi}_{J-1}^n. \tag{2.4.12b}$$

It remains to prove Thm. 2.10. Therefore, we will first show in Lem. 2.13, that no eigenvalue of $\mathbf{A}^{-1}\mathbf{B} + \mathbf{I}$ has an absolute value of one. Then we will show the asserted splitting of the eigenvalues for $\mathbf{M} = 0$ and argue, that due to the continuity of the eigenvalues the border $|\lambda| = 1$ cannot be crossed.

LEMMA 2.13. *For $|z| \neq 1$ the matrix $\mathbf{A}^{-1}\mathbf{B} + \mathbf{I}$ has no eigenvalue λ with $|\lambda| = 1$.*

PROOF. Assume that $\lambda = a + bi$ with $|\lambda| = 1$ were an eigenvalue of the discrete problem (2.4.3). Then $\hat{\boldsymbol{\varphi}}_j = \lambda^j \hat{\boldsymbol{\varphi}}_0$ is a solution of (2.4.3). For the differences the following

expressions are valid:

$$\Delta^+\hat{\boldsymbol{\varphi}}_j = \lambda^{j-1}\hat{\boldsymbol{\varphi}}_0\left(\lambda^2-\lambda\right) \tag{2.4.13a}$$

$$\Delta^-\hat{\boldsymbol{\varphi}}_j = \lambda^{j-1}\hat{\boldsymbol{\varphi}}_0\left(\lambda-1\right) \tag{2.4.13b}$$

$$\Delta^+\Delta^-\hat{\boldsymbol{\varphi}}_j = \lambda^{j-1}\hat{\boldsymbol{\varphi}}_0\left(\lambda^2-2\lambda+1\right). \tag{2.4.13c}$$

Defining $g(z)=\frac{z-1}{z+1}$ and inserting $\hat{\boldsymbol{\varphi}}_j=\lambda^j\hat{\boldsymbol{\varphi}}_0$ in equation (2.4.3) yields

$$\begin{aligned} i\frac{2h^2}{k}g(z)\lambda\hat{\boldsymbol{\varphi}}_0 &= \left(-\mathbf{N}(\lambda^2-2\lambda+1)+i\mathbf{M}\frac{h}{2}(\lambda^2-1)+h^2\mathbf{V}\lambda\right)\hat{\boldsymbol{\varphi}}_0 \\ &= \left(-\mathbf{N}(2a^2+2abi-2(a+bi))+i\mathbf{M}\frac{h}{2}(-2b^2+2abi)+h^2\mathbf{V}\lambda\right)\hat{\boldsymbol{\varphi}}_0 \qquad (2.4.14) \\ &= \left(-\mathbf{N}(\lambda(a-1))+i\mathbf{M}\frac{h}{2}2b(b-ai)+h^2\mathbf{V}\lambda\right)\hat{\boldsymbol{\varphi}}_0, \end{aligned}$$

i.e.

$$i\frac{2h^2}{k}g(z)\hat{\boldsymbol{\varphi}}_0=\left(-\mathbf{N}(a-1)+\mathbf{M}hb+h^2\mathbf{V}\right)\hat{\boldsymbol{\varphi}}_0. \tag{2.4.15}$$

Equation (2.4.15) is an eigenvalue equation for the matrix $-\mathbf{N}(a-1)+\mathbf{M}hb+h^2\mathbf{V}$, which is as a sum of Hermitian matrices again Hermitian, and has therefore only real eigenvalues. Thus $i\frac{2h^2}{k}g(z)$ must be real. We examine this expression further:

$$g(z)=\frac{z-1}{z+1}=\frac{|z|^2-1+2i\,\mathrm{Im}(z)}{|z+1|^2}. \tag{2.4.16}$$

Thus it is obvious, that $g(z)\in i\mathbb{R}$ if and only if $|z|=1$. □

To understand the eigenvalue-splitting for the general case, i.e. equation (2.4.3), we shall now use a perturbation argument and consider the special case $\mathbf{M}=\mathbf{0}$. Then equation (2.4.3) reads

$$2i\frac{h^2}{k}\frac{z-1}{z+1}\hat{\boldsymbol{\varphi}}_j=-\mathbf{N}\Delta^+\Delta^-\hat{\boldsymbol{\varphi}}_j+h^2\mathbf{V}\hat{\boldsymbol{\varphi}}_j. \tag{2.4.17}$$

Exchanging the space index $j\to -j$ yields the identical equation. Thus, both problems have the same solutions and the eigenvalues of $\mathbf{A}^{-1}\mathbf{B}+\mathbf{I}$ are in both cases the same. Since decaying solutions are increasing for $j\to -j$ and vice versa, the eigenvalues must split in d yielding for decaying and d yielding increasing solutions for $|z|\neq 1$ and $j\to\infty$.

To (2.4.17) we add the term $i\epsilon\frac{h}{2}\mathbf{M}(\Delta^++\Delta^-)\boldsymbol{\varphi}$ for $0\le\epsilon\le 1$. Then Lem. 2.13 shows that no eigenvalue λ can have an absolute value one. Since these eigenvalues are continuous

in ϵ(cf. [**HJ99a**]), d eigenvalues must remain inside the unit circle when ϵ varies from 0 to 1 and d eigenvalues stay outside.

This finishes the proof of Thm. 2.10.

4.1. Stability. In Sec. 3.1 we showed that the l^2-norm of the whole-space problem is constant in time. For the interior scheme with the DTBC the l^2-norm is bounded by the l^2-norm of the whole-space problem, because the DTBCs cut off the exterior parts of the solution:

$$||\boldsymbol{\varphi}^n||^2_{l^2(0,J)} \leq ||\boldsymbol{\varphi}^n||^2_{l^2(-\infty,\infty)} = ||\boldsymbol{\varphi}^0||^2_{l^2(-\infty,\infty)}. \tag{2.4.18}$$

Thus the interior scheme with DTBCs constructed from exact convolution coefficients is stable. Since we compute the convolution coefficients numerically we consider (2.3.4) for $j = 0, \ldots, J$. Then there remain some boundary terms due to the summation by parts rule:

$$\begin{aligned}\frac{h^2}{k}\triangle_t^+||\boldsymbol{\varphi}^n||^2_{l^2} = \mathrm{Im}\Big((\boldsymbol{\varphi}_0^{n+\frac{1}{2}})^H\mathbf{N}_L\triangle^+\boldsymbol{\varphi}_0^{n+\frac{1}{2}} + (\boldsymbol{\varphi}_0^{n+\frac{1}{2}})^H\mathbf{M}_{SL}^H\boldsymbol{\varphi}_0^{n+\frac{1}{2}} \\ +(\boldsymbol{\varphi}_J^{n+\frac{1}{2}})^H\mathbf{N}_R\triangle^-\boldsymbol{\varphi}_J^{n+\frac{1}{2}} + (\boldsymbol{\varphi}_J^{n+\frac{1}{2}})^H\mathbf{M}_{SR}^H\boldsymbol{\varphi}_J^{n+\frac{1}{2}}\Big).\end{aligned} \tag{2.4.19}$$

As differences we rather consider $\triangle^+$ and $\triangle^-$ than $\triangle^0_{\frac{h}{2}}$. This is possible, since at the boundaries the coefficient matrix $\mathbf{N}$ is already constant. Thus, if (2.4.19) is non-positive the Crank-Nicolson scheme with DTBCs is stable. Unfortunately, inserting the DTBCs for the differential terms does not help to show this, because we know too little of the properties of the convolution matrices. Instead, we will perform in Sec. 5 a numerical evaluation of the boundary terms in (2.4.19) and thus numerically test the stability for an example.

5. Numerical examples

In this section we will present two numerical examples. The first example is a simple 4×4 system of Schrödinger equations without any external potential. For this example we show the typical behaviour of the convolution coefficients and of the numerical solution. The second example investigates the previous system of Schrödinger equations with a piecewise step function as a potential.

5.1. Example 1: The 4×4 system of free Schrödinger equations. To start with, we use a 4-band structure for light holes, which yields a 4×4 system of Schrödinger equations. We do not consider any external potential, but only the four free coupled Schrödinger equations (2.3.1) with

$$\mathbf{N} = \begin{pmatrix} \gamma & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \gamma \end{pmatrix}$$

$$\mathbf{M}_1 = -\frac{1}{2}\frac{\gamma_3}{\gamma_1 + 2\gamma_2}\sqrt{3}i \begin{pmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \qquad \mathbf{M}_2 = -i\mathbf{M}_1$$

$$\mathbf{U}_{11} = \frac{1}{2}\frac{1}{\gamma_1 + 2\gamma_2} \begin{pmatrix} \gamma_1 + \gamma_2 & 0 & -\sqrt{3}\gamma_2 & 0 \\ 0 & \gamma_1 - \gamma_2 & 0 & -\sqrt{3}\gamma_2 \\ -\sqrt{3}\gamma_2 & 0 & \gamma_1 - \gamma_2 & 0 \\ 0 & -\sqrt{3}\gamma_2 & 0 & \gamma_1 + \gamma_2 \end{pmatrix}$$

$$\mathbf{U}_{22} = \frac{1}{2}\frac{1}{\gamma_1 + 2\gamma_2} \begin{pmatrix} \gamma_1 + \gamma_2 & 0 & \sqrt{3}\gamma_2 & 0 \\ 0 & \gamma_1 - \gamma_2 & 0 & \sqrt{3}\gamma_2 \\ \sqrt{3}\gamma_2 & 0 & \gamma_1 - \gamma_2 & 0 \\ 0 & \sqrt{3}\gamma_2 & 0 & \gamma_1 + \gamma_2 \end{pmatrix}$$

$$\mathbf{U}_{12} + \mathbf{U}_{21} = \frac{1}{\gamma_1 + 2\gamma_2}\sqrt{3}\gamma_2 i \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}$$

$k_1 = 2.3$,$\gamma = \frac{\gamma_1 - 2\gamma_2}{\gamma_1 + 2\gamma_2}$, $\gamma_1 = 4 \cdot 6.85$,$\gamma_2 = 4 \cdot 2.1$,$\gamma_3 = 4 \cdot 2.9$.

At first we will investigate the convolution coefficients $\mathbf{D}^n_{k,l}$. Their real and imaginary parts are given in Fig. 2.1 and Fig. 2.2 respectively. We observe, that the overall behaviour of the diagonal elements is equivalent to that in the scalar case: the real parts decay rapidly, but the imaginary parts alternate without visible decay. This behaviour cannot be found for any non-diagonal element: four non-diagonal elements show an opposite behaviour: for $d^n_{1,2}$,$d^n_{2,1}$,$d^n_{3,4}$ and $d^n_{4,3}$ the real parts alternate, whereas the imaginary parts

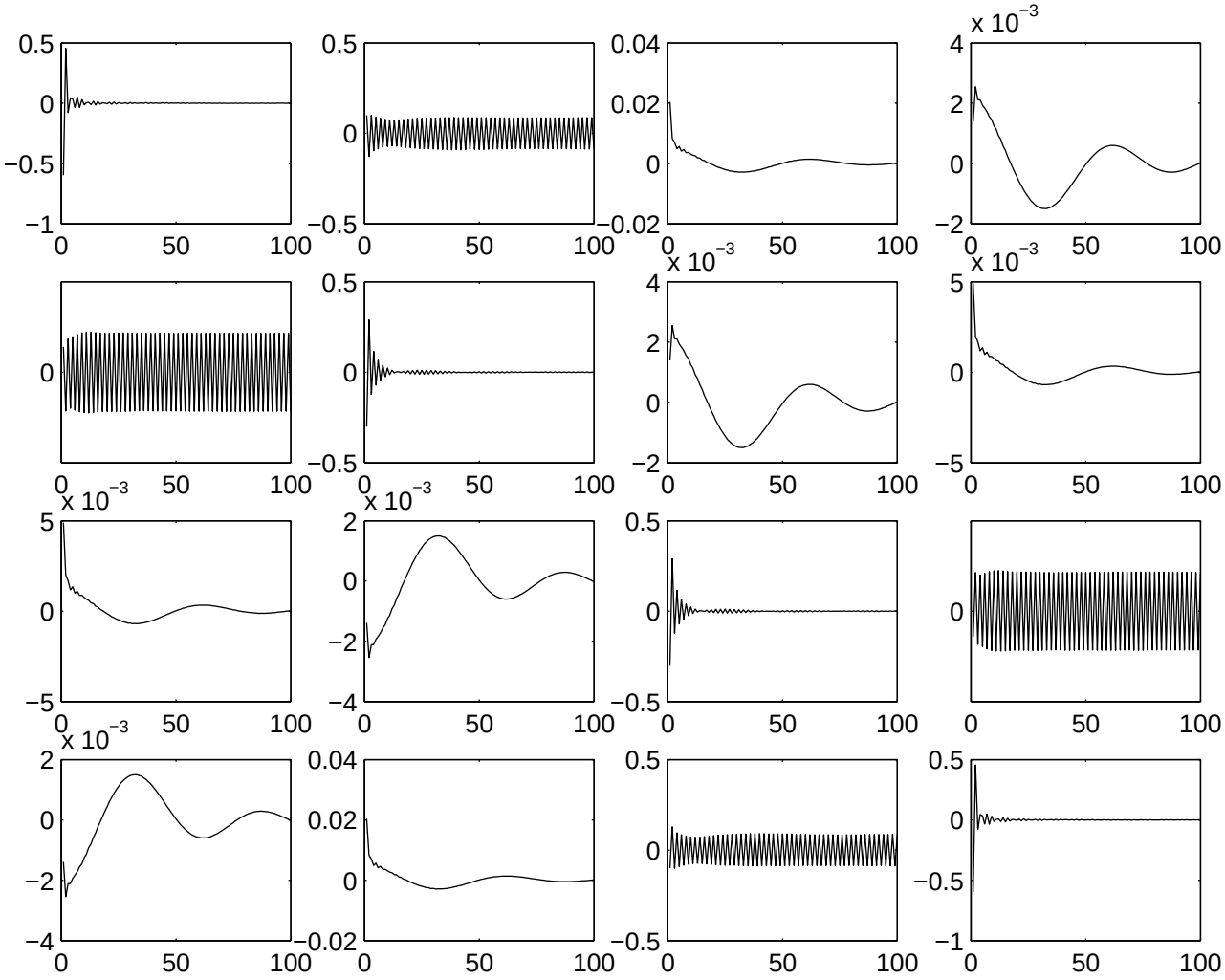


Figure 2.1: Example 1: Real parts of the convolution coefficients $\mathbf{D}^n_{k,l}$

decay rapidly. For the other non-diagonal elements none of the previous behaviour can be observed. They have a much smaller absolute value and decay comparatively slowly. The asymptotic behaviour of the diagonal and four non-diagonal elements with alternating imaginary and respectively real parts, suggest the use of summed convolution coefficients to avoid subtractive cancellation. That this is successful can be seen in the summed convolution coefficients $\mathbf{S}^n_{k,l}$. Their real and imaginary parts are given in Fig. 2.3 and Fig. 2.4 respectively. The eight slowly decaying elements remain unchanged in their qualitative

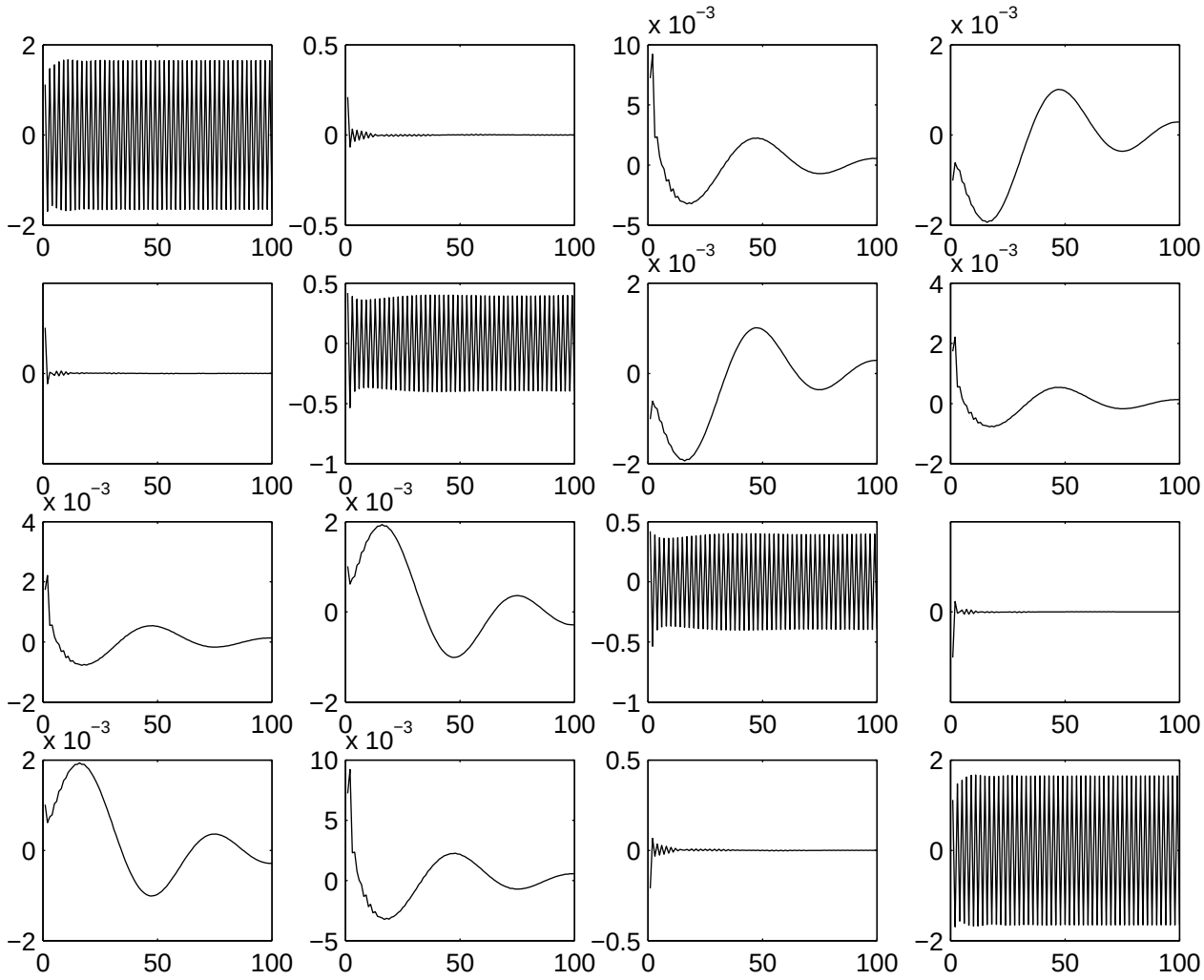


Figure 2.2: Example 1: Imaginary parts of the convolution coefficients $\mathbf{D}^n_{k,l}$

behaviour, but for the diagonal elements and the four non-diagonal elements with alternating real parts the real parts as well as imaginary parts decay rapidly. Therefore we will use the summed convolutions coefficients $\mathbf{S}^n_{k,l}$ for the numerical solution in both examples. The time dependent behaviour of the numerical solution is illustrated in Fig. 2.5 for the first and second component. The fourth and fifth component are nearly equal to zero and thus we will not show them. As initial condition we use a Gaussian wave packet in the second component $\boldsymbol{\varphi}(x,0) = \left((2\pi\sigma^2)^{\frac{1}{4}} \exp\left(ik_r x\right) - \frac{(x-x_0)^2}{\sigma^2}\right) \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$ with $\sigma = 3$, $x_0 = -2\sigma$ and $k_r = \sqrt{6.99}$. We observe, that the right-travelling wave packet after a short time $t = 0.5$

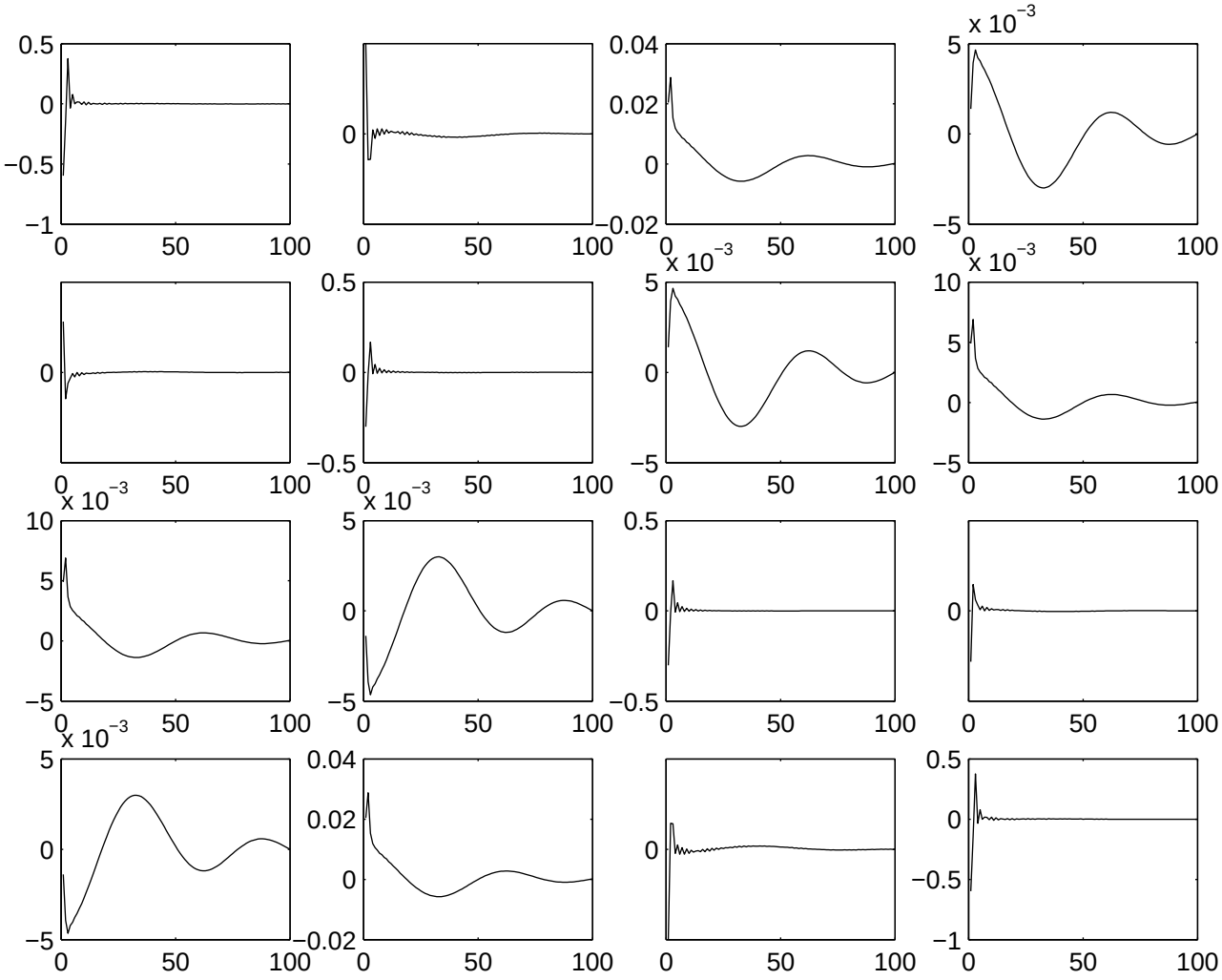


Figure 2.3: Example 1: Real parts of the summed convolution coefficients $\mathbf{S}^n_{k,l}$

decomposes in two wave packets with different velocities. Fig. 2.6 shows the l^2-norm of the solution. After about 130 time steps ($t = 6.5$) the first wave packet has left the computational domain, the second is just about to start leaving it. After 1000 time steps ($t = 50$) nearly all the density left the computational domain. If we disregard the dispersion, the l^2-norm of the solution should be equal to the concatenation of two error functions (*erf*),

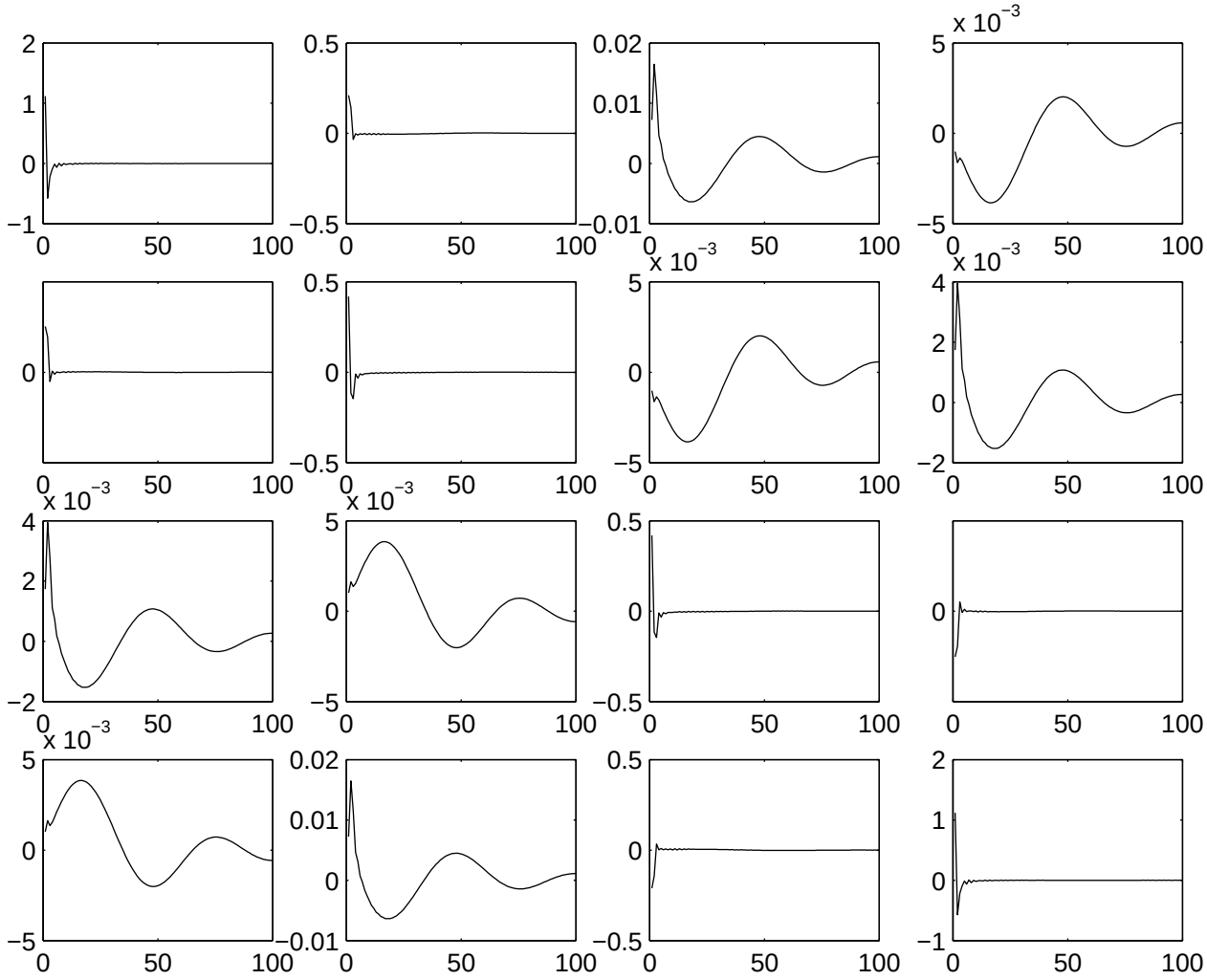


Figure 2.4: Example 1: Imaginary parts of the summed convolution coefficients $\mathbf{S}^n_{k,l}$

one for each wave packet, while it leaves the computational domain. On the r.h.s. of Fig 2.6 we illustrate how good the l^2-norm of the solution can be approximated by two error

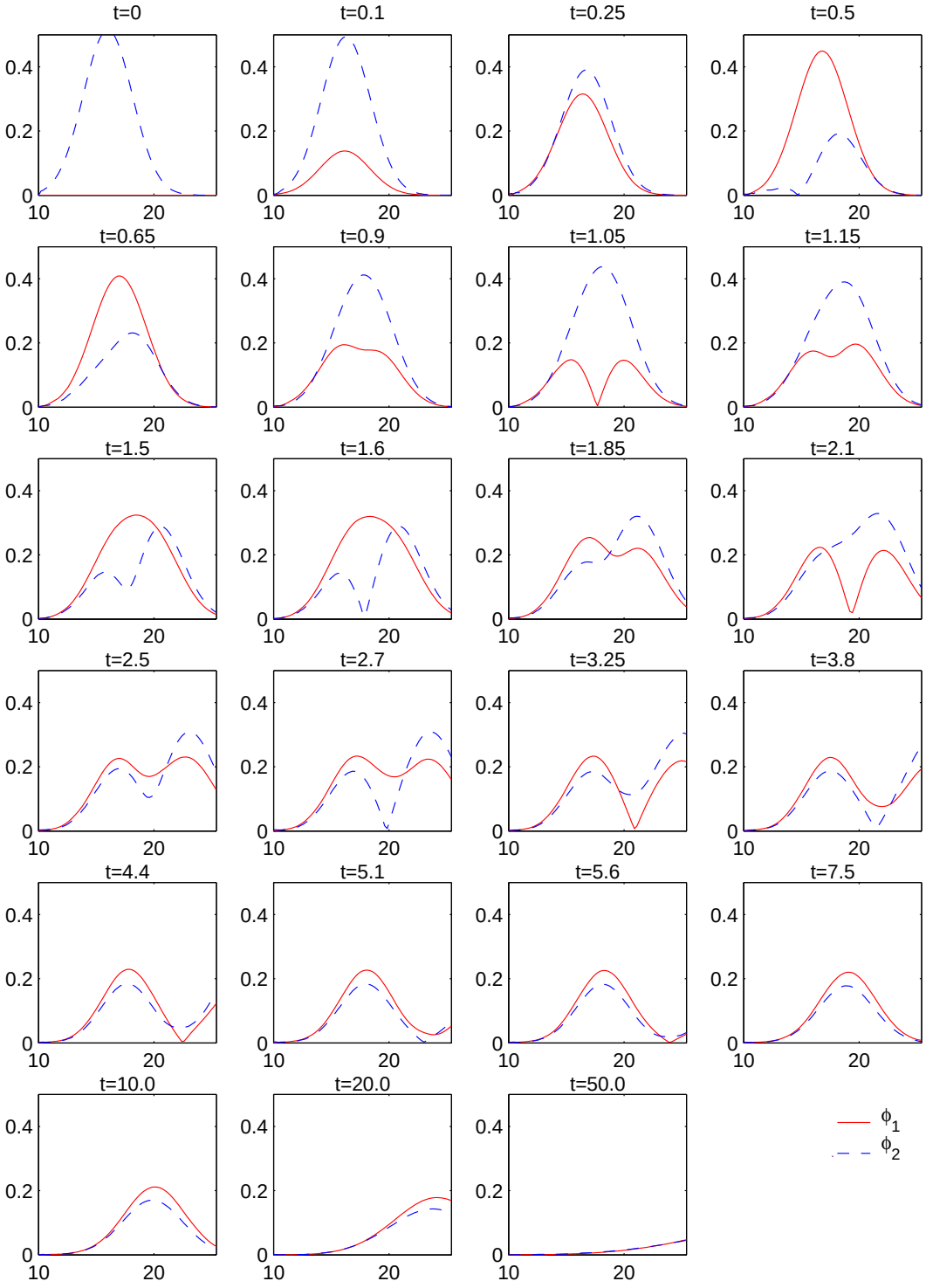


Figure 2.5: Example 1: Time dependent behaviour of $\varphi_{1,j}$ (solid) and $\varphi_{2,j}$ (dashed).

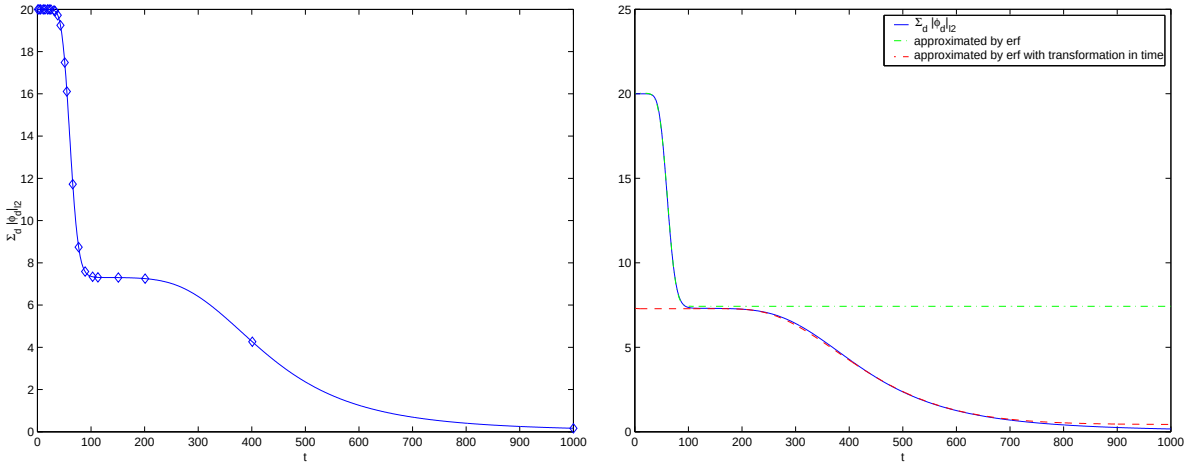


Figure 2.6: Example 1: Overall density $\sum_{l=1}^{d} |\varphi_l|_{l^2}$ with marked points for the time dependent output in Fig. 2.5 (left) and approximated by the error function erf (right)

functions. For the second erf we had to use a transformation in time to get a good fit due to the advanced dispersion.

At the beginning of the computation the density oscillates between the two first components. With advancing time this becomes less. Fig. 2.7 compares the l^2 norms of the

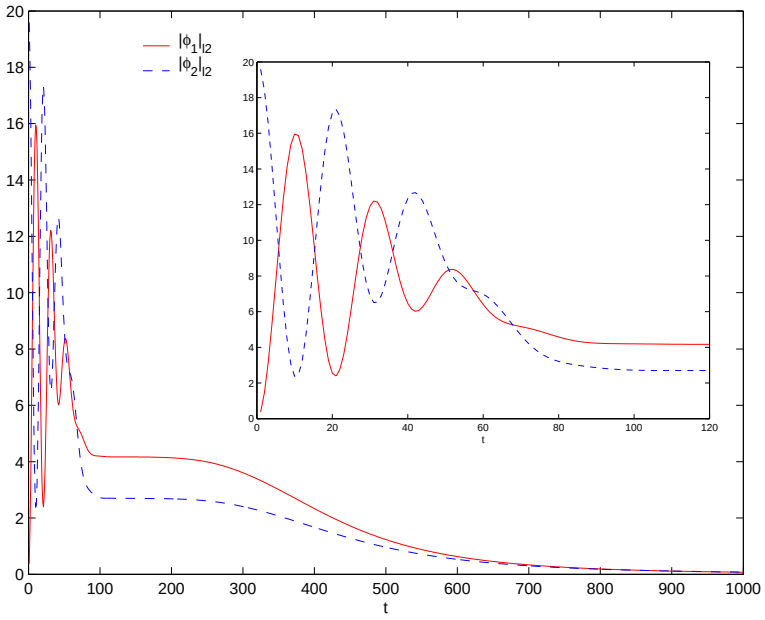


Figure 2.7: Example 1: l^2-norm of the components one and two

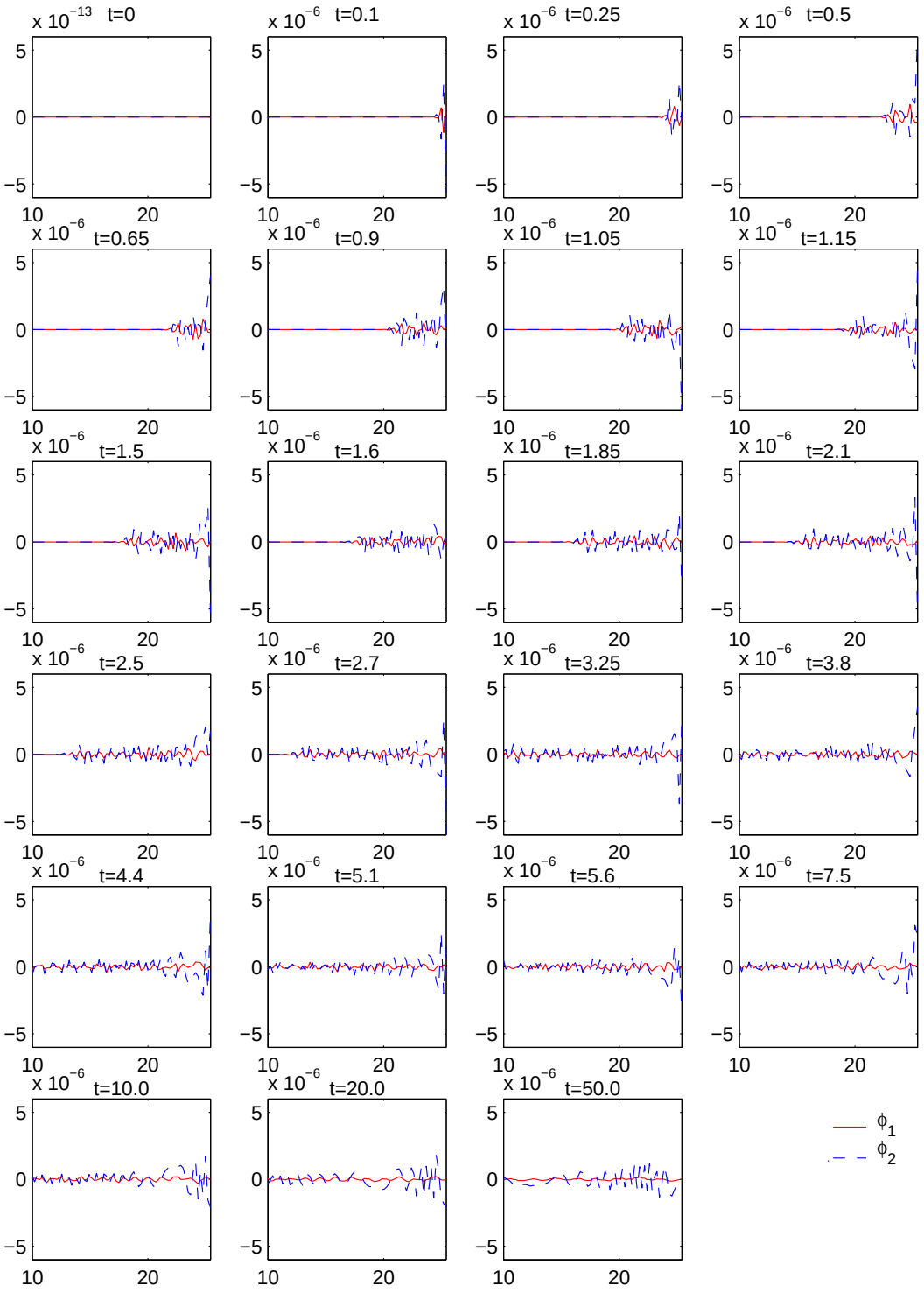


Figure 2.8: Example 1: Error in the time dependent behaviour of $\varphi_{1,j}$ (solid) and $\varphi_{2,j}$ (dashed) compared to a reference solution.

first and second component. We observe, that this oscillation ceases after about 70 time steps (t=3.5), which is about the same time the local maximum of the first wave package leaves the computational domain. We made sure, that there is no connection between these events, by regarding the solution on a much bigger computational domain ($x_{\max} = 41$). This solution served also as a reference solution to investigate the error caused by the DTBC. Fig. 2.8 shows this error for the same time steps as considered before. We observe, that the error is larger than in the parabolic case. Since no approximation was introduced, this error is due to numerical problems. One problem that occurs, is the inversion of the matrices P_2, P_4 respectively (cf. 2.4.9). Numerically they are sometimes badly conditioned - where *sometimes* means for about 10 of 4096 sampling points.

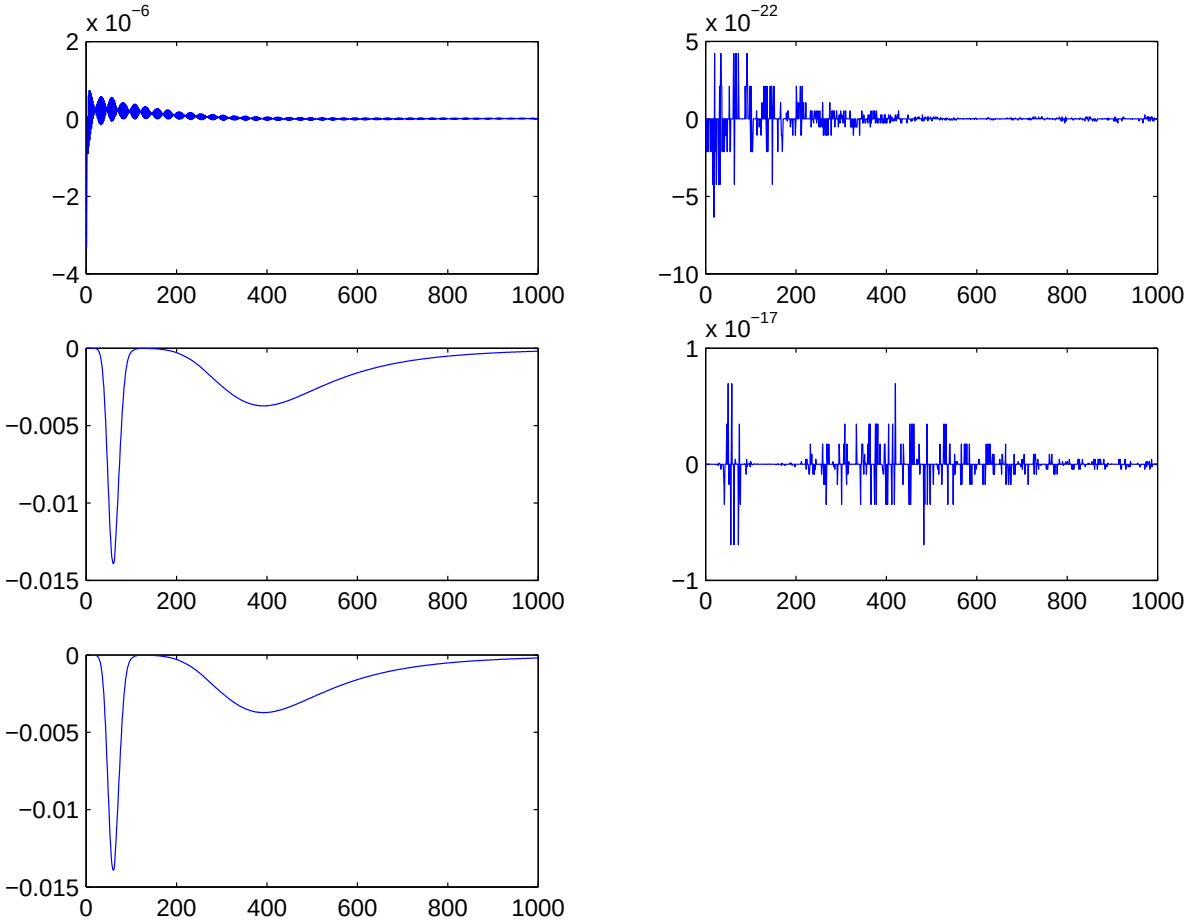


Figure 2.9: Example 1: Time dependent behaviour of the separate stability terms and the time dependent change in the the l^2-norm at the bottom

At last we will numerically check the stability for this example. In (2.4.19) we showed, that for the stability the imaginary part of

(2.5.1)
$$(\boldsymbol{\varphi}_0^{n+\frac{1}{2}})^H \mathbf{N}_L \triangle^+ \boldsymbol{\varphi}_0^{n+\frac{1}{2}} + (\boldsymbol{\varphi}_0^{n+\frac{1}{2}})^H \mathbf{M}_{SL}^H \boldsymbol{\varphi}_0^{n+\frac{1}{2}} + (\boldsymbol{\varphi}_J^{n+\frac{1}{2}})^H \mathbf{N}_R \triangle^- \boldsymbol{\varphi}_J^{n+\frac{1}{2}} + (\boldsymbol{\varphi}_J^{n+\frac{1}{2}})^H \mathbf{M}_{SR}^H \boldsymbol{\varphi}_J^{n+\frac{1}{2}}$$

must be less or equal to zero. For the current example we give in Fig. 2.9 the terms in (2.5.1) in order of appearance from top left to bottom right. The last plot is the sum of the preceding ones. We observe, that the third term dominates and thus ensures the stability. If we change the initial condition in that way, that the wave packages travel to the left, i.e. choosing a negative k_r, exchanges the behaviour of the left and right boundary terms.

5.2. Example 2: The quantum well structure with double barrier. More interesting than the system of free Schrödinger equations, is an additional external potential. We will again consider Example 1 with an additional double barrier near the right boundary, which yields a quantum well structure. The considered double barrier structure is defined by

(2.5.2)
$$\mathbf{e}(x) = \begin{cases} 0, & x \leq 22 \\ \frac{25}{2}, & 22 < x \leq 22.5 \\ \frac{5}{2}, & 22.5 < x \leq 23 \\ 0, & 23 < x \leq 23.5 \\ \frac{25}{2}, & 23.5 < x \leq 24 \\ 0, & 24 < x \end{cases}$$

and can be seen in Fig. 2.10 and 2.11, where it is scaled to fit the more important data.

Fig. 2.10 shows the time dependent behaviour of $\varphi_{1,j}$ (solid) and $\varphi_{2,j}$ (dashed). We concentrate again on the first two components, since there is less to observe in component three and four. We enlarged the computational domain on the l.h.s. to show some nice effects later. As in Example 1 the initial Gaussian wave packet moves to the right and fragments in two. When the faster wave package reaches the first barrier, it is partly reflected and partly transmitted. Fig. 2.11 shows a closer view of the barrier area. With advancing time some part of the density accumulates between the barriers and is slowly transmitted through the second barrier, then leaving the domain of computation. The part of the density, which is reflected at the first barrier moves on to the left and after a time,

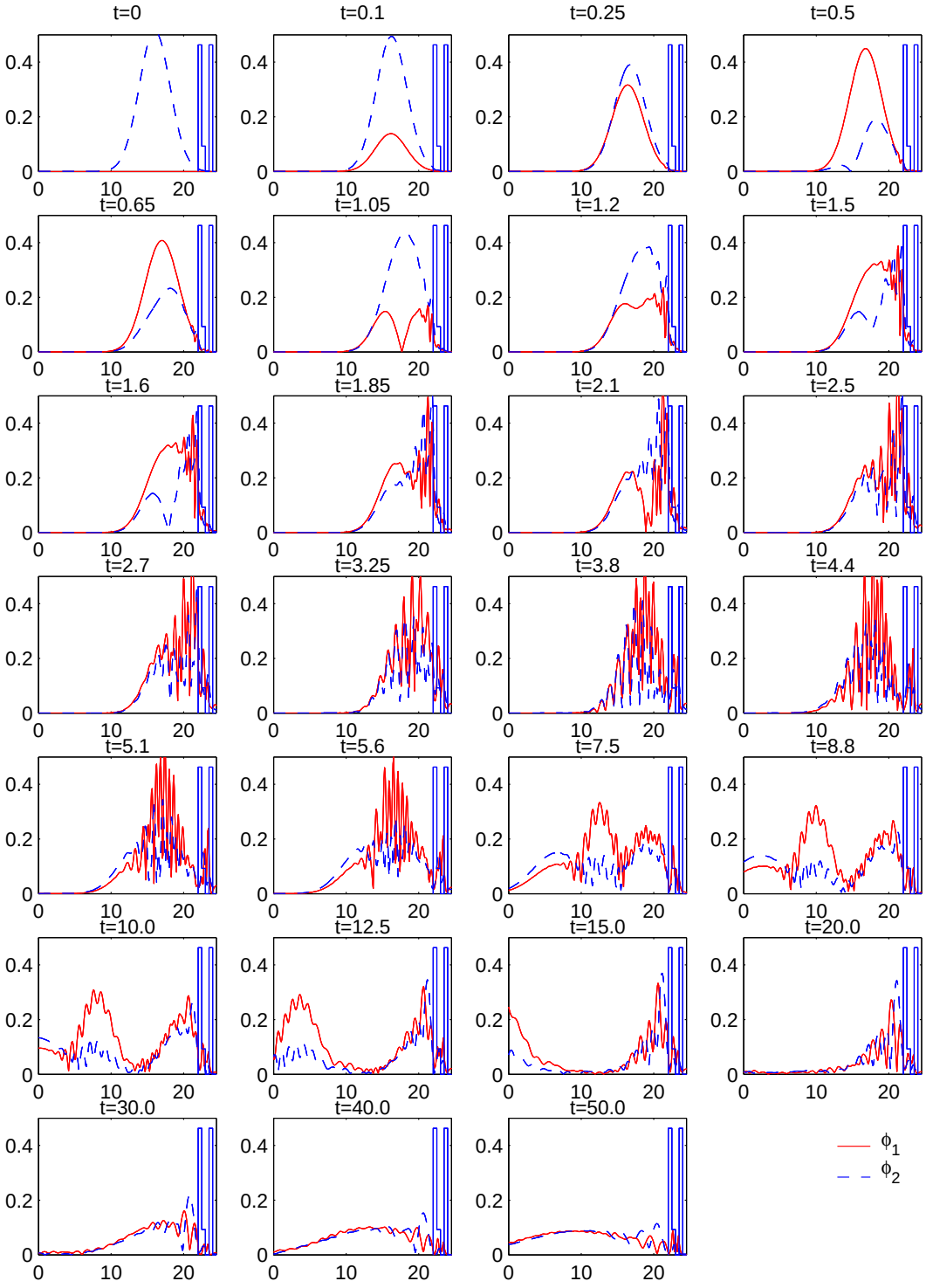


Figure 2.10: Example 2: Time dependent behaviour of $\varphi_{1,j}$ (solid) and $\varphi_{2,j}$ (dashed).

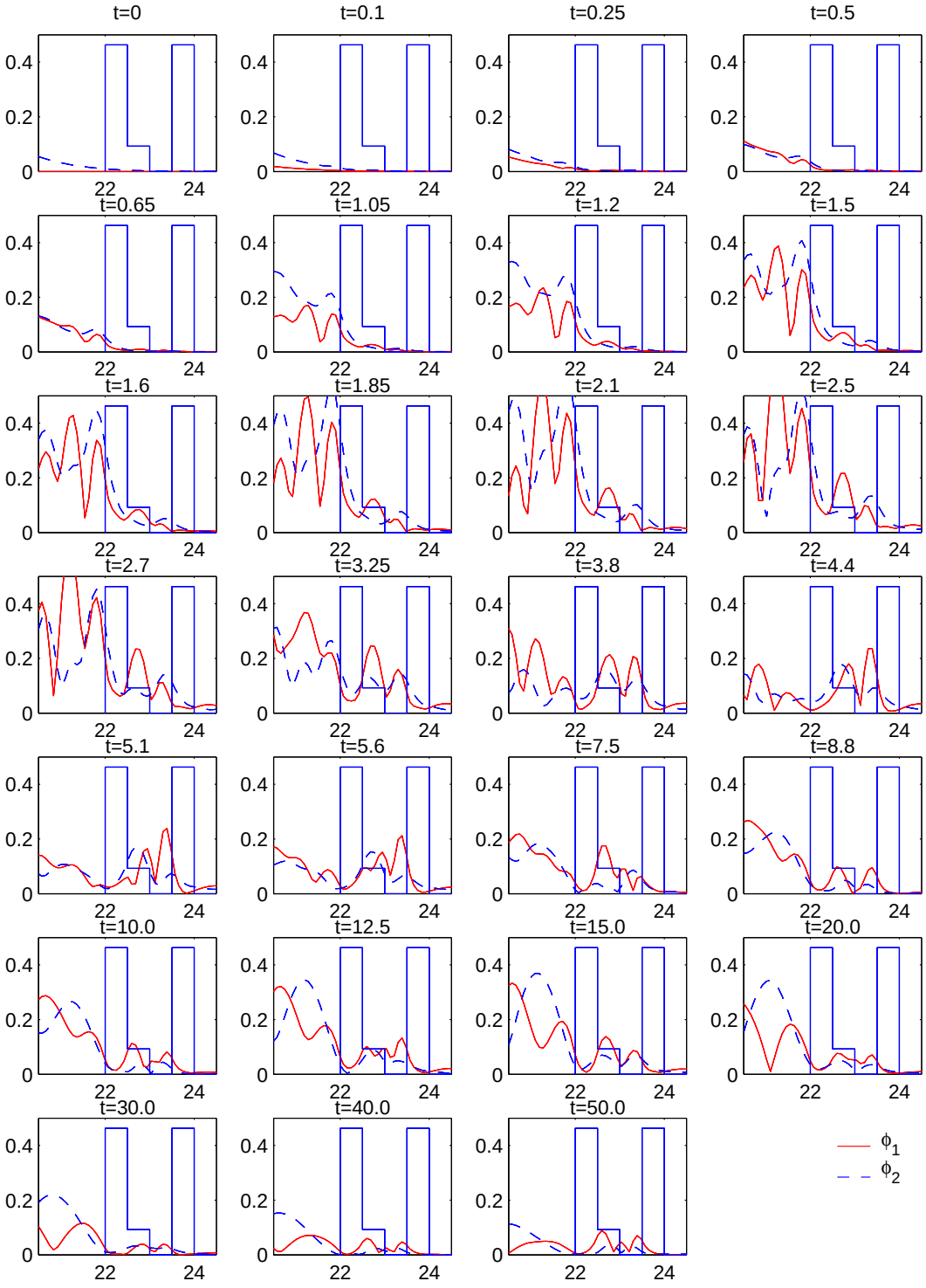


Figure 2.11: Example 2: Zoom of barrier zone of the time dependent behaviour of $\varphi_{1,j}$ (solid) and $\varphi_{2,j}$ (dashed).

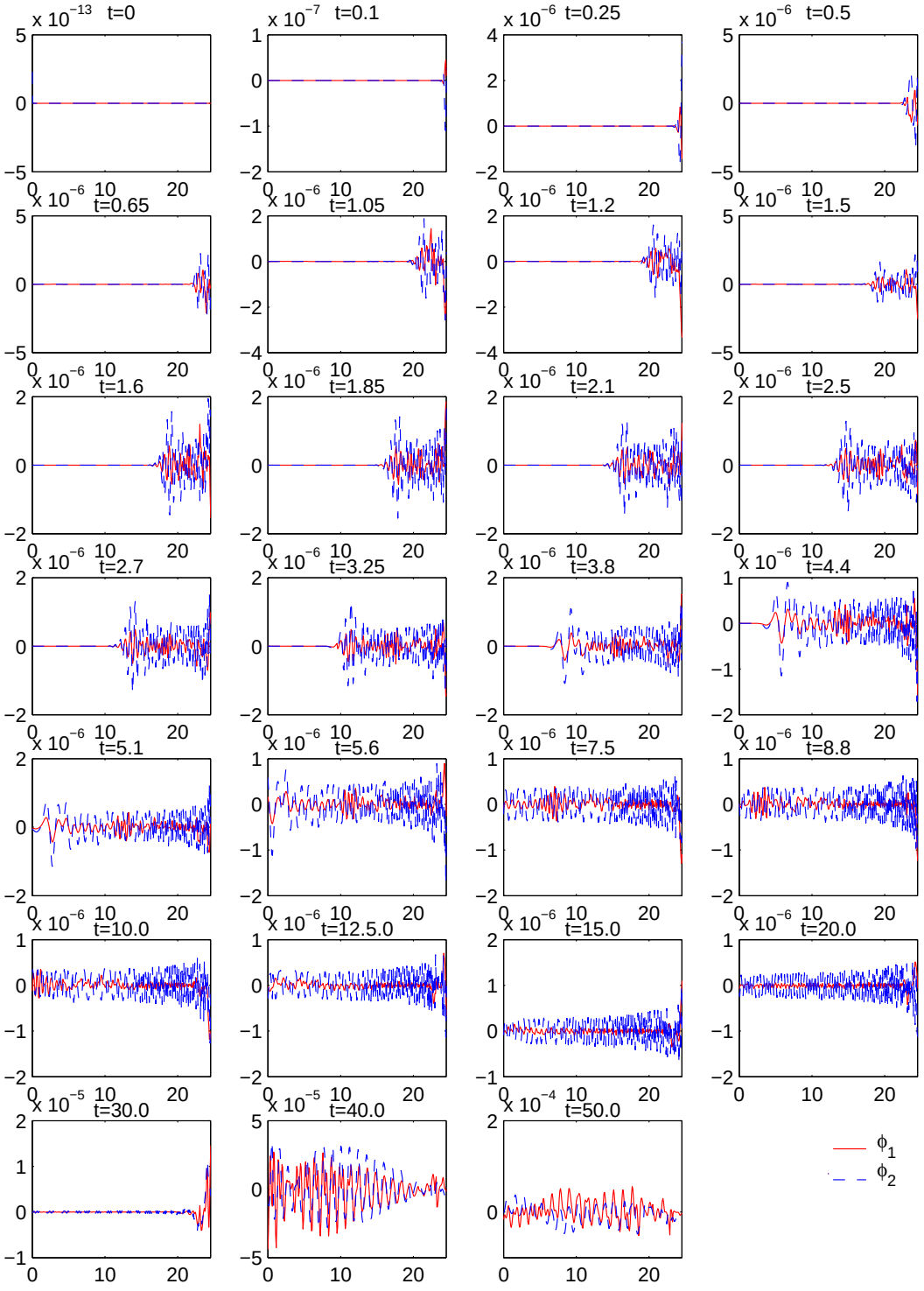


Figure 2.12: Example 2: Error of the time dependent behaviour of $\varphi_{1,j}$ (solid) and $\varphi_{2,j}$ (dashed) compared to a reference solution.

where the wave packages superpose each other, it moves again in form of a Gaussian wave package to the left boundary of the computational domain. The slower wave package seems not to recompose smoothly. Fig. 2.12 shows the error compared to a reference solution on a larger domain ($x_{\max} = 51$). Again 10^{-6} is the order of magnitude of the error.

Conclusion and perspectives

In this dissertation we generalised the approach of discrete transparent boundary conditions as they were used in [**Arn98**],[**Ehr01**] and [**EA01**] for the scalar parabolic and the scalar Schrödinger equation to systems of parabolic or Schrödinger-type equations. To construct the TBC the exterior problem is solved. To this end the partial differential (difference) equation is Laplace-transformed ($\mathcal{Z}$-transformed), yielding an ordinary differential (difference) equation. The main additional problem arising from systems of equations is the fact, that in general the Laplace-transformed ($\mathcal{Z}$-transformed) solution cannot be inverse transformed explicitly as it is possible for scalar equations. At this point it is possible to approximate the image function in such a way, that an explicit inverse Laplace-transformation ($\mathcal{Z}$-transformation) of the solution exists. This was e.g. done in [**Lil92**] and [**Hag94**].

In this dissertation we decided to perform the inverse $\mathcal{Z}$-transformation numerically, involving new sources for numerical errors and new problems, particularly the detection of a suitable radius for inverse $\mathcal{Z}$-transforming the convolution coefficients. We discussed the error of the numerical inverse $\mathcal{Z}$-transformation. It is composed of the approximation and the roundoff error. Whereas the approximation error decays for increasing radius, the roundoff error grows. Additionally the numerical error depends on the coefficient parameters of the considered example and the number of sampling points. Still we observed for the numerical examples, that the dependence on the number of sampling points is stronger than the dependence on the model parameters.

For both types of equations (parabolic and Schrödinger type) we gave examples and computed a numerical solution with discrete transparent boundary conditions and compared it to a reference solution, that was obtained by enlarging the computational domain. For the fluid stochastic Petri nets as well as for the $k \cdot p$ Schrödinger equation the DTBC worked well. It was possible to show the stability of the used difference schemes.

A real drawback is the numerical effort of the DTBCs. The number of terms in the convolution grows with each time step. Therefore it is necessary to look for good approximations. The simple cut-off of the convolution yields bad results and cannot be recommended. We showed for the fluid stochastic Petri nets the use of approximated convolution coefficients as proposed in [**AES03**] for scalar equations. These worked well for all considered examples, but of course involve a reduction of accuracy. This approximation should also be implemented for the system of Schrödinger equations, which was not realised in the scope of this dissertation.

Another interesting prospect is the construction of TBCs for systems of linear hyperbolic equations. We are confident, that the DTBCs for systems of hyperbolic equations will work as well as for systems of parabolic equations.

Appendix

Proof of Theorem 1.13

In this section we will prove Thm. 1.13, which yields an explicit formulation for the coefficients in the DTBC generated by the ansatz method. Theorem 1.13 will be proved by induction over the number $\mathcal{N}$ of different zeros. Before we proceed to the main proof, we will first show the following identity:

LEMMA 2.14. *For $M \geq 2$ the following identity holds*

$$\sum_{k=1}^{M} \binom{M}{k} (-1)^{k+l+1} k^l = \delta_{l,0} = \begin{cases} 1, & l = 0 \\ 0, & l \neq 0 \end{cases}, \quad l = 0, \ldots, M-1. \tag{2.5.3}$$

Proof: We first consider $l = 0$ and use the binomial theorem

$$\begin{aligned} \sum_{k=1}^{M} \binom{M}{k} (-1)^{k+1} k^0 &= \binom{M}{0} + \sum_{k=0}^{M} \binom{M}{k} (-1)^{k+1} \\ &= 1 - \sum_{k=0}^{M} \binom{M}{k} (-1)^k 1^{M-k} = 1 - (1-1)^M = 1. \end{aligned}$$

For $l \in \{1, \ldots, M-1\}$ we use induction over M and first verify the *basis* $M = 2, l = 1$:

$$\sum_{k=1}^{2} \binom{2}{k} (-1)^{k+2} k = \binom{2}{1} (-1) + \binom{2}{2} (-1)^2 2 = -2 + 2 = 0. \tag{2.5.4}$$

For the induction *step* $M \to M+1$ we consider

$$\begin{aligned} \sum_{k=1}^{M+1} \binom{M+1}{k} (-1)^{k+l+1} k^l &= \sum_{k=1}^{M+1} \left\{ \binom{M}{k} + \binom{M}{k-1} \right\} (-1)^{k+l+1} k^l \\ &= \sum_{k=1}^{M} \binom{M}{k} (-1)^{k+l+1} k^l + \sum_{k=1}^{M+1} \binom{M}{k-1} (-1)^{k+l+1} k^l \\ &\overset{(\star)}{=} \sum_{k=1}^{M} \binom{M}{k} (-1)^{k+l+1} k^l - \sum_{k=1}^{M} \binom{M}{k} (-1)^{k+l+1} k^l = 0. \end{aligned}$$

At $(\star)$ we used the identity

$$\sum_{k=1}^{M+1} \binom{M}{k-1} (-1)^{k+l+1} k^l = -\sum_{k=1}^{M} \binom{M}{k} (-1)^{k+l+1} k^l,$$

that we will show now using the induction hypothesis for $\omega < l \leq M$ at $(\star\star)$

$$\begin{aligned}
\sum_{k=1}^{M+1} \binom{M}{k-1} (-1)^{k+l+1} k^l &= \sum_{k=0}^{M} \binom{M}{k} (-1)^{k+l} (k+1)^l = \sum_{k=0}^{M} \binom{M}{k} (-1)^{k+l} \sum_{\omega=0}^{l} \binom{l}{\omega} k^\omega 1^{l-\omega} \\
&= \sum_{\omega=0}^{l} \binom{l}{\omega} (-1)^{l-\omega-1} \sum_{k=0}^{M} \binom{M}{k} (-1)^{k+\omega+1} k^\omega \\
&= \sum_{\omega=0}^{l} \binom{l}{\omega} (-1)^{l-\omega-1} \left\{ (-1)^{\omega+1} 0^\omega + \sum_{k=1}^{M} \binom{M}{k} (-1)^{k+\omega+1} k^\omega \right\} \\
&\overset{(\star\star)}{=} \binom{l}{0} (-1)^{l-1} 0 + \binom{l}{l} (-1)^{-1} \left\{ (-1)^{l+1} 0 + \sum_{k=1}^{M} \binom{M}{k} (-1)^{k+l+1} k^l \right\} \\
&= 0 - \sum_{k=1}^{M} \binom{M}{k} (-1)^{k+l+1} k^l.
\end{aligned}$$

□

After the preceeding preparations we can now prove

Theorem 1.13 (1.7.22) *is solved by*

$$\hat{\ell}_k^{(S)} = (-1)^{k+1} \sum_{1 \leq s_1 < \ldots < s_k \leq S} \varkappa_{s_1} \cdot \ldots \cdot \varkappa_{s_k}, \qquad k = 1, \ldots, S.$$

These $\hat{\ell}_k^{(S)}$ *obey the recursion formula*

$$\hat{\ell}_k^{(M)} = \hat{\ell}_k^{(M-1)} - \varkappa_M \hat{\ell}_{k-1}^{(M-1)}, \qquad k = 1, \ldots, S$$

$$\text{with } \hat{\ell}_0^{(M)} = -1, \quad \hat{\ell}_m^{(M)} = 0 \text{ if } \begin{cases} m < 0 \text{ or} \\ m > M \end{cases}, \quad M = 1, \ldots, S$$

and the γ_i*-level recursion*

$$\hat{\ell}_i^{(\gamma_1 + \ldots + \gamma_{\mathcal{N}})} = \sum_{k=0}^{\gamma_{\mathcal{N}}} \binom{\gamma_{\mathcal{N}}}{k} (-1)^k \varkappa_{\mathcal{N}}^k \hat{\ell}_{i-k}^{(\gamma_1 + \ldots + \gamma_{\mathcal{N}-1})}, \qquad i = 1, \ldots, S.$$

Proof:

First we will show that the recursion (1.7.26) yields the $\hat{\ell}_k^{(S)}$ given in (1.7.25). Then an induction over the number $\mathcal{N}$ of different zeros will show that these $\hat{\ell}_k^{(S)}$ in fact solve

(1.7.22).

For the induction over M we formulate the *induction hypothesis*: The recursion

$$\hat{\ell}_k^{(M)} = \hat{\ell}_k^{(M-1)} - \varkappa_M \hat{\ell}_{k-1}^{(M-1)} \qquad , \quad k = 1, \dots, S$$

$$\text{with } \hat{\ell}_0^{(M)} = -1, \quad \hat{\ell}_m^{(M)} = 0 \text{ if } \begin{cases} m < 0 \text{ or} \\ m > M \end{cases} \qquad , \quad M = 1, \dots, S$$

can be described by the explicit formula

$$\hat{\ell}_k^{(S)} = (-1)^{k+1} \sum_{1 \le s_1 < \dots < s_k \le S} \varkappa_{s_1} \cdot \ldots \cdot \varkappa_{s_k}, \qquad k = 1, \dots, S.$$

We first verify the *basis* for $M = 2$:

recursion	explicit
$\hat{\ell}_1^{(2)} = \hat{\ell}_1^{(1)} - \varkappa_2 \hat{\ell}_0^{(1)} = \varkappa_1 + \varkappa_2$	$\hat{\ell}_1^{(2)} = (-1)^2 \sum_{k=1}^{2} \varkappa_k = \varkappa_1 + \varkappa_2$
$\hat{\ell}_2^{(2)} = \hat{\ell}_2^{(1)} - \varkappa_2 \hat{\ell}_1^{(1)} = -\varkappa_1 \varkappa_2$	$\hat{\ell}_2^{(2)} = (-1)^3 \varkappa_1 \varkappa_2$

$$\left(\quad \square_{\text{recursion basis}} \right)$$

For the induction we consider the *step* $M \to M+1$:

$$\begin{aligned} \hat{\ell}_k^{(M+1)} &= \hat{\ell}_k^{(M)} - \varkappa_{M+1} \hat{\ell}_{k-1}^{(M)} \overset{IH}{=} (-1)^{k+1} \sum_{1 \le s_1 < \dots < s_k \le M} \varkappa_{s_1} \cdot \ldots \cdot \varkappa_{s_k} - \varkappa_{M+1} (-1)^k \sum_{1 \le s_1 < \dots < s_{k-1} \le M} \varkappa_{s_1} \cdot \ldots \cdot \varkappa_{s_{k-1}} \\ &= (-1)^{k+1} \left\{ \sum_{1 \le s_1 < \dots < s_k \le M} \varkappa_{s_1} \cdot \ldots \cdot \varkappa_{s_k} + \sum_{1 \le s_1 < \dots < s_{k-1} \le M} \overbrace{\varkappa_{s_1} \cdot \ldots \cdot \varkappa_{s_{k-1}} \cdot \varkappa_{M+1}}^{k\text{factors}} \right\} \\ &= (-1)^{k+1} \sum_{1 \le s_1 < \dots < s_k \le M+1} \varkappa_{s_1} \cdot \ldots \cdot \varkappa_{s_k} \qquad \text{for } k = 1, \dots, M. \end{aligned}$$

For the remaining case $k = M+1$ holds

$$\begin{aligned} \hat{\ell}_{M+1}^{(M+1)} &= \hat{\ell}_{M+1}^{(M)} - \varkappa_{M+1} \hat{\ell}_M^{(M)} = -\varkappa_{M+1} (-1)^{M+1} \sum_{1 \le s_1 < \dots < s_M \le M} \varkappa_{s_1} \cdot \ldots \cdot \varkappa_{s_{M+1}} \\ &= (-1)^{M+2} \sum_{1 \le s_1 < \dots < s_M \le M} \overbrace{\varkappa_{s_1} \cdot \ldots \cdot \varkappa_{s_M} \cdot \varkappa_{M+1}}^{M+1\text{factors}} \\ &= (-1)^{M+2} \sum_{1 \le s_1 < \dots < s_{M+1} \le M+1} \varkappa_{s_1} \cdot \ldots \cdot \varkappa_{s_{M+1}} \end{aligned}$$

. $(\square_{\text{recursion}})$

The γ_i-level recursion follows by application of the simple recursion γ_i times.

In the following we will refer to the (variable) number of states as $S_{\mathcal{N}}$ where $\mathcal{N}$ gives the number of different zeros with $\sum_{i=1}^{\mathcal{N}} \nu_i = S_{\mathcal{N}}$. With induction over $\mathcal{N}$ we will show the *hypothesis:*

$$\boldsymbol{\mathcal{A}}^{(S_{\mathcal{N}})}\hat{\boldsymbol{\ell}}^{(S_{\mathcal{N}})} = \mathbf{b}^{(S_{\mathcal{N}})}, \text{ for } \mathcal{N} \in \mathbb{N}. \tag{IH}$$

First we verify the *basis* $\mathcal{N} = 1$ with one zero of multiplicity $S_1 = \nu_1$:
Let be $\varkappa_1 = \ldots = \varkappa_{S_1} =: \alpha$. Then

$$\hat{\ell}_k^{S_1} \;=\; (-1)^{k+1} \sum_{1\le s_1<\ldots<s_k\le S_1} \varkappa_{s_1} \cdot \ldots \cdot \varkappa_{s_k} \;=\; (-1)^{k+1} \sum_{1\le s_1<\ldots<s_k\le S_1} \alpha^k \;=\; (-1)^{k+1} \binom{S_1}{k} \alpha^k, \quad k = 1, \ldots, S_1.$$

Now we can show $\boldsymbol{\mathcal{A}}^{(S_1)}\hat{\boldsymbol{\ell}}^{(S_1)} = \mathbf{b}^{(S_1)}$ or equivalently $\boldsymbol{\mathcal{A}}_1^{(S_1)}\hat{\boldsymbol{\ell}}^{(S_1)} = \mathbf{b}_1^{(S_1)}$:

$$\begin{aligned}
\left(\boldsymbol{\mathcal{A}}_1^{(S_1)}\hat{\boldsymbol{\ell}}^{(S_1)}\right)_p &= \sum_{q=1}^{S_1} (J-q)^{p-1}\alpha^{S_1-q}\hat{\ell}\,{}_q^{(S_1)} = \sum_{q=1}^{S_1} (J-q)^{p-1}\alpha^{S_1-q}(-1)^{q+1}\binom{S_1}{q}\alpha^q \\
&= \alpha^{S_1}\sum_{q=1}^{S_1}(-1)^{q+1}\binom{S_1}{q}\sum_{l=0}^{p-1}\binom{p-1}{l}J^{p-1-l}(-q)^l \\
&= \alpha^{S_1}\sum_{l=0}^{p-1}\binom{p-1}{l}J^{p-1-l}\sum_{q=1}^{S_1}(-1)^{q+l+1}\binom{S_1}{q}q^l \\
&\overset{Lem.2.14}{=} \alpha^{S_1}\sum_{l=0}^{p-1}\binom{p-1}{l}J^{p-1-l}\delta_{l,0} \qquad \text{for } p = 1,\ldots,S_1, \quad l = 1,\ldots,p-2 \\
&= \alpha^{S_1}J^{p-1} = \left(\mathbf{b}_1^{(S_1)}\right)_p, \qquad p = 1,\ldots,S_1
\end{aligned}$$

$(\square_{\text{basis}})$

For the induction *step* we have to show $\boldsymbol{\mathcal{A}}^{(S_{\mathcal{N}+1})}\hat{\boldsymbol{\ell}}^{(S_{\mathcal{N}+1})} = \mathbf{b}^{(S_{\mathcal{N}+1})}$. Therefore we consider

$i = 1, \ldots, \mathcal{N}$ and $p = 0, \ldots, \nu_i - 1$:

$$\begin{aligned}
\left(\boldsymbol{\mathcal{A}}_i^{(S_{\mathcal{N}+1})}\hat{\boldsymbol{\ell}}^{(S_{\mathcal{N}+1})}\right)_{p+1} &= \sum_{n=1}^{S_{\mathcal{N}+1}} \alpha_i^{S_{\mathcal{N}+1}-n}(J-n)^p \hat{\ell}_n^{(S_{\mathcal{N}+1})} \\
&\overset{(1.7.27)}{=} \sum_{n=1}^{S_{\mathcal{N}+1}} \alpha_i^{S_{\mathcal{N}+1}-n}(J-n)^p \sum_{k=0}^{\nu_{\mathcal{N}+1}} \binom{\nu_{\mathcal{N}+1}}{k} (-1)^k \alpha_{\mathcal{N}+1}^k \hat{\ell}_{n-k}^{(S_{\mathcal{N}})} \\
&= \sum_{k=0}^{\nu_{\mathcal{N}+1}} \binom{\nu_{\mathcal{N}+1}}{k} (-1)^k \alpha_{\mathcal{N}+1}^k \underbrace{\sum_{n=1}^{S_{\mathcal{N}+1}} \alpha_i^{S_{\mathcal{N}+1}-n}(J-n)^p \hat{\ell}_{n-k}^{(S_{\mathcal{N}})}}_{=:d}
\end{aligned}$$

Setting $\tilde{n} = n - k$ we get

$$\begin{aligned}
d &= \sum_{\tilde{n}=1-k}^{S_{\mathcal{N}+1}-k} \alpha_i^{S_{\mathcal{N}+1}-\tilde{n}-k}(J-\tilde{n}-k)^p \hat{\ell}_{\tilde{n}}^{(S_{\mathcal{N}})} = \alpha_i^{-k} \sum_{n=1-k}^{S_{\mathcal{N}+1}-k} (J-n-k)^p \alpha_i^{S_{\mathcal{N}+1}-n} \hat{\ell}_n^{(S_{\mathcal{N}})} \\
&= \alpha_i^{-k} \left(\sum_{n=1-k}^{0} (J-n-k)^p \alpha_i^{S_{\mathcal{N}+1}-n} \hat{\ell}_n^{(S_{\mathcal{N}})} + \sum_{n=1}^{S_{\mathcal{N}}} (J-n-k)^p \alpha_i^{S_{\mathcal{N}+1}-n} \hat{\ell}_n^{(S_{\mathcal{N}})} \right. \\
&\qquad \left. + \sum_{n=S_{\mathcal{N}}+1}^{S_{\mathcal{N}+1}-k} (J-n-k)^p \alpha_i^{S_{\mathcal{N}+1}-n} \hat{\ell}_n^{(S_{\mathcal{N}})} \right) \\
&= \alpha_i^{-k} \left((J-k)^p \alpha_i^{S_{\mathcal{N}+1}} (\delta_{k,0} - 1) + \sum_{n=1}^{S_{\mathcal{N}}} (J-n-k)^p \alpha_i^{S_{\mathcal{N}+1}-n} \hat{\ell}_n^{(S_{\mathcal{N}})} + 0 \right).
\end{aligned}$$

The last equality is true since $\hat{\ell}_n^{S_{\mathcal{N}}} = 0$ for $n < 0$ and $n > S_{\mathcal{N}}$. Furthermore we notice that the first sum is empty for $k = 0$. This leads to the $\delta_{k,0}$ term. Setting $\tilde{J} = J - k$ and using the induction hypothesis we obtain

$$\begin{aligned}
d &= \alpha_i^{-k} \left((J-k)^p \alpha_i^{S_{\mathcal{N}+1}} (\delta_{k,0} - 1) + \sum_{n=1}^{S_{\mathcal{N}}} (\tilde{J}-n)^p \alpha_i^{S_{\mathcal{N}}-n} \alpha_i^{\nu_{\mathcal{N}+1}} \hat{\ell}_n^{(S_{\mathcal{N}})} \right) \\
&\overset{IH}{=} \alpha_i^{-k} \left((J-k)^p \alpha_i^{S_{\mathcal{N}+1}} (\delta_{k,0} - 1) + \alpha_i^{\nu_{\mathcal{N}+1}} \tilde{J}^p \alpha_i^{S_{\mathcal{N}}} \right) \\
&= (J-k)^p \alpha_i^{S_{\mathcal{N}+1}-k} \delta_{k,0}
\end{aligned}$$

We insert d in the original equation and get

$$\left(\boldsymbol{\mathcal{A}}_i^{(S_{\mathcal{N}+1})}\hat{\boldsymbol{\ell}}^{(S_{\mathcal{N}+1})}\right)_{p+1} = \sum_{k=0}^{\nu_{\mathcal{N}+1}} \binom{\nu_{\mathcal{N}+1}}{k} (-1)^k \alpha_{\mathcal{N}+1}^k \alpha_i^{S_{\mathcal{N}+1}-k} (J-k)^p \delta_{k,0} = \alpha_i^{S_{\mathcal{N}+1}} J^p = (\mathbf{b})_{p+1}.$$

So far we could prove that the hypothesis is valid for $0 < i \leq \mathcal{N}$. We now regard $i = \mathcal{N}+1$ with $0 \leq p < \nu_{\mathcal{N}+1} - 1$:

$$\begin{aligned}\left(\boldsymbol{\mathcal{A}}_{\mathcal{N}+1}^{(S_{\mathcal{N}+1})}\hat{\boldsymbol{\ell}}^{(S_{\mathcal{N}+1})}\right)_{p+1} &= \sum_{n=1}^{S_{\mathcal{N}+1}}(J-n)^p\alpha_{\mathcal{N}+1}^{S_{\mathcal{N}+1}-n}\hat{\ell}_n^{(S_{\mathcal{N}+1})} \\ &\overset{(1.7.27)}{=} \sum_{n=1}^{S_{\mathcal{N}+1}}(J-n)^p\alpha_{\mathcal{N}+1}^{S_{\mathcal{N}+1}-n}\sum_{k=0}^{\nu_{\mathcal{N}+1}}\binom{\nu_{\mathcal{N}+1}}{k}(-1)^k\alpha_{\mathcal{N}+1}^k\hat{\ell}_{n-k}^{(S_{\mathcal{N}})}.\end{aligned}$$

Including the summand for $n = 0$ enables us to change summation limits later.

$$\begin{aligned}\left(\boldsymbol{\mathcal{A}}_{\mathcal{N}+1}^{(S_{\mathcal{N}+1})}\hat{\boldsymbol{\ell}}^{(S_{\mathcal{N}+1})}\right)_{p+1} &= -J^p\alpha_{\mathcal{N}+1}^{S_{\mathcal{N}+1}}\sum_{k=0}^{\nu_{\mathcal{N}+1}}\binom{\nu_{\mathcal{N}+1}}{k}(-1)^k\alpha_{\mathcal{N}+1}^k\hat{\ell}_{-k}^{(S_{\mathcal{N}})} \\ &\quad + \sum_{k=0}^{\nu_{\mathcal{N}+1}}\binom{\nu_{\mathcal{N}+1}}{k}(-1)^k\alpha_{\mathcal{N}+1}^k\sum_{n=0}^{S_{\mathcal{N}+1}}(J-n)^p\alpha_{\mathcal{N}+1}^{S_{\mathcal{N}+1}-n}\hat{\ell}_{n-k}^{(S_{\mathcal{N}})} \\ &= -J^p\alpha_{\mathcal{N}+1}^{S_{\mathcal{N}+1}}(-1) + \sum_{k=0}^{\nu_{\mathcal{N}+1}}\binom{\nu_{\mathcal{N}+1}}{k}(-1)^k\alpha_{\mathcal{N}+1}^k\sum_{n=k}^{S_{\mathcal{N}}+k}(J-n)^p\alpha_{\mathcal{N}+1}^{S_{\mathcal{N}+1}-n}\hat{\ell}_{n-k}^{(S_{\mathcal{N}})}.\end{aligned}$$

For all other n the sum is zero, because $\hat{\ell}_n^{S_{\mathcal{N}}} = 0$ for $n < 0$ and $n > S_{\mathcal{N}}$. We set $d = n - k$.

$$\begin{aligned}\left(\boldsymbol{\mathcal{A}}_{\mathcal{N}+1}^{(S_{\mathcal{N}+1})}\hat{\boldsymbol{\ell}}^{(S_{\mathcal{N}+1})}\right)_{p+1} &= J^a\alpha_{\mathcal{N}+1}^{S_{\mathcal{N}+1}} + \sum_{k=0}^{\nu_{\mathcal{N}+1}}\binom{\nu_{\mathcal{N}+1}}{k}(-1)^k\alpha_{\mathcal{N}+1}^k\sum_{d=0}^{S_{\mathcal{N}}}(J-d-k)^p\alpha_{\mathcal{N}+1}^{S_{\mathcal{N}+1}-d-k}\hat{\ell}_d^{(S_{\mathcal{N}})} \\ &= J^a\alpha_{\mathcal{N}+1}^{S_{\mathcal{N}+1}} + \sum_{k=0}^{\nu_{\mathcal{N}+1}}\binom{\nu_{\mathcal{N}+1}}{k}(-1)^k\sum_{d=0}^{S_{\mathcal{N}}}\alpha_{\mathcal{N}+1}^{S_{\mathcal{N}+1}-d}\hat{\ell}_d^{(S_{\mathcal{N}})}\sum_{\omega=0}^{p}\binom{p}{\omega}(J-d)^{p-\omega}(-k)^\omega \\ &= J^a\alpha_{\mathcal{N}+1}^{S_{\mathcal{N}+1}} - \sum_{d=0}^{S_{\mathcal{N}}}\alpha_{\mathcal{N}+1}^{S_{\mathcal{N}+1}-d}\hat{\ell}_d^{(S_{\mathcal{N}})}\sum_{\omega=0}^{p}\binom{p}{\omega}(J-d)^{p-\omega}\sum_{k=0}^{\nu_{\mathcal{N}+1}}\binom{\nu_{\mathcal{N}+1}}{k}(-1)^{k+\omega+1}k^\omega \\ &= J^a\alpha_{\mathcal{N}+1}^{S_{\mathcal{N}+1}} - X = J^a\alpha_{\mathcal{N}+1}^{S_{\mathcal{N}+1}} = (\mathbf{b})_{p+1}^{(S_{\mathcal{N}+1})}, \qquad p = 0,\ldots,\nu_{\mathcal{N}+1}-1\end{aligned}$$

To realize that $X = 0$ we discern the two cases

- $\nu_{\mathcal{N}+1} = 1$

$$X = \sum_{d=0}^{S_{\mathcal{N}}}\alpha_{\mathcal{N}+1}^{S_{\mathcal{N}+1}-d}\hat{\ell}_d^{(S_{\mathcal{N}})}\sum_{\omega=0}^{p}\binom{p}{\omega}(J-d)^{p-\omega}\left(\binom{1}{0}(-1)^{\omega+1}0^\omega + \binom{1}{1}(-1)^\omega 1^\omega\right) = 0$$

- $\nu_{\mathcal{N}+1} \geq 2$

$$X \overset{\text{Lem.2.14}}{=} \sum_{d=0}^{S_{\mathcal{N}}} \alpha_{\mathcal{N}+1}^{S_{\mathcal{N}+1}-d} \hat{\ell}_d^{(S_{\mathcal{N}})} \sum_{\omega=0}^{p} \binom{p}{\omega} (J-d)^{p-\omega} \left(\binom{\nu_{\mathcal{N}+1}}{0} (-1)^{\omega+1} 0^{\omega} + \delta_{\omega,0} \right)$$

$$= \sum_{d=0}^{S_{\mathcal{N}}} \alpha_{\mathcal{N}+1}^{S_{\mathcal{N}+1}-d} \hat{\ell}_d^{(S_{\mathcal{N}})} (-1+1) = 0.$$

□

Glossary

Here we will give an overview of the used notation.

Variables in Petri nets		
$\boldsymbol{\pi}(t, \boldsymbol{z})$	density function	
$\mathbf{Q}$	generator matrix	8
$R(k, s)$	fluid parameters	12
$\mathbf{M}_k(x)$	expectation of the flow rate	13
$\mathbf{\Sigma}_k^2(x)$	variance of the flow rate	13
$\mathbf{\Phi}^n$	matrix of convolution coefficients	41
$\mathbf{\Psi}^n$	matrix of summed convolution coefficients	45
Variables in the Schrödinger equation		
$\boldsymbol{D}^n$	matrix of convolution coefficients	83
$\boldsymbol{S}^n$	matrix of summed convolution coefficients	84
Error notation		
ϵ_{approx}	approximation error	50
ϵ_{round}	roundoff error	50
ϵ_m	machine accuracy	51
Number fields		
$\mathbb{B}$	boolean values	
$\mathbb{C}$	complex numbers	
$\mathbb{N}$	natural numbers	
$\mathbb{R}$	real numbers	
Complex numbers		
$\mathrm{Im}(z)$	imaginary part of complex variable z	
$\mathrm{Re}(z)$	real part of complex variable z	

Functions of random variables

$P(X > x)$	probability of the random variable X to be larger than x
$E(X)$	expectation of a random variable X
$\mathrm{Var}(X)$	variance of a random variable X

Grid parameters

h	step size in space
k	step size in time

Matrix properties and functions

$\mathbf{v}^T, \quad \mathbf{M}^T$	transpose of a vector or matrix
$\mathbf{M}^H$	Hermitian matrix (transpose conjugate)
$i(\mathbf{M})$	inertia of matrix $\mathbf{M}$, $i(\mathbf{M}) = \big(i_+(\mathbf{M}), i_-(\mathbf{M}), i_0(\mathbf{M})\big)$ 24
$\mathrm{diag}(\mathbf{v})$	diagonal matrix with the diagonal elements $\mathbf{v}$
γ	geometric multiplicity
ν	algebraic multiplicity
$\rho(\mathbf{M})$	spectral radius of matrix $\mathbf{M}$, $\rho(\mathbf{M}) = \max\{\lvert\lambda\rvert : \lambda \in \sigma(\mathbf{M})\}$
$\sigma(\mathbf{M})$	spectrum of matrix $\mathbf{M}$, $\sigma(\mathbf{M}) = \{\lambda : \lambda \text{ eigenvalue of } \mathbf{M}\}$

Bibliography

[ABC+95] M. Ajmone Marsan, G. Balbo, G. Chiola, S. Donatelli, and G. Francheschinis. *Modelling with Generalized Stochastic Petri Nets*. John Wiley & Sons, 1995.

[AES03] A. Arnold, M. Ehrhardt, and I. Sofronov. Discrete transparent boundary conditions for the Schrödinger equation: Fast calculation, approximation, and stability. *Comm.Math.Sci.*, 1(3):501–556, 2003.

[Ajm90] M. Ajmone Marsan. Stochastic Petri Nets: an elementary Introduction. In G. Rozenberg, editor, *Advances in Petri Nets 1989*, volume 424 of *Lecture Notes in Computer Science*, pages 1–29. Springer-Verlag, 1990.

[Arn98] A. Arnold. Numerically Absorbing Boundary Conditions for Quantum Evolution Equations. *VLSI Design*, 6:313–319, 1998.

[Bas88] G. Bastard. *Wave Mechanics Applied to Semiconductor Heterostructures*. Hasted Press, 1988.

[BB97] A. Bultheel and M. Van Barel. *Linear algebra, rational approximation and orthogonal polynomials*. Studies in Computational Mathematics 6. North-Holland, 1997.

[BGG+99] A. Bobbio, S. Garg, M. Gribaudo, A. Horváth, M. Sereno, and M. Telek. Modeling Software Systems with Rejuvenation, Restoration and Checkpointing through Fluid Stochastic Petri Nets. In *Proc. Eighth International Workshop on Petri Nets and Performance Models - PNPM'99*, Zaragoza, Spain, 1999.

[BKKR00] U. Bandelow, H.-Chr. Kaiser, Th. Koprucki, and J. Rehberg. Spectral properties of $k \cdot p$ Schrödinger operators in one space dimension. *Numer. Funct. Anal. Optimization*, 21(3-4):379–409, 2000.

[Car96] M. Cardona. *Fundamentals of Semiconductors*. Springer-Verlag, Berlin, 1996.

[CC92] C. Y.-P. Chao and S. L. Chuang. Spin-orbit-coupling effects on the valence-band structure of strained semiconductor quantum wells. *Phys. Rev. B*, 46(7):4110–4122, 1992.

[Che73] C.-T. Chen. A generalization of the inertia theorem. *SIAM J. Appl. Math.*, 25(2):158–161, 1973.

[CHMM78] A. J. Chorin, T. J. Hughes, M. F. McCracken, and J. E. Marsden. Product formulas and numerical algorithms. *Comm. Pure Appl. Math.*, 31:205–256, 1978.

[Chu91] S. L. Chuang. Efficient band-structure calculations of strained quantum wells. *Phys. Rev. B*, 43(12):9649–9661, 1991.

[Chu95] S. L. Chuang. *Physics of optoelectronic Devices*. Wiley & Sons, New York, 1995.

[CM65] D.R. Cox and H.D. Miller. *The Theory of Stochastic Processes*. Chapman and Hall, 1965.

[CNT97] G. Ciardo, D. Nicol, and K.S. Trivedi. Discrete-event Simulation of Fluid Stochastic Petri Nets. In *Proc. Seventh International Workshop on Petri Nets and Performance Models - PNPM'97*, pages 217–225, Saint Malo, France, June 3–6 1997. IEEE-CS Press.

[CS63] D. Carlson and H. Schneider. Inertia theorems for matrices: The semidefinite case. *J. Math. Anal. Appl.*, 6:430–446, 1963.

[Doe67] G. Doetsch. *Anleitung zum praktischen Gebrauch der Laplace-Transformation und der Z-Transformation*. R. Oldenburg Verlag München, Wien, 3rd edition, 1967.

[EA01] M. Ehrhardt and A. Arnold. Discrete transparent boundary conditions for the Schrödinger equation. *Rivista di Matematica della Università di Parma*, 6:57–108, 2001.

[Ehr97] M. Ehrhardt. Discrete transparent boundary conditions for parabolic equations. *Z. Angew. Math. Mech.*, 77(S2):543–544, 1997.

[Ehr01] M. Ehrhardt. *Discrete Artificial Boundary Conditons*. PhD thesis, Technische Universität Berlin, 2001.

[Ela96] S.N. Elaydi. *An Introduction to Difference Equations*. Springer-Verlag, 1996.

[EM79] B. Engquist and A. Majda. Radiation boundary conditions for acoustic and elastic wave calculations. *Comm. Pure Appl. Math.*, 32:313–357, 1979.

[Fel68] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume I. John Wiley & Sons, Inc., 1968.

[GLS90] G. Gripenberg, S.-O. Londen, and O. Staffans. *Volterra Integral and Functional Equations*. Cambridge University Press, 1990.

[GSB99] M. Gribaudo, M. Sereno, and A. Bobbio. Fluid Stochastic Petri Nets: An Extended Formalism to Include non-Markovian Models. In *Proc. Eighth International Workshop on Petri Nets and Performance Models - PNPM'99*, Zaragoza, Spain, 1999.

[Hac89] W. Hackbusch. *Integralgleichungen, Theorie und Numerik*. Studienbücher Mathematik. Teubner, 1989.

[Hag94] T. Hagstrom. Open boundary conditions for a parabolic system. *Math. Comput. Modelling*, 20(10-11):55–68, 1994.

[Hal82] L. Halpern. Absorbing boundary conditions for the discretization schemes for the one–dimensional wave equation. *Math. Comp.*, 38:415–429, 1982.

[HBSS98] I. Harari, P.E. Barbone, M. Slavutin, and R. Shalom. Boundary infinite elements for the Helmholtz equation in exterior domains. *Int. J. Numer. Methods Eng.*, 41:1105–1131, 1998.

[Hen79] P. Henrici. Fast Fourier Methods in Computational Complex Analysis. *SIAM Review*, 21(4):481–527, 1979.

[HJ99a] R. A. Horn and Ch. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1999.

[HJ99b] R. A. Horn and Ch. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1999.

[HKNT98] G. Horton, V. G. Kulkarni, D. M. Nicol, and K. S. Trivedi. Fluid Stochastic Petri Nets: Theory, Applications and Solution. *European Journal of Operations Research*, 105(1):184–201, February 1998. Also published as ICASE report no. 96-5.

[JSSB93] J. Douglas Jr., J.E. Santos, D. Sheen, and L.S. Bennethum. Frequency domain treatment of one–dimensional scalar waves. *Math. Mod. and Meth. in Appl. Sc.*, 3:171–194, 1993.

[JT72] M.A. Jenkins and J.F. Traub. Algorithm 419: Zeros of a Complex Polynomial. *Comm. ACM*, 15(2):97–99, 1972.

[Kan82] E. O. Kane. Energy Band Theory. In W. Paul, editor, *Handbook on Semiconductors*, volume 1, chapter 4a, pages 193–217. North-Holland, Amsterdam, New York, Oxford, 1982.

[KL89] H.-O. Kreiss and J. Lorenz. *Initial-Boundary Value Problems and the Navier-Stokes Equations*, volume 136 of *Pure and Applied Mathematics*. Academic Press, 1989.

[Kle76] L. Kleinrock. *Queueing Systems Vol II: Computer Applications*. John Wiley, 1976.

[Kyt95] P.K. Kythe. *An introduction to boundary element methods*. CRC Press, Boca Raton, FL., 1995.

[Lev00] M. Levy. *Parabolic equation methods for electromagnetic wave propagation*, volume 45 of *IEEE: Electromagnetic waves series*. The institution of electrical Engineers, 2000.

[Lil92] G. Lill. *Diskrete Randbedingungen an künstlichen Rändern*. PhD thesis, Technische Hochschule Darmstadt, 1992.

[LS71] J.N. Lyness and G. Sande. Algorithm 413: ENTCAF and ENTCRE: Evaluation of Normalized Taylor Coefficients of an Analytic Function. *Comm. of the ACM*, 14(10):669–675, 1971.

[Lub88] C. Lubich. Convolution Quadrature and Discretized Operational Calculus II. *Numer. Math.*, 52:413–425, 1988.

[May89] B. Mayfield. *Non–local boundary conditions for the Schrödinger equation*. PhD thesis, University of Rhode Island, Providence, RI, 1989.

[New71] G. F. Newell. *Applications of Queueing Theory*. Chapman and Hall, Ltd. (London), 1971.

[OS62] A. Ostrowski and H. Schneider. Some theorems on the inertia of general matrices. *J. Math. Anal. Appl.*, 4:72–84, 1962.

[Paz83] A. Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations*, volume 44 of *Applied Mathematical Science*. Springer-Verlag, 1983.

[Pet62] C. A. Petri. *Kommunikation mit Automaten*. PhD thesis, Universität Bonn, 1962.

[PW84] M.H. Protter and H.F. Weinberger. *Maximum Principles in Differential Equations*. Springer-Verlag, 1984.

[QSS00] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics*, volume 37 of *Texts in Applied Mathematics*. Springer-Verlag, 2000.

[RC00] C.W. Rowley and T. Colonius. Discretely nonreflecting boundary conditions for linear hyperbolic systems. *J. Comp. Phys.*, 157:500–538, 2000.

[Ris84] H. Risken. *The Fokker-Planck Equation*. Springer-Verlag, 1984.

[SB90] J. Stoer and R. Bulirsch. *Numerische Mathematik 2*. Springer-Verlag, Berlin, 1990.

[Sin93] J. Singh. *Physics of semiconductors and their heterostructures*. McGraw–Hill, New York, 1993.

[Str89] J. C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations*. Wadsworth & Brooks/Cole, 1989.

[TK93] K. S. Trivedi and V. G. Kulkarni. FSPNs: Fluid Stochastic Petri Nets. In *Proc. 14th Int. Conf. on the Application and Theory of Petri Nets*, pages 24–31, Chicago, 1993.

[Wag85] L. Wagatha. On boundary conditions for the numerical simulation of wave propagation. *Appl. Num. Math.*, 1:309–314, 1985.

[Wil82] J.C. Wilson. Derivation of boundary conditions for the artificial boundaries associated with the solution of certain time dependent problems by Lax–Wendroff type difference schemes. *Proc. Edinb. Math. Soc.*, 1982.

[Wol99] K. Wolter. *Performance and Dependability Modelling with Second Order Fluid Stochastic Petri Nets*. PhD thesis, Technische Universität Berlin, 1999.

[WP98] P. White and J. Powell. Spatial invasion of pine beetles into lodgepole forests: a numerical approach. *SIAM J. Sci. Comput.*, 20(1):164–184, 1998.

[WZ96] D. M. Wood and A. Zunger. Successes and failures of the $k \cdot p$ method: A direct assessment for $GaAs/AlAs$ quantum structures. *Phys. Rev. B*, 53(12):7949–7963, 1996.

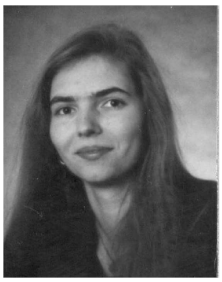
[WZ01] K. Wolter and A. Zisowsky. On Markov reward modelling with FSPNs. *Performance Evaluation*, 44:165–186, 2001.

[WZH02] K. Wolter, A. Zisowsky, and G. Hommel. Performance models for a hybrid reactor system. In E. Schnieder and S. Engell, editors, *Modelling, Analysis, and Design of Hybrid Systems*, volume 279 of *Lecture Notes in Computer Science*, pages 193–210. Springer-Verlag, 2002.

[Zis98] A. Zisowsky. Entwurf und Implementierung eines Verfahrens für die transiente Analyse fluider stochastischer Petri-Netze. Master's thesis, Technische Universität Berlin, 1998.

Index

Curriculum Vitae



Name	Andrea Zisowsky
Date of Birth	03.02.1972
Place of Birth	Berlin
Citizenship	german

Education

08/79 – 07/84	Havelmüller-Grundschule in Berlin-Reinickendorf
09/84 – 06/91	Gabriele-von-Bülow-Gymnasium in Berlin-Reinickendorf
10/91 – 10/98	Study of technomathematics at the TU Berlin
10/98	Master of science 'Diplom' in technomathematics

Professional Experience

05/96 – 09/98	Student assistant in the working group TimeNET, departement of informatics, TU Berlin.
11/98 – 02/01	Assistant in the DFG-project *Analysis und Numerik von kinetischen Quantentransportgleichungen*, TU Berlin.
03/01 – 02/03	Assistant in the DFG-project *Numerik und Asymptotik mikroskopischer und makroskopischer Gleichungen für Quantensysteme*, Saarbrücken.
since 03/03	Assistant in the DFG Research Center Berlin *Mathematics for key technologies*, member of the junior research group *Applied Analysis*.

Berlin, July 2003