universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

"Space–time finite element methods"

verfasst von / submitted by

Paul Stocker, BSc MSc

angestrebter akademischer Grad / in partial fulfillment of the requirements for the degree of

Doktor der Naturwissenschaften (Dr.rer.nat)

Wien, 2021 / Vienna, 2021

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 796 605 405
Studienrichtung lt. Studienblatt / field of study as it appears on the student record sheet:	Mathematik
Betreut von / Supervisor:	Univ.-Prof. Ilaria Perugia, PhD

I would like to extend my sincere thanks to my adviser, Prof. Ilaria Perugia, for her clear guidance, continuous support and for sharing her expertise in countless meetings. Also for the many conferences, workshops, and other great opportunities I was able to partake in during my PhD.

Thanks also to my co-adviser, Prof. Joachim Schöberl, for his help and advice. I also had the great pleasure of working with Prof. Andrea Moiola, whom I would like to thank for the great time I had visiting him in Pavia. Moreover, I wish to thank Prof. Lise-Marie Imbert-Gérard for our collaboration, as well as Marcel Braukhoff for his input and all the fruitful discussions, and Christoph Wintersteiger for our collaborate coding sessions.

Special thanks to my parents for their support, wise counsel and sympathetic ear.

Paul Oswald Stocker

Abstract

Space–time finite element methods approximate solutions of time dependent partial differential equations (PDEs) with a discrete set of functions that live on a mesh of space and time. They allow for space–time adaptive meshing and are naturally high-order methods. However, compared to time-stepping methods, they are inherently expensive due to time being treated as an additional dimension of the mesh and of the approximation spaces. In the literature space–time methods for linear hyperbolic and parabolic problems are well studied. However, much less work has been devoted to nonlinear equations.

In this thesis, we explore ways to improve the efficiency of space–time finite element methods for the wave equation using Trefftz methods combined with tent-pitching. Then, we introduce a novel space–time method for a class of nonlinear parabolic PDEs known as cross-diffusion systems.

Trefftz methods are high-order Galerkin schemes in which all discrete functions are elementwise solution of the PDE to be approximated. We present a space–time Trefftz discontinuous Galerkin (DG) method for approximating the acoustic wave equation semi-explicitly on tent pitched meshes. Tent pitched meshes are meshes that comply with the causality property of the PDE. They allow to solve the equation elementwise, allowing locally optimal advances in time. Trefftz methods are only viable when the PDE is linear and its coefficients are piecewise constant.

For the discretisation of the acoustic wave equation with piecewise smooth wavespeed, we introduce a 'quasi-Trefftz' discontinuous Galerkin method, where the discrete functions are elementwise approximate PDE solutions. We show that the new discretisation possesses the same good approximation properties as the classical Trefftz one, and prove stability and high-order convergence of the DG scheme. We introduce polynomial basis functions for the new discrete spaces and describe a simple algorithm to compute them.

Cross-diffusion systems are systems of nonlinear parabolic PDEs that are used to describe dynamical processes in several application, including chemical concentrations and cell biology. We present a space–time approach to the proof of existence of bounded weak solutions of cross-diffusion systems, making use of the system entropy to study long-term behavior and to show nonnegativity of the solution, even when a maximum principle is not available. This approach naturally gives rise to a novel space–time Galerkin method for the numerical approximation of cross-diffusion systems that conserves their entropy structure. We prove existence and convergence of the discrete solutions, and present numerical results for the porous medium, the Fisher-KPP, and the Maxwell-Stefan problem.

All these methods have been implemented in Netgen/NGSolve. The source code is available online at https://github.com/PaulSt.

Kurzfassung

Raum-Zeit-Finite-Elemente-Methoden approximieren Lösungen von zeitabhängigen partiellen Differentialgleichungen (PDE) mithilfe einer diskreten Menge an Funktionen, die über einem Netz von Raum und Zeit konstruiert werden. Die Methoden ermöglichen adaptive Raum-Zeit-Netze und sind naturgemäß von hoher Konvergenzordnung. Verglichen mit Zeitschrittverfahren sind sie jedoch aufwändiger, da Zeit als eine weitere Dimension des Netzes und der Approximationsräume behandelt wird. Lineare hyperbolische und parabolische Probleme sind in der Literatur bereits umfangreich behandelt. Nicht-lineare Gleichungen wurden in dieser Hinsicht bisher kaum besprochen.

In dieser Arbeit werden Möglichkeiten zur Effizienzsteigerung von Raum-Zeit-Finite-Elemente-Methoden besprochen, unter Verwendung von Trefftz Methoden und in Kombination mit zeltförmigen Netzen. Weiters wird eine neue Raum-Zeit-Methode für eine bestimmte Klasse an nicht-linearen parabolischen PDEs, bekannt als Kreuz-Diffusionssysteme, vorgestellt.

Trefftz-Methoden sind Galerkin-Methoden hoher Konvergenzordnung, in denen alle diskreten Funktionen auf jedem Element des Netzes bereits eine Lösung der betrachteten PDE sind. Zur Approximation der akustische Wellengleichung präsentieren wir eine Raum-Zeit Trefftz unstetige Galerkin-Methode (DG), die sich auf zeltförmigen Netzen explizit lösen lässt. Zeltförmige Netze unterteilen das Raum-Zeit-Gebiet in zeltförmige Elemente, welche die Kausalität der PDE berücksichtigen. Sie erlauben die numerische Lösung elementweise zu berechnen, mit lokal optimalem Fortschritt in der Zeit. Trefftz-Methoden sind ausschließlich auf lineare PDEs mit stückweise konstanten Koeffizienten anwendbar.

Um die Trefftz-DG-Methode auf die akustische Wellengleichung mit stückweise stetigem Koeffizienten zu erweitern, stellen wir eine „quasi-Trefftz“-Methode vor, in der die diskreten Funktionen elementweise Approximationen der PDE Lösungen sind. Wir zeigen, dass die neue Diskretisierung die gleichen guten Approximationseigenschaften der klassischen Trefftz-Methode hat und Stabilität und Konvergenz von hoher Ordnung aufweist. Weiters konstruieren wir polynomiale Basisfunktionen für die Diskretisierung anhand eines simplen Algorithmus.

Kreuz-Diffusionssysteme sind Systeme von nichtlinearen parabolischen PDEs, welche die Entwicklung von Dichten oder Konzentrationen in Mehrkomponentensystemen beschreiben. Wir präsentieren eine Raum-Zeit Variante des Beweises der Existenz von beschränkten schwachen Lösungen, in dem wir die Entropie des Systems benutzen, um das Langzeitverhalten der Lösungen zu untersuchen, und zeigen darüber hinaus, dass sie nichtnegativ ist, auch wenn das Maximumprinzip nicht anwendbar ist. Diese Herangehensweise führt auf natürliche Art zu einer neuen Raum-Zeit Galerkin Methode zur Diskretisierung von Kreuz-Diffusionssystemen, welche die Entropie-Struktur des Systems erhält. Wir zeigen Existenz und Konvergenz der diskreten Lösung und präsentieren numerische Resultate für die poröse Medium-Gleichung, die Fisher-KPP-Gleichung und das Maxwell-Stefan-Problem.

Alle diese Methoden wurden in Netgen/NGSolve implementiert. Der Code ist online verfügbar unter https://github.com/PaulSt.

Contents

List of Notation

We use standard multi-index notation for partial derivatives and monomials, adapted for space–time fields: for $\boldsymbol{i} = (\boldsymbol{i}_{\mathbf{x}}, i_t) = (i_{x_1}, \ldots, i_{x_n}, i_t) \in \mathbb{N}_0^{n+1}$, $D^{\boldsymbol{i}} f := \partial_{x_1}^{i_{x_1}} \cdots \partial_{x_n}^{i_{x_n}} \partial_t^{i_t} f$ and $(\mathbf{x}, t)^{\boldsymbol{i}} = \mathbf{x}^{\boldsymbol{i}_{\mathbf{x}}} t^{i_t} = x_1^{i_{x_1}} \cdots x_n^{i_{x_n}} t^{i_t}$. We use the canonical basis of $\mathbb{R}^{n+1}$, namely $\{\mathbf{e}_k \in \mathbb{R}^{n+1}, 1 \leq k \leq n+1, (\mathbf{e}_k)_l = \delta_{kl}\}$.

The Laplacian Δ, gradient ∇ and divergence $\nabla\cdot$ are considered with respect to the space variable $\boldsymbol{x}$ only, i.e.

$$\begin{aligned} \nabla f(\mathbf{x}, t) &= (\partial_{x_1} f, \ldots, \partial_{x_n} f) \\ \Delta f(\mathbf{x}, t) &= \partial_{x_1}^2 f + \cdots + \partial_{x_n}^2 f \\ \nabla \cdot \boldsymbol{\sigma}(\mathbf{x}, t) &= \partial_{x_1} \sigma_1 + \cdots + \partial_{x_n} \sigma_n \end{aligned}$$

For any vector distribution w in $\mathbb{R}^{n+1}$ a space–time divergence $\mathrm{div}_{(\mathbf{x},t)}$ and a space–time matrix-valued curl-operator $\mathrm{curl}_{(\mathbf{x},t)}$ are defined as

$$\mathrm{div}_{(\mathbf{x},t)} w = \sum_{j=1}^{n+1} D^{\mathbf{e}_j} w_j, \qquad \text{and} \qquad (\mathrm{curl}_{(\mathbf{x},t)} w)_{jk} = D^{\mathbf{e}_k} w_j - D^{\mathbf{e}_j} w_k, \ 1 \leq j, k \leq n+1.$$

For a domain $Q_T \subset \mathbb{R}^{n+1}$ in space and time the Sobolev spaces with regularity index $k \in \mathbb{N}$ and summability index $p \in \mathbb{N}$ are given by

$$W^{k,p}(Q_T) = \{f \in L^p : \ D^{\boldsymbol{i}} f \in L^p(Q_T), \ \forall |\boldsymbol{i}| \leq k\}.$$

The corresponding (semi-)norms are defined as

$$|f|_{W^{k,p}(\Omega)} = \left(\sum_{|\boldsymbol{i}|=k} \int_\Omega |D^{\boldsymbol{i}} f|^p dx \right)^{\frac{1}{p}} \qquad \text{and} \qquad \|f\|_{W^{k,p}(\Omega)} = \left(\sum_{j=1}^{k} |f|^p_{W^{k,p}(\Omega)} \right)^{\frac{1}{p}}.$$

In the special case of $p = 2$ we write for the resulting Hilbert space $H^k(Q_T) := W^{k,2}(Q_T)$, endowed with the inner product

$$(u, v)_{H^k(Q_T)} = \sum_{|\boldsymbol{i}|=0}^{k} (D^{\boldsymbol{i}} u, D^{\boldsymbol{i}} v)_{L^2(Q_T)}.$$

To describe different regularity in space and time we make use of Bochner spaces over a time interval $[0, T]$. For a separable real Hilbert space H the Bochner space $L^2([0,T]; H)$ consist of classes of measurable functions $f : [0, T] \to H$, i.e. $f(y) \in H$ for almost all $y \in [0, T]$ such that

$$\left(\int_{[0,T]} \|f(y)\|_H^2 \ dy \right)^{\frac{1}{2}} \leq \infty.$$

The Bochner Sobolev space is then given by

$$H^k([0,T]; H) = \{f \in L^2([0,T]; H) : \ \partial_y^{\boldsymbol{i}} f \in L^2([0,T]; H), \ \forall |\boldsymbol{i}| \leq k\}.$$

Chapter 1

Introduction

Finite element methods (FEM) are widely used numerical method to approximate the solution of partial differential equations (PDE). They are based on a variational formulation of the PDE problem and on a discretization of the functional space. The main idea to construct a discrete approximation space is to subdivide the domain and to build local basis functions, which are then used to approximate the solution. Main advantages of the FEM are its high flexibility, accuracy, and solid mathematical foundation. Flexibility comes from the possibility of using unstructured meshes and freedom in choosing the discrete spaces. Accuracy is usually provided by two parameters: the mesh size h and the degree of the discrete space p, giving high order approximation.

The idea origins from Walther Ritz, who, in 1908, used a discrete subspace of the infinite dimensional space in which the solution lies, in order to approximate the solution by minimizing the energy of the PDE [96]. In 1915, Boris Galerkin then applied this idea to the variational formulation of the PDE [37]. The notion of using finite elements to construct the discrete approximation space was first applied to the variational setting by Courant in [22]. This marks the starting points for the FEM for static problems. In order to treat time dependent problems, the classic approach is to use finite element methods to discretize space and then use time stepping schemes to advance in time. These methods are still popular today. However, in 1969 John Argyris first proposed using finite elements in time and space in [5]. This marks the starting point of space–time FEM philosophy: *building the mesh and the basis functions in space and time.*

This thesis explores space–time finite elements for parabolic and hyperbolic problems, investigating the advantages and battling the drawbacks. The main advantages of using FEM in space–time carry over from the static FEM: high flexibility, accuracy, and solid mathematical foundation. A drawback compared to simple time stepping is that the use of approximation spaces based on piecewise "total degree" polynomials in both space and time leads to a higher number of degrees of freedom. Furthermore, the possibility of using unstructured meshes in space–time is a two edged sword as it requires us to mesh the full space–time domain. On the upside, *hp*-refinement is made possible in space–time, allowing for straightforward higher order approximation.

1.1 Hyperbolic problems

We will focus first on hyperbolic problems, considering the acoustic wave equation. Space–time finite element methods for linear wave propagation go as far back as [51], and have been used in combination with discontinuous Galerkin (DG) methods e.g. in [27, 36, 73, 82]. DG methods are based on discontinuous piecewise polynomial functions, and so-called numerical fluxes which impose continuity constraints at mesh inter-element boundaries. To construct an efficient method we make use the of the flexebility of the space–time approach by using Trefftz functions and tent–pitching.

Tent pitching techniques generate a space–time mesh, which complies with the causality prop-

erties of the hyperbolic PDE. The resulting mesh consists of tent shaped objects, each advancing locally optimally in time, with the PDE being explicitly solvable in each of them. DG methods pair well with tent pitched meshes as DG upwind fluxes provide a natural way to advanced the solution across element faces, see [36, 75, 82, 94, 114]. A way of constructing these meshes can, for instance, be found in [31, 45, 110, 113]. Though the tent pitching strategy pairs well with DG methods, also other methods are applicable in combination with tent pitched meshes. In [28, 43, 45, 46, 113], schemes for the semi-discretization of different hyperbolic equations on tents are presented, which map tents to a domain where space and time are separable. Similar to the Trefftz-DG method, these schemes are able to solve in 3+1 dimensions, without building four dimensional elements. Friedrichs theory is used in [44] in order to derive a conforming methods, and to prove its convergence properties. We point out that tent pitching is not the only way to deal with the time step restriction of locally refined meshes. A stabilization for a conforming space–time finite element method on Cartesian (in time) meshes is presented in [106]. Classical time-stepping schemes can still be applied successfully by splitting the domain into a coarse-mesh and a fine-mesh region, then explicit time stepping in the coarse-mesh region is combined with local implicit or explicit time stepping in the fine-mesh region. A fully explicit scheme can be found in [47, 48].

The *Trefftz methods* are a class of Galerkin schemes for the approximation of linear partial differential equations. Their distinctive property is that the restrictions to mesh elements of all test and trial functions are particular solutions of the underlying PDE. The variational formulation weakly enforces interelement continuity and initial/boundary conditions. They are named after the seminal work of Erich Trefftz [109]. The main advantage of Trefftz schemes over more classical ones is the higher accuracy for comparable numbers of degrees of freedom.

Trefftz methods have proved particularly successful for wave propagation in time-harmonic regime; see e.g. [50] for a survey of the scalar case. Trefftz methods are often formulated in a DG framework. DG methods are a popular choice for time-domain wave propagation, due to their flexibility, efficiency and simplicity; see e.g. [3, 34, 56, 82, 85].

Trefftz discretisations of time-dependent PDEs are intrinsically *space–time* methods (as opposed to space semi-discretisations and time-stepping): for the test and trial functions to be solution of the PDE they need to be functions of both space and time variables. Trefftz DG schemes developed for time-domain (acoustic, electromagnetic and elastic) wave problems include interior penalty (IP-DG) [8], hybrid DG (involving Lagrange multipliers on mesh interfaces) [91, 112], and versions related to the "ultra-weak variational formulation" [10, 30, 66–68, 81, 90]. In all cases, a sensible choice of the DG numerical fluxes allows to write space–time Trefftz DG schemes as simply as standard "DG-in-space+time-stepping" schemes. In particular, there is no need to solve huge global space–time linear systems but implicit (and, on suitable meshes, even explicit, [90]) time-stepping is possible. Numerical experiments on a wide range of academic test cases have shown excellent properties in terms of approximation and convergence rates [8, 10, 30, 66–68, 90, 91, 112], dissipation [8, 30, 66, 90], dispersion [30, 66], conditioning [67], and even parallelism [90].

Since a sufficiently rich family of local exact solutions of the PDE is needed *Trefftz schemes require PDE coefficients to be elementwise constant.* However, many relevant wave propagation problems take place in a smoothly varying medium: classical examples are well-known in aeroacoustics, underwater acoustics, plasma physics, biomedical imaging, etc. Approximation of smooth coefficients by piecewise-constant ones is not a viable strategy because it immediately spoils high-order convergence, which is one of the strongest reasons to opt for a Trefftz approach. In the case of time-harmonic acoustic wave propagation (Helmholtz equation) Trefftz methods were adapted to smoothly varying coefficients with the introduction of *generalized plane waves* (GPWs) in [53]. GPWs are not exact PDE solutions but rather "solutions up to a given order", in the sense of Taylor polynomials. GPWs extend the accuracy property of Trefftz schemes to a much wider setting (provably for h-convergence [52], so far only numerically for p-convergence). The critical point to construct GPWs relies on the choice of an ansatz, mimicking the oscillatory behavior of plane waves, while allowing for more degrees of freedom. However this is, in a sense, due more to the nature of the Helmholtz equation than to the GPW idea in itself.

1.2 Parabolic problems

We will consider a class of parabolic problems, including linear and nonlinear problems, known as cross-diffusion systems.

Cross-diffusion systems are commonly used to describe dynamical processes appearing in modeling, for example, population dynamics, ion transport through nanopores, tumor growth models, and multicomponent gas mixtures. The challenge in the analysis of these systems is that the diffusion matrix is not necessarily symmetric nor positive semi-definite, and thus no maximum principle is available. Following the *boundedness-by-entropy* method introduced in [58], the remedy is to make use of the entropy structure of the system. For a textbook version see [59]; see also [20, 63]. Introducing the entropy function, a transformation of the solution, allows us to examine long-term behavior and show that the solution is nonnegative and bounded. The key difference of our work to the existing literature is that we do not make use of time stepping, but instead consider time and space altogether. This naturally leads to a novel space–time Galerkin method for the numerical approximation of cross-diffusion systems. The space–time approach entails test and trail spaces, as well as the mesh, where time is included as additional dimension. This provides an easy way to increase the approximation degree simultaneously in space and time, and makes space–time *hp*-refinement possible.

Existing numerical schemes for cross-diffusion systems rely on time stepping methods. An entropy/energy conserving time-stepping algorithm for thermomechanical problems was developed in [92] being of second order in time. In [69], assuming existence of sufficient regular strong solutions on some time interval $[0, T]$ of a scalar diffusion equation, Runge-Kutta methods were studied using maximal regularity. Although maximal regularity also applies to a certain type of cross-diffusion systems [93], Runge-Kutta methods were only applied to very restrictive class; an example (semi-discrete Runge-Kutta scheme) can be found in [61]. In [57], an entropy diminishing/mass conserving fully discrete variational formulation for a cross-diffusion system was presented.

Maxwell-Stefan systems, see [78, 102], describe multicomponent diffusive fluxes in non-dilute solutions or gas mixtures, and are a prime example for the cross-diffusion systems considered here. The first result on global solutions for the Maxwell-Stefan equations close to the equilibrium is given in [41]. The global existence of solutions close to equilibrium and the large time convergence to this equilibrium can be found in [39, Chapter 9], [40, 49], and [93, Chapter 12]. The proof of existence of local classical solutions to the Maxwell-Stefan equations can be found in [12]. For a textbook on this topic, see [93]. The fact that the Maxwell-Stefan equations satisfy the assumptions made in this work, see (H1)-(H2) in Section 4.1, is due to [62], where the entropy structure of the Maxwell-Stefan system was used to prove the existence of globally bounded weak solutions. An entropy structure was also identified for a generalized Maxwell-Stefan system coupled to the Poisson equation in [60], where the existence of global weak solutions was proven as well. The unconditional convergence to the unique equilibrium for given mass was shown in [49, 77] without reaction terms. Those results were extended to also include reaction terms using mass-action kinetics in [23], whenever a detailed balance equilibrium exists. The heat equation can be recovered from the Maxwell-Stefan equation as a relaxation limit [97].

Numerical schemes for the Maxwell-Stefan equations in the literature rely on time-stepping. A finite differences approximation can be found in [72, 74]. Fast solvers for explicit finite-difference schemes were studied in [38]. A posteriori estimates for finite elements in the stationary case are given in [19]. In [89], a mass conserving finite volume scheme was presented. Existence of solutions for a mixed finite element scheme under some restrictions on the coefficients was proven in [79]. The scheme of [26] was proven to also conserve the L^∞ bound by making use of a maximum principle. A scheme using finite elements in space and implicit Euler in time was used to approximate a Poisson-Maxwell-Stefan system in [60]. That scheme, which is based on a formulation in entropy variables, admits solutions that conserve the mass as well as the entropy structure. As a by product, the solution satisfy an L^∞ bound. Another scheme that is mass conserving and conserves the L^∞ bounds of the solutions was presented in [14]. A finite volume

scheme that conserves mass and the non-negativity of the solutions is presented in [18] along side a proof of its convergence.

On simultaneous space–time finite element approaches for parabolic problems, there is a rich literature on the linear case, focusing on the heat equation. In continuous space–time methods, due to the different orders of derivatives present, it is typical to choose a Petrov-Galerkin method, see [2, 6, 103]. In [115], an unconditionally stable formulation for the finite element method on anisotropic spaces is derived using a Hilbert-type transform, with the goal of a finite element-boundary element coupling. In [108], bubble functions are used to derive a method that is stable with respect to small values of the diffusion coefficient. A space–time wavelet method was presented in [101]. Other recent developments include space–time discontinuous Galerkin methods, with at least a discontinuity in the test functions in time, see [25, 70, 95, 98]. For space–time multi-grid methods see [88, 104, 105]. We also point to [24, 71, 83].

For nonlinear parabolic equations, the existing literature on space–time methods is much sparser. The adaptive finite element scheme introduced in [32] for linear parabolic problems was extended in [33] to the scalar version of the nonlinear reaction-diffusion equation treated in this work. A space–time discontinuous Galerkin method for scalar nonlinear convection and diffusion was introduced in [111]. A space–time method for nonlinear PDEs using adaptive wavelets was introduced in [1].

1.3 Outline of the dissertation

Chapter 2 follows [90]. In this chapter, we present original numerical results for the space–time Trefftz discontinuous Galerkin method studied in [81], confirming the known, and conjectured, properties of the method numerically. This implementation combines a Trefftz discontinuous Galerkin method with tent pitched meshes, highlighting the importance of both techniques in order to produce an efficient method. We point out that we can solve the problem in $n + 1$ dimensions, $n \in \{1, 2, 3\}$, without the need for $n + 1$ dimensional elements, since the Trefftz method only requires the computation of integrals at interelement boundaries. Furthermore, we introduce a way of recovering the solution of the second order system, and address in detail the issue of inhomogeneous materials.

Chapter 3 follows [54]. Inspired by the generalized plane waves idea, we propose an extension of the space–time Trefftz discontinuous Galerkin scheme for the acoustic wave equation of [81] to the smoothly-varying wavespeed case. Since the Galerkin basis functions are solution of the PDE up to a given order (with respect to the mesh size), the scheme is referred to as quasi-Trefftz discontinuous Galerkin method. A remarkable outcome is that test and trial basis functions can be taken as polynomials, and their coefficients can be computed with a simple iteration, which is initialized by assigning their values at a given time. Their computation uses the first Taylor-expansion terms of the function $G(\mathbf{x}) = c^{-2}(\mathbf{x})$, c being the problem wavespeed. The definition, the algorithm for computing the basis construction, and the analysis of the polynomial quasi-Trefftz discrete space properties are the main novel contributions of this chapter.

Chapter 4 follows [15]. Here, we develop a space–time approach to the boundedness-by-entropy method, presented in [58], to prove existence of bounded weak solutions of cross-diffusion systems, making use of the system entropy to examine long-term behavior and to show that the solution is nonnegative, even when a maximum principle is not available. The main tool for the proof will be the method of compensated compactness, which is a special technique applying the classical div-curl lemma [107]. This approach naturally gives rise to a novel space-time Galerkin method for the numerical approximation of cross-diffusion systems that conserves their entropy structure. We prove existence and convergence of the discrete solutions, and present numerical results for the heat equation, the porous medium, the Fisher-KPP, and the Maxwell-Stefan problem.

Chapter 2

Trefftz–DG method for the wave equation

This chapter proceeds as follows. First, we introduce the Trefftz-DG method for the acoustic wave equation with piecewise-constant wavespeed in Section 2.2, starting by stating the model problem, defining the Trefftz spaces and finishing the section by formulating the method, as it was introduced in [81]. We continue in Section 2.2.4 by reviewing different strategies of discretizing the Trefftz spaces. In Section 2.3, we discuss some numerical details on how to evolve the solution elementwise on a tent pitched mesh, and in Section 2.4, we show a way to recover the second order solution from the first order formulation. Finally, we present numerical results[1], which were obtained by implementation of the method in Netgen/NGSolve [99, 100], in Section 2.5.

2.1 Acoustic wave equation

We consider the initial boundary value problem (IBVP), given by the homogeneous acoustic wave equation in first order formulation:

$$\begin{cases} \nabla \cdot \boldsymbol{\sigma} + c^{-2}\partial_t v = 0 & \text{in } Q_T, \\ \nabla v + \partial_t \boldsymbol{\sigma} = \mathbf{0} & \text{in } Q_T, \\ v(\cdot,0) = v_0,\ \boldsymbol{\sigma}(\cdot,0) = \boldsymbol{\sigma}_0 & \text{on } \Omega, \\ v = g_D & \text{on } \Gamma_D \times [0,T], \\ \mathbf{n}_\Omega^x \cdot \boldsymbol{\sigma} = g_N & \text{on } \Gamma_N \times [0,T], \\ \frac{\partial}{c} v - \boldsymbol{\sigma} \cdot \mathbf{n}_\Omega^x = g_R & \text{on } \Gamma_R \times [0,T]. \end{cases} \tag{2.1}$$

Here

$n \in \mathbb{N}$	is the physical space dimension,
$\Omega \subset \mathbb{R}^n$	is an open, bounded, Lipschitz polytope,
$T > 0$	is the final time,
$Q_T = \Omega \times (0,T)$	is the space–time cylinder,
$(v, \boldsymbol{\sigma}) : Q_T \to \mathbb{R} \times \mathbb{R}^n$	are the unknown fields (e.g. acoustic pressure and velocity),
$\mathbf{n}_\Omega^x \in \mathbb{R}^n$	is the outward pointing normal unit vector on $\partial\Omega$,
$\Gamma_\mathrm{D}, \Gamma_\mathrm{N}, \Gamma_\mathrm{R}$	are a partition of $\partial\Omega$, one or two of them may be empty,
$v_0 \in L^2(\Omega), \boldsymbol{\sigma}_0 \in L^2(\Omega)^n$	are the initial conditions,
g_D, g_N, g_R	are Dirichlet, Neumann and Robin boundary data, respectively,

[1] The code is available online at https://github.com/PaulSt/NGSTrefftz

$\vartheta \in L^\infty(\Gamma_R \times [0,T])$	is a (uniformly positive) impedance parameter,
$\nabla, \nabla\cdot$	are the gradient and divergence operators in the space variable $\mathbf{x}$,
∂_t	is the time derivative,
$0 < c \in L^\infty(\Omega)$	wavespeed, assumed piecewise-constant and independent of time.

If the initial condition $\boldsymbol{\sigma}_0$ is the gradient of a scalar field u_0, i.e. $\boldsymbol{\sigma}_0 = -\nabla u_0$, then the first order system is equivalent to the second order system obtained by setting $v = \partial_t u$ and $\boldsymbol{\sigma} = -\nabla u$:

$$\begin{cases} -\Delta u + c^{-2}\partial_t^2 u = 0 & \text{in } Q_T, \\ \partial_t u(\cdot,0) = v_0,\ u(\cdot,0) = u_0 & \text{on } \Omega, \\ \partial_t u = g_D & \text{on } \Gamma_D \times [0,T], \\ -\mathbf{n}_\Omega^x \cdot \nabla u = g_N & \text{on } \Gamma_N \times [0,T], \\ \frac{\vartheta}{c}\partial_t u + \nabla u \cdot \mathbf{n}_\Omega^x = g_R & \text{on } \Gamma_R \times [0,T]. \end{cases} \tag{2.2}$$

The well-posedness of IBVP (3.1) in Bochner spaces is briefly discussed in [9, §2.2]. In two space dimensions $d = 2$, the regularity of the solution in corner-weighted Sobolev space of Kondrat'ev type is investigated in detail in [65, 76] and used in the convergence analysis of DG schemes in, e.g., [9, 84, 86].

Remark 2.1.1. *If a source term is present, i.e. if the right-hand side of the IBVP is non-zero, then one can use Duhamel's principle to construct a particular solution and reduce the problem back to the homogeneous one, see [35, Sect. 2.4.2].*

2.2 Trefftz–DG method

2.2.1 Space–time meshes

The mesh $\mathcal{T}_h(Q_T)$ of the space–time domain Q_T is assumed to consist of non-overlapping Lipschitz polytopes, where $h = \max_{K\in\mathcal{T}_h(Q_T)} h_K$, with h_K being the anisotropic diameter defined in (2.8). For each mesh face $F = \partial K_1 \cap \partial K_2$, for $K_1, K_2 \in \mathcal{T}_h(Q_T)$, we assume that it either lies below the characteristic speed $1/c$, or is vertical (parallel to the time axis). In more rigorous terms: Let $(\mathbf{n}_F^x, n_F^t)$ be the normal vector to F with $n_F^t \geq 0$, then either

$$\begin{aligned} c|\mathbf{n}_F^x| < n_F^t \quad & \text{and we call the face } \textit{space-like}, \text{ or} \\ n_F^t = 0 \quad & \text{and we call the face } \textit{time-like}. \end{aligned}$$

Notice, however, that no CFL-condition or any other time step size restriction is imposed on the time-like faces.

A mesh with space-like faces only is called a tent pitched mesh. In the numerical experiments presented below, we generate tent pitched meshes by using the algorithm presented in [45] (see also [113]). The mesh is buit by progressively advancing in time, stacking tent-shaped objects on top or each other, each of them union of $(n+1)$-simplices. The main idea is that the tent height is chosen such that the differential equation is explicitly solvable in each tent. Therefore, the local maximal time advance at a spatial point has to respect the causality constraint, which corresponds to a *local* CFL-condition. This allows us to advance the solution tent by tent, not necessarily having to solve a global system. For independent tents, i.e. tents that are not on top of each other, the computations can be done in parallel. We remark that, in the numerical tests, we observe no stability issues with tents pitched very close to the limit of the causality condition.

We use the following notation for the mesh skeleton and its parts:

$$\mathcal{F}_h := \bigcup_{K\in\mathcal{T}_h} \partial K,$$
$$\mathcal{F}_h^{\text{space}} := \bigcup\{F \subset \mathcal{F}_h \text{ space-like face}\}, \qquad \mathcal{F}_h^{\text{time}} := \bigcup\{F \subset \mathcal{F}_h \text{ time-like face}\},$$

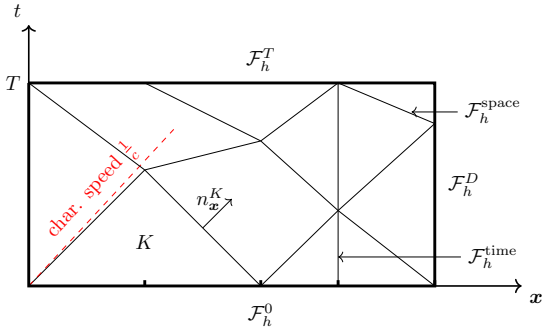

Figure 2.1: A tent pitched mesh with faces below the characteristic speed and a time-like faces, which are contained inside two tents.

$$\mathcal{F}_h^0 := \Omega \times \{0\}, \qquad \mathcal{F}_h^T := \Omega \times \{T\},$$
$$\mathcal{F}_h^{\mathrm{D}} := \Gamma_{\mathrm{D}} \times [0,T], \qquad \mathcal{F}_h^{\mathrm{N}} := \Gamma_{\mathrm{N}} \times [0,T], \qquad \mathcal{F}_h^{\mathrm{R}} := \Gamma_{\mathrm{R}} \times [0,T].$$

This classification of the faces is represented in Figure 2.1.

2.2.2 Trefftz spaces

By definition, Trefftz functions in the kernel of the considered differential operator. For the first order wave equation, we define the local and global Trefftz space as

$$\begin{aligned}\boldsymbol{T}(K) := \big\{(w,\boldsymbol{\tau}) \in L^2(K)^{1+n} \text{ s.t. } \boldsymbol{\tau}|_{\partial K} \in L^2(\partial K)^n,\ \partial_t w,\ \nabla\cdot\boldsymbol{\tau} \in L^2(K),\\ \partial_t\boldsymbol{\tau},\ \nabla w \in L^2(K)^{1+n},\ \nabla\cdot\boldsymbol{\tau} + c^{-2}\partial_t w = 0,\ \nabla w + \partial_t\boldsymbol{\tau} = \mathbf{0}\ \big\}\end{aligned}$$

$$\boldsymbol{T}(\mathcal{T}_h) := \Big\{(w,\boldsymbol{\tau}) \in L^2(Q_T)^{1+n},\ \text{ s.t. } (w|_K, \boldsymbol{\tau}|_K) \in \boldsymbol{T}(K)\ \forall K \in \mathcal{T}_h\Big\},$$

respectively. Note that, by assuming that the solution is in $\boldsymbol{T}(\mathcal{T}_h)$, we require additional smoothness on the solution, as in general we only have that $\partial_t v + \nabla\cdot\boldsymbol{\sigma} \in L^2(K)$, for all $K \in \mathcal{T}_h$.

We derive the Trefftz-DG method for any choice of discrete test and trial space with a Trefftz property, which we denote by $\boldsymbol{T}_p(\mathcal{T}_h)$. A possible choice for a polynomial $\boldsymbol{T}_p(\mathcal{T}_h)$ is given in Section 2.2.4 below.

2.2.3 The method

Following [81], we derive the Trefftz-DG method for the IBVP in (2.1). The method is derived from a local weak formulation, obtained by multiplying the two equations in (2.1) by test and trial functions w and $\boldsymbol{\tau}$, respectively, and integrating by parts on each element K of the mesh $\mathcal{T}_h(Q_T)$. Then, adding the two equations gives

$$\begin{aligned}&-\int_K v\left(\nabla\cdot\boldsymbol{\tau} + c^{-2}\partial_t w\right) + \boldsymbol{\sigma}\cdot\left(\partial_t\boldsymbol{\tau} + \nabla w\right)\ dV\\ &+\int_{\partial K} v\left(\boldsymbol{\tau}\cdot\mathbf{n}_K^x + c^{-2} w n_K^t\right) + \boldsymbol{\sigma}\cdot\left(w\cdot\mathbf{n}_K^x + \boldsymbol{\tau} n_K^t\right)\ dS = 0.\end{aligned} \tag{2.3}$$

By choosing Trefftz test functions $(w,\boldsymbol{\tau}) \in \boldsymbol{V}_p(K)$, the volume integrals over K vanishes. We are left with:

$$\int_{\partial K} \hat{v}_{hp}\left(\boldsymbol{\tau}\cdot\mathbf{n}_K^x + c^{-2} w n_K^t\right) + \hat{\boldsymbol{\sigma}}_{hp}\cdot\left(w\cdot\mathbf{n}_K^x + \boldsymbol{\tau} n_K^t\right)\ dS = 0. \tag{2.4}$$

Typical for DG methods, the continuity of the numeric solution on inter-element boundaries is enforced within the bilinear form of the method. To this end, the trace of the solution $(v, \boldsymbol{\sigma})$ in the boundary integral has been replaced by the numeric fluxes $(\hat{v}_{hp}, \hat{\boldsymbol{\sigma}}_{hp})$, which we define below.

We use standard DG notation for averages $\{\!\{\cdot\}\!\}$, space normal jumps $[\![\cdot]\!]_{\mathbf{N}}$ and time (full) jumps $[\![\cdot]\!]_t$ of piecewise-continuous scalar and vector fields on internal faces: on $F = \partial K_1 \cap \partial K_2$, for $K_1, K_2 \in \mathcal{T}_h$,

$$\begin{aligned} \{\!\{w\}\!\} &:= \frac{w_{|K_1} + w_{|K_2}}{2}, & \{\!\{\boldsymbol{\tau}\}\!\} &:= \frac{\boldsymbol{\tau}_{|K_1} + \boldsymbol{\tau}_{|K_2}}{2}, \\ [\![w]\!]_{\mathbf{N}} &:= w_{|K_1}\mathbf{n}^x_{K_1} + w_{|K_2}\mathbf{n}^x_{K_2}, & [\![\boldsymbol{\tau}]\!]_{\mathbf{N}} &:= \boldsymbol{\tau}_{|K_1}\cdot\mathbf{n}^x_{K_1} + \boldsymbol{\tau}_{|K_2}\cdot\mathbf{n}^x_{K_2}, \\ [\![w]\!]_t &:= w_{|K_1}n^t_{K_1} + w_{|K_2}n^t_{K_2} & [\![\boldsymbol{\tau}]\!]_t &:= \boldsymbol{\tau}_{|K_1}n^t_{K_1} + \boldsymbol{\tau}_{|K_2}n^t_{K_2} \end{aligned}$$

Across time-like faces, the information is passed by using centered fluxes with jump penalization, whereas, across space-like faces, the information is passed upward in time, resembling an up-wind scheme. More precisely, the fluxes on the inter-element faces are chosen as

$$\hat{v}_{hp} = \begin{cases} v^-_{hp} \\ v_0 \\ \{\!\{v_{hp}\}\!\} + \beta[\![\boldsymbol{\sigma}_{hp}]\!]_N \\ v_{hp} \\ g_D \\ v_{hp} + \beta(\boldsymbol{\sigma}_{hp}\cdot\mathbf{n}^x_\Omega - g_N) \\ (1-\delta)v_{hp} + \frac{\delta c}{\vartheta}(\boldsymbol{\sigma}\cdot\mathbf{n}^x_\Omega + g_R) \end{cases} \qquad \hat{\boldsymbol{\sigma}}_{hp} = \begin{cases} \boldsymbol{\sigma}^-_{hp} & \text{on } \mathcal{F}_h^{\text{space}}, \\ \boldsymbol{\sigma}_0 & \text{on } \mathcal{F}_h^0, \\ \{\!\{\boldsymbol{\sigma}_{hp}\}\!\} + \alpha[\![v_{hp}]\!]_N & \text{on } \mathcal{F}_h^{\text{time}}, \\ \boldsymbol{\sigma}_{hp} & \text{on } \mathcal{F}_h^T, \\ \boldsymbol{\sigma}_{hp} - \alpha(v_{hp} - g_D)\mathbf{n}^x_\Omega & \text{on } \mathcal{F}_h^D, \\ g_N\mathbf{n}^x_\Omega & \text{on } \mathcal{F}_h^N, \\ (1-\delta)(\frac{\vartheta}{c}v_{hp} - g_R)\mathbf{n}^x_\Omega + \delta\boldsymbol{\sigma}_{hp} & \text{on } \mathcal{F}_h^R. \end{cases}$$

where α, β, and δ are penalty parameters, which will be chosen constant (notice that they are needed on time-like and boundary faces only). By w^+ and w^- we denote the trace of a function w on space-like faces from the adjacent element at higher and lower times, respectively. Notice that these fluxes are consistent. Moreover, the fluxes on the Robin faces satisfy the following "cross consistency" property:

$$\frac{\vartheta}{c}\hat{v}_{hp} - \hat{\boldsymbol{\sigma}}_{hp}\cdot\mathbf{n}^x_\Omega = g_R.$$

Finally, we plug the definition of the fluxes into (2.4) and sum over all elements $K \in \mathcal{T}_h(Q_T)$. Then the Trefftz-DG method for the wave equation reads:

We recap the Trefftz-DG method introduced in [81]. Let $\boldsymbol{T}_p(\mathcal{T}_h)$ be a closed (e.g. finite-dimensional) subspace of the Trefftz space $\boldsymbol{T}(\mathcal{T}_h)$. We consider the following variational formulation:

$$\begin{aligned} &\text{Seek} && (v_{hp}, \boldsymbol{\sigma}_{hp}) \in \boldsymbol{T}_p(\mathcal{T}_h) \\ &\text{such that} && \mathcal{A}(v_{hp}, \boldsymbol{\sigma}_{hp}; w, \boldsymbol{\tau}) = \ell(w, \boldsymbol{\tau}) \quad \forall (w, \boldsymbol{\tau}) \in \boldsymbol{T}_p(\mathcal{T}_h), \\ &\text{where} \end{aligned} \tag{2.5}$$

$$\begin{aligned} \mathcal{A}(v_{hp},\boldsymbol{\sigma}_{hp}; w, \boldsymbol{\tau}) := &\int_{\mathcal{F}_h^{\text{space}}} \left(c^{-2}v^-_{hp}[\![w]\!]_t + \boldsymbol{\sigma}^-_{hp}\cdot[\![\boldsymbol{\tau}]\!]_t + v^-_{hp}[\![\boldsymbol{\tau}]\!]_{\mathbf{N}} + \boldsymbol{\sigma}^-_{hp}\cdot[\![w]\!]_{\mathbf{N}}\right) \mathrm{d}S \\ &+ \int_{\mathcal{F}_h^T} (c^{-2}v_{hp}w + \boldsymbol{\sigma}_{hp}\cdot\boldsymbol{\tau})\,\mathrm{d}\mathbf{x} \\ &+ \int_{\mathcal{F}_h^{\text{time}}} \left(\{\!\{v_{hp}\}\!\}[\![\boldsymbol{\tau}]\!]_{\mathbf{N}} + \{\!\{\boldsymbol{\sigma}_{hp}\}\!\}\cdot[\![w]\!]_{\mathbf{N}} + \alpha[\![v_{hp}]\!]_{\mathbf{N}}\cdot[\![w]\!]_{\mathbf{N}} + \beta[\![\boldsymbol{\sigma}_{hp}]\!]_{\mathbf{N}}[\![\boldsymbol{\tau}]\!]_{\mathbf{N}}\right) \mathrm{d}S \\ &+ \int_{\mathcal{F}_h^{\mathrm{D}}} \left(\boldsymbol{\sigma}_{hp}\cdot\mathbf{n}^x_\Omega\, w + \alpha v_{hp}w\right) \mathrm{d}S + \int_{\mathcal{F}_h^{\mathrm{N}}} \left(v_{hp}(\boldsymbol{\tau}\cdot\mathbf{n}^x_\Omega) + \beta(\boldsymbol{\sigma}_{hp}\cdot\mathbf{n}^x_\Omega)(\boldsymbol{\tau}\cdot\mathbf{n}^x_\Omega)\right) \mathrm{d}S \\ &+ \int_{\mathcal{F}_h^{\mathrm{R}}} \Big(\frac{(1-\delta)\vartheta}{c}v_{hp}w + (1-\delta)v_{hp}(\boldsymbol{\tau}\cdot\mathbf{n}^x_\Omega) + \delta(\boldsymbol{\sigma}_{hp}\cdot\mathbf{n}^x_\Omega)w + \frac{\delta c}{\vartheta}(\boldsymbol{\sigma}_{hp}\cdot\mathbf{n}^x_\Omega)(\boldsymbol{\tau}\cdot\mathbf{n}^x_\Omega)\Big) \mathrm{d}S \end{aligned}$$

$$\ell(w,\boldsymbol{\tau}) := \int_{\mathcal{F}_h^0} (c^{-2}v_0 w + \boldsymbol{\sigma}_0\cdot\boldsymbol{\tau})\,\mathrm{d}\mathbf{x} + \int_{\mathcal{F}_h^{\mathrm{D}}} g_D\big(\alpha w - \boldsymbol{\tau}\cdot\mathbf{n}_\Omega^x\big)\,\mathrm{d}S$$
$$+ \int_{\mathcal{F}_h^{\mathrm{N}}} g_N\big(\beta\,\boldsymbol{\tau}\cdot\mathbf{n}_\Omega^x - w\big)\,\mathrm{d}S + \int_{\mathcal{F}_h^{\mathrm{R}}} g_R\Big((1-\delta)w - \frac{\delta c}{\vartheta}\,\boldsymbol{\tau}\cdot\mathbf{n}_\Omega^x\Big)\,\mathrm{d}S.$$

On a tent pitched mesh, as the one in Figure 2.1, the method is semi-explicit, meaning that the solution on each tent only depends on the tents below, allowing to solve each tent explicitly, and tents independent from each other in parallel; details are given in Section 2.3.1 below. The situation where also vertical faces are present, is needed, for instance, in the case of piecewise constant wavespeed, is discussed in Section 2.3.2 below. Note that the method only includes integrals over element boundaries, thus only quadrature on n dimensional simplices is needed.

2.2.4 Choice of discrete Trefftz spaces

So far, we have not specified what discretization of the Trefftz space $\boldsymbol{T}_p(K) \subset \boldsymbol{T}(K)$ to use. We introduce the straightforward choice, given by all polynomials in space–time that fulfill the first order wave equation. For an element $K \subset \mathbb{R}^{n+1}$ in the mesh $\mathcal{T}_h(Q_T)$, we define the local polynomial Trefftz space as

$$\mathbb{T}^p(K) := \mathbb{P}^p(\mathbb{R}^{n+1})^{n+1} \cap \boldsymbol{T}(K),$$

where we denote by $\mathbb{P}^p(K)$ the space of polynomials on K of degree $\leq p$. In general, it is possible to choose different polynomial degrees p in different elements. Here, we choose a uniform p, as this is consistent with the numerical examples below. The global Trefftz-DG space on the whole mesh is then given by $\mathbb{T}^p(\mathcal{T}_h) := \prod_{K\in\mathcal{T}_h} \mathbb{T}^p(K)$. The dimension of the elemental Trefftz space is given by

$$\dim \mathbb{T}^p(K) = (n+1)\binom{p+n}{n} = \mathcal{O}_{p\to\infty}(p^n),$$

where we recall that $\binom{a}{b} = \frac{a!}{b!(a-b)!}$ for $b \leq a \in \mathbb{N}_0$. Notice that, for the total degree polynomial space, one has $\dim(\mathbb{P}^p(\mathbb{R}^{n+1})^{n+1}) = \mathcal{O}_{p\to\infty}(p^{n+1})$.

Let us now assume that the first order problem is derived from a second order problem. Then it is natural to derive the vector valued Trefftz space for the first order problem from a scalar Trefftz space for the second order problem. We now detail this approach as it is the one we use for the numerical results presented in Section 2.5. Let us start by defining the polynomial Trefftz space for the second order problem:

$$\mathbb{U}^p(K) := \left\{u \in \mathbb{P}^p(K) : -\Delta u + c^{-2}\partial_t^2 u = 0\right\}.$$

We are able to construct a basis for this space using the recursion formula introduced in [81, Remark 13]. We recall it here, for completeness. We need some multi-index notation: for $\boldsymbol{\alpha} \in \mathbb{N}_0^n$ we denote $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_n$ and $\boldsymbol{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1}\ldots x_n^{\alpha_n}$. Furthermore, let $\boldsymbol{e}_m := (0,\ldots,0,1,0,\ldots,0) \in \mathbb{N}_0^n$ with 1 in the m-th entry. Consider a space–time polynomial

$$u(\boldsymbol{x},t) = \sum_{\substack{\boldsymbol{\alpha}\in\mathbb{N}_0^n,\ k\in\mathbb{N}_0,\\ |\boldsymbol{\alpha}|+k\leq p}} a_{k,\boldsymbol{\alpha}}\boldsymbol{x}^{\boldsymbol{\alpha}}t^k.$$

We want to compute the coefficients $a_{k,\boldsymbol{\alpha}}$ such that the polynomial is Trefftz. This is done by inserting the polynomial into the second order wave equation and collecting terms of equal power to find that

$$a_{k,\boldsymbol{\alpha}} = \frac{c^2}{k(k-1)}\sum_{m=1}^{n}(\alpha_m+1)(\alpha_m+2)a_{k-2,\boldsymbol{\alpha}+2\boldsymbol{e}_m} \tag{2.6}$$

has to hold for the polynomial to be Trefftz. To start the recursion, we need to choose polynomial bases (in the space variables only) for $k=0$ and $k=1$, respectively. More precisely, we start by

choosing polynomial basis functions $\{\tilde{b}_1, \dots, \tilde{b}_{\binom{p+n}{n}}\}$ for the space $\mathbb{P}^p(\mathbb{R}^n)$ and $\{\hat{b}_1, \dots, \hat{b}_{\binom{p-1+n}{n}}\}$ for $\mathbb{P}^{p-1}(\mathbb{R}^n)$. Then we can introduce a basis for $\mathbb{U}^p(K)$ such that either $u(\cdot,0) = \tilde{b}_\ell$ and $\partial_t u(\cdot,0) = 0$, or $u(\cdot,0) = 0$ and $\partial_t u(\cdot,0) = \hat{b}_\ell$ for some ℓ. Hence, we can construct the basis for $\mathbb{U}^p(K)$ out of two sets of polynomial basis functions of $\mathbb{P}^p(K)$ and $\mathbb{P}^{p-1}(K)$. This lets us determine the dimension as

$$\dim \mathbb{U}^p(K) = \binom{p+n}{n} + \binom{p-1+n}{n} = \frac{2p+n}{p}\binom{p-1+n}{n}.$$

Then, a Trefftz space for solving the second order system using a first order formulation can be derived as in [81, §6.2]

$$\begin{aligned}\mathbb{W}^p(\mathcal{T}_h) := \big\{(w,\boldsymbol{\tau}) \in \mathbf{H}(\mathcal{T}_h) : w|_K = \partial_t u,\ \boldsymbol{\tau}|_K = -\nabla u,\ u \in \mathbb{P}^{p+1}(K),\\ \text{with } -\Delta u + c^{-2}\partial_t^2 u = 0 \text{ in } K, \forall K \in (\mathcal{T}_h)\big\}.\end{aligned} \tag{2.7}$$

With a basis given by

$$\mathbb{U}^{p+1}(K) = \operatorname{span}\{b_j,\ j \in \mathcal{I}\} \text{ by setting } \mathbb{W}^p(K) = \operatorname{span}\{(\partial_t b_j, -\nabla b_j),\ j \in \mathcal{I}\}.$$

We have that

$$\dim \mathbb{W}^p(K) = \dim \mathbb{U}^{p+1}(K) - 1 = \frac{2p+n+2}{p+1}\binom{n+p}{n}$$

and furthermore $\mathbb{W}^p(K) \subset \mathbb{T}^p(K)$. A recursion formula, similar to (2.6), can also be derived for $\mathbb{T}^p(K)$, however the numerical results in Section 2.5 are centered around $\mathbb{W}^p(K)$.

Remark 2.2.1. *It is sufficient to compute the coefficients only once for $c = 1$ and then fix the wavenumber by a coordinate transform. Furthermore, for numerical stability, it is convenient to shift the basis functions to the center of the element and scale them by its anisotropic diameter, which is defined by*

$$h_K := \sup_{(\boldsymbol{x},t),(\boldsymbol{y},s)\in K} (|\boldsymbol{x}-\boldsymbol{y}|^2 + c^2|t-s|^2)^{1/2} \tag{2.8}$$

for a mesh element K. For reference coordinates $(\hat{\boldsymbol{x}}, \hat{t})$, the coordinate transform given by

$$(\boldsymbol{x}, t) = (h_K\hat{\boldsymbol{x}}, h_K c^{-1}\hat{t})$$

transforms the Trefftz basis $\hat{u}(\hat{\boldsymbol{x}}, \hat{t})$ of wavespeed 1 to Trefftz basis functions $\hat{u}(x,t)$ of arbitrary wavespeed c. In the case of Trefftz functions for the first order system $(\hat{v}, \hat{\boldsymbol{\sigma}})$, we need to choose

$$v(\boldsymbol{x},t) = c\hat{v}(\hat{\boldsymbol{x}},\hat{t}), \quad \boldsymbol{\sigma}(\boldsymbol{x},t) = \hat{\boldsymbol{\sigma}}(\hat{\boldsymbol{x}},\hat{t}).$$

2.3 Evolution within a tent

The tent pitched mesh allows us to solve local tents explicitly. This is due to the fact that the slope of the mesh faces is below the characteristic speed $1/c$, thus the local solution on a tent can be computed once the solution on its inflow boundary is known. In Section 2.3.1, we discuss how to evolve the solution within a tent with constant wavespeed inside the tent itself. The case where the wavespeed changes within a tent is considered in Section 2.3.2. Notice that, in the constant wavespeed case, tents coincide with mesh elements, while in the latter case tents on the interface contain more than one mesh element.

2.3.1 Constant wavespeed

Let us denote the bottom and top faces of the tent, respectively, by

$$T_h^{\text{bot}} \subset (\mathcal{F}_h^{\text{space}} \cup \mathcal{F}_h^0) \text{ and } T_h^{\text{top}} \subset (\mathcal{F}_h^{\text{space}} \cup \mathcal{F}_h^T).$$

Furthermore, tent faces on the boundary are denoted by $T_h^D \subset \mathcal{F}_h^D$ for Dirichlet, $T_h^N \subset \mathcal{F}_h^N$ for Neumann boundaries, and $T_h^R \subset \mathcal{F}_h^N$ for Robin boundaries.

Since the solution is explicit on each tent, we only need to solve a local system of size $\dim(\boldsymbol{T}_p(K)) \times \dim(\boldsymbol{T}_p(K))$. The system is derived from (2.5) and is given by the following equation

$$\begin{aligned}
&\int_{T_h^{\text{top}}} c^{-2} v_{hp} w n_K^t + \boldsymbol{\sigma}_{hp} \cdot \boldsymbol{\tau} n_K^t + v_{hp} \boldsymbol{\tau} \cdot \mathbf{n}_K^x + \boldsymbol{\sigma}_{hp} \cdot (w \mathbf{n}_K^x)\, dS \\
&\quad + \int_{T_h^D} (\boldsymbol{\sigma}_{hp} \cdot \mathbf{n}_\Omega^x + \alpha v_{hp}) w \, dS + \int_{T_h^N} v_{hp} (\boldsymbol{\tau} \cdot \mathbf{n}_\Omega^x) + \beta (\boldsymbol{\sigma} \cdot \mathbf{n}_\Omega^x)(\boldsymbol{\tau} \cdot \mathbf{n}_\Omega^x)\, dS \\
&\quad + \int_{\mathcal{F}_h^{\text{R}}} \Big(\frac{(1-\delta)\vartheta}{c} v_{hp} w + (1-\delta) v_{hp} (\boldsymbol{\tau} \cdot \mathbf{n}_\Omega^x) + \delta (\boldsymbol{\sigma}_{hp} \cdot \mathbf{n}_\Omega^x) w + \frac{\delta c}{\vartheta} (\boldsymbol{\sigma}_{hp} \cdot \mathbf{n}_\Omega^x)(\boldsymbol{\tau} \cdot \mathbf{n}_\Omega^x) \Big)\, \mathrm{d}S \\
&= - \int_{T_h^{\text{bot}}} c^{-2} v_{\text{bot}} w n_K^t + \boldsymbol{\sigma}_{\text{bot}} \cdot \boldsymbol{\tau} n_K^t + v_{\text{bot}} \boldsymbol{\tau} \cdot \mathbf{n}_K^x + \boldsymbol{\sigma}_{\text{bot}} \cdot \mathbf{n}_K^x w \, dS \\
&\quad + \int_{T_h^D} g_D (\alpha w - \boldsymbol{\tau} \cdot \mathbf{n}_\Omega^x)\, dS + \int_{T_h^N} g_N (\beta \boldsymbol{\tau} \cdot \mathbf{n}_\Omega^x - w)\, dS + \int_{\mathcal{F}_h^{\text{R}}} g_R \Big((1-\delta) w - \frac{\delta c}{\vartheta} \boldsymbol{\tau} \cdot \mathbf{n}_\Omega^x \Big)\, \mathrm{d}S
\end{aligned} \tag{2.9}$$

where, in the case $T_h^{\text{bot}} \subset \mathcal{F}_h^0$, $(v_{\text{bot}}, \boldsymbol{\sigma}_{\text{bot}}) = (v_0, \boldsymbol{\sigma}_0)$, and in the case $T_h^{\text{bot}} \subset \mathcal{F}_h^{space}$, $(v_{\text{bot}}, \boldsymbol{\sigma}_{\text{bot}})$ on a given face is the previously computed solution in the tent sharing that face in lower time.

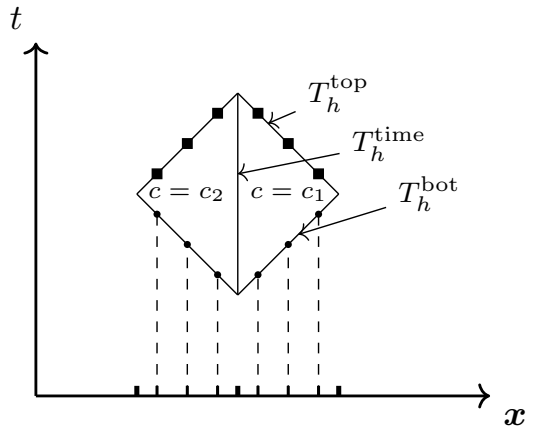

Figure 2.2: The spatial integration points are mapped to the faces of the tent. The solution is determined using the known input on the bottom integration points (dots), and is evaluated on the top integration points (squares).

For the numerical integration, we only need an integration rule for n-simplices, in order to integrate over the boundary of the tent. We can define an integration rule on the spatial mesh once, which we can then map to the faces of the tent. This idea is visualized in Figure 2.2. After solving on the tent, we need to evaluate $(v_{hp}, \boldsymbol{\sigma}_{hp})$ in the integration points on T_h^{top}, and store these values for the next tent. On each spatial integration point, we only need to store the most recent results, leading to a total storage of: (total number of integration points) $\cdot\, (n+1)$.

2.3.2 Piecewise constant wavespeed

Recall that we assume that the wavespeed is constant in time and piecewise constant in space. In this case, we always consider initial spatial meshes that are aligned with the discontinuities of the wavespeed. To treat such a jump within a space–time tent, we need to incorporate the jump terms from our DG formulation (2.5). This involves integrating on the time-like inter-element boundary

contained inside the tent, denoted by $T_h^{\text{time}} \subset \mathcal{F}_h^{\text{time}}$. According to (2.5), one has to add to the left-hand side of (2.9) the term

$$\int_{T_h^{\text{time}}} (\{\!\{v_{hp}\}\!\}[\![\boldsymbol{\tau}]\!]_N + \{\!\{\boldsymbol{\sigma}_{hp}\}\!\} \cdot [\![w]\!]_N + \alpha [\![v_{hp}]\!]_N \cdot [\![w]\!]_N + \beta [\![\boldsymbol{\sigma}_{hp}]\!]_N [\![\boldsymbol{\tau}]\!]_N) \; dS.$$

Since the tent now includes two mesh elements, the system matrix is now of size $2\dim(\boldsymbol{T}_p(K)) \times 2\dim(\boldsymbol{T}_p(K))$. The extension to interfaces between more than two materials follows.

2.4 Recovery of the solution of the second order equation

In the case where the problem comes from a second order formulation we can substitute $v = \partial_t u$ and $\boldsymbol{\sigma} = -\nabla u$ to write the method in terms of test and trial functions from $\mathbb{U}_{p+1}(K)$. Then the method (2.5) reads:

$$\begin{aligned} &\text{find } u_{hp} \in \mathbb{U}_{p+1}(\mathcal{T}_h) \quad \text{s.t.} \\ &\hat{\mathcal{A}}(u_{hp}; v) = \hat{\ell}(v) \qquad \forall v \in \mathbb{U}_{p+1}(\mathcal{T}_h), \end{aligned} \tag{2.10}$$

with

$$\hat{\mathcal{A}}(u_{hp}; v) := \mathcal{A}(\partial_t u_{hp}, -\nabla u_{hp}; \partial_t v, -\nabla v) \text{ and } \quad \hat{\ell}(v) := \ell(\partial_t v, -\nabla v).$$

The constant basis function does not contribute to formulation (2.10), as only derivatives of the unknown u_{hp} are present. Thus, this formulation produces the same results as (2.5) with $\boldsymbol{v}_p(\mathcal{T}_h) = \mathbb{W}_p(\mathcal{T}_h)$. In order to fix the constants and recover the solution to the second order wave equation, we modify the original formulation by adding the additional terms

$$\int_{\mathcal{F}_h^{\text{space}}} -[\![u_{hp}]\!]_t v^+ \; dS + \int_{\mathcal{F}_h^0} u_{hp} v \; dS$$

to the bilinear form $\hat{\mathcal{A}}(u_{hp}; v)$, and

$$\int_{\mathcal{F}_h^0} u_0 v \; dS$$

to the right hand side $\hat{\ell}(v)$, where and $u_0(x) = u(x, 0)$. Note that these terms preserve the consistency of the formulation.

Therefore, when evolving the solution inside a single tent, we need to add $\int_{T_h^{\text{top}}} u_{hp} v$ and $\int_{T_h^{\text{bot}}} u_{\text{bot}} v$ to the left- and right-hand side, respectively, of the formulation discussed in Section 2.3.

2.5 Numerical tests

In this section we present numerical test results in one, two, and three spatial dimensions. If not otherwise stated, we use the following settings for the numerical examples. We consider the problem (3.1) with initial and Dirichlet boundary conditions such that the analytical solution is $(v, \boldsymbol{\sigma}) = (\partial_t u, -\nabla u)$, where u is the standing wave

$$u(\boldsymbol{x}, t) = \cos(\pi x_1) \cos(\pi x_2) \cos(\pi x_3) \sin(\pi t c \sqrt{n})/(\sqrt{n}\pi), \tag{2.11}$$

given here in 3+1 dimensions, and set the wavespeed $c = 1$. An example is plotted in 1+1 dimensions in Figure 2.3. The penalty parameters are chosen as $\alpha = \beta = 0.5$. We measure the error

$$e(v, \boldsymbol{\sigma}; v_{hp}, \boldsymbol{\sigma}_{hp}) = \left(c^{-2} \|v(\cdot, T) - v_{hp}(\cdot, T)\|^2_{L^2(\Omega)} + \|\boldsymbol{\sigma}(\cdot, T) - \boldsymbol{\sigma}_{hp}(\cdot, T)\|^2_{L^2(\Omega)} \right)^{\frac{1}{2}},$$

at final time T, which we choose at $T = 1$. The tent pitched meshes are produced by the algorithm presented in [45]. In this algorithm, the height of the tents is limited by the slope of the edges, and not by the slope of the faces. All timings were performed on a server with two Intel(R) Xeon(R) CPU E5-2687W v4, with 12 cores each.

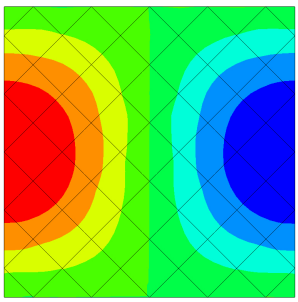

Figure 2.3: Approximation of the standing wave on a 1+1 dimensional space–time tent pitched mesh.

2.5.1 Approximation properties of Trefftz spaces

We compare the approximation properties of the Trefftz space to the first-order derivatives $\mathbb{Y}^p(\mathcal{T}_h)$ of the full polynomial space given by

$$\mathbb{Y}^p(\mathcal{T}_h) := \big\{(w,\boldsymbol{\tau}) \in \mathbf{H}(\mathcal{T}_h) : \ w|_K = \partial_t u, \ \boldsymbol{\tau}|_K = -\nabla u, \ u \in \mathbb{P}^{p+1}(K), \ \forall K \in (\mathcal{T}_h)\big\}, \qquad p \in \mathbb{N}_0.$$

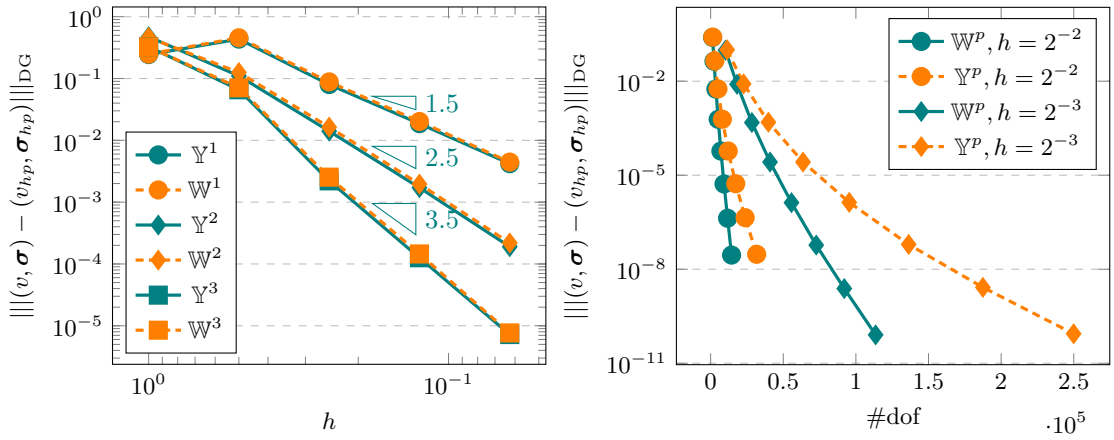

Figure 2.4: Comparison between Trefftz functions $\mathbb{W}^p$ and full polynomial space $\mathbb{Y}^p$, in terms of order (left) and local dofs (right) for fixed $h = 2^{-2}, 2^{-3}$.

The DG method used with the full polynomial space $\mathbb{Y}^p$ requires the computation of integrals also in space–time volumes, as opposed to the Trefftz-DG method (2.5) where the volume integrals cancel. Therefore, we need to add back in the volume term to the left-hand side of the formulation (2.5), giving the new left-hand side:

$$\begin{aligned}\tilde{\mathcal{A}}(v_{hp},\boldsymbol{\sigma}_{hp};w,\boldsymbol{\tau}) :=& \\ &- \sum_{K\in\mathcal{T}_h} \int_K v_{hp}\left(\nabla\cdot\boldsymbol{\tau} + c^{-2}\partial_t w\right) + \boldsymbol{\sigma}_{hp}\cdot\left(\partial_t\boldsymbol{\tau} + \nabla w\right)\, dV \\ &+ \mathcal{A}(v_{hp},\boldsymbol{\sigma}_{hp};w,\boldsymbol{\tau}).\end{aligned}$$

We now need to solve $\tilde{\mathcal{A}}(v_{hp},\boldsymbol{\sigma}_{hp};w,\boldsymbol{\tau}) = \ell(w,\boldsymbol{\tau})$, $\forall (w,\boldsymbol{\tau}) \in \mathbb{P}^p(\mathcal{T}_h)^{n+1}$ for $(v_{hp},\boldsymbol{\sigma}_{hp}) \in \mathbb{Y}^p(\mathcal{T}_h)^{n+1}$. We discuss the DG method with volume integral in more detail in Section 3.2.3.

For the numerical results, we have taken as spatial domain the unit square $(0,1)^2$ partitioned into uniform, triangular, unstructured meshes in with mesh width h. The time domain is partitioned by a uniform mesh of size $h_t \approx h$. The results obtained with the Trefftz-DG method

and with the non-Trefftz–DG method on these Cartesian meshes are shown in Figure 2.4. The h convergence in the DG norm is optimal, according to [81, Thm. 3], with a rate of $\mathcal{O}(h^{p+1/2})$. In terms of polynomial degree p, both choices exhibit exponential convergence speed. The benefits of the Trefftz space becomes clear when comparing errors versus the global number of degrees of freedom, as seen in the right plot in Figure 2.4.

2.5.2 Comparing space–time meshing strategies

In Section 2.3, we have seen how to advance the solution element wise on a tent pitched mesh. We now compare this approach to solving the full system on a Cartesian (in time) space–time slab. To solve the full system we use a block Jacobi solver. When comparing the timing of the two strategies, we only consider solving the tents sequentially, i.e. on a single thread. For this comparison, we choose a quasi-uniform mesh of the unit square in space and the final time equal to the mesh size, i.e. one CFL-conforming time step on the Cartesian mesh. As a spatial domain we take the unit square, meshed with

As spatial mesh, we take a uniform, triangular, unstructured mesh of the unit square $(0,1)^2$ with mesh width 2^{-3}. On top of which we construct both, the tent pitched and the Cartesian, mesh. For the Cartesian mesh in time, we choose the hight of the element $h_t \approx 2^{-3}$.

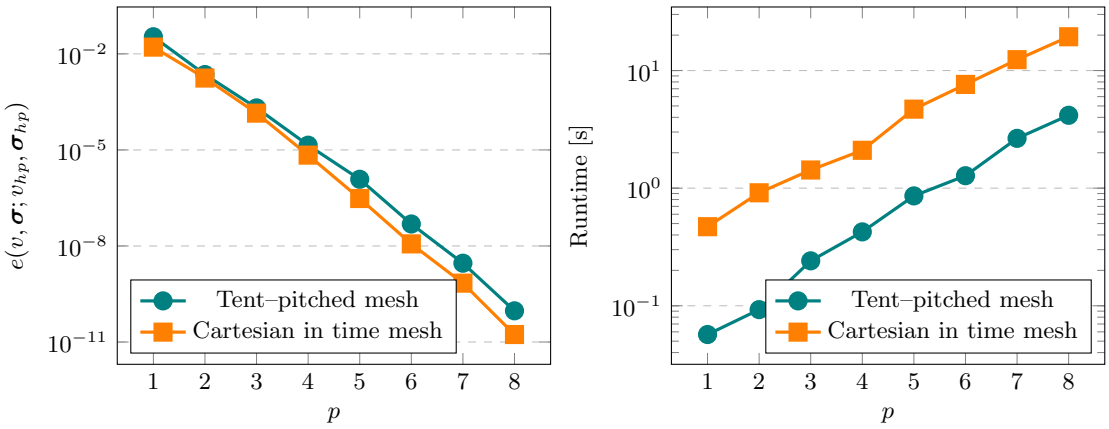

Figure 2.5: Comparison of the Trefftz-DG method on Cartesian (in time) meshes and tent pitched meshes.

The results in Figure 2.5 on the left show that the error between the two mesh types differs slightly. As the two space–time meshes are constructed on top of the same spatial mesh, a single space–time tent element will contain several spatial elements, giving it a larger diameter than the space–time right triangular prism in the Cartesian mesh. This explains the difference in error and convergence rate. On the right in Figure 2.5 we compare the runtime, where sequential tent pitching is about a magnitudes faster. This shows that no parallelization of the tent pitching method is necessary to outperform the implicit solver.

2.5.3 Choice of spatial basis functions

As we have seen in Section 2.2.4, the recursion formula (2.6) for the derivation of the Trefftz basis functions, can be initialized with an arbitrary choice of polynomial basis functions in space variable only.

In the following, we compare three different choices for the initial polynomial basis functions: monomials, Legendre, and Chebychev polynomials. In all cases, the basis functions are shifted to the center of each element and scaled, as described in Remark 2.2.1. We compare them in 1+1 dimensions, on the space–time unit square. The mesh considered is a Cartesian mesh of spatial size $h = 2^{-2}$. The problem is solved globally using formulation (2.5).

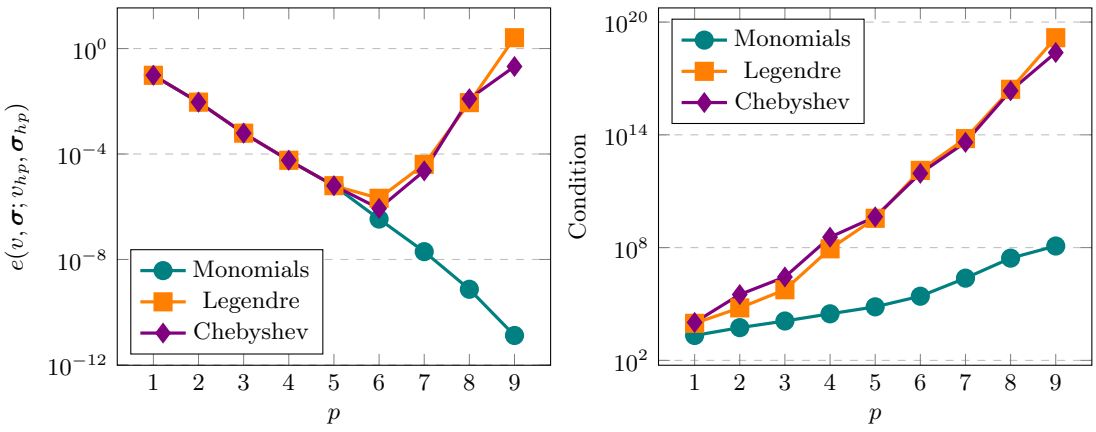

Figure 2.6: Different types of initial polynomial basis functions. Comparison of the error (left) and the conditioning of the system global matrix (right).

The results in Figure 2.6 show that all choices behave the same for low degrees. However, for higher degrees, Legendre and Chebychev polynomials fail to approximate the solution, due to the bad conditioning of the system matrix, compared to the monomials. The good properties of the two sets of basis functions do not carry over when developed in the recursion.

2.5.4 Tent pitching in 3 space dimensions

For this example, we choose as a spatial domain Ω the unit cube $(0, 1)^3$. As discussed in Section 2.3, we solve the tent pitched mesh elementwise, and in parallel. The initial quasi-uniform spatial mesh consists of tetrahedrons of maximal size h. We then use tent pitching in $3 + 1$ dimensions, until the algorithm stops at time $T = 1$, where we compute the error. The results of this are shown in Figure 2.7.

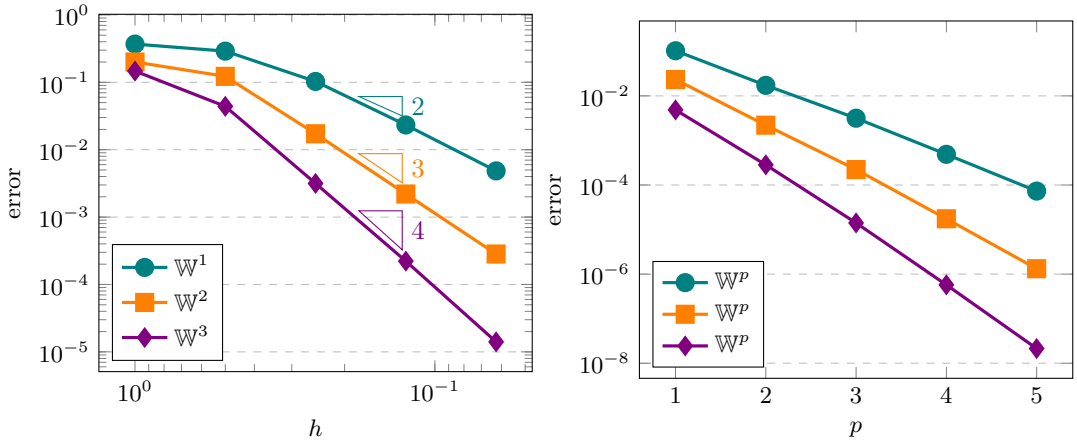

Figure 2.7: Tent pitching 3+1 dimensions. Convergence comparison with respect to the maximum mesh size on the left, and with respect to the Trefftz polynomial degree on the right.

In Figure 2.7 on the left, we plot the error in terms of h for different values of polynomial degree p of $\mathbb{W}^p(Q_T)$. Note that here we are plotting the error in $L^2(\Omega \times T)$ norm. Contrary to the DG norm, the L^2 norm is mesh independent. We observe the rate $\mathcal{O}(h^{p+1})$. We also consider convergence in terms of degree p of the Trefftz space $\mathbb{W}_p(\mathcal{T}_h)$, and report the results in Figure 2.7, right plots. For our analytic solution, we can observe exponential convergence.

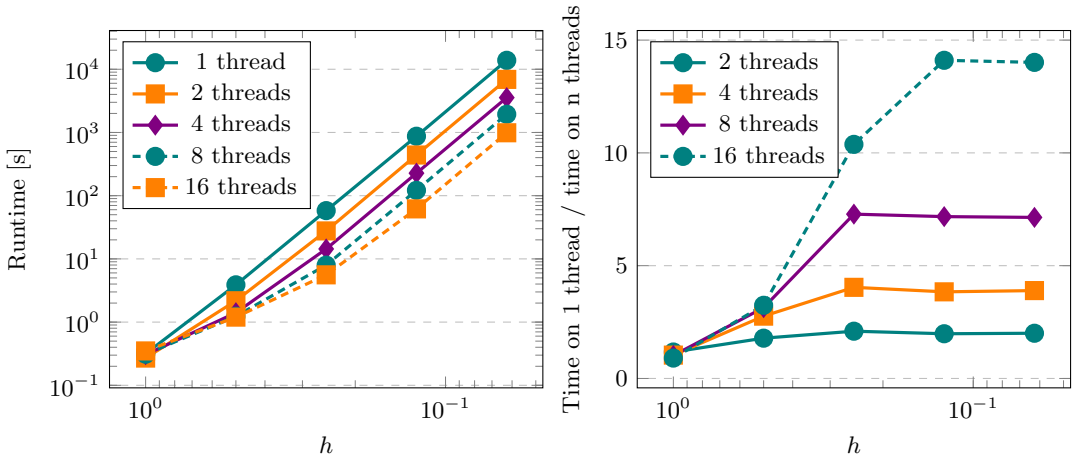

Figure 2.8: Comparison of computational time for different number of threads.

In Figure 2.8, we compare computational times for different number of threads used for the parallel processing of the tents. For large mesh sizes we observe limited speedup, as there are not enough elements to be processed in parallel. With decreasing meshsize the speedup is almost proportional to the number of threads. However, one has to take into account that some overhead is involved.

2.5.5 Dissipation of energy

For smooth enough functions $(w, \boldsymbol{\tau})$ the energy at a fixed time $\hat{t}$ is given by

$$E(w, \boldsymbol{\tau}) = \frac{1}{2} \int_{\Omega \times \hat{t}} \left(c^{-2} w^2 + |\boldsymbol{\tau}|^2 \right) \, dS.$$

In [81, §5.3], the Trefftz-DG method was shown to be dissipative, which we can also observe in numerical examples. We test on a model problem with analytical solution

$$u(x, t) = \sin(\pi x) \sin(\pi t),$$

on the domain $[0, 1] \times [0, T]$. We solve using the tent pitching algorithm.

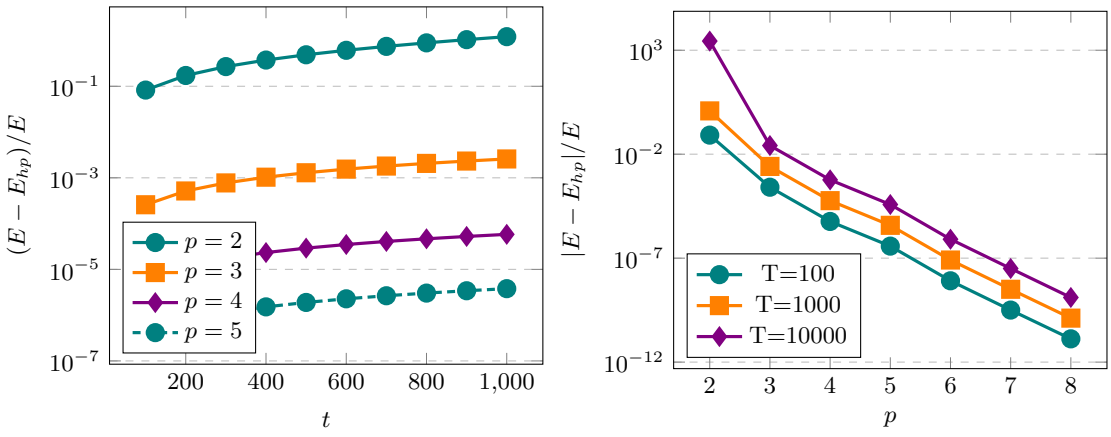

Figure 2.9: Error in energy over time and for different order of Trefftz polynomials.

The space mesh considered is a uniform partition of the interval $[0,1]$ into 5 elements. We measure the relative error in the energy given by

$$\frac{E(\partial_t u, -\nabla u) - E(u_{hp}, \boldsymbol{\sigma}_{hp})}{E(\partial_t u, -\nabla u)}.$$

In Figure 2.9 on the left, we can see that the error in energy increases in time. As the energy of the analytical solution is constant, and we are plotting the error without absolute value, we deduce that the energy of the numerical solution is decreasing in time. The results actually suggest that the energy of the numerical solution decreases linearly in time. In Figure 2.9 on the right, we compare the error in the energy at three different times $T = 100, 1000, 10000$, plotting it against the degree of Trefftz polynomials in a range from 2 to 8. We observe exponential convergence for increasing order. Furthermore, greater times T seem to affect the error only by a multiplicative factor.

2.5.6 Non-uniformly refined spatial meshes

Now that we have verified the convergence of the Trefftz-DG method with tent pitching initialized on quasi-uniform spatial meshes, we test the advantages of the method on a non-uniformly refined spatial mesh.

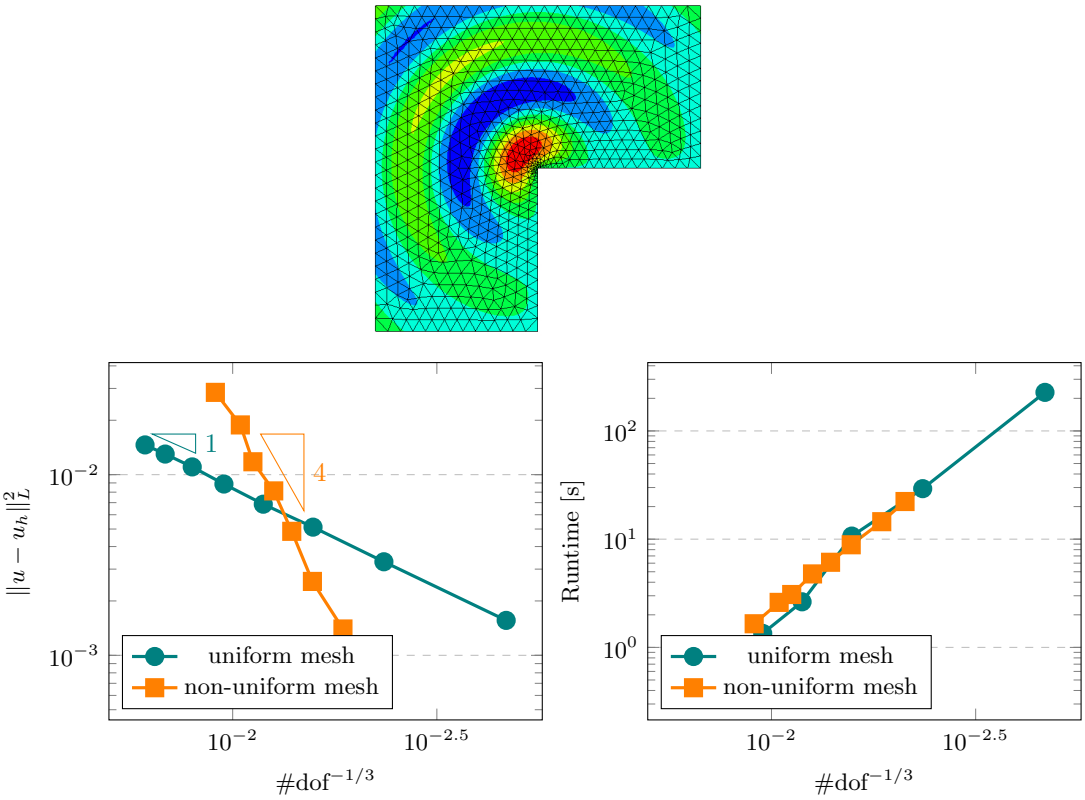

Figure 2.10: The convergence rates on uniform and non-uniform meshes and timings (bottom) with the singular initial condition (top) for Trefftz functions of order $p = 3$.

In this test, the refinement is applied to resolve a singular solution at the reentrant corner of an L-shaped domain, given by $\Omega = [-1,1]^2 \setminus ([0,1] \times [-1,0])$. The mesh refinement strategy used takes the diameter of a spatial mesh elements K as

$$h_K = h_{\max} r^{1-\mu},$$

where r is the distance of K to the reentrant corner, fixing a minimal mesh size of $h_{\min} = h_{\max}^{1/\mu}$. Motivated by the theoretical results in [4], we choose $\mu = \frac{1}{3}$.

We consider a model problem with solution given, in polar coordinates, by

$$u(r, \phi, t) = \cos(at)\sin(\nu\phi)J_\nu(ar), \tag{2.12}$$

where J_ν denotes the Bessel function of the first kind. We consider $\nu = 2/3$, so that ∇u is singular at the origin. We solve up to time $T = 1$ for $a = 10$. To avoid numerically integrate the singularity, we use the method to reconstruct the second order solution $u_{hp} \in \mathbb{U}^p(Q_T)$, introduced in Section 2.4, to measure the error given by $\|u(\cdot, T) - u_{hp}(\cdot, T)\|_{L^2(\Omega)}$.

The comparison between results obtained with uniform and non-uniform mesh refinement are shown in Figure 2.10 for Trefftz functions of degree $p = 3$. We compare the two different meshing strategies by plotting them against (global dof)$^{-1/3}$. For the uniformly refined meshes, the convergence rate is bounded by the smoothness of the solution $u \in H^{5/3-\varepsilon}(Q_T)$, for $\varepsilon > 0$. We observe a convergence rate of $\mathcal{O}(h^1)$. Using the non-uniformly refined meshes, we are able to recover optimal convergence for the third order Trefftz polynomials, as seen in Figure 2.10.

mesh	$h_{\max}$	total #dofs	L2-error	dof-rate	runtime [s]
uniform	0.07	3.2×10^5	1.8×10^{-2}	-	0.6234
	0.05	8.8×10^5	1.2×10^{-2}	1.2638	1.5916
	0.03	3.9×10^6	6.7×10^{-3}	1.1022	7.9157
	0.01	1.0×10^8	2.1×10^{-3}	1.0662	255.3233
non-uniform	0.12	1.1×10^6	2.2×10^{-2}	-	3.276
	0.10	2.0×10^6	8.3×10^{-3}	5.1308	4.6809
	0.08	3.8×10^6	3.0×10^{-3}	4.7041	8.0069
	0.06	9.8×10^6	8×10^{-4}	4.2104	23.4588

Table 2.1: Convergence rates and runtime comparison for a singular solution on the L-shapes domain, comparing uniform meshing and meshes refined towards the singularity.

Table 2.1, gives a closer look on some of the properties already visualized in Figure 2.10 and also shows the runtime (in seconds). For the computations we used 24 threads. In Figure 2.10 on the bottom right, we compare the runtime with the degrees of freedom. We observe that the uniform and the non-uniform mesh take about the same time for comparable numbers of degrees of freedom. Thus, no significant locking, due to the spatial refinement, occurs.

2.5.7 Wave propagation in an heterogeneous material

In the following example we investigate the reflection of a wave at an interface of two different materials. This experimental setup was also performed in [7, 64]. We consider the space–time domain $Q_T = [0, 2]^2 \times (0, 1]$, and problem (2.5) with homogeneous Dirichlet boundary conditions. The wavespeed is the piecewise constant function given by

$$c(x_1, x_2) = \begin{cases} 1 & x_1 \leq 1.2, \\ 3 & x_1 > 1.2. \end{cases}$$

As initial condition, we take a Gaussian wave given by

$$u_0(\boldsymbol{x}) = \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}_0\|^2/\delta^2\right), \; v_0(\boldsymbol{x}) = 0,$$

where we choose $\boldsymbol{x}_0 = (1, 1)$ and $\delta = 0.01$. The computations are performed with polynomial degree $p = 4$.

Snapshots of the solution are shown in Figure 2.12. In the Snapshots, the right part of the domain has spatial mesh sizes up to 0.03, whereas in the left part we choose as spatial mesh size

of 0.01, in order to better capture the steeper wavefront in the slower traveling material. First, we see that the initial condition unfolds in the left homogeneous part of the medium. At $T = 0.2$, the wave crosses over into the material with higher wave velocity. In the next snapshot we can see that the wave splits into a part traveling to the right with a higher velocity and shallow wavefront, and a part reflected at the interface traveling backwards to the left. Finally, at the time $T = 0.4$, we can also observe the weaker Huygens wave, which traveled parallel to the interface, before traveling back towards the left.

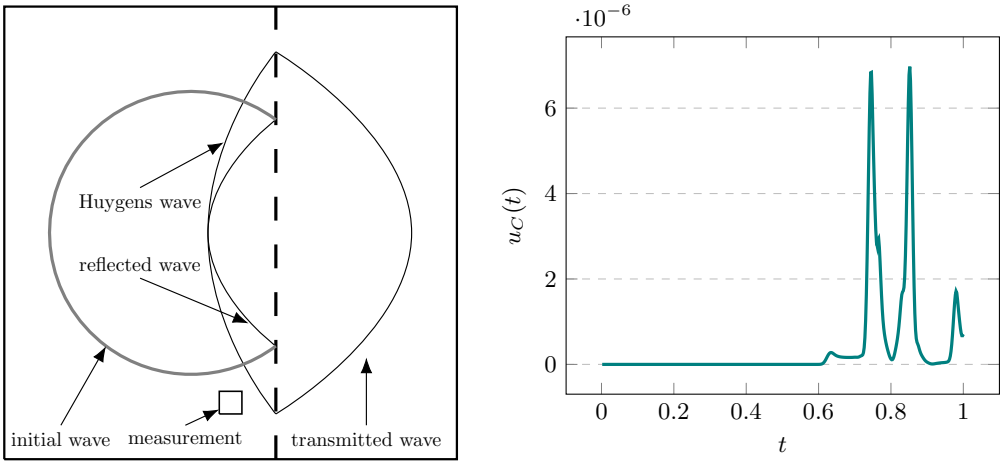

Figure 2.11: Sketch of the expected wave pattern (left) and measured output quantity (right).

In Figure 2.11 on the left we present a sketch of the actions described above, also indicating a region where we measured the output

$$u_C(t) = \|u(\cdot, t)\|_{L^1(\Omega_C)}.$$

The domain of measurement was chosen $\Omega_C = [1 - \varepsilon_C, 1 + \varepsilon_C] \times [0.25 - \varepsilon_C, 0.25 + \varepsilon_C]$, with $\varepsilon_C = 2^{-7}$. The measurement over time is presented in Figure 2.11 on the right and shows that we are able to distinguish the three incoming waves. We can see the very weak Huygens wave arriving first, followed by the initial wave and the reflected one.

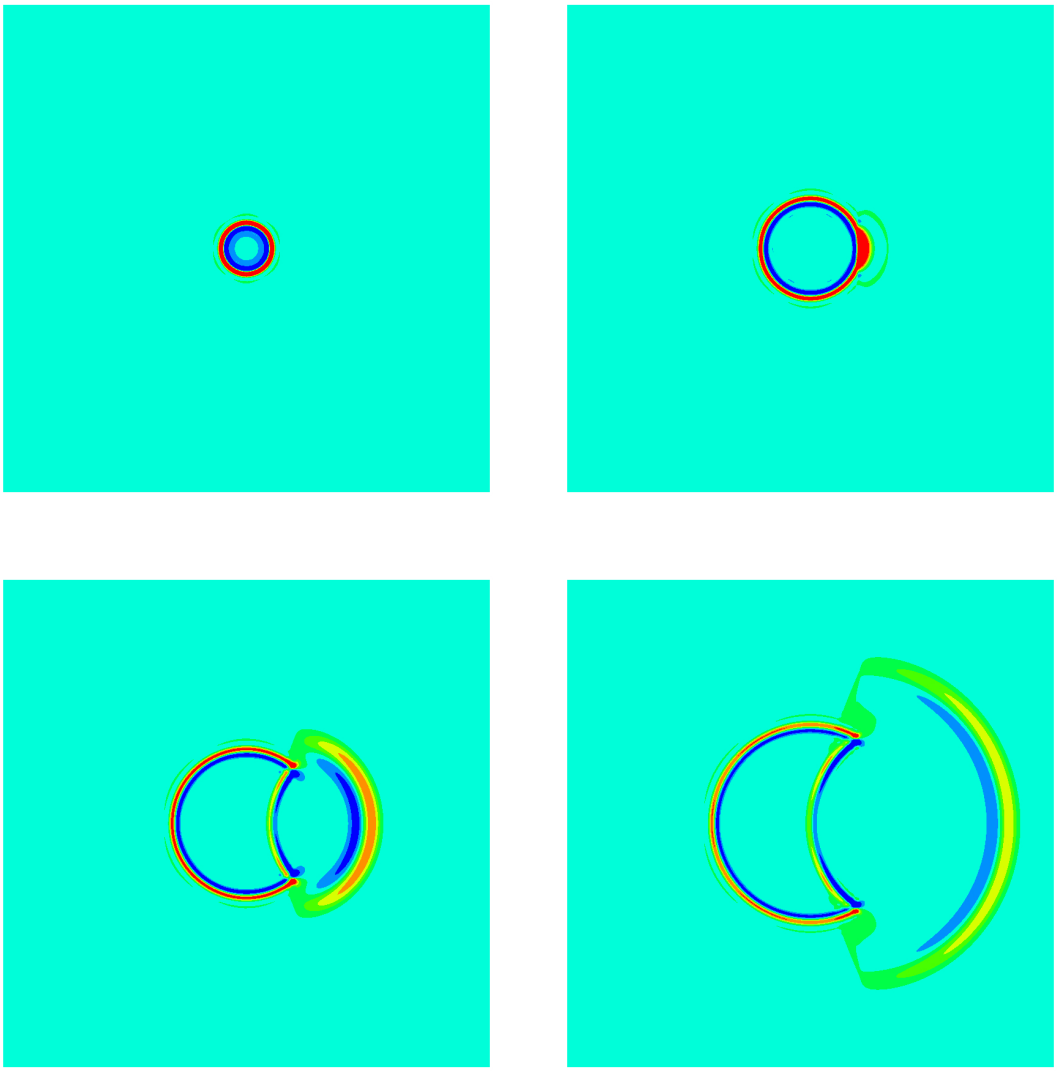

Figure 2.12: Wave traveling through inhomogeneous material, shown at times $T = 0.1, 0.2, 0.3, 0.4$.

Chapter 3

Quasi–Trefftz DG method for the wave equation with piecewise-smooth coefficient

We give several preliminary definitions and notation concerning the initial boundary value problems to be discretised in §3.1, the space–time meshes in §3.2.1, the numerical parameters needed for the definition of the DG scheme in §3.2.2, the mesh-dependent norms used in the error analysis in §3.2.4, along with Taylor polynomials, wave operators and anisotropic weighted norms in §3.3.1.

The variational formulation of the quasi-Trefftz DG method is introduced in §3.2.3. Remark 3.2.1 compares this formulation with some closely related ones appeared in [9, 81, 82, 90]. Well-posedness, stability and quasi-optimality of the quasi-Trefftz DG scheme are described in §3.2.5, heavily relying on [81]. This section also briefly lists several related results such as sharper bounds under more restrictive assumptions, energy dissipation, as well as error bounds on interfaces and partial cylinders.

The polynomial, local, discrete, quasi-Trefftz space $\mathbb{QU}^p(K)$ for the (smooth-wavenumber) second-order wave equation is defined in §3.3.2, p standing for the polynomial degree of the basis functions on a mesh element K. Proposition 3.3.2 shows that for an appropriate choice of p all wave equation solutions are approximated by this space with high orders in the (space–time) element size. This is a fully-explicit, high-order, h-convergence result; on the other hand p-convergence results (i.e. regarding convergence for increasing polynomial degrees) on general elements are not available, neither for the Trefftz DG for the constant-coefficients wave equation [81], nor for the GPW-based DG scheme for the Helmholtz equation [52]. These best-approximation estimates lead to convergence bounds for the quasi-Trefftz DG scheme in §3.3.3. In particular, for a suitable choice of the numerical parameters entering the DG formulation, we obtain the same orders of convergence as in the constant-coefficient case [81, §6], even if we require stronger solution regularity (see Remark 3.3.7).

For the method to be practical, of course one needs to be able to explicitly compute the basis functions: we describe a family of bases in §3.3.4. Given any basis of the classical polynomial space in n real variables, Algorithms 1 and 2 give a simple recipe for the computation of a corresponding quasi-Trefftz basis (in $n = 1$ and $n > 1$ space dimensions, respectively).

Sections 3.3.2–3.3.4 focus on quasi-Trefftz schemes for solutions of the second-order wave equation. However the DG scheme (3.7) applies more generally to the first-order acoustic wave equation. Thus §3.3.5 briefly describes another quasi-Trefftz discrete space, suited for the acoustic first-order system, together with the recipe for the computation of its basis. In the constant-coefficient case the two classes of discrete spaces were proposed and analysed in [81, §6.1–6.2].

In Section 3.4 we illustrate the results of several numerical experiments for the implementation

of the quasi-Trefftz DG method in Netgen/NGSolve[1] [99,100]. In particular, we briefly discuss the dependence on the penalty parameters and the orders of convergence, we compare the proposed method against standard polynomial and Trefftz DG schemes, and consider both prismatic and "tent-pitched" meshes, corresponding to implicit and semi-explicit time-stepping respectively.

3.1 Acoustic wave equation with variable coefficient

We consider the IBVP given in Equations (2.1) and (2.2), now with wavespeed $0 < c \in L^\infty(\Omega)$ assumed to be piecewise-smooth in space. We recall the IBVP:

$$\begin{cases} \nabla \cdot \boldsymbol{\sigma} + c^{-2}\partial_t v = 0 & \text{in } Q_T, \\ \nabla v + \partial_t \boldsymbol{\sigma} = \mathbf{0} & \text{in } Q_T, \\ v(\cdot,0) = v_0,\ \boldsymbol{\sigma}(\cdot,0) = \boldsymbol{\sigma}_0 & \text{on } \Omega, \\ v = g_D & \text{on } \Gamma_D \times [0,T], \\ \boldsymbol{n}_\Omega^x \cdot \boldsymbol{\sigma} = g_N & \text{on } \Gamma_N \times [0,T], \\ \frac{\vartheta}{c} v - \boldsymbol{\sigma} \cdot \boldsymbol{n}_\Omega^x = g_R & \text{on } \Gamma_R \times [0,T], \end{cases} \tag{3.1}$$

With the same notation and assumptions, the corresponding IBVP for the second-order (scalar) wave equations reads as

$$\begin{cases} -\Delta u + c^{-2}\partial_t^2 u = 0 & \text{in } Q_T, \\ \partial_t u(\cdot,0) = v_0,\ u(\cdot,0) = u_0 & \text{on } \Omega, \\ \partial_t u = g_D & \text{on } \Gamma_D \times [0,T], \\ -\boldsymbol{n}_\Omega^x \cdot \nabla u = g_N & \text{on } \Gamma_N \times [0,T], \\ \frac{\vartheta}{c}\partial_t u + \nabla u \cdot \boldsymbol{n}_\Omega^x = g_R & \text{on } \Gamma_R \times [0,T]. \end{cases} \tag{3.2}$$

3.2 Discontinuous Galerkin discretisation

In this section we closely follow [81, §3–5]; we extend assumptions, notation, definitions and results to the case of piecewise-smooth c and non-Trefftz discretisations.

3.2.1 Mesh assumptions and notation

The space–time domain Q_T is subdivided in a non-overlapping mesh $\mathcal{T}_h$, where every element $K \in \mathcal{T}_h$ is an $n+1$-dimensional Lipschitz polytope. We assume that $c|_K \in C^\infty(K)$ for all $K \in \mathcal{T}_h$, and that each face ($F = \overline{K_1} \cap \overline{K_2}$ with positive n-dimensional measure for some $K_1, K_2 \in \mathcal{T}_h$) is an n-dimensional polytope. We denote by $(\mathbf{n}_F^x, n_F^t) \in \mathbb{R}^{n+1}$ the unit normal vector orthogonal to a mesh face F, with $n_F^t \geq 0$ and $|\mathbf{n}_F^x|^2 + (n_F^t)^2 = 1$. We assume that each face F is either

$$\begin{aligned} &\text{space-like, i.e. } |\mathbf{n}_F^x| \sup_{(\mathbf{x},t)\in F} c(\mathbf{x}) \leq n_F^t, \\ &\text{or time-like, i.e. } n_F^t = 0. \end{aligned} \tag{3.3}$$

A space-like face F lies below (i.e. in the past of) the cone of dependance of each of its points; its slope (when seen as the graph the function $\mathbf{x} \mapsto t$ such that $(\mathbf{x},t) \in F$) is bounded by $1/c(\mathbf{x})$. A time-like face is a union of segments parallel to the time axis. The class of meshes includes both Cartesian-product meshes such as those of [9] ($\mathbf{n}_F^x = 0$, $n_F^t = 1$ on all space-like faces) and tent-pitched meshes such as those of [82,90] $\left(n_F^t \approx \frac{c}{\sqrt{1+c^2}}\right)$; see two examples plotted in Figure 3.5.

[1] The code is available online at https://github.com/PaulSt/NGSTrefftz

We choose a "centre point" $(\mathbf{x}_K, t_K) \in K$ for each mesh element $K \in \mathcal{T}_h$, for example the barycentre, which will be used in the proof of the approximation estimates and to define the basis functions. We define a radius and a "weighted radius" of each element as

$$r_K := \sup_{(\mathbf{x},t)\in K} |(\mathbf{x},t) - (\mathbf{x}_K, t_K)|, \qquad r_{K,c} := \sup_{(\mathbf{x},t)\in K} \left|\big(\mathbf{x}, c(\mathbf{x})t\big) - \big(\mathbf{x}_K, c(\mathbf{x}_K)t_K\big)\right|, \tag{3.4}$$

with $|\cdot|$ the Euclidean distance in $\mathbb{R}^{n+1}$.

We introduce a piecewise-constant function γ defined on $\mathcal{F}_h^{\text{space}} \cup \mathcal{F}_h^0 \cup \mathcal{F}_h^T$, measuring how close to characteristic cones the space-like mesh faces are:

$$\gamma := \frac{\|c\|_{C^0(F)} \, |\mathbf{n}_F^x|}{n_F^t} \text{ on } F \subset \mathcal{F}_h^{\text{space}}, \qquad \gamma := 0 \text{ on } \mathcal{F}_h^0 \cup \mathcal{F}_h^T. \tag{3.5}$$

We define a "space-like interface" as a connected union of space-like faces $\Sigma \subset \mathcal{F}_h^{\text{space}} \cup \mathcal{F}_h^0 \cup \mathcal{F}_h^T$ that is the graph of a Lipschitz-continuous function $f_\Sigma : \overline{\Omega} \to [0,T]$. By (3.3), the Lipschitz constant of f_Σ in $\mathbf{x} \in \Omega$ will be at most $c^{-1}(\mathbf{x})$. The unit normal vector on Σ is denoted $(\mathbf{n}_\Sigma^x, n_\Sigma^t)$.

3.2.2 DG flux and penalisation parameters

We fix three "numerical flux parameter" functions on portions of the mesh skeleton, and two "volume penalisation coefficients":

$$\alpha \in L^\infty\left(\mathcal{F}_h^{\text{time}} \cup \mathcal{F}_h^{\text{D}}\right), \qquad \beta \in L^\infty\left(\mathcal{F}_h^{\text{time}} \cup \mathcal{F}_h^{\text{N}}\right), \qquad \delta \in L^\infty\left(\mathcal{F}_h^{\text{R}}\right), \qquad \mu_1, \mu_2 \in L^\infty(Q_T).$$

We assume that all these are uniformly positive and bounded:

$$\alpha, \beta, \delta, \mu_1, \mu_2 > 0, \qquad \|\delta\|_{L^\infty(\mathcal{F}_h^{\text{R}})} < 1,$$

$$\left\|\alpha^{-1}\right\|_{L^\infty(\mathcal{F}_h^{\text{time}} \cup \mathcal{F}_h^{\text{D}})}, \left\|\beta^{-1}\right\|_{L^\infty(\mathcal{F}_h^{\text{time}} \cup \mathcal{F}_h^{\text{N}})}, \left\|\delta^{-1}\right\|_{L^\infty(\mathcal{F}_h^{\text{R}})}, \left\|\mu_1^{-1}\right\|_{L^\infty(Q_T)}, \left\|\mu_2^{-1}\right\|_{L^\infty(Q_T)} < \infty.$$

We also define the values

$$\begin{aligned} \mu_{K+} &:= \max\left\{ \|\mu_1\|_{L^\infty(K)}, \|\mu_2\|_{L^\infty(K)} \right\}, \\ \mu_{K-} &:= \max\left\{ \left\|\mu_1^{-1}\right\|_{L^\infty(K)}, \left\|\mu_2^{-1}\right\|_{L^\infty(K)} \right\}, \quad \forall K \in \mathcal{T}_h. \end{aligned} \tag{3.6}$$

3.2.3 DG formulation

Let $\mathbf{V}_{hp}(\mathcal{T}_h)$ be a closed (e.g. finite-dimensional) subspace of the broken Sobolev space

$$\mathbf{H}(\mathcal{T}_h) := \prod_{K\in\mathcal{T}_h} \left(H^1(K) \times H^1(K)^n\right).$$

We consider the following variational formulation:

$$\begin{aligned} &\text{Seek} \quad (v_{hp}, \boldsymbol{\sigma}_{hp}) \in \mathbf{V}_{hp}(\mathcal{T}_h) \\ &\text{such that} \quad \mathcal{A}(v_{hp}, \boldsymbol{\sigma}_{hp}; w, \boldsymbol{\tau}) = \ell(w, \boldsymbol{\tau}) \quad \forall (w, \boldsymbol{\tau}) \in \mathbf{V}_{hp}(\mathcal{T}_h), \\ &\text{where} \end{aligned} \tag{3.7}$$

$$\begin{aligned} \mathcal{A}(v_{hp},\boldsymbol{\sigma}_{hp}; w, \boldsymbol{\tau}) := &- \sum_{K\in\mathcal{T}_h} \int_K \left(v_{hp}\Big(\nabla\cdot\boldsymbol{\tau} + c^{-2}\partial_t w\Big) + \boldsymbol{\sigma}_{hp}\cdot\Big(\partial_t\boldsymbol{\tau} + \nabla w\Big)\right) \mathrm{d}V \\ &+ \int_{\mathcal{F}_h^{\text{space}}} \left(c^{-2} v_{hp}^- [\![w]\!]_t + \boldsymbol{\sigma}_{hp}^- \cdot [\![\boldsymbol{\tau}]\!]_t + v_{hp}^- [\![\boldsymbol{\tau}]\!]_{\mathbf{N}} + \boldsymbol{\sigma}_{hp}^- \cdot [\![w]\!]_{\mathbf{N}}\right) \mathrm{d}S \\ &+ \int_{\mathcal{F}_h^T} (c^{-2} v_{hp} w + \boldsymbol{\sigma}_{hp}\cdot\boldsymbol{\tau}) \, \mathrm{d}\mathbf{x} \end{aligned}$$

$$
\begin{aligned}
&+ \int_{\mathcal{F}_h^{\mathrm{time}}} \big(\{\!\{ v_{hp} \}\!\} [\![\boldsymbol{\tau}]\!]_{\mathbf{N}} + \{\!\{ \boldsymbol{\sigma}_{hp} \}\!\} \cdot [\![w]\!]_{\mathbf{N}} + \alpha [\![v_{hp}]\!]_{\mathbf{N}} \cdot [\![w]\!]_{\mathbf{N}} + \beta [\![\boldsymbol{\sigma}_{hp}]\!]_{\mathbf{N}} [\![\boldsymbol{\tau}]\!]_{\mathbf{N}} \big) \, \mathrm{d}S \\
&+ \int_{\mathcal{F}_h^{\mathrm{D}}} \big(\boldsymbol{\sigma}_{hp} \cdot \mathbf{n}_\Omega^x \, w + \alpha v_{hp} w \big) \, \mathrm{d}S + \int_{\mathcal{F}_h^{\mathrm{N}}} \big(v_{hp} (\boldsymbol{\tau} \cdot \mathbf{n}_\Omega^x) + \beta (\boldsymbol{\sigma}_{hp} \cdot \mathbf{n}_\Omega^x)(\boldsymbol{\tau} \cdot \mathbf{n}_\Omega^x) \big) \, \mathrm{d}S \\
&+ \int_{\mathcal{F}_h^{\mathrm{R}}} \Big(\frac{(1-\delta)\vartheta}{c} v_{hp} w + (1-\delta) v_{hp} (\boldsymbol{\tau} \cdot \mathbf{n}_\Omega^x) + \delta (\boldsymbol{\sigma}_{hp} \cdot \mathbf{n}_\Omega^x) w + \frac{\delta c}{\vartheta} (\boldsymbol{\sigma}_{hp} \cdot \mathbf{n}_\Omega^x)(\boldsymbol{\tau} \cdot \mathbf{n}_\Omega^x) \Big) \, \mathrm{d}S \\
&+ \sum_{K \in \mathcal{T}_h} \int_K \mu_1 c^2 \big(\nabla \cdot \boldsymbol{\sigma}_{hp} + c^{-2} \partial_t v_{hp} \big) \big(\nabla \cdot \boldsymbol{\tau} + c^{-2} \partial_t w \big) \, \mathrm{d}V \\
&+ \sum_{K \in \mathcal{T}_h} \int_K \mu_2 \big(\partial_t \boldsymbol{\sigma}_{hp} + \nabla v_{hp} \big) \cdot \big(\partial_t \boldsymbol{\tau} + \nabla w \big) \, \mathrm{d}V, \\
\ell(w, \boldsymbol{\tau}) &:= \int_{\mathcal{F}_h^0} (c^{-2} v_0 w + \boldsymbol{\sigma}_0 \cdot \boldsymbol{\tau}) \, \mathrm{d}\mathbf{x} + \int_{\mathcal{F}_h^{\mathrm{D}}} g_D \big(\alpha w - \boldsymbol{\tau} \cdot \mathbf{n}_\Omega^x \big) \, \mathrm{d}S \\
&+ \int_{\mathcal{F}_h^{\mathrm{N}}} g_N \big(\beta \, \boldsymbol{\tau} \cdot \mathbf{n}_\Omega^x - w \big) \, \mathrm{d}S + \int_{\mathcal{F}_h^{\mathrm{R}}} g_R \Big((1-\delta) w - \frac{\delta c}{\vartheta} \, \boldsymbol{\tau} \cdot \mathbf{n}_\Omega^x \Big) \, \mathrm{d}S.
\end{aligned}
$$

Noting that all terms involving c are integrated by parts with respect to the time variable only, the derivation in [81, §4] shows that the formulation (3.7) is consistent:

$$
\text{if } (v, \boldsymbol{\sigma}) \in \mathbf{H}(\mathcal{T}_h) \text{ is solution of (3.1), then } \quad \mathcal{A}(v, \boldsymbol{\sigma}; w, \boldsymbol{\tau}) = \ell(w, \boldsymbol{\tau}) \quad \forall (w, \boldsymbol{\tau}) \in \mathbf{H}(\mathcal{T}_h).
$$

Remark 3.2.1. *The differences between (3.7) and [81, equation (7)] (equiv. equation (2.5)) are the following: (i) we allow position-dependent and possibly discontinuous wavespeed c; (ii) we allow fields that are not local solution of the PDE system (i.e. our method is not Trefftz); (iii) as a consequence we have a volume term in $\mathcal{A}(\cdot,\cdot)$, ensuring consistency; (iv) we have a further stabilisation/penalisation volume term (the term involving μ_1, μ_2). This term can be understood as a Galerkin–least squares (GLS) correction.*

The formulation of [81] has been studied also in [90] and extended to the non-Trefftz case, with piecewise-constant coefficient, in [9] (with tensor-product and sparse polynomial bases). With appropriate choices of the numerical flux parameters the present formulation is a special case of that in [82] (see the comparison in [81, Rem. 4]).

Although the variational problem (3.7) couples the discrete solution on all mesh elements in Q_T, the structure of the terms on $\mathcal{F}_h^{\mathrm{space}}$ (the space-like part of the mesh skeleton) allows to compute the solution $(v_{hp}, \boldsymbol{\sigma}_{hp})$ by solving a sequence of smaller linear systems; see [81, p. 398]. E.g., if the elements of a quasi-uniform mesh can be grouped in N "time-slabs" $\Omega \times (t_{i-1}, t_i)$, with $0 = t_0 < t_1 < \cdots < t_N = T$, $t_i - t_{i-1} \approx T/N$, then the discrete solution on each time-slab can be computed from the solution on the previous time-slab, solving N linear systems of size $O(\dim \mathbf{V}_{hp}(\mathcal{T}_h)/N)$ each. This is equivalent to an implicit time-stepping.

3.2.4 Mesh-dependent norms

We define two mesh- and flux-dependent norms on $\mathbf{T}(\mathcal{T}_h)$:

$$
\begin{aligned}
|||(w, \boldsymbol{\tau})|||_{\mathrm{DG}}^2 := \; & \frac{1}{2} \left\| \left(\frac{1-\gamma}{n_F^t} \right)^{1/2} c^{-1} [\![w]\!]_t \right\|_{L^2(\mathcal{F}_h^{\mathrm{space}})}^2 + \frac{1}{2} \left\| \left(\frac{1-\gamma}{n_F^t} \right)^{1/2} [\![\boldsymbol{\tau}]\!]_t \right\|_{L^2(\mathcal{F}_h^{\mathrm{space}})^n}^2 \\
& + \frac{1}{2} \left\| c^{-1} w \right\|_{L^2(\mathcal{F}_h^0 \cup \mathcal{F}_h^T)}^2 + \frac{1}{2} \left\| \boldsymbol{\tau} \right\|_{L^2(\mathcal{F}_h^0 \cup \mathcal{F}_h^T)^n}^2 \\
& + \left\| \alpha^{1/2} [\![w]\!]_{\mathbf{N}} \right\|_{L^2(\mathcal{F}_h^{\mathrm{time}})^n}^2 + \left\| \beta^{1/2} [\![\boldsymbol{\tau}]\!]_{\mathbf{N}} \right\|_{L^2(\mathcal{F}_h^{\mathrm{time}})}^2 \\
& + \left\| \alpha^{1/2} w \right\|_{L^2(\mathcal{F}_h^{\mathrm{D}})}^2 + \left\| \beta^{1/2} \boldsymbol{\tau} \cdot \mathbf{n}_\Omega^x \right\|_{L^2(\mathcal{F}_h^{\mathrm{N}})}^2
\end{aligned}
\tag{3.8}
$$

$$
\begin{aligned}
&+ \left\| \Big(\frac{(1-\delta)\vartheta}{c} \Big)^{1/2} w \right\|^2_{L^2(\mathcal{F}_h^{\mathrm{R}})} + \left\| \Big(\frac{\delta c}{\vartheta} \Big)^{1/2} \boldsymbol{\tau} \cdot \mathbf{n}_\Omega^x \right\|^2_{L^2(\mathcal{F}_h^{\mathrm{R}})} \\
&+ \sum_{K \in \mathcal{T}_h} \left(\left\| \mu_1^{1/2} (c \nabla \cdot \boldsymbol{\tau} + c^{-1} \partial_t w) \right\|^2_{L^2(K)} + \left\| \mu_2^{1/2} (\partial_t \boldsymbol{\tau} + \nabla w) \right\|^2_{L^2(K)^n} \right); \\
|||(w,\boldsymbol{\tau})|||^2_{\mathrm{DG}^+} &:= |||(w,\boldsymbol{\tau})|||^2_{\mathrm{DG}} \\
&+ 2 \left\| \Big(\frac{n_F^t}{1-\gamma} \Big)^{1/2} c^{-1} w^- \right\|^2_{L^2(\mathcal{F}_h^{\mathrm{space}})} + 2 \left\| \Big(\frac{n_F^t}{1-\gamma} \Big)^{1/2} \boldsymbol{\tau}^- \right\|^2_{L^2(\mathcal{F}_h^{\mathrm{space}})^n} \\
&+ \left\| \beta^{-1/2} \{\!\{ w \}\!\} \right\|^2_{L^2(\mathcal{F}_h^{\mathrm{time}})} + \left\| \alpha^{-1/2} \{\!\{ \boldsymbol{\tau} \}\!\} \right\|^2_{L^2(\mathcal{F}_h^{\mathrm{time}})^n} \\
&+ \left\| \alpha^{-1/2} \boldsymbol{\tau} \cdot \mathbf{n}_\Omega^x \right\|^2_{L^2(\mathcal{F}_h^{\mathrm{D}})} + \left\| \beta^{-1/2} w \right\|^2_{L^2(\mathcal{F}_h^{\mathrm{N}})} \\
&+ \sum_{K \in \mathcal{T}_h} \left(\left\| \mu_1^{-1/2} c^{-1} w \right\|^2_{L^2(K)} + \left\| \mu_2^{-1/2} \boldsymbol{\tau} \right\|^2_{L^2(K)} \right).
\end{aligned}
$$

These are *norms* on the broken Sobolev space $\mathbf{H}(\mathcal{T}_h)$ defined on the mesh $\mathcal{T}_h$. Indeed, $|||(w,\boldsymbol{\tau})||| = 0$ for $(w,\boldsymbol{\tau}) \in \mathbf{H}(\mathcal{T}_h)$ implies that $(w,\boldsymbol{\tau})$ is solution of the IBVP with zero initial and boundary conditions, so $(w,\boldsymbol{\tau}) = (0,\mathbf{0})$ by the well-posedness of the IBVP itself; see [67, Lemma 4.1].

As in [81, §5.3], we define the energy of a field $(w,\boldsymbol{\tau}) \in \mathbf{H}(\mathcal{T}_h)$ on a space-like interface Σ as

$$
\mathcal{E}(\Sigma; w, \boldsymbol{\tau}) := \int_\Sigma \Big(w \boldsymbol{\tau} \cdot \mathbf{n}_\Sigma^x + \frac{1}{2} (c^{-2} w^2 + |\boldsymbol{\tau}|^2) n_\Sigma^t \Big) \, \mathrm{d}S. \tag{3.9}
$$

3.2.5 Well-posedness, stability, quasi-optimality

Integration by part on a mesh element gives for any field $(w,\boldsymbol{\tau}) \in \mathbf{H}(\mathcal{T}_h)$

$$
\int_K w \Big(\nabla \cdot \boldsymbol{\tau} + c^{-2} \partial_t w \Big) + \boldsymbol{\tau} \cdot \Big(\partial_t \boldsymbol{\tau} + \nabla w \Big) \, \mathrm{d}V - \int_{\partial K} w \boldsymbol{\tau} \cdot \mathbf{n}_K^x + \frac{1}{2} \Big(c^{-2} w^2 + |\boldsymbol{\tau}|^2 \Big) n_K^t \, \mathrm{d}S = 0. \tag{3.10}
$$

The results of [81, §5.2] hold also in the current, slightly extended, setting and are summarised in the following theorem.

Theorem 3.2.2. *The bilinear form $\mathcal{A}$ is coercive in $|||\cdot|||_{\mathrm{DG}}$ norm and continuous in $|||\cdot|||_{\mathrm{DG}^+} - |||\cdot|||_{\mathrm{DG}}$ norms, and the linear functional ℓ is continuous:*

$$
\begin{aligned}
&\mathcal{A}(w,\boldsymbol{\tau}; w,\boldsymbol{\tau}) \geq |||(w,\boldsymbol{\tau})|||^2_{\mathrm{DG}}, && (3.11) \\
&|\mathcal{A}(v,\boldsymbol{\sigma}; w,\boldsymbol{\tau})| \leq C_c |||(v,\boldsymbol{\sigma})|||_{\mathrm{DG}^+} |||(w,\boldsymbol{\tau})|||_{\mathrm{DG}}, \quad \textit{where} \\
&\qquad C_c := \begin{cases} 2, & \textit{if } \mathcal{F}_h^{\mathrm{R}} = \emptyset, \\ 2 \max \Big\{ \big\| \frac{1-\delta}{\delta} \big\|^{1/2}_{L^\infty(\mathcal{F}_h^{\mathrm{R}})}, \big\| \frac{\delta}{1-\delta} \big\|^{1/2}_{L^\infty(\mathcal{F}_h^{\mathrm{R}})} \Big\} & \textit{if } \mathcal{F}_h^{\mathrm{R}} \neq \emptyset, \end{cases} && (3.12) \\
&|\ell(w,\boldsymbol{\tau})| \leq \Big(2 \left\| c^{-1} v_0 \right\|^2_{L^2(\mathcal{F}_h^0)} + 2 \left\| \boldsymbol{\sigma}_0 \right\|^2_{L^2(\mathcal{F}_h^0)} + 2 \left\| \alpha^{1/2} g_D \right\|^2_{L^2(\mathcal{F}_h^{\mathrm{D}})} \\
&\qquad + 2 \left\| \beta^{1/2} g_N \right\|^2_{L^2(\mathcal{F}_h^{\mathrm{N}})} + \left\| (c/\vartheta)^{1/2} g_R \right\|^2_{L^2(\mathcal{F}_h^{\mathrm{R}})} \Big)^{1/2} |||(w,\boldsymbol{\tau})|||_{\mathrm{DG}^+}.
\end{aligned}
$$

The variational problem (3.7) admits a unique solution $(v_{hp},\boldsymbol{\sigma}_{hp}) \in \mathbf{V}_{hp}(\mathcal{T}_h)$, for any choice of $\mathbf{V}_{hp}(\mathcal{T}_h)$. The discrete solution satisfies the error bound

$$
\boxed{|||(v - v_{hp}, \boldsymbol{\sigma} - \boldsymbol{\sigma}_{hp})|||_{\mathrm{DG}} \leq (1 + C_c) \inf_{(w,\boldsymbol{\tau}) \in \mathbf{V}_{hp}(\mathcal{T}_h)} |||(v - w, \boldsymbol{\sigma} - \boldsymbol{\tau})|||_{\mathrm{DG}^+}.} \tag{3.13}
$$

Moreover, if $g_D = g_N = 0$ (or the corresponding parts $\mathcal{F}_h^{\mathrm{D}}, \mathcal{F}_h^{\mathrm{N}}$ of the boundary are empty) then

$$|||(v_{hp}, \boldsymbol{\sigma}_{hp})|||_{\mathrm{DG}} \leq \left(2 \left\| c^{-1} v_0 \right\|^2_{L^2(\mathcal{F}_h^0)} + 2 \left\| \boldsymbol{\sigma}_0 \right\|^2_{L^2(\mathcal{F}_h^0)} + \left\| (c/\vartheta)^{1/2} g_R \right\|^2_{L^2(\mathcal{F}_h^{\mathrm{R}})} \right)^{1/2}.$$

Of the differences between the methods in §3.2.3 and in [81] listed in Remark 3.2.1: (i) is unimportant for the proof of Theorem 3.2.2 as the terms involving c are integrated by parts in time only; (ii) does not affect the theorem as the Trefftz property is replaced by the presence of the first volume term in (3.7); the term described in (iii) is taken care by the identity (3.10); the term of (iv) coincides with the new term in the $|||\cdot|||_{\mathrm{DG}}$ norm.

The error bound (3.13) slightly differs from the quasi-optimality result (Céa lemma) in classical FEM analysis in that the norm ($|||\cdot|||_{\mathrm{DG}^+}$) at the right-hand side is stronger than that ($|||\cdot|||_{\mathrm{DG}}$) at the left-hand side. This mismatch is typical of the DG formulation employed here, not only for hyperbolic equations (as in [9, 67, 81]) but also for the analogous discretisation of the Helmholtz equation, see [50, §2.2.1].

Under more restrictive assumptions, slightly stronger results are possible. On a mesh made of Cartesian-product elements, or more generally if $\mathbf{n}_F^x = \mathbf{0}$ on all faces $F \subset \mathcal{F}_h^{\mathrm{space}}$, then the coercivity inequality (3.11) is an equality: $\mathcal{A}(w, \boldsymbol{\tau}; w, \boldsymbol{\tau}) = |||(w, \boldsymbol{\tau})|||^2_{\mathrm{DG}}$ for all $(w, \boldsymbol{\tau}) \in \mathbf{T}(\mathcal{T}_h)$. If $g_D = g_N = 0$ (or the corresponding parts $\mathcal{F}_h^{\mathrm{D}}, \mathcal{F}_h^{\mathrm{N}}$ of the boundary are empty) then the $|||(w, \boldsymbol{\tau})|||_{\mathrm{DG}^+}$ norm at the right-hand side of the bound on $|\ell(w, \boldsymbol{\tau})|$ can be substituted by $|||(w, \boldsymbol{\tau})|||_{\mathrm{DG}}$.

The bound (3.13) allows to control the DG error only in the $|||\cdot|||_{\mathrm{DG}}$ norm, which involves jumps on internal faces. However a simple adaptation of the proof allows to control the L^2 norm of the traces on space-like interfaces of the error. Let Σ be a space-like interface, as defined in §3.2.1. Assume that $\mathbf{V}_{hp}(\mathcal{T}_h) = \{(w, \boldsymbol{\tau}) \in \mathbf{V}_{hp}(\mathcal{T}_h), \mathrm{supp}\,(w, \boldsymbol{\tau}) \subset \{(\mathbf{x}, t), 0 \leq t \leq f_\Sigma(\mathbf{x})\}\} \oplus \{(w, \boldsymbol{\tau}) \in \mathbf{V}_{hp}(\mathcal{T}_h), \mathrm{supp}\,(w, \boldsymbol{\tau}) \subset \{(\mathbf{x}, t), f_\Sigma(\mathbf{x}) \leq t \leq T\}\}$ (i.e. that the discrete functions are indeed discontinuous across Σ). Then [81, Prop. 1] gives that

$$\mathcal{E}(\Sigma; v - v_{hp}^-, \boldsymbol{\sigma} - \boldsymbol{\sigma}_{hp}^-) \leq \frac{5}{2} \left\| (1-\gamma)^{-1} \right\|_{L^\infty(\Sigma)} (1 + C_c)^2 \inf_{(w,\boldsymbol{\tau}) \in \mathbf{V}_{hp}(\mathcal{T}_h)} |||(v - w, \boldsymbol{\sigma} - \boldsymbol{\tau})|||^2_{\mathrm{DG}^+}.$$

If $g_D = g_N = 0$ and $\Gamma_{\mathrm{R}} = \emptyset$ the IBVP (3.1) preserves energy: $\mathcal{E}(\mathcal{F}_h^T; v, \boldsymbol{\sigma}) = \mathcal{E}(\mathcal{F}_h^0; v, \boldsymbol{\sigma})$. The DG scheme dissipates energy: $\mathcal{E}(\mathcal{F}_h^T; v_{hp}, \boldsymbol{\sigma}_{hp}) \leq \mathcal{E}(\mathcal{F}_h^0; v_{hp}, \boldsymbol{\sigma}_{hp})$; the dissipation can be quantified in terms of the jumps on the faces and the residual in the mesh elements with the same technique of [81, §5.3] and [9, Rem. 5.7].

For any space-like interface Σ, the results of Theorem 3.2.2 can be localised to the partial space–time cylinder $Q_\Sigma = \{(\mathbf{x}, t), \mathbf{x} \in \Omega, 0 < t < f_\Sigma(\mathbf{x})\}$, by proceeding as in [9, §5.3].

3.3 Quasi-Trefftz space

In this section, we present the first extension of Generalized Plane Waves to a time-dependent problem, focusing on two main aspects: properties of the resulting function spaces, and explicit construction of the basis fucntions.

3.3.1 Definitions and notation

We recall the Leibniz product rule for multi-indices:

$$D^{\boldsymbol{i}}(f\tilde{f}) = \sum_{\boldsymbol{j} \in \mathbb{N}_0^{n+1},\, \boldsymbol{j} \leq \boldsymbol{i}} \binom{\boldsymbol{i}}{\boldsymbol{j}} D^{\boldsymbol{j}} f D^{\boldsymbol{i}-\boldsymbol{j}} \tilde{f}, \quad \text{where} \quad \binom{\boldsymbol{i}}{\boldsymbol{j}} := \frac{\boldsymbol{i}!}{\boldsymbol{j}!(\boldsymbol{i}-\boldsymbol{j})!} = \binom{i_{x_1}}{j_{x_1}} \cdots \binom{i_{x_n}}{j_{x_n}} \binom{i_t}{j_t}, \tag{3.14}$$

$\boldsymbol{i}! := i_{x_1}! \cdots i_{x_n}! i_t!$ and $\boldsymbol{j} \leq \boldsymbol{i}$ means that the inequality holds component-wise. The length of a multi-index is $|\boldsymbol{i}| = |\boldsymbol{i}_{\mathbf{x}}| + i_t := i_{x_1} + \cdots + i_{x_n} + i_t$. For any field $f \in C^m(K)$, denote the Taylor

polynomial of order $m+1$ (and polynomial degree at most m) centered at $(\mathbf{x}_K, t_K)$ by

$$T_K^{m+1}[f](\mathbf{x},t) := \sum_{|\boldsymbol{i}|\le m} \frac{1}{\boldsymbol{i}!}(\mathbf{x}-\mathbf{x}_K)^{\boldsymbol{i}_\mathbf{x}}(t-t_K)^{i_t} D^{\boldsymbol{i}} f(\mathbf{x}_K, t_K).$$

It follows that

$$D^{\boldsymbol{i}} T_K^{m+1}[f](\mathbf{x}_K,t_K) = \begin{cases} D^{\boldsymbol{i}} f(\mathbf{x}_K,t_K) & \text{if } |\boldsymbol{i}|\le m, \\ 0 & \text{if } |\boldsymbol{i}|>m. \end{cases} \tag{3.15}$$

Lagrange's form of the Taylor remainder [17, Cor. 3.19] is the following: if f has $m+1$ continuous derivatives in a neighbourhood of the segment S with extremes $(\mathbf{x}_K,t_K)$ and $(\mathbf{x},t)$, then

$$\exists (\mathbf{x}_*,t_*)\in S \quad \text{such that} \quad f(\mathbf{x},t) - T_K^{m+1}[f](\mathbf{x},t) = \sum_{|\boldsymbol{j}|=m+1} \frac{1}{\boldsymbol{j}!}(\mathbf{x}-\mathbf{x}_K)^{\boldsymbol{j}_\mathbf{x}}(t-t_K)^{j_t} D^{\boldsymbol{j}} f(\mathbf{x}_*,t_*). \tag{3.16}$$

For an elementwise-smooth, positive, spatial function $G:\Omega\to\mathbb{R}$ (representing c^{-2}) we denote the variable-coefficient second-order wave operator

$$(\Box_G f)(\mathbf{x},t) := \Delta f(\mathbf{x},t) - G(\mathbf{x})\partial_t^2 f(\mathbf{x},t).$$

We denote the partial derivatives of G evaluated at an element centre as

$$g_{\boldsymbol{i}_\mathbf{x}} := \frac{1}{\boldsymbol{i}_\mathbf{x}!} D^{(\boldsymbol{i}_\mathbf{x},0)} G(\mathbf{x}_K), \quad \text{so that} \quad G(\mathbf{x}) = \sum_{\boldsymbol{i}_\mathbf{x}\in\mathbb{N}_0^n} (\mathbf{x}-\mathbf{x}_K)^{\boldsymbol{i}_\mathbf{x}} g_{\boldsymbol{i}_\mathbf{x}}. \tag{3.17}$$

We underline the value of a particular partial derivative that will come into play in the definition of the quasi-Trefftz space: for $f\in C^{|\boldsymbol{i}|+2}(K)$,

$$(D^{\boldsymbol{i}}\Box_G f)(\mathbf{x}_K,t_K) = \sum_{k=1}^n D^{\boldsymbol{i}+2\mathbf{e}_k} f(\mathbf{x}_K,t_K) - \sum_{\boldsymbol{j}_\mathbf{x}\le\boldsymbol{i}_\mathbf{x}} \frac{\boldsymbol{i}_\mathbf{x}!}{\boldsymbol{j}_\mathbf{x}!} g_{\boldsymbol{i}_\mathbf{x}-\boldsymbol{j}_\mathbf{x}} D^{(\boldsymbol{j}_\mathbf{x},i_t+2)} f(\mathbf{x}_K,t_K). \tag{3.18}$$

This is obtained using Leibniz formula (3.14) and noting that only terms with $j_t = i_t$ contribute since G is independent of time.

We use standard notation for local C^m norms and seminorms, introduce wavespeed-weighted (dimensionally-homogeneous) seminorms C_c^m, and extend local spaces to global spaces in the piecewise-smooth case: for $m\in\mathbb{N}_0$

$$\begin{gathered} \|f\|_{C^0(K)} := \sup_{(\mathbf{x},t)\in K} |f(\mathbf{x},t)|, \quad |f|_{C^m(K)} := \max_{|\boldsymbol{i}|=m} \left\|D^{\boldsymbol{i}} f\right\|_{C^0(K)}, \quad |f|_{C_c^m(K)} := \max_{|\boldsymbol{i}|=m} \left\|c^{-i_t} D^{\boldsymbol{i}} f\right\|_{C^0(K)}, \\ C^m(\mathcal{T}_h) := \prod_{K\in\mathcal{T}_h} C^m(K), \qquad |f|_{C^m(\mathcal{T}_h)} := \max_{K\in\mathcal{T}_h} |f|_K|_{C^m(K)}, \quad |f|_{C_c^m(\mathcal{T}_h)} := \max_{K\in\mathcal{T}_h} |f|_K|_{C_c^m(K)}. \end{gathered} \tag{3.19}$$

3.3.2 Local quasi-Trefftz space and approximation properties

We define the "quasi-Trefftz" space for the second-order wave equation on a mesh element $K\in\mathcal{T}_h$ as

$$\boxed{\mathbb{QU}^p(K) := \left\{ f\in\mathbb{P}^p(K) \mid D^{\boldsymbol{i}}\Box_G f(\mathbf{x}_K,t_K) = 0,\ \forall \boldsymbol{i}\in\mathbb{N}_0^{n+1}, |\boldsymbol{i}|<p-1 \right\}, \qquad p\in\mathbb{N}.} \tag{3.20}$$

This is the space of degree-p space–time polynomials f, such that the Taylor polynomial of their image by the wave operator $\Box_G f$ vanishes at the element centre $(\mathbf{x}_K,t_K)$ up to order $p-2$. From (3.18), the space $\mathbb{QU}^p(K)$ is well-defined if $G\in C^{\max\{p-2,0\}}$ in a neighbourhood of $\mathbf{x}_K$. For $p=1$, we simply have $\mathbb{QU}^1(K) = \mathbb{P}^1(K)$.

Remark 3.3.1. *Compare the above definition to the 'standard' Trefftz space. We define the polynomial Trefftz space for the second order wave equation with constant wavespeed inside the mesh element K as*

$$\mathbb{U}^p(K) := \left\{u \in \mathbb{P}^p(K) : -\Delta u + c^{-2}\partial_t^2 u = 0\right\}.$$

For constant wavespeed the quasi-Trefftz space is equal to this space, see Remark 3.3.11.

The next proposition shows that smooth solutions of the wave equation are approximated in $\mathbb{QU}^p(K)$ with optimal convergence rate with respect to the element radius r_K (recall $r_K \leq \operatorname{diam} K$ from the definition (3.4)). By "optimal" we mean that the rate is equal to the rate offered by the full polynomial space $\mathbb{P}^p(K)$. We give two approximation estimates: one in classical C^m seminorms and one in their weighted version C_c^m defined in (3.19), with $r_{K,c}$ in place of r_K. We will use the latter bound in the convergence analysis of the DG scheme in §3.3.3.

Proposition 3.3.2. *Let $u \in C^{p+1}(K)$ be solution of $\Box_G u = 0$, with $G \in C^{\max\{p-2,0\}}(K)$.*

Then the Taylor polynomial $T_K^{p+1}[u] \in \mathbb{QU}^p(K)$.

Moreover, if K is star-shaped with respect to $(\mathbf{x}_K, t_K)$, with r_K and $r_{K,c}$ are as defined in (3.4) while $q \in \mathbb{N}_0$ satisfies $q \leq p$, then

$$\begin{aligned}
\inf_{P\in\mathbb{QU}^p(K)} |u-P|_{C^q(K)} &\leq \frac{(n+1)^{p+1-q}}{(p+1-q)!} r_K^{p+1-q}\, |u|_{C^{p+1}(K)}\,, \\
\inf_{P\in\mathbb{QU}^p(K)} |u-P|_{C_c^q(K)} &\leq \frac{(n+1)^{p+1-q}}{(p+1-q)!} r_{K,c}^{p+1-q}\, |u|_{C_c^{p+1}(K)}\,.
\end{aligned} \tag{3.21}$$

Proof. Since $T_K^{p+1}[u]$ is polynomial of degree p, in order to show that it belongs to $\mathbb{QU}^p(K)$ we only need to verify that $D^{\boldsymbol{i}}\Box_G T_K^{p+1}[u](\mathbf{x}_K, t_K) = 0$, for all $|\boldsymbol{i}| < p-1$. From the identity (3.18), this quantity is a linear combination of the partial derivatives of order at most equal to $|\boldsymbol{i}|+2 \leq p$ of the Taylor polynomial at $(\mathbf{x}_K, t_K)$, which according to (3.15) coincide with the corresponding partial derivatives of u:

$$D^{\boldsymbol{i}}\Box_G T_K^{p+1}[u](\mathbf{x}_K, t_K) = D^{\boldsymbol{i}}\Box_G u(\mathbf{x}_K, t_K).$$

Since $\Box_G u = 0$ in K, these partial derivatives vanish, hence $T_K^{p+1}[u] \in \mathbb{QU}^p(K)$.

We prove the inequality in (3.21) involving the weighted norms $C_c^m(K)$ using the norm definition (3.19), the identity $D^{\boldsymbol{i}} T_K^{p+1}[u] = T_K^{p+1-|\boldsymbol{i}|}[D^{\boldsymbol{i}}u]$ for $|\boldsymbol{i}| \leq p$ from [80, eq. (3.5)], and Taylor's theorem (3.16):

$$\begin{aligned}
\inf_{P\in\mathbb{QU}^p(K)} |u-P|_{C_c^q(K)} &\leq \left|u - T_K^{p+1}[u]\right|_{C_c^q(K)} \\
&= \max_{\boldsymbol{i}\in\mathbb{N}_0^{n+1},\, |\boldsymbol{i}|=q} \left\| c^{-i_t} D^{\boldsymbol{i}}(u - T_K^{p+1}[u]) \right\|_{C^0(K)} \\
&= \max_{\boldsymbol{i}\in\mathbb{N}_0^{n+1},\, |\boldsymbol{i}|=q} \left\| c^{-i_t} (D^{\boldsymbol{i}}u - T_K^{p+1-q}[D^{\boldsymbol{i}}u]) \right\|_{C^0(K)} \\
&\overset{(3.16)}{\leq} \max_{\boldsymbol{i}\in\mathbb{N}_0^{n+1},\, |\boldsymbol{i}|=q} \sum_{|\boldsymbol{j}|=p+1-q} \frac{1}{\boldsymbol{j}!} \left\| \left((\mathbf{x}, ct) - (\mathbf{x}_K, ct_K)\right)^{\boldsymbol{j}} c^{-i_t-j_t} D^{\boldsymbol{i}+\boldsymbol{j}} u(\mathbf{x}, t) \right\|_{C^0(K)} \\
&\leq \frac{(n+1)^{p+1-q}}{(p+1-q)!} r_{K,c}^{p+1-q}\, |u|_{C_c^{p+1}(K)}\,.
\end{aligned}$$

In the last step we used $\sum_{|\boldsymbol{j}|=p+1-q} \frac{1}{\boldsymbol{j}!} = \frac{(n+1)^{p+1-q}}{(p+1-q)!}$, [80, p. 198]. The first bound in (3.21) follows from the same chain of inequalities, after dropping all powers of c. □

Bound (3.21) gives approximation rates with respect to the mesh size (h-convergence) but is unsuitable for proving convergence for increasing polynomial degrees (p-convergence): while the coefficient in the bound is infinitesimal for $p \to \infty$, in general, the seminorm $|u|_{C^{p+1}(K)}$ is not bounded in the same limit.

Remark 3.3.3. *In general, unlike full polynomial spaces, quasi-Trefftz spaces with increasing p are not nested, i.e. $\mathbb{QU}^p(K) \not\subset \mathbb{QU}^{p+1}(K)$. To see this: consider $f(\mathbf{x},t) = x_1^2 + t^2 \in \mathbb{P}^2$, $G(\mathbf{x}) = 1 + x_1$ in a neighbourhood K of $(\mathbf{x}_K, t_K) = (\mathbf{0}, 0)$, then f satisfies $\Box_G f = 2 - 2(1 + x_1) = -2x_1$, so $\Box_G f(\mathbf{x}_K, t_K) = 0$ and $\partial_x \Box_G f(\mathbf{x}_K, t_K) = -2$, therefore $f \in \mathbb{QU}^2(K) \setminus \mathbb{QU}^3(K)$.*

On the other hand, $\mathbb{QU}^1(K) = \mathbb{P}^1(K) \subset \mathbb{QU}^2(K) = \{f \in \mathbb{P}^2(K) : \Delta f - G(\mathbf{x}_K, t_K)\partial_t^2 f = 0\}$.

Remark 3.3.4. *One could define a more general version of the space $\mathbb{QU}^p$ by imposing the vanishing of the derivatives up to an arbitrary order:*

$$\mathbb{QU}^{p,q}(K) := \left\{ f \in \mathbb{P}^p(K) \mid D^{\boldsymbol{i}} \Box_G f(\mathbf{x}_K, t_K) = 0,\ \forall |\boldsymbol{i}| < q - 1 \right\}, \qquad p, q \in \mathbb{N}.$$

For these spaces we have the inclusions $\mathbb{QU}^{p,q+1}(K) \subset \mathbb{QU}^{p,q}(K) \subset \mathbb{QU}^{p+1,q}(K)$ and $\mathbb{QU}^{p,p}(K) = \mathbb{QU}^p(K)$. However, let us now motivate why the choice $q = p$ is preferable.

- *For $q < p$ the space $\mathbb{QU}^{p,q}(K)$ is larger than $\mathbb{QU}^p(K)$, but since $\mathbb{QU}^{p,q}(K) \subset \mathbb{P}^p(K)$ it does not offer better h-convergence rates than those showed in Proposition 3.3.2 for $\mathbb{QU}^p(K)$. Moreover, it does not serve as a generalisation of a Trefftz space any more. Indeed, in the case of constant G the inclusion $\mathbb{U}^p(K) := \{f \in \mathbb{P}^p(K) : \Box_G f = 0 \text{ in } K\} \subset \mathbb{QU}^{p,q}(K)$ is always true, nevertheless the identity $\mathbb{U}^p(K) = \mathbb{QU}^{p,q}(K)$ holds if and only if $q \geq p$.*

- *For $q > p$ the space $\mathbb{QU}^{p,q}(K)$ is too small and loses his favorable approximation properties. Indeed, take $n = 1$, $i_x = p - 1$, $i_t = 0$. Then, for a solution f of $\Box_G f = 0$ we have that*

$$\begin{aligned}
\partial_x^{i_x} \partial_t^{i_t} \Box_G T_K^{p+1}[f](x_K, t_K) &\overset{(3.18)}{=} \left(\partial_x^{p+1} T_K^{p+1}[f] - \sum_{j_x=0}^{i_x} \frac{i_x!}{j_x!} g_{i_x - j_x} \partial_x^{j_x} \partial_t^2 T_K^{p+1}[f] \right)(x_K, t_K) \\
&= -\sum_{j_x=0}^{p-2} \frac{i_x!}{j_x!} g_{i_x - j_x} \partial_x^{j_x} \partial_t^2 T_K^{p+1}[f](x_K, t_K) \qquad \left(T_K^{p+1}[f] \in \mathbb{P}^p(K)\right) \\
&\overset{(3.15)}{=} -\sum_{j_x=0}^{p-2} \frac{i_x!}{j_x!} g_{i_x - j_x} \partial_x^{j_x} \partial_t^2 f(x_K, t_K) \\
&\overset{(3.18)}{=} \Big(\underbrace{\partial_x^{i_x} \partial_t^{i_t} \Box_G f}_{=0} - \partial_x^{p+1} f + g_0 \partial_x^{p-1} \partial_t^2 f \Big)(x_K, t_K) \\
&= \partial_x^{p-1} \big((g_0 - G) \partial_t^2 f \big)(x_K, t_K) \qquad (\partial_x^2 f = G \partial_t^2 f) \\
&\neq 0, \text{ in general.}
\end{aligned}$$

  Therefore, $T_K^{p+1}[f] \notin \mathbb{QU}^{p,p+1}(K)$, which contradicts the essential property used in the proof of Proposition 3.3.2 to prove the approximation properties of the space.

  Moreover, for $q > p$ the dimension of $\mathbb{QU}^{p,q}(K)$ depends on the function G (is equal to $\dim \mathbb{U}^p(K)$ for constant G and smaller in general), while we see in the following that $\dim \mathbb{QU}^p(K)$ is independent of G.

Hence, the choice $q = p$ yields the smallest subspace of $\mathbb{P}^p(K)$ in this class that offers the same h-convergence rates of $\mathbb{P}^p(K)$ itself, when approximating solutions of $\Box_G u = 0$.

3.3.3 Global quasi-Trefftz space and DG convergence bounds

We use the local spaces $\mathbb{QU}^p(K)$ to define a discrete space for the DG scheme of §3.2.3. Recall that $\mathbb{QU}^p(K)$ was constructed for the second-order scalar wave equation, while the DG scheme addresses the first-order system. A global quasi-Trefftz discrete space can be defined as

$$\boxed{\mathbb{QW}^p(\mathcal{T}_h) := \left\{ (w, \boldsymbol{\tau}) \in \mathbf{H}(\mathcal{T}_h) :\ w|_K = \partial_t u,\ \boldsymbol{\tau}|_K = -\nabla u,\ u \in \mathbb{QU}^{p+1}(K) \right\}, \quad p \in \mathbb{N}_0,} \tag{3.22}$$

where in each element K the local space $\mathbb{QU}^{p+1}(K)$ is defined with $G(\mathbf{x}) = c^{-2}(\mathbf{x})$. The elements of $\mathbb{QW}^p(\mathcal{T}_h)$ are vector polynomials of degree at most p.

Following again [81], for each element $K \in \mathcal{T}_h$ we introduce a notation for the space-like and the time-like parts of its boundary and two related coefficients:

$$\begin{aligned}
\partial^{\text{space}}K :=& \partial K \cap (\mathcal{F}_h^{\text{space}} \cup \mathcal{F}_h^0 \cup \mathcal{F}_h^T), \qquad \partial^{\text{time}}K := \partial K \cap (\mathcal{F}_h^{\text{time}} \cup \mathcal{F}_h^{\text{D}} \cup \mathcal{F}_h^{\text{N}} \cup \mathcal{F}_h^{\text{R}}),\\
\xi_K^{\text{time}} :=& \max\Big\{ \|2c\alpha\|_{L^\infty(\partial K \cap (\mathcal{F}_h^{\text{time}} \cup \mathcal{F}_h^{\text{D}}))} + \|c/\beta\|_{L^\infty(\partial K \cap (\mathcal{F}_h^{\text{time}} \cup \mathcal{F}_h^{\text{N}}))},\\
&\quad \|2\beta/c\|_{L^\infty(\partial K \cap (\mathcal{F}_h^{\text{time}} \cup \mathcal{F}_h^{\text{N}}))} + \|1/(c\alpha)\|_{L^\infty(\partial K \cap (\mathcal{F}_h^{\text{time}} \cup \mathcal{F}_h^{\text{D}}))},\\
&\quad \|(1-\delta)\vartheta\|_{L^\infty(\partial K \cap (\mathcal{F}_h^{\text{R}}))}, \qquad \|\delta/\vartheta\|_{L^\infty(\partial K \cap (\mathcal{F}_h^{\text{R}}))} \Big\},\\
\xi_K^{\text{space}} :=& \left\| n_K^t \big(2(1-\gamma)^{-1} + 1\big) \right\|_{L^\infty(\partial^{\text{space}}K)},
\end{aligned} \tag{3.23}$$

with γ as defined in (3.5). The dimensionless coefficients $\xi_K^\bullet$ measure the impact of the choices made in a concrete implementation of the DG scheme – in terms of the numerical flux parameters and the element shapes – on the convergence bounds of Theorem 3.3.5. If $c \in C^0(\Omega)$ and

$$\alpha = \beta^{-1} = c^{-1}, \qquad \delta = \vartheta^2/(1+\vartheta^2), \tag{3.24}$$

then $\xi_K^{\text{time}} = 3$ while ξ_K^{space} only depends on the maximal slope of the space-like faces of K and on c. If all faces of K are either aligned with or perpendicular to the time axis, i.e. $n_t \in \{0, 1\}$, then $\xi_K^{\text{space}} = 3$ as well.

We measure mesh regularity by fixing a dimensionless parameter $\eta > 0$ such that

$$r_{K,c}\Big(|\partial K^{\text{space}}|\, \|c\|_{C^0(K)}^{-1} + |\partial K^{\text{time}}|\Big) \le \eta |K| \qquad \forall K \in \mathcal{T}_h. \tag{3.25}$$

Remark 3.3.9 gives more details about η in the case of cuboidal elements.

The next theorem gives error bounds for the quasi-Trefftz space–time DG scheme (3.7).

Theorem 3.3.5. *Let $u \in C^1(\overline{Q_T}) \cap C^{p+2}(\mathcal{T}_h)$, for some $p \in \mathbb{N}_0$, be solution of the IBVP (3.2) with $G \in C^0(\Omega) \cap C^{\max\{p-1,0\}}(\mathcal{T}_h)$ and $(v, \boldsymbol{\sigma}) = (\partial_t u, -\nabla u)$ be the corresponding solution to the IBVP (3.1). Let $(v_{hp}, \boldsymbol{\sigma}_{hp})$ be the solution of the DG formulation (3.7) with the discrete space $\mathbf{V}_{hp}(\mathcal{T}_h) = \mathbb{QW}^p(\mathcal{T}_h)$. Assume that each mesh element K is star-shaped with respect to its centre point $(\mathbf{x}_K, t_K)$. Then,*

$$\begin{aligned}
&\frac{1}{2}\left\|c^{-1}(v - v_{hp})\right\|_{L^2(\mathcal{F}_h^T)} + \frac{1}{2}\left\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_{hp}\right\|_{L^2(\mathcal{F}_h^T)^n}\\
&\le |||(v, \boldsymbol{\sigma}) - (v_{hp}, \boldsymbol{\sigma}_{hp})|||_{\text{DG}}\\
&\le (1 + C_c)\frac{(n+1)^{p+1}}{p!}\Bigg(\sum_{K \in \mathcal{T}_h} |K| \Bigg[\Big(\eta \|c\|_{C^0(K)} \max\{\xi_K^{\text{space}}, \xi_K^{\text{time}}\} + \mu_{K-} r_{K,c}\Big) \frac{(n+1)^3 r_{K,c}}{(p+1)^2}\\
&\qquad\qquad + 4n\mu_{K+} \|c\|_{C^0(K)}^2 \Bigg] r_{K,c}^{2p} |u|_{C_c^{p+2}(K)}^2 \Bigg)^{1/2}.
\end{aligned} \tag{3.26}$$

The values of C_c, $\mu_{K\pm}$, $\xi_K^\bullet$ and η are defined in equations (3.12), (3.6), (3.23) and (3.25), respectively. If the numerical flux parameters are set according to (3.24) and all space-like faces are perpendicular to the time axis, then $\xi_K^{\text{space}} = \xi_K^{\text{time}} = 3$.

If moreover the volume penalty parameters are chosen as

$$\mu_1|_K = \mu_2|_K = r_{K,c} \|c\|_{C^0(K)}^{-1} \qquad \forall K \in \mathcal{T}_h, \tag{3.27}$$

then the right-hand side of the estimate (3.26) can be bounded by

$$(1 + C_c)\frac{|Q_T|^{1/2}(n+1)^{p+1}}{p!} \sup_{K \in \mathcal{T}_h} \left(\|c\|_{C^0(K)}^{1/2} \left[\frac{(\eta \xi_K + 1)(n+1)^3}{(p+1)^2} + 4n \right]^{1/2} r_{K,c}^{p+1/2} |u|_{C_c^{p+2}(K)} \right) \tag{3.28}$$

with $\xi_K = \max\{\xi_K^{\text{space}}, \xi_K^{\text{time}}\}$.

Proof. Recalling that the $|||\cdot|||_{\mathrm{DG}^+}$ norm in (3.8) differs from the analogous one in [81] only by the presence of the volume terms, the first step in the proof of [81, Thm. 2] allows to control the $|||\cdot|||_{\mathrm{DG}^+}$ norm of any (possibly discontinuous) $(w,\boldsymbol{\tau}) \in C^1(\mathcal{T}_h)$ by local norms:

$$
\begin{aligned}
|||(w,\boldsymbol{\tau})|||_{\mathrm{DG}^+}^2 \leq \sum_{K\in\mathcal{T}_h} \Bigg[& \xi_K^{\mathrm{space}} \Big(\left\| c^{-1} w \right\|_{L^2(\partial^{\mathrm{space}} K)}^2 + \|\boldsymbol{\tau}\|_{L^2(\partial^{\mathrm{space}} K)^n}^2 \Big) \\
& + \xi_K^{\mathrm{time}} \Big(\left\| c^{-1/2} w \right\|_{L^2(\partial^{\mathrm{time}} K)}^2 + \left\| c^{1/2} \boldsymbol{\tau} \right\|_{L^2(\partial^{\mathrm{time}} K)^n}^2 \Big) + \left\| \mu_1^{-1/2} c^{-1} w \right\|_{L^2(K)}^2 \\
& + \left\| \mu_2^{-1/2} \boldsymbol{\tau} \right\|_{L^2(K)^n}^2 + \left\| \mu_1^{1/2} (c\nabla\cdot\boldsymbol{\tau} + c^{-1}\partial_t w) \right\|_{L^2(K)}^2 + \left\| \mu_2^{1/2} (\partial_t \boldsymbol{\tau} + \nabla w) \right\|_{L^2(K)^n}^2 \Bigg] \\
\leq \sum_{K\in\mathcal{T}_h} \Bigg[& \Big(\xi_K^{\mathrm{space}} |\partial^{\mathrm{space}} K| + \xi_K^{\mathrm{time}} \|c\|_{C^0(K)} |\partial^{\mathrm{time}} K| + \mu_{K-}|K| \Big) \Big(\left\| c^{-1} w \right\|_{C^0(K)}^2 + n \|\boldsymbol{\tau}\|_{C^0(K)^n}^2 \Big) \\
& + |K| \mu_{K+} \Big(\left\| c^{-1} \partial_t w \right\|_{C^0(K)}^2 + n \|\nabla w\|_{C^0(K)^n}^2 + n \|\partial_t \boldsymbol{\tau}\|_{C^0(K)^n}^2 + \|c\nabla\cdot\boldsymbol{\tau}\|_{C^0(K)}^2 \Big) \Bigg].
\end{aligned}
$$

Now we use the quasi-optimality (3.13), the relation between the discrete spaces $\mathbb{QU}^{p+1}(K)$ and $\mathbb{QW}^p(\mathcal{T}_h)$, the assumption $(v,\boldsymbol{\sigma}) = (\partial_t u, -\nabla u)$, the local best-approximation bound (3.21), the definition of η (3.25):

$$
\begin{aligned}
& \frac{1}{(1+C_c)^2} |||(v,\boldsymbol{\sigma}) - (v_{hp},\boldsymbol{\sigma}_{hp})|||_{\mathrm{DG}}^2 \\
& \overset{(3.13)}{\leq} \inf_{(w_{hp},\boldsymbol{\tau}_{hp})\in\mathbb{QW}^p(\mathcal{T}_h)} |||(v,\boldsymbol{\sigma}) - (w_{hp},\boldsymbol{\tau}_{hp})|||_{\mathrm{DG}^+}^2 \\
& \overset{(3.22)}{=} \inf_{u_{hp}\in\prod_{K\in\mathcal{T}_h}\mathbb{QU}^{p+1}(K)} |||\big(\partial_t(u-u_{hp}), -\nabla(u-u_{hp})\big)|||_{\mathrm{DG}^+}^2 \\
& \leq \sum_{K\in\mathcal{T}_h} \inf_{u_{hp}\in\mathbb{QU}^{p+1}(K)} \Bigg[\Big(\xi_K^{\mathrm{space}} |\partial^{\mathrm{space}} K| + \xi_K^{\mathrm{time}} \|c\|_{C^0(K)} |\partial^{\mathrm{time}} K| + \mu_{K-}|K| \Big) (n+1) \, |u-u_{hp}|_{C_c^1(K)}^2 \\
& \qquad\qquad + 4n|K| \mu_{K+} \|c\|_{C^0(K)}^2 \, |u-u_{hp}|_{C_c^2(K)}^2 \Bigg] \\
& \overset{(3.21)}{\leq} \frac{(n+1)^{2p}}{(p!)^2} \sum_{K\in\mathcal{T}_h} \Bigg[\Big(\xi_K^{\mathrm{space}} |\partial^{\mathrm{space}} K| + \xi_K^{\mathrm{time}} \|c\|_{C^0(K)} |\partial^{\mathrm{time}} K| + \mu_{K-}|K| \Big) \frac{(n+1)^3 r_{K,c}^2}{(p+1)^2} \\
& \qquad\qquad + 4n|K| \mu_{K+} \|c\|_{C^0(K)}^2 \Bigg] r_{K,c}^{2p} \, |u|_{C_c^{p+2}(K)}^2 \\
& \overset{(3.25)}{\leq} \frac{(n+1)^{2p}}{(p!)^2} \sum_{K\in\mathcal{T}_h} |K| \Bigg[\Big(\eta \|c\|_{C^0(K)} \max\{\xi_K^{\mathrm{space}}, \xi_K^{\mathrm{time}}\} + \mu_{K-} r_{K,c} \Big) \frac{(n+1)^3 r_{K,c}}{(p+1)^2} \\
& \qquad\qquad + 4n \mu_{K+} \|c\|_{C^0(K)}^2 \Bigg] r_{K,c}^{2p} \, |u|_{C_c^{p+2}(K)}^2 .
\end{aligned}
$$

Under assumption (3.27) the last expression is bounded by

$$
\begin{aligned}
& \frac{(n+1)^{2p}}{(p!)^2} \sum_{K\in\mathcal{T}_h} |K| \, \|c\|_{C^0(K)} \left[(\eta \max\{\xi_K^{\mathrm{space}}, \xi_K^{\mathrm{time}}\} + 1) \frac{(n+1)^3}{(p+1)^2} + 4n \right] r_{K,c}^{2p+1} \, |u|_{C_c^{p+2}(K)}^2 \\
& \leq \frac{|Q_T|(n+1)^{2p}}{(p!)^2} \sup_{K\in\mathcal{T}_h} \Bigg(\|c\|_{C^0(K)} \left[(\eta \max\{\xi_K^{\mathrm{space}}, \xi_K^{\mathrm{time}}\} + 1) \frac{(n+1)^3}{(p+1)^2} + 4n \right] r_{K,c}^{2p+1} \, |u|_{C_c^{p+2}(K)}^2 \Bigg),
\end{aligned}
$$

where we used that $\sum_{K\in\mathcal{T}_h} |K| = |Q_T|$. Estimates (3.26) and (3.28) follow taking square roots. □

Theorem 3.3.5 immediately extends to quasi-Trefftz discrete spaces $\mathbf{V}_{hp}(\mathcal{T}_h)$ with different polynomial degrees in each mesh elements.

Remark 3.3.6 (Relevance of μ_2)**.** *When the discrete space is taken as $\mathbf{V}_{hp}(\mathcal{T}_h) = \mathbb{QW}^p(\mathcal{T}_h)$, the choice of the parameter μ_2 is immaterial because of the vanishing of the term in $\mathcal{A}(\cdot;\cdot)$ it multiplies. On the other hand, the assertion of Theorem 3.3.5 holds also for the space $\mathbf{V}_{hp}(\mathcal{T}_h) = \mathbb{QT}^p(\mathcal{T}_h)$ (defined below in §3.3.5), and in this case μ_2 needs to be chosen as in (3.27).*

Remark 3.3.7 (Error analysis in C^m and Sobolev spaces)**.** *In Proposition 3.3.2 and Theorem 3.3.5 we study approximation properties of quasi-Trefftz functions and convergence rates of the corresponding DG scheme using $C^m(K)$-type spaces. With piecewise-constant wavespeed c, Trefftz best-approximation estimates in Sobolev spaces require every element K to be star-shaped with respect to a space–time ellipsoid, [81, §6.1.2]. The parameters defining this ellipsoid have an important role in the approximation bounds. This ellipsoid is defined as the image of a ball under an affine transformation of K whose pull-back transforms the wave equation into its copy with unit speed (see the proof of [81, Corollary 3]). In the setting considered here instead c varies smoothly in K, thus the shape defined by mapping the sphere is in general not an ellipsoid but a more complicated set. Moreover, the non-affine transformation would not preserve polynomials, preventing a straightforward extension of the theory of [81, §6.1.2, 6.2.3] to the smooth wavespeed case. A precise approximation theory for quasi-Trefftz spaces and Sobolev norms is beyond the scope of this work.*

A study using classical Sobolev norms $H^m(K)$, possibly weighted with the wavespeed similarly to (3.19) above and [81, eq. (37)], will be the subject of future work. This would allow to treat less regular solutions.

Remark 3.3.8 (Rate optimality)**.** *Despite the use of C^m spaces in the analysis, the convergence rates for a given polynomial degree are optimal: compare the term $r_{K,c}^{p+1/2}$ in (3.28) and the term $h_K^{m_K+1/2}$ in [81, eq. (46)] (with $h_K \approx 2r_{K,c}$, $m_K = p$). On the other hand, the solution regularity required is stronger.*

A more sophisticated treatment of corner singularity in polygonal domains using weighted Sobolev spaces is done in [9].

Remark 3.3.9 (Value of η for a cuboid)**.** *Condition (3.25) is a condition on the "chunkiness" of the mesh elements. For instance, if all elements are translated copies of the Cartesian product $(0, L_{\mathbf{x}})^n \times (0, L_t)$ between a space segment/square/cube and a time interval, the centres $(\mathbf{x}_K, t_K)$ are their barycentres, and c is constant in K, then*

$$|K| = L_{\mathbf{x}}^n L_t, \qquad |\partial K^{\text{space}}| = 2L_{\mathbf{x}}^n, \qquad |\partial K^{\text{time}}| = 2nL_{\mathbf{x}}^{n-1}L_t, \qquad r_{K,c} = \frac{1}{2}\sqrt{nL_{\mathbf{x}}^2 + c^2L_t^2}$$

and η can be taken as $\eta_{cuboid} := 2n^{\frac{3}{2}}\left(\frac{L_{\mathbf{x}}}{cL_t} + \frac{cL_t}{L_{\mathbf{x}}}\right)$ since

$$2n^{\frac{3}{2}}\left(\frac{L_{\mathbf{x}}}{cL_t}+\frac{cL_t}{L_{\mathbf{x}}}\right) \geq \frac{n^{\frac{3}{2}}(L_{\mathbf{x}}+cL_t)^2}{cL_{\mathbf{x}}L_t} \geq \frac{\frac{1}{2}\sqrt{nL_{\mathbf{x}}^2+c^2L_t^2}(2L_{\mathbf{x}}^n + 2ncL_{\mathbf{x}}^{n-1}L_t)}{cL_{\mathbf{x}}^nL_t}$$
$$= \frac{r_{K,c}(|\partial K^{\text{space}}| + c|\partial K^{\text{time}}|)}{c|K|}.$$

Thus η is minimal for "isotropic" cuboids with $L_{\mathbf{x}} = cL_t$.

3.3.4 Basis functions

Here, we describe the construction of basis functions for $\mathbb{QU}^p(K)$; those for $\mathbb{QW}^p(\mathcal{T}_h)$ can then obtained simply by taking their appropriate partial derivatives.

To construct a basis of the quasi-Trefftz space $\mathbb{QU}^p(K)$ space, we first choose two polynomial bases in the space variable only:

$$\left\{\widehat{b}_J\right\}_{J=1,\dots,\binom{p+n}{n}} \text{ basis for } \mathbb{P}^p(\mathbb{R}^n), \quad \text{and} \quad \left\{\widetilde{b}_J\right\}_{J=1,\dots,\binom{p-1+n}{n}} \text{ basis for } \mathbb{P}^{p-1}(\mathbb{R}^n).$$

Their total cardinality is

$$N(n,p) := \binom{p+n}{n} + \binom{p-1+n}{n} = \frac{(p-1+n)!\,(2p+n)}{n!\,p!}.$$

We define the following $N(n,p)$ elements of $\mathbb{QU}^p(K)$:

$$\left\{ b_J \in \mathbb{QU}^p(K) \mid \begin{array}{l} b_J(\cdot,t_K) = \widehat{b}_J \ \text{ and } \ \partial_t b_J(\cdot,t_K) = 0 \text{ for } J \le \binom{p+n}{n}, \\ b_J(\cdot,t_K) = 0 \ \text{ and } \ \partial_t b_J(\cdot,t_K) = \widetilde{b}_{J-\binom{p+n}{n}} \text{ for } \binom{p+n}{n} < J \end{array} \right\}_{J=1,\dots,N(n,p).} \tag{3.29}$$

In the rest of this section we show that the elements b_J are well-defined, that they can be computed with a simple algorithm, and that they constitute a basis of $\mathbb{QU}^p(K)$.

Constructing the J-th basis function of (3.29) is equivalent to finding the set of coefficients $(a_{\boldsymbol{k}} = a_{\boldsymbol{k}_{\mathbf{x}},k_t})_{\boldsymbol{k}\in\mathbb{N}_0^{n+1},|\boldsymbol{k}|\le p}$ such that the polynomial

$$b_J(\mathbf{x},t) := \sum_{\boldsymbol{k}\in\mathbb{N}_0^{n+1},|\boldsymbol{k}|\le p} a_{\boldsymbol{k}} (\mathbf{x}-\mathbf{x}_K)^{\boldsymbol{k}_{\mathbf{x}}} (t-t_K)^{k_t}$$

fulfills the following system of equations

$$\begin{cases} D^{\boldsymbol{i}} \Box_G b_J(\mathbf{x}_K,t_K) = 0 & J = 1,\dots,N(n,p),\ \boldsymbol{i}\in\mathbb{N}_0^{n+1},\ |\boldsymbol{i}| < p-1, \\ b_J(\cdot,t_K) = \widehat{b}_J \text{ and } \partial_t b_J(\cdot,t_K) = 0 & J = 1,\dots,\binom{p+n}{n}, \\ b_J(\cdot,t_K) = 0 \text{ and } \partial_t b_J(\cdot,t_K) = \widetilde{b}_J & J = \binom{p+n}{n}+1,\dots,N(n,p). \end{cases} \tag{3.30}$$

The second and the third sets of equations assign the values of all $a_{\boldsymbol{k}_{\mathbf{x}},0}$ and $a_{\boldsymbol{k}_{\mathbf{x}},1}$. From (3.18) and since $D^{\boldsymbol{i}}((\mathbf{x}-\mathbf{x}_K)^{\boldsymbol{k}_{\mathbf{x}}}(t-t_K)^{k_t}) \neq 0$ at $(\mathbf{x}_K,t_K)$ if and only if $\boldsymbol{i} = \boldsymbol{k}$, the first equation in (3.30) corresponding to the $\big(\partial_{\mathbf{x}}^{\boldsymbol{i}_{\mathbf{x}}}\partial_t^{i_t}\Box_G b_J\big)(\mathbf{x}_K,t_K)$ term reads

$$\sum_{l=1}^{n} (\boldsymbol{i}_{\mathbf{x}} + 2\mathbf{e}_l)!\, i_t!\; a_{\boldsymbol{i}_{\mathbf{x}}+2\mathbf{e}_l,i_t} - \sum_{\boldsymbol{j}_{\mathbf{x}}\le\boldsymbol{i}_{\mathbf{x}}} \boldsymbol{i}_{\mathbf{x}}!(i_t+2)! g_{\boldsymbol{i}_{\mathbf{x}}-\boldsymbol{j}_{\mathbf{x}}} a_{\boldsymbol{j}_{\mathbf{x}},i_t+2} = 0. \tag{3.31}$$

This equation is used to compute the element with $\boldsymbol{j} = \boldsymbol{i}$ in the sum, since $g_{\mathbf{0}} = G(\mathbf{x}_K) > 0$ its coefficient is non-zero:

$$a_{\boldsymbol{i}_{\mathbf{x}},i_t+2} = \sum_{l=1}^{n} \frac{(i_{x_l}+2)(i_{x_l}+1)}{(i_t+2)(i_t+1)g_{\mathbf{0}}} a_{\boldsymbol{i}_{\mathbf{x}}+2\mathbf{e}_l,i_t} - \sum_{\boldsymbol{j}_{\mathbf{x}}<\boldsymbol{i}_{\mathbf{x}}} \frac{g_{\boldsymbol{i}_{\mathbf{x}}-\boldsymbol{j}_{\mathbf{x}}}}{g_{\mathbf{0}}} a_{\boldsymbol{j}_{\mathbf{x}},i_t+2}, \tag{3.32}$$

where the strict inequality $\boldsymbol{j} < \boldsymbol{i}$ in the summation means that $\boldsymbol{j} \le \boldsymbol{i}$ and $\boldsymbol{j} \neq \boldsymbol{i}$. We compute these values iteratively. We need to make sure that at every step of the iterations we use values already computed.

We note that the parameter function $G = c^{-2}$ enters the computation of b_J only through its Taylor coefficients at $(\mathbf{x}_K,t_K)$, i.e. the $g_{\boldsymbol{i}}$ defined in (3.17).

One could also write the equations (3.31) as a linear system, where the right-hand side vector is given by the known values of $a_{\boldsymbol{k}_{\mathbf{x}},0}$, $a_{\boldsymbol{k}_{\mathbf{x}},1}$, and solve it. However the recursive implementation appears simpler.

The next two sections describe in detail the iterative algorithm to compute the coefficients in the cases $n = 1$ and $n > 1$, respectively. Possible choices of the space-only bases $\left\{\widehat{b}_J\right\}$, $\left\{\widetilde{b}_J\right\}$ are described in the numerical results section 3.4.

Proposition 3.3.10. *The polynomials $\{b_J\}_{J=1,\dots,N(n,p)}$ defined by (3.30) (and computable with the algorithms of §3.3.4–3.3.4) constitute a basis for the space $\mathbb{QU}^p(K)$.*

Proof. The algorithms described in §3.3.4–3.3.4 show that (3.30) uniquely defines the $N(n,p)$ polynomials b_J. The first set of conditions in (3.30) ensures that $b_J \in \mathbb{QU}^p(K)$. The traces on

$\{t = t_K\}$ of these polynomials ensure that they are linearly independent: if $\sum_{J=1}^{N(n,p)} c_J b_J = 0$ then, by (3.30) $\sum_{J=1}^{\binom{p+n}{n}} c_J \widehat{b}_J = 0$ and $\sum_{J=\binom{p+n}{n}+1}^{N(n,p)} c_J \widetilde{b}_J = 0$, so $c_J = 0$ for all J because $\left\{\widehat{b}_J\right\}$ and $\left\{\widetilde{b}_J\right\}$ are assumed to be linearly independent.

Relation (3.32) holds not only for the elements of the basis, but for the monomial expansion of any element of $\mathbb{QU}^p(K)$. Then Algorithms 1 and 2 (which simply apply (3.32) following a precise ordering of the multi-indices $\boldsymbol{i}$) show that the Taylor coefficients in $(\mathbf{x}_K, t_K)$ of any $f \in \mathbb{QU}^p(K)$ are uniquely determined by the coefficients $a_{\boldsymbol{i}_\mathbf{x},0}$ and $a_{\boldsymbol{i}_\mathbf{x},1}$, and hence by $f(\cdot, t_K)$ and $\partial_t f(\cdot, t_K)$. Since $f(\cdot, t_K) \in \mathbb{P}^p(\mathbb{R}^n)$ and $\partial_t f(\cdot, t_K) \in \mathbb{P}^{p-1}(\mathbb{R}^n)$, f is linear combination of the $\{b_J\}$. Thus this set spans $\mathbb{QU}^p(K)$ and we conclude the proof. □

From the proposition it follows that the conditions in the definition (3.20) of $\mathbb{QU}^p(K)$ are linearly independent:

$$\dim\left(\mathbb{P}^p(K)\right) - \#\{\boldsymbol{i} \in \mathbb{N}_0^{n+1} \mid |\boldsymbol{i}| \leq p-2\} = \binom{p+n+1}{n+1} - \binom{p+n-1}{n+1} = \dim\left(\mathbb{QU}^p(K)\right).$$

Remark 3.3.11. *Proposition 3.3.10 implies that*

$$\dim\left(\mathbb{QU}^p(K)\right) = N(n,p) = \begin{cases} 2p+1 & n=1 \\ (p+1)^2 & n=2 \\ \frac{1}{6}(p+2)(p+1)(2p+3) & n=3 \end{cases} = \mathcal{O}_{p\to\infty}(p^n).$$

For large polynomial degrees p, the dimension of the quasi-Trefftz space is much smaller than the dimension of the full space–time polynomial space of the same degree: $\dim(\mathbb{P}^p(K)) = \binom{p+n+1}{n+1} = \mathcal{O}_{p\to\infty}(p^{n+1})$ *(recall that $K \subset \mathbb{R}^{n+1}$).*

Proposition 3.3.2 shows that both spaces $\mathbb{QU}^p(K)$ and $\mathbb{P}^p(K)$ have comparable h-approximation properties when the function to be approximated is solution of the (variable-coefficient) wave equation. This is the main advantage offered by Trefftz and quasi-Trefftz schemes: same approximation power for much fewer degrees of freedom.

The dimension of $\mathbb{QU}^p(K)$ is equal to the dimension of the Trefftz space of the same degree for the constant-coefficient wave equation [81, §6.2.1], for the Laplace equation (i.e. the space of harmonic polynomials in $\mathbb{R}^{n+1}$ of degree $\leq p$) and for the Helmholtz equation [50, §3] (the space of circular/spherical and plane waves in $\mathbb{R}^{n+1}$ with the same approximation order).

Similarly, $\dim(\mathbb{QW}^p(K)) = \dim(\mathbb{QU}^{p+1}(K)) - 1 = \mathcal{O}_{p\to\infty}(p^n)$.

The construction of the basis functions: the case $n = 1$

We first describe the one-dimensional case for the sake of clarity.

For each basis function b_J we need to compute the coefficients a_{k_x,k_t}, $k_x, k_t \in \mathbb{N}_0$, $k_x + k_t \leq p$; they are represented by the dots constituting a triangular shape in the plan of indices $(k_x, k_t) \in \mathbb{N}_0^2$, as represented on Figure 3.1. We recall that the coefficients $\{a_{i_x,0}, 0 \leq i_x \leq p\}$, and $\{a_{i_x,1}, 0 \leq i_x \leq p-1\}$, represented by the shaded area in the figure, are known from the choice of the "Cauchy data" bases $\left\{\widehat{b}_J\right\}$, $\left\{\widetilde{b}_J\right\}$. Formula (3.32) allows to compute a_{i_x,i_t+2} from similar coefficients a_{k_x,i_t+2} with $k_x < i_x$ and from a_{i_x+2,i_t}. This suggests to proceed "diagonally": i.e. to compute the values a_{k_x,k_t} for $k_x + k_t = \ell$ increasingly from $\ell = 2$ to $\ell = p$. On each of these diagonals (in gray in the figure) we compute the values of $a_{k_x,\ell-k_x}$ for decreasing k_x. This means that we perform a double loop: in terms of the graphical representation, the external loop moves away from the origin ($\nearrow$) and the inner loop moves from the k_x axis to the k_t axis ($\nwarrow$). This procedure is described in Algorithm 1.

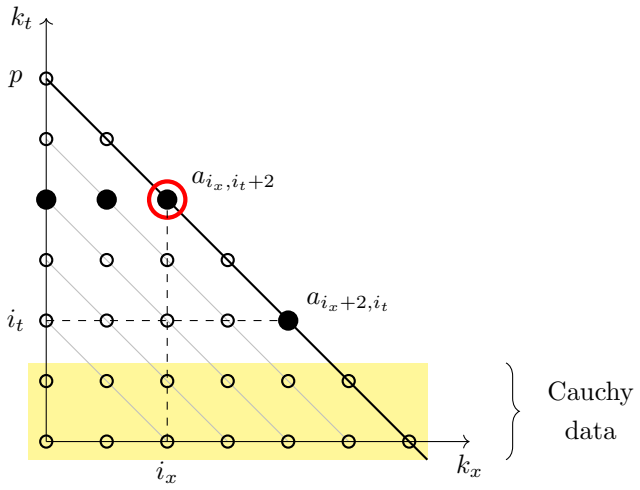

Figure 3.1: Graphical representation of the algorithm used to compute a quasi-Trefftz basis function in the case $n = 1$ (and $p = 6$), see §3.3.4. The function b_J is defined by the coefficients a_{k_x,k_t} corresponding to the small circles ∘. The coefficients corresponding to the dots in the shaded area ($k_t \in \{0, 1\}$) are given by the "Cauchy data", the second and third set of equations in (3.30). The first equation in (3.30) for $(i_x, i_t) = (2, 2)$ relates the four nodes depicted with a black dot • and is used to compute a_{i_x,i_t+2} (explicitly with (3.32)), corresponding to the node surrounded by a red circle ◯. All these coefficients (in the non-shaded region) are computed with formula (3.33) in a double loop: first across diagonals ↗, and then along diagonals ↖.

ALGORITHM
Data: $(g_m)_{m\in\mathbb{N}_0}$, x_K, t_K, p.
Choose polynomial bases $\left\{\widehat{b}_J\right\}$, $\left\{\widetilde{b}_J\right\}$, fixing coefficients $a_{k_x,0}$, $a_{k_x,1}$.
For each $J = 1, \dots, N(n, p)$ (i.e. for each basis function), we construct b_J as follows:
for $\ell = 2$ *to* p *(loop across diagonals ↗)* **do**
  for $i_t = 0$ *to* $\ell - 2$ *(loop along diagonals ↖)* **do**
    set $i_x = \ell - i_t - 2$ and compute

$$a_{i_x,i_t+2} = \frac{(i_x+2)(i_x+1)}{(i_t+2)(i_t+1)g_0} a_{i_x+2,i_t} - \sum_{j_x=0}^{i_x-1} \frac{g_{i_x-j_x}}{g_0} a_{j_x,i_t+2} \tag{3.33}$$

  end
end

$$b_J(x,t) = \sum_{0<k_x+k_t\le p} a_{k_x,k_t}(x-x_K)^{k_x}(t-t_K)^{k_t}$$

Algorithm 1: The algorithm for the construction of b_J in the case $n = 1$, §3.3.4.

The construction of the basis functions: the case $n > 1$

Algorithm 2 extends Algorithm 1 to the general case $n > 1$. The main novelty is that for each value of $\ell = |\boldsymbol{i}_\mathbf{x}| + i_t + 2$ and of i_t there are several coefficients $a_{\boldsymbol{i}_\mathbf{x},i_t}$ to be computed, exactly one for each $\boldsymbol{i}_\mathbf{x} \in \mathbb{N}_0^n$ with $|\boldsymbol{i}_\mathbf{x}| = \ell - 2 - i_t$, thus a further inner loop over $\boldsymbol{i}_\mathbf{x}$ is needed. Each coefficient of the innermost loop can be computed independently of the others.

Figure 3.2 depicts the dependence between these coefficients, represented as integer-coordinate points in the $(\boldsymbol{i}_\mathbf{x}, i_t)$ space for $n = 2$. The general coefficient, indicated by the red diamond, is computed with (3.34) as linear combination of the coefficients corresponding to the black dots. The

structure of the algorithm ensures that, when a coefficient $a_{\boldsymbol{i}_{\mathbf{x}},i_t}$ is computed, all the coefficients needed for the right-hand side of (3.34) have already been computed.

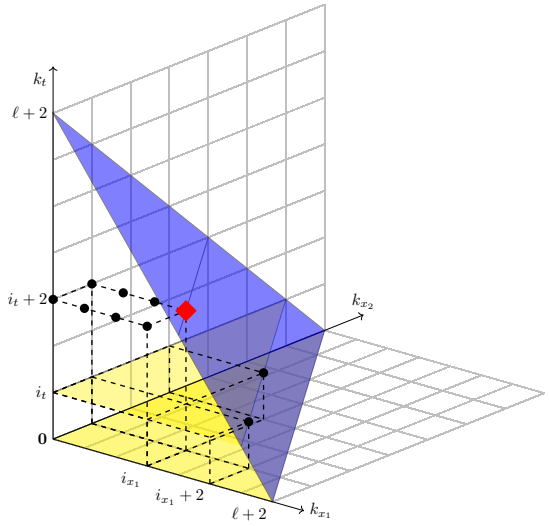

Figure 3.2: Representation of Algorithm 2 to compute the coefficient $a_{\boldsymbol{i}_{\mathbf{x}},i_t}$ of a quasi-Trefftz basis function b_J for $n=2$, $p\geq 5$, $\ell=5$ and $(\boldsymbol{i}_{\mathbf{x}},i_t)=(3,1,1)$. The coefficient $a_{3,1,3}$, represented by the large red diamond ♦, is computed with formula (3.34) from the nine coefficients indicated by the black dots •. The two yellow triangles in the planes $k_t=0$ and $k_t=1$ indicate the coefficients whose values are given by the "initial conditions" $\widehat{b}_J$ and $\widetilde{b}_J$ in the second and third equation of (3.30). To ensure that the right-end side of (3.34) is well-defined for every $(\boldsymbol{i}_{\mathbf{x}},i_t)$, Algorithm 2 computes the coefficients first looping through triangles parallel to the one depicted in blue (which corresponds to stage $\ell=5$ of the loop), then through horizontal planes, and finally along the horizontal segments determined by the intersection between the two planes.

ALGORITHM

Data: $(g_m)_{m\in\mathbb{N}_0}$, $\mathbf{x}_K$, t_K, p.

Choose polynomial bases $\left\{\widehat{b}_J\right\}$, $\left\{\widetilde{b}_J\right\}$, fixing coefficients $a_{\boldsymbol{k}_{\mathbf{x}},0}$, $a_{\boldsymbol{k}_{\mathbf{x}},1}$.

For each $J=1,\dots,N(n,p)$ (i.e. for each basis function), we construct b_J as follows:

for $\ell=2$ *to* p *(loop across* $\{|\boldsymbol{i}_{\mathbf{x}}|+i_t=\ell-2\}$ *hyperplanes,* $\nearrow$*)* **do**

  for $i_t=0$ *to* $\ell-2$ *(loop across constant-time hyperplanes* $\uparrow$*)* **do**

    for $\boldsymbol{i}_{\mathbf{x}}$ *with* $|\boldsymbol{i}_{\mathbf{x}}|=\ell-i_t-2$ **do**

$$a_{\boldsymbol{i}_{\mathbf{x}},i_t+2}=\sum_{l=1}^{n}\frac{(i_{x_l}+2)(i_{x_l}+1)}{(i_t+2)(i_t+1)g_{\mathbf{0}}}a_{\boldsymbol{i}_{\mathbf{x}}+2\mathbf{e}_l,i_t}-\sum_{\boldsymbol{j}_{\mathbf{x}}<\boldsymbol{i}_{\mathbf{x}}}\frac{g_{\boldsymbol{i}_{\mathbf{x}}-\boldsymbol{j}_{\mathbf{x}}}}{g_{\mathbf{0}}}a_{\boldsymbol{j}_{\mathbf{x}},i_t+2}\qquad(3.34)$$

    end

  end

end

$$b_J(\mathbf{x},t)=\sum_{|\boldsymbol{k}_{\mathbf{x}}|+k_t\leq p}a_{\boldsymbol{k}_{\mathbf{x}},k_t}(\mathbf{x}-\mathbf{x}_K)^{\boldsymbol{k}_{\mathbf{x}}}(t-t_K)^{k_t}$$

Algorithm 2: The algorithm for the construction of b_J in the general case.

3.3.5 Quasi-Trefftz discrete spaces for the first-order problem

The quasi-Trefftz space $\mathbb{QW}^p(\mathcal{T}_h)$ was defined in (3.22) from derivatives of solutions to the second-order wave equation. Thus it offers high-order approximation properties only for solutions $(v, \boldsymbol{\sigma})$ to IBVPs (3.1) related to a solution u of the second-order IBVPs (3.2) by the relation $(v, \boldsymbol{\sigma}) = (\partial_t u, -\nabla u)$. We briefly describe a larger discrete space suitable to approximate the general first-order IBVP (3.1).

For $p \in \mathbb{N}_0$ and any mesh element $K \in \mathcal{T}_h$, we set

$$\boxed{\begin{aligned}\mathbb{QT}^p(K) &:= \left\{(w,\boldsymbol{\tau}) \in \mathbb{P}^p(K)^{n+1} \,\middle|\, \begin{array}{ll} D^{\boldsymbol{i}}(\nabla w + \partial_t \boldsymbol{\tau})(\mathbf{x}_K, t_K) = \mathbf{0} & \\ D^{\boldsymbol{i}}(\nabla\cdot\boldsymbol{\tau} + G\partial_t w)(\mathbf{x}_K, t_K) = 0 & \end{array} \quad \forall \boldsymbol{i} \in \mathbb{N}_0^{n+1}, |\boldsymbol{i}| < p\right\},\\ \mathbb{QT}^p(\mathcal{T}_h) &:= \prod_{K\in\mathcal{T}_h} \mathbb{QT}^p(K).\end{aligned}} \tag{3.35}$$

It is easy to check that $\mathbb{QW}^p(\mathcal{T}_h) \subset \mathbb{QT}^p(\mathcal{T}_h)$. This implies that the convergence results of Theorem 3.3.5 hold also with $\mathbb{QT}^p(\mathcal{T}_h)$ in place of $\mathbb{QW}^p(\mathcal{T}_h)$.

Given any basis $\left\{\widetilde{b}_J(\mathbf{x})\right\}_{J=1,\dots,\binom{p+n}{n}}$ of $\mathbb{P}^p(\mathbb{R}^n)$, we can define a basis for $\mathbb{QT}^p(K)$ as

$$\left\{\mathbf{b}_{J,l}(\mathbf{x},t) \in \mathbb{QT}^p(K) \;\text{ such that }\; \begin{array}{l}\mathbf{b}_{J,0}(\mathbf{x},t_K) = \left(\widetilde{b}_J(\mathbf{x}), \mathbf{0}\right),\\ \mathbf{b}_{J,l}(\mathbf{x},t_K) = \left(0, \widetilde{b}_J(\mathbf{x})\mathbf{e}_l\right),\ l = 1,\dots,n\end{array}\right\}_{J=1,\dots,\binom{p+n}{n};\, l=0,\dots,n}.$$

To compute explicitly a basis element $\mathbf{b}_{J,l}$ from $\widetilde{b}_J$, we expand it in monomials:

$$\mathbf{b}_{J,l}(\mathbf{x},t) = \sum_{\boldsymbol{k}\in\mathbb{N}_0^{n+1},\,|\boldsymbol{k}|\le p} \mathbf{a}_{\boldsymbol{k}}(\mathbf{x}-\mathbf{x}_K)^{\boldsymbol{k}_{\mathbf{x}}}(t-t_K)^{k_t}, \quad l = 0,\dots,n, \text{ for } \left\{\mathbf{a}_{\boldsymbol{k}} = \mathbf{a}_{\boldsymbol{k}}(J,l)\right\}_{|\boldsymbol{k}|\le p} \in \mathbb{R}^{n+1}.$$

We index the components of the field $\mathbf{b}(\mathbf{x},t) = \mathbf{b}_{J,l}(\mathbf{x},t)$ from $b^0(\mathbf{x},t)$ to $b^n(\mathbf{x},t)$, and write similarly $\mathbf{a}_{\boldsymbol{k}} = (a^0_{\boldsymbol{k}}, \dots, a^n_{\boldsymbol{k}})$. Space–time multi-indices are split as previously in space and time parts $\boldsymbol{k} = (\boldsymbol{k}_{\mathbf{x}}, k_t)$. Then the conditions corresponding respectively to $D^{\boldsymbol{i}}\left(\partial_{x_\lambda} b^0 + \partial_t b^\lambda\right)(\mathbf{x}_K, t_K)$ for λ from 1 to n, namely the components of the vector-valued constraint, and $D^{\boldsymbol{i}}\Big(\sum_{\lambda=1}^n \partial_{x_\lambda} b^\lambda + c^{-2}\partial_t b^0\Big)(\mathbf{x}_K, t_K)$, namely the scalar-valued constraint, in the definition (3.35) of $\mathbb{QT}^p(K)$ can be written in terms of coefficients as

$$\begin{aligned} 0 =& (\boldsymbol{i}_{\mathbf{x}} + \mathbf{e}_\lambda)! i_t! a^0_{(\boldsymbol{i}_{\mathbf{x}}+\mathbf{e}_\lambda, i_t)} + \boldsymbol{i}_{\mathbf{x}}!(i_t+1)! a^\lambda_{(\boldsymbol{i}_{\mathbf{x}}, i_t+1)}, \qquad \lambda = 1,\dots,n,\\ 0 =& \sum_{\lambda=1}^n (\boldsymbol{i}_{\mathbf{x}} + \mathbf{e}_\lambda)! i_t! a^\lambda_{(\boldsymbol{i}_{\mathbf{x}}+\mathbf{e}_\lambda, i_t)} + \sum_{\boldsymbol{j}_{\mathbf{x}} \le \boldsymbol{i}_{\mathbf{x}}} \boldsymbol{i}_{\mathbf{x}}!(i_t+1)! g_{\boldsymbol{i}_{\mathbf{x}}-\boldsymbol{j}_{\mathbf{x}}} a^0_{(\boldsymbol{j}_{\mathbf{x}}, i_t+1)}. \end{aligned}$$

Then $\mathbf{b} = \mathbf{b}_{J,l} \in \mathbb{QT}^p(K)$ if and only if its coefficients satisfy the recurrence relations

$$\begin{aligned} a^0_{(\boldsymbol{i}_{\mathbf{x}}, i_t+1)} &= -\sum_{\lambda=1}^n \frac{i_{x_\lambda}+1}{g_{\mathbf{0}}(i_t+1)} a^\lambda_{(\boldsymbol{i}_{\mathbf{x}}+\mathbf{e}_\lambda, i_t)} - \sum_{\boldsymbol{j}_{\mathbf{x}} < \boldsymbol{i}_{\mathbf{x}}} \frac{g_{\boldsymbol{i}_{\mathbf{x}}-\boldsymbol{j}_{\mathbf{x}}}}{g_{\mathbf{0}}} a^0_{(\boldsymbol{j}_{\mathbf{x}}, i_t+1)},\\ a^\lambda_{(\boldsymbol{i}_{\mathbf{x}}, i_t+1)} &= -\frac{i_{x_\lambda}+1}{i_t+1} a^0_{(\boldsymbol{i}_{\mathbf{x}}+\mathbf{e}_\lambda, i_t)}, \qquad \lambda = 1,\dots,n. \end{aligned}$$

The coefficients $a^\lambda_{(\boldsymbol{k}_{\mathbf{x}},0)}$, $\lambda = 0,\dots,n$, $|\boldsymbol{k}_{\mathbf{x}}| \le p$, are known from the comparison with the space-only basis element $\widetilde{b}_J$. All the other coefficients $a^\lambda_{\boldsymbol{k}}$ can be computed with a double loop: first over $|\boldsymbol{k}| = 1,\dots,p$, and then over $k_t = 1,\dots,|\boldsymbol{k}|$, similarly to Algorithms 1–2. The procedure is described in Algorithm 3.

It is possible to verify that the $\mathbf{b}_{J,l}$ constitute a basis of $\mathbb{QT}^p(K)$ following the lines of the proof of Proposition 3.3.10. It follows that

$$\dim\big(\mathbb{QT}^p(K)\big) = (n+1)\binom{p+n}{n} = \frac{(n+1)(p+1)}{2p+2+n}\big(\dim\big(\mathbb{QW}^p(K)\big)+1\big) = \mathcal{O}_{p\to\infty}(p^n).$$

ALGORITHM
Data: $(g_m)_{m\in\mathbb{N}_0}$, $\mathbf{x}_K$, t_K, p.
Choose polynomial basis $\{\widetilde{b}_J\}$, fixing coefficients $a^\lambda_{\boldsymbol{k}_\mathbf{x},0}$.
For each $J = 1,\dots,N(n,p)$ and $l = 0,\dots,n$, we construct $\mathbf{b}_{J,l}$ as follows:
for $\ell = 1$ *to* p *(loop across* $\{|\boldsymbol{i}_\mathbf{x}| + i_t = \ell - 1\}$ *hyperplanes,* $\nearrow$*)* **do**
  for $i_t = 0$ *to* $\ell - 1$ *(loop across constant-time hyperplanes* $\uparrow$*)* **do**
    for $\boldsymbol{i}_\mathbf{x}$ *with* $|\boldsymbol{i}_\mathbf{x}| = \ell - i_t - 1$ **do**

$$a^0_{(\boldsymbol{i}_\mathbf{x},i_t+1)} = -\sum_{\lambda=1}^{n}\frac{i_{x_\lambda}+1}{g_\mathbf{0}(i_t+1)}a^\lambda_{(\boldsymbol{i}_\mathbf{x}+\mathbf{e}_\lambda,i_t)} - \sum_{\boldsymbol{j}_\mathbf{x}<\boldsymbol{i}_\mathbf{x}}\frac{g_{\boldsymbol{i}_\mathbf{x}-\boldsymbol{j}_\mathbf{x}}}{g_\mathbf{0}}a^0_{(\boldsymbol{j}_\mathbf{x},i_t+1)},$$

$$a^\lambda_{(\boldsymbol{i}_\mathbf{x},i_t+1)} = -\frac{i_{x_\lambda}+1}{i_t+1}a^0_{(\boldsymbol{i}_\mathbf{x}+\mathbf{e}_\lambda,i_t)}, \qquad \lambda = 1,\dots,n.$$

    end
  end
end

$$\mathbf{b}_{J,l}(\mathbf{x},t) = \sum_{\boldsymbol{k}\in\mathbb{N}_0^{n+1},\,|\boldsymbol{k}|\le p}\mathbf{a}_{\boldsymbol{k}}(\mathbf{x}-\mathbf{x}_K)^{\boldsymbol{k}_\mathbf{x}}(t-t_K)^{k_t}$$

Algorithm 3: The algorithm for the construction of $\mathbf{b}_{J,l}$ in the general case.

3.4 Numerical tests

We present some numerical test results in one and two space dimensions. Except for the last example, we consider the initial boundary value problem (3.2) with Dirichlet boundary conditions only, i.e. $\Gamma_\mathrm{N} = \Gamma_\mathrm{R} = \emptyset$. We test our method with the following wavespeed parameters $G = c^{-2}$, exact solutions u, and space–time domain Q_T:

$$n=1,\quad G(x) = x+1, \qquad u(x,t) = \mathrm{Ai}(-x-1)\cos(t), \qquad Q_T = (0,5)^2, \tag{3.36a}$$

$$n=2,\quad G(x_1,x_2) = x_1+x_2+1, \qquad u(x_1,x_2,t) = \mathrm{Ai}(-x_1-x_2-1)\cos(\sqrt{2}t), \qquad Q_T = (0,1)^3, \tag{3.36b}$$

$$n=2,\quad G(x_1,x_2) = (x_1+x_2+1)^{-2}, \quad u(x_1,x_2,t) = (x_1+x_2+1)^a \mathrm{e}^{-\sqrt{2}\sqrt{a(a-1)}t}, \quad Q_T = (0,1)^3. \tag{3.36c}$$

Here Ai is the Airy function, which fulfills $\mathrm{Ai}''(x) = x\mathrm{Ai}(x)$, and we choose $a = 2.5$ in (3.36c). The corresponding wavespeeds $c(\mathbf{x})$ range respectively in the intervals $[\sqrt{1/6},1] \approx [0.41,1]$, $[\sqrt{1/3},1] \approx [0.58,1]$ and $[1,3]$ for the three problems (3.36). Then, the solution of the first-order wave equation is given by $(v,\boldsymbol{\sigma}) = (\partial_t u, -\nabla u)$.

To construct the quasi-Trefftz basis we pre-compute coefficients of G's Taylor expansion (3.17) at the centre of each mesh element. We choose a monomial basis (scaled according to the element size) for $\left\{\widehat{b}_J\right\}$, and $\left\{\widetilde{b}_J\right\}$, as input of Algorithms 1–2. This is motivated by experiments described in [90, §6.3], where monomials, chosen as initial basis for the construction of the standard Trefftz space, outperformed Legendre and Chebyshev basis. Remarkably, if space–time mesh elements share the same centre in space, namely $\mathbf{x}_K$, then the coefficients of the quasi-Trefftz basis functions are identical on both elements, therefore they can be computed once and used on both elements.

The section continues as follows. In §3.4.1 we compare different choices for the penalisation coefficients. The quasi-Trefftz discretisation is compared against a full polynomial space and a standard Trefftz space in §3.4.2. In §3.4.3 we use a special type of space–time meshes allowing for semi-explicit time-stepping: tent-pitched meshes. Finally, we show snapshots of the numerical approximation of a Gaussian pulse traveling through a heterogeneous medium in §3.4.4.

3.4.1 Volume penalisation and numerical flux parameters

In this experiment we consider different combinations of the numerical flux parameters α, β and the volume penalisation coefficient μ_1. We recall that for $\mathbb{QW}^p(\mathcal{T}_h)$ the choice of the parameter μ_2 is irrelevant (see Remark 3.3.6). Furthermore, we use Dirichlet boundary conditions, thus δ does not appear. We compare the choices for the parameters given in (3.27) and (3.24) against setting them to zero. We fix $p = 4$ and a sequence of Cartesian meshes in 1+1 dimensions with square space–time mesh elements $K = (\mathbf{x}_K - \frac{h}{2}, \mathbf{x}_K + \frac{h}{2}) \times (t_K - \frac{h}{2}, t_K + \frac{h}{2})$, and compare against the exact solution (3.36a) (which can be seen in Figure 3.5).

$\mu_1 = 0$, $p = 4$, problem (3.36a)								
	$\alpha = 0,\ \beta = 0$		$\alpha = c^{-1},\ \beta = 0$		$\alpha = 0,\ \beta = c$		$\alpha = c^{-1},\ \beta = c$	
h	DG-error	rate	DG-error	rate	DG-error	rate	DG-error	rate
2^{-3}	2.0×10^{-6}	0.	2.5×10^{-6}	0.	2.7×10^{-6}	0.	3.1×10^{-6}	0.
2^{-4}	8.9×10^{-8}	4.50	1.1×10^{-7}	4.49	1.2×10^{-7}	4.49	1.4×10^{-7}	4.49
2^{-5}	3.9×10^{-9}	4.50	4.8×10^{-9}	4.50	5.4×10^{-9}	4.49	6.1×10^{-9}	4.49
2^{-6}	1.7×10^{-10}	4.50	2.1×10^{-10}	4.50	2.4×10^{-10}	4.50	2.7×10^{-10}	4.50

Table 3.1: Errors committed by the quasi-Trefftz DG method for different combinations of the numerical flux parameters and vanishing volume penalisation coefficient.

$\mu_1\|_K = r_{K,c} \, \|c\|^{-1}_{L^\infty(K)}$, $p = 4$, problem (3.36a)								
	$\alpha = 0,\ \beta = 0$		$\alpha = c^{-1},\ \beta = 0$		$\alpha = 0,\ \beta = c$		$\alpha = c^{-1},\ \beta = c$	
h	DG-error	rate	DG-error	rate	DG-error	rate	DG-error	rate
2^{-3}	2.1×10^{-6}	0.	2.5×10^{-6}	0.	2.5×10^{-6}	0.	3.1×10^{-6}	0.
2^{-4}	9.0×10^{-8}	4.53	1.1×10^{-7}	4.51	1.1×10^{-7}	4.50	1.4×10^{-7}	4.50
2^{-5}	4.0×10^{-9}	4.51	4.8×10^{-9}	4.50	4.8×10^{-9}	4.50	6.1×10^{-9}	4.50
2^{-6}	1.7×10^{-10}	4.50	2.1×10^{-10}	4.50	2.1×10^{-10}	4.50	2.7×10^{-10}	4.50

Table 3.2: Errors committed by the quasi-Trefftz DG method for different combinations of the numerical flux parameters and positive volume penalisation coefficient.

The results are shown in Tables 3.1 and 3.2. The errors are measured in the $|||\cdot|||_{\mathrm{DG}}$ norm (3.8). We observe optimal convergence in all cases, despite vanishing jump- or volume-penalisation term. Even though the volume penalisation term is needed for the well-posedness proof in Theorem 3.2.2, in this example it is not necessary for the discrete problem to be well-posed and for the numerics to converge with optimal rate. In this example, the choices suggested by the analysis (shown in the last column of Table 3.2) result in a slightly larger error: this is because some of the terms on time-like faces in the $|||\cdot|||_{\mathrm{DG}}$ norm vanish when α or β are set to zero. Similar behaviors were observed for the wave equation in [67, Fig. 6], for the Helmholtz equation in [42, Fig. 7–8] (concerning the flux parameters) and in [55, §5.1] (concerning the volume penalisation parameter).

The $\mathcal{O}(h^{p+1/2})$ convergence rates observed coincide with those proved in the bound (3.28). If, instead of using the $|||\cdot|||_{\mathrm{DG}}$ norm, we measure the error at final time only, specifically in the $L^2(\Omega \times \{T\})$ norm for both the v and the $\boldsymbol{\sigma}$ components, we obtain $\mathcal{O}(h^{p+1})$ convergence rates (we

do not report the values here), i.e. they are half a power higher than those in the $|||\cdot|||_{\mathrm{DG}}$ norm. The same half-order difference has been observed for the non-Trefftz version of the same method and $c = 1$ in Table 1 of [9]; see also the considerations after Proposition 6.5 therein. Moreover, being the $L^2(\Omega \times \{T\})$ norm parameter-independent, the errors are slightly smaller for the flux parameter values suggested in (3.24).

3.4.2 Approximation properties of quasi-Trefftz spaces

We compare the numerical error for different choices of the discretisation spaces: the quasi-Trefftz space $\mathbb{QW}^p(\mathcal{T}_h)$ of (3.22), the first-order derivatives $\mathbb{Y}^p(\mathcal{T}_h)$ of the full polynomial space, and the Trefftz space $\mathbb{W}^p(\mathcal{T}_h)$, respectively defined by

$$\mathbb{Y}^p(\mathcal{T}_h) := \big\{(w,\boldsymbol{\tau}) \in \mathbf{H}(\mathcal{T}_h) :\ w|_K = \partial_t u,\ \boldsymbol{\tau}|_K = -\nabla u,\ u \in \mathbb{P}^{p+1}(K),\ \forall K \in (\mathcal{T}_h)\big\}, \qquad p \in \mathbb{N}_0,$$
$$\mathbb{W}^p(\mathcal{T}_h) := \big\{(w,\boldsymbol{\tau}) \in \mathbf{H}(\mathcal{T}_h) :\ w|_K = \partial_t u,\ \boldsymbol{\tau}|_K = -\nabla u,\ u \in \mathbb{P}^{p+1}(K),$$
$$-\Delta u + c^{-2}(\mathbf{x}_K)\partial_t^2 u = 0 \text{ in } K, \forall K \in (\mathcal{T}_h)\big\}.$$

Here we recall the definition of the space $\mathbb{W}^p(\mathcal{T}_h)$, already given in (2.7). This is the Trefftz space for the approximated IBVP in which the wavespeed c is substituted by an elementwise-constant approximant. We have $\mathbb{W}^p(\mathcal{T}_h), \mathbb{QW}^p(\mathcal{T}_h) \subset \mathbb{Y}^p(\mathcal{T}_h)$ and $\dim \mathbb{W}^p(\mathcal{T}_h) = \dim \mathbb{QW}^p(\mathcal{T}_h)$.

We consider the problems (3.36b) and (3.36c) in 2+1 dimensions and set initial and boundary conditions accordingly. We use meshes that are Cartesian product between a spatial, quasi-uniform, unstructured, triangular mesh in $(0,1)^2$ with spatial meshwidth h, and a uniform mesh in time with time-step $h_t \approx h$. Therefore all elements are right triangular prisms and all their sides have comparable lengths. We set the volume penalisation and numerical flux parameters to the values in (3.27) and (3.24), respectively. The errors are measured in $|||\cdot|||_{\mathrm{DG}}$ norm. The results are displayed in Figures 3.3 and 3.4.

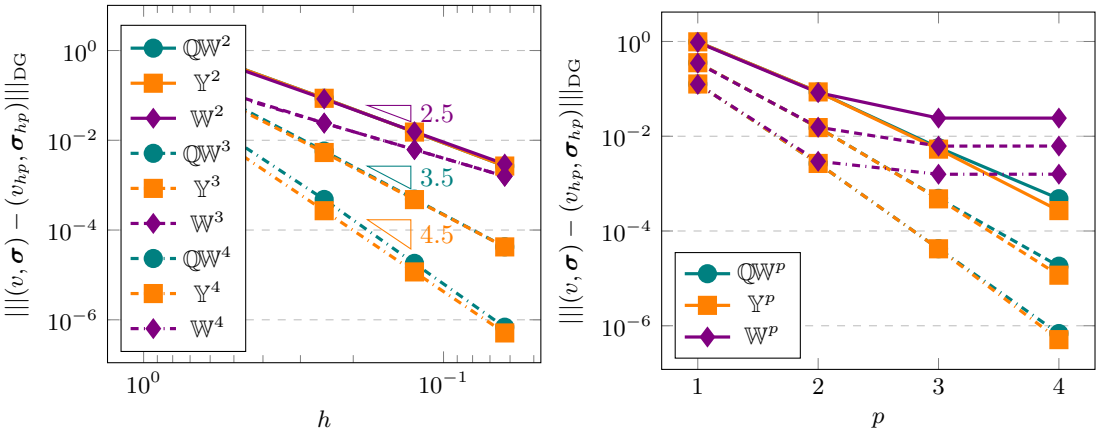

Figure 3.3: Comparison of different approximation spaces for problem (3.36c) as described in Section 3.4.2. Left panel: h-convergence. Right panel: p-convergence; the three sets of curves correspond to $h = 2^{-2}, 2^{-3}, 2^{-4}$.

Figure 3.3 focuses on (3.36c). The left panel plots the error against the mesh size for different values of p: the quasi-Trefftz space and the full polynomial space show the same, optimal, rate of convergence $\mathcal{O}(h^{p+1/2})$. The full polynomial space has a slightly smaller error throughout. The standard Trefftz space, however, does not achieve convergence with the same rate, but the rate is instead limited by roughly $\mathcal{O}(h^2)$; this is due to the low-order (piecewise-constant) approximation of c in the construction of the basis functions. The right panel of Figure 3.3 shows the error against the polynomial degree p for mesh sizes $h = 2^{-2}, 2^{-3}, 2^{-4}$. We observe exponential convergence for both the quasi-Trefftz space and the full polynomial space. As expected, the standard Trefftz

space does not lead to convergence in p because the approximation of c does not improve with p-refinement.

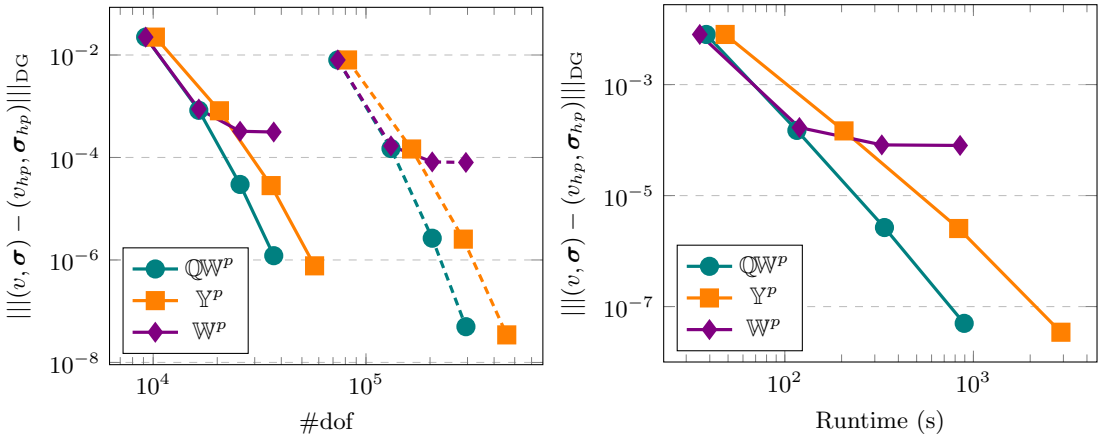

Figure 3.4: Left panel: comparison of different approximation spaces in terms of numbers of degrees of freedom for problem (3.36b), as described in §3.4.2. The continuous lines correspond to a mesh with $h = 2^{-3}$ and the dashed ones to $h = 2^{-4}$. The nodes in each line correspond to polynomial degrees $p = 1, 2, 3, 4$. Right panel: the same errors (for $h = 2^{-4}$) plotted against the computational time, including calculating the basis functions, assembly and solve. Both plots show that the quasi Trefftz space $\mathbb{QW}^p(\mathcal{T}_h)$ allows more efficient computations than the full polynomial space $\mathbb{Y}^p(\mathcal{T}_h)$.

In Figure 3.4 we switch to problem (3.36b) and plot (in the left panel) the error against the global number of degrees of freedom, on a fixed mesh, for increasing polynomial degrees p. The continuous and dashed lines correspond to two different mesh sizes, $h = 2^{-3}$ and $h = 2^{-4}$ respectively. The right panel plots the same error against the computational time. These plots illustrate the power of the quasi-Trefftz approach compared to the full polynomial approach, as discussed in Remark 3.3.11: for comparable numbers of degrees of freedom the quasi-Trefftz method can achieve much higher accuracy. In this example the accuracy improvement is up to about one and a half orders of magnitude, as observed when comparing the errors and the number of degrees of freedom for $\mathbb{QW}^4(\mathcal{T}_h)$ and $\mathbb{Y}^3(\mathcal{T}_h)$.

3.4.3 Tent-pitched meshes

The meshes used in all numerical examples in §3.4.1–3.4.2 are Cartesian products between a mesh in space and one in time. Thus the numerical solution has to be computed simultaneously for all the elements corresponding to the same time interval; this is analogous to an implicit time-stepping scheme.

We now discuss an alternative space–time meshing strategy: tent pitching. We call a mesh "tent-pitched" if all interior faces are space-like according to the definition in (3.3). This implies that the numerical solution in a given element K can be computed only from the numerical solutions on the elements that are adjacent to K and lying "before" K, thanks to the causality constraint (represented in the DG formulation (3.7) by the use of the v_{hp}^- and $\boldsymbol{\sigma}_{hp}^-$ traces on $\mathcal{F}_h^{\mathrm{space}}$). The solution can be computed independently, and in parallel, in several mesh elements and the solution procedure resembles an explicit time-stepping.

An example of a 1+1-dimensional tent-pitched mesh on $Q_T = (0,5) \times (0,5)$ can be seen in Figure 3.5. This mesh is constructed for the wavespeed $c(x) = (1+x)^{-1/2}$ of problem (3.36a), thus the tents in the right part of the domain are allowed to be "taller" than those on the left, i.e. to have longer extension in the time direction, without violating the causality constraint of having slope bounded by $c^{-1}(x)$.

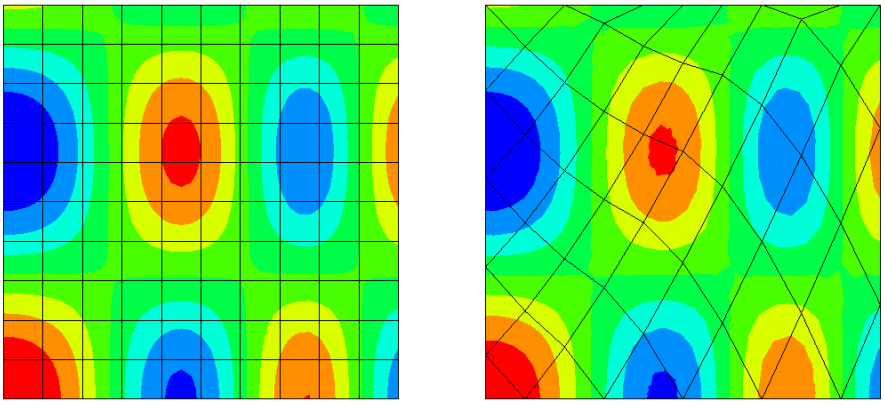

Figure 3.5: A Cartesian-product mesh (left) and a tent-pitched mesh (right) on the domain $Q_T = (0,5)^2$. Both show the solution u of problem (3.36a).

The algorithm used to produce the tent-pitched mesh used here can be found in [45]. A closer look into the implementation of Trefftz functions on tent-pitched meshes is given in [90].

To optimize storage during the computations on a tent-pitched mesh we only need to store the solution furthest in time. Therefore, in this section we measure the error at the final-time term in the definition of the DG norm (3.8):

$$\text{error}(T) = \left(\left\| \sqrt{G(\cdot)}(v(\cdot,T) - v_{hp}(\cdot,T)) \right\|^2_{L^2(\Omega)} + \|\boldsymbol{\sigma}(\cdot,T) - \boldsymbol{\sigma}_{hp}(\cdot,T)\|^2_{L^2(\Omega)} \right)^{1/2},$$

with final time $T = 1$. We use tent-pitched meshes in 2+1 dimensions to approximate problem (3.36b).

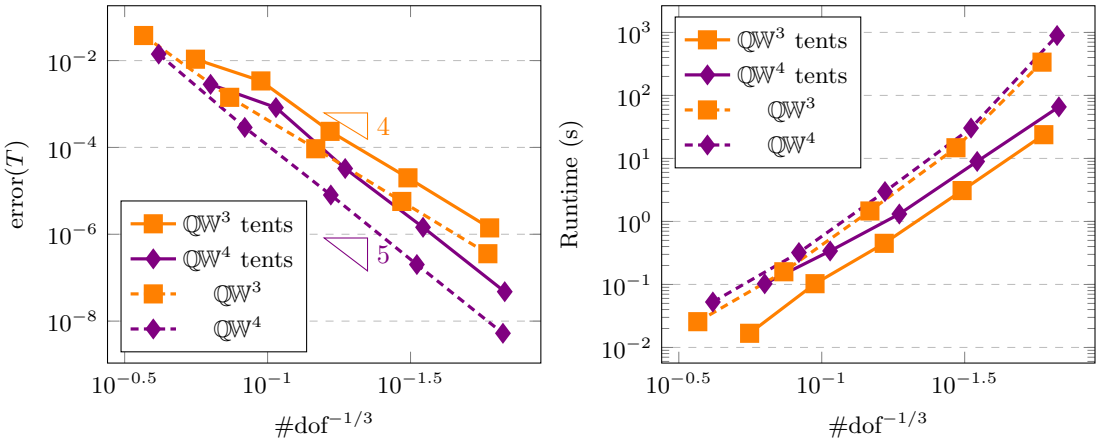

Figure 3.6: The final-time error (left) and the computational time (right) for the sequential solution of Problem (3.36b) on tent-pitched (continuous lines) and Cartesian meshes (dashed lines).

We first compare the error committed on tent-pitched meshes against that on Cartesian-product meshes (of the same kind of those in §3.4.2). For a fair comparison we plot the error in terms of the number of degrees of freedom, for varying mesh sizes. On the left panel of Figure 3.6 we observe optimal convergence rates of $\mathcal{O}(\#\text{dof}^{-(p+1)/3})$ for both meshing strategies, which corresponds to $\mathcal{O}(h^{p+1})$. The Cartesian-product mesh outperforms the tent-pitched mesh in terms of efficiency per degrees of freedom, due to the fact that we need more tent elements to cover the same space–time volume. However, in terms of computational time, shown in the right

panel of Figure 3.6, the tents perform better since they do not require the solution of any large linear system, even though in this comparison the solution is only solved sequentially without any parallelisation.

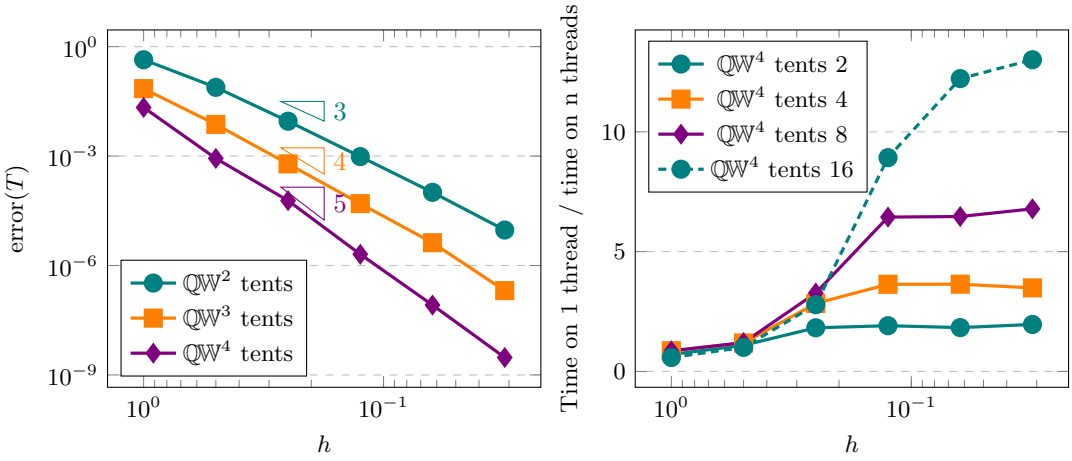

Figure 3.7: Left panel: the error on tent-pitched meshes for problem (3.36c). Right panel: the speedup in the computational time for 2 to 16 threads.

Next we study the effect of parallelisation. We measure the speedup obtained by increasing the number of threads, i.e. the maximum number of elements on which the solution is computed independently in parallel. Now we consider problem (3.36c); the final-time error in terms of the mesh size is shown in the left panel of Figure 3.7. The speedup in the computational time for 2, 4, 8 and 16 threads is shown in the right panel of Figure 3.7. We observe that the speedup factor is quite close to the number of threads. The figure shows that increasing the number of threads is beneficial only for moderate mesh sizes, as otherwise there are not enough independent tents. All timings were performed on a server with two Intel(R) Xeon(R) CPU E5-2687W v4, with 12 cores each.

3.4.4 Gaussian pulse in a non-homogenous medium

We illustrate the propagation of a vertical Gaussian pulse traveling through a medium with wavespeed varying along the x_2-direction, $G(x_1, x_2) = 1 + x_2$. The initial conditions are given by

$$\boldsymbol{\sigma}_0(x_1, x_2) = \Big(-\frac{2x_1}{\delta^2} \mathrm{e}^{-\frac{x_1^2}{\delta^2}}, 0 \Big), \qquad v_0(x) = 0 \qquad \text{on } \Omega = (0,1)^2,$$

setting $\delta = 2^{-5}$. We choose homogeneous Neumann boundary conditions, a tent-pitched mesh as discussed in the previous section, spatial mesh size $h = 2^{-7}$ and polynomial degree $p = 3$. Snapshots of the solution are shown in Figure 3.8. At $T = 0$ the initial condition is constant in x_2-direction. In the next snapshot, at $T = 0.25$, we can see the expected effects of the variable wavespeed: at the top of the domain, the wave travels faster than at the bottom. At $T = 0.5$ the wavefront on the top side reaches the right border. In the last image, at $T = 0.75$, we can see the wave being reflected from the right boundary. Boundary effects due to the homogeneous Neumann boundary conditions at the top and bottom of the domain can also be observed.

In Figure 3.9 we plot the energy (3.9) for different spatial mesh sizes $h = 2^{-5}, 2^{-6}, 2^{-7}$. The energy is computed at constant times t multiple of 0.0025 as $\mathcal{E}(t; w, \boldsymbol{\tau}) := \frac{1}{2} \int_\Omega (c^{-2} w^2 + |\boldsymbol{\tau}|^2) \, \mathrm{d}S$, by forcing the tent pitched mesh into slabs. As observed in §3.2.5, the method is dissipative. For $h = \delta = 2^{-5}$ there are not enough elements to resolve the wave front with sufficient accuracy, and the energy dissipates very quickly. For the two finer meshes the energy loss behaves much better; in particular for $h = 2^{-7}$ only 0.076% of the initial energy is lost at the final time $T = 1$.

Figure 3.8: Snapshots of the solution of the problem described in Section 3.4.4 at times $0, 0.25, 0.5, 0.75$.

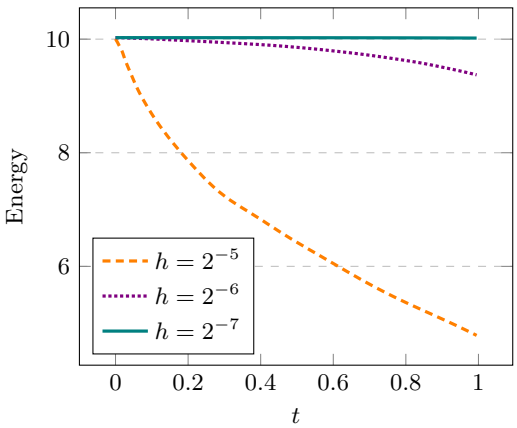

Figure 3.9: The energy (3.9) of the numerical solution shown in Figure 3.8.

Chapter 4

Entropy structure preserving method for cross-diffusion systems

The structure of this chapter is as follows. In Section 4.1, we state the problem and make the necessary assumptions for the existence of an entropy function. In Section 4.2, we present the space–time Galerkin method on a regularized formulation of the problem in the entropy variable unknown, and state our two main results in proposition 4.2.2 and proposition 4.2.3, namely existence and convergence of discrete solutions, respectively. Existence of discrete solutions is proven in Section 4.2.1. The proof of convergence will be split into two parts, first showing convergence with respect to mesh size in Section 4.2.2, then proving convergence as the regularization parameter goes to zero in Section 4.2.3. In Section 4.2.4, we are then able to prove existence of a weak solution of the continuous problem. Numerical tests for the porous medium, the Fisher-KPP, and the Maxwell-Stefan problem are presented in Section 4.3. All numerical results[1] were obtained using the finite element software Netgen/NGSolve, see [99,100]. Additionally, in Section 4.4, we reformulate the Maxwell-Stefan system with implicitly given currents in terms of the concentrations, and test it numerically.

4.1 General setting

Let $\Omega \subset \mathbb{R}^n$ be a bounded domain, and $\rho_0 \in L^\infty(\Omega)^N$, $N \geq 1$, a vector-valued function. We consider the following nonlinear reaction-diffusion system in the vector-valued unknown $\rho(t) = (\rho_1, \dots, \rho_N)(\cdot, t) : \Omega \to \mathbb{R}^N$:

$$\begin{cases} \partial_t \rho - \nabla \cdot (A(\rho)\nabla\rho) = f(\rho) & \text{in } \Omega,\ t > 0, \\ (A(\rho)\nabla\rho) \cdot \nu = 0 & \text{on } \partial\Omega,\ t > 0, \\ \rho(0) = \rho_0 & \text{in } \Omega. \end{cases} \tag{4.1}$$

Here, $A(\rho) \in \mathbb{R}^{N\times N}$ is the diffusion matrix, $f(\rho) : \mathbb{R}^N \to \mathbb{R}^N$ represents the reactions, and ν is the outward pointing unit normal vector at $\partial\Omega$; moreover, for $1 \leq i \leq N$,

$$(\nabla \cdot (A(\rho)\nabla\rho))_i = \sum_{\mu=1}^{n}\sum_{j=1}^{N} \frac{\partial}{\partial x_\mu}\left(A_{ij}(\rho)\frac{\partial \rho_j}{\partial x_\mu}\right), \quad ((A(\rho)\nabla\rho)\cdot\nu)_i = \sum_{\mu=1}^{n}\sum_{j=1}^{N} A_{ij}(\rho)\frac{\partial \rho_j}{\partial x_\mu}\nu_\mu.$$

We make the following hypotheses, which are similar to the assumptions made by A. Jüngel in [58].

(H1) $A \in C^0(\overline{\mathcal{D}}; \mathbb{R}^{N\times N})$ and $f \in C^0(\overline{\mathcal{D}}; \mathbb{R}^N)$, for a bounded domain $\mathcal{D} \subset (0, \infty)^N$.

[1] The code is available online at https://github.com/PaulSt/CrossDiff

(H2) There exists a convex function $s \in C^2(\mathcal{D}, [0,\infty)) \cap C^0(\overline{\mathcal{D}})$, with $s' : \mathcal{D} \to \mathbb{R}^N$ invertible and $u := (s')^{-1} \in C^1(\mathbb{R}^N, \mathcal{D})$, such that the following two conditions are satisfied:

(H2a) There exists a constant $\gamma > 0$ such that

$$z \cdot s''(\rho) A(\rho) z \geq \gamma |z|^2 \qquad \forall z \in \mathbb{R}^N,\ \rho \in \mathcal{D}.$$

Note that $s''(\rho)$ is matrix-valued, with $(s''(\rho))_{k\ell} = \frac{\partial}{\partial \rho_k}(s'(\rho))_\ell = \frac{\partial^2}{\partial \rho_k \partial \rho_\ell} s(\rho)$.

(H2b) There exists a constant $C_f \geq 0$ such that

$$f(\rho) \cdot s'(\rho) \leq C_f \qquad \forall \rho \in \mathcal{D}.$$

A discussion on when it is posssible to find a convex function s such that $(H2)$ is satisfied for cross-diffusion equations can be found in [21] (see [21, Lemma 22]).

Definition 4.1.1. *Let $T > 0$. We call $\rho \in L^2(0,T; H^1(\Omega)^N) \cap H^1(0,T; (H^1(\Omega)')^N)$ a weak solution of (4.1) if*

$$\int_0^T \langle \phi, \partial_t \rho \rangle dt + \sum_{i,j=1}^N \int_0^T \int_\Omega \nabla \phi_i \cdot A_{ij}(\rho) \nabla \rho_j dx dt = \int_0^T \int_\Omega \phi \cdot f(\rho) dx dt \tag{4.2}$$

for all $\phi \in L^2(0,T; H^1(\Omega)^N)$, with $\rho(0) = \rho_0$, where $\langle \cdot, \cdot \rangle$ denotes the duality product between $H^1(\Omega)^N$ and $(H^1(\Omega)')^N$.

4.2 Space–time Galerkin method

Let the time $T \in (0,\infty)$ be fixed. We denote by $Q_T = (0,T) \times \Omega$ the space–time cylinder for a domain $\Omega \subset \mathbb{R}^n$, $n \geq 1$. The main idea for a space–time numerical scheme is to perform integration by parts in (4.2) in the time variable, and to use the embedding

$$C([0,T]; L^2(\Omega)^N) \subset L^2(0,T; H^1(\Omega)^N) \cap H^1(0,T; (H^1(\Omega)')^N), \tag{4.3}$$

which can be proved exactly as in [35, Chapter 5.9, Theorem 3]. We arrive at the following lemma, which will be proved in section 4.2.4 below (see Remark 4.2.9).

Lemma 4.2.1. *Let $T > 0$. A function $\rho \in L^2(0,T; H^1(\Omega)^N) \cap H^1(0,T; (H^1(\Omega)')^N)$ is a weak solution of (4.1) if and only if*

$$\overbrace{\int_\Omega \phi(T) \cdot \rho(T) dx - \int_\Omega \phi(0) \cdot \rho_0 dx - \int_0^T \int_\Omega \partial_t \phi \cdot \rho dx dt}^{a(\rho,\phi;\rho_0) :=} + \sum_{i,j=1}^N \int_0^T \int_\Omega \nabla \phi_i \cdot A_{ij}(\rho) \nabla \rho_j dx dt = \int_0^T \int_\Omega \phi \cdot f(\rho) dx dt \tag{4.4}$$

for all $\phi \in H^1(Q_T)^N$. Here, we use the notation $\phi(t) := \mathrm{tr}(\phi)(t,\cdot)$, where tr denotes the trace operator $\mathrm{tr} : H^1(Q_T)^N \to L^2(\{0,T\} \times \Omega)^N$.

The next step is to introduce the following regularized problem: find $w \in H^1(Q_T)^N$ such that

$$\varepsilon(\phi, w)_{H^1(Q_T)^N} + a(\rho, \phi; \rho_0) + \sum_{i,j=1}^N \int_0^T \int_\Omega \nabla \phi_i \cdot A_{ij}(\rho) \nabla \rho_j dx dt = \int_0^T \int_\Omega \phi \cdot f(\rho) dx dt \tag{4.5}$$

for all $\phi \in H^1(Q_T)^N$, where w is the so-called entropy variable, which satisfies $\rho = u(w)$. Here, we have denoted by $(\cdot,\cdot)_{H^1(Q_T)^N}$ the standard $H^1(Q_T)^N$ inner product.

Next, we discretize equation (4.5). Let $\{\boldsymbol{V}_h\}_{h>0}$ be a family of finite dimensional spaces, parametrized by $h > 0$, such that, for every h, $\boldsymbol{V}_h \subset C^0(\overline{Q}_T)^N$. We make the following approximability assumption on the family of spaces $\{\boldsymbol{V}_h\}_{h>0}$.

(H3) For all $v \in H^1(Q_T)^N$,

$$\lim_{h\to 0} \inf_{v_h \in \boldsymbol{V}_h} \|v - v_h\|_{H^1(Q_T)^N} = 0.$$

Finally, we consider the following space–time Galerkin scheme in the entropy variable unknown: Find $w_h^\varepsilon \in \boldsymbol{V}_h$ such that, by setting $\rho_h^\varepsilon := u(w_h^\varepsilon)$, it holds true that

$$\epsilon(\phi, w_h^\varepsilon)_{H^1(Q_T)^N} + a(\rho_h^\varepsilon, \phi; \rho_0) + \sum_{i,j=1}^N \int_0^T \int_\Omega \nabla\phi_i \cdot A_{ij}(\rho_h^\varepsilon)\nabla(\rho_h^\varepsilon)_j dxdt = \int_0^T \int_\Omega \phi \cdot f(\rho_h^\varepsilon) dxdt \tag{4.6}$$

for all $\phi \in \boldsymbol{V}_h$. The first term in (4.6) can be interpreted as a stabilization term for the Galerkin scheme, with parameter $\varepsilon > 0$. This is used to obtain a control of the entropy variable. Note that we want to find a solution $w_h^\varepsilon \in \boldsymbol{V}_h$. Due to the nonlinearity of u, we expect that $\rho = u(w_h^\varepsilon) \notin \boldsymbol{V}_h$.

The following two propositions will be proven in section 4.2.1 and section 4.2.4, respectively.

Proposition 4.2.2 (Existence of discrete solutions)**.** *Assume that $\rho_0 : \Omega \to \overline{\mathcal{D}}$ is measurable. Then there exists a solution $w_h^\varepsilon \in \boldsymbol{V}_h$ of method (4.6). Moreover, every solution $w_h^\varepsilon \in \boldsymbol{V}_h$ of (4.6), for $\varepsilon, h > 0$, satisfies the entropy estimate*

$$\epsilon\|w_h^\varepsilon\|^2_{H^1(Q_T)^N} + \int_\Omega s(\rho_h(T))dx + \gamma \int_{Q_T} |\nabla\rho_h^\varepsilon|^2 dxdt \le \int_\Omega s(\rho_0)dx + C_f|\Omega|T, \tag{4.7}$$

where $\rho = u(w_h^\varepsilon)$, $|\Omega|$ is the volume of Ω, and γ and C_f are as in Assumption (H2).

Proposition 4.2.3 (Convergence)**.** *Assume that $\rho_0 : \Omega \to \overline{\mathcal{D}}$ is measurable, and let $w_h^\epsilon \in \boldsymbol{V}_h$ be a solution of (4.6) for $\epsilon, h > 0$. Then there exist a weak solution*

$$\rho \in L^2(0,T; H^1(\Omega)^N) \cap H^1(0,T; (H^1(\Omega)')^N) \cap L^\infty((0,T)\times\Omega)^N$$

of (4.1) and sequences $h_i, \epsilon_i \to 0$, as $i \to \infty$, such that

$$u(w_{h_i}^{\epsilon_i}) \to \rho \qquad \text{in } L^r(Q_T)^N, \text{ as } i \to \infty$$

for all $r \in [1, \infty)$. Moreover, ρ satisfies the entropy estimate

$$\int_\Omega s(\rho(\tau))dx + \gamma \int_0^\tau \int_\Omega |\nabla\rho|^2 dxdt \le \int_\Omega s(\rho_0)dx + C_f|\Omega|\tau \tag{4.8}$$

for all $\tau \in (0,T]$, where $|\Omega|$ is the volume of Ω, and γ and C_f are as in Assumption (H2).

4.2.1 Existence of a solution of the numerical scheme

Proof of Proposition 4.2.2. The idea is to use Leray-Schauder fixed-point theorem for the mapping $\Phi : \boldsymbol{V}_h \to \boldsymbol{V}_h$, $v \mapsto w$, where w denotes the unique solution of (4.6) for $\rho = u(v)$. Since A, f, u are continuous, so is Φ. Since $\boldsymbol{V}_h$ has finite dimension, Φ is also compact. Then by the Leray-Schauder fixed-point theorem, we obtain that Φ admits a fixed-point if we can show that the set

$$\{w \in \boldsymbol{V}_h : w = \sigma\Phi(w),\ \sigma \in [0,1]\}$$

is bounded.

Let $w = \sigma\Phi(w)$ for $\sigma \in (0,1]$ and choose $\phi := w$. Then (4.6) entails

$$\frac{\epsilon}{\sigma}\|w\|^2_{H^1(Q_T)^N} + \int_\Omega w(T)\cdot\rho(T)dx - \int_\Omega w(0)\cdot\rho_0 dx - \int_0^T\int_\Omega \partial_t w\cdot\rho dxdt + \sum_{i,j=1}^N \int_0^T\int_\Omega \nabla w_i \cdot A_{ij}(\rho)\nabla\rho_j dxdt = \int_0^T\int_\Omega w\cdot f(\rho)dxdt.$$

Using that $\rho = u(w)$ and $\partial_t(s(u(w))) = s'(u(w)) \cdot \partial_t(u(w)) = w \cdot \partial_t(u(w))$, we have

$$\begin{aligned}\partial_t w \cdot \rho = \partial_t w \cdot u(w) &= \partial_t(w \cdot u(w)) - w \cdot \partial_t(u(w)) = \partial_t(w \cdot u(w) - s(u(w)))\\ &= \partial_t(w \cdot \rho - s(\rho)).\end{aligned}$$

Thus, by the fundamental theorem of calculus,

$$\begin{aligned}&\int_\Omega w(T) \cdot \rho(T) dx - \int_\Omega w(0) \cdot \rho_0 dx - \int_0^T \int_\Omega \partial_t w \cdot \rho dx dt\\ &\qquad = -\int_\Omega \big(s(\rho(0)) + w(0) \cdot (\rho_0 - \rho(0))\big) dx + \int_\Omega s(\rho(T)) dx.\end{aligned}$$

Note that, by definition, $s'(\rho) = s'(u(w)) = w$. The convexity of s then implies that

$$s(\rho(0)) + w(0) \cdot (\rho_0 - \rho(0)) = s(\rho(0)) + s'(\rho(0)) \cdot (\rho_0 - \rho(0)) \le s(\rho_0)$$

and hence,

$$\int_\Omega w(T) \cdot \rho(T) dx - \int_\Omega w(0) \cdot \rho_0 dx - \int_0^T \int_\Omega \partial_t w \cdot \rho dx dt \ge \int_\Omega s(\rho(T)) dx - \int_\Omega s(\rho_0) dx.$$

The next step is to use (H2a) in combination with $w = s'(\rho)$, which yields that

$$\begin{aligned}\sum_{i,j=1}^N \nabla w_i \cdot A_{ij}(\rho) \nabla \rho_j &= \sum_{i,j=1}^N \nabla (s'(\rho))_i \cdot A_{ij}(\rho) \nabla \rho_j\\ &= \sum_{i,j,k=1}^N \nabla \rho_k \cdot (s''(\rho))_{ki} A_{ij}(\rho) \nabla \rho_j \ge \gamma |\nabla \rho|^2,\end{aligned}$$

where $|\nabla \rho|^2 := \sum_{\ell=1}^n |\frac{\partial}{\partial x_\ell} \rho|^2$. Moreover, due to (H2b) and $w = s'(\rho)$, we have

$$w \cdot f(\rho) = s'(\rho) \cdot f(\rho) \le C_f.$$

Therefore, we can conclude the entropy estimate

$$\frac{\epsilon}{\sigma} \|w\|^2_{H^1(Q_T)^N} + \int_\Omega s(\rho(T)) dx + \gamma \int_{Q_T} |\nabla \rho|^2 dx dt \le \int_\Omega s(\rho_0) dx + C_f |\Omega| T.$$

Hence, $\|w\|^2_{H^1(Q_T)^N}$ is uniformly bounded, because $\sigma \le 1$. Thus, the Leray-Schauder theorem is applicable and yields that Φ has a fixed point, and therefore the scheme (4.6) admits a solution. Using these calculations for $\sigma = 1$, it follows that every solution has to satisfy the entropy inequality (4.7). □

4.2.2 Convergence of the numerical scheme as $h \to 0$

We will show that, for a fixed $\epsilon > 0$, the numerical scheme (4.6) converges as $h \to 0$.

Proposition 4.2.4 (Convergence of the scheme for fixed $\epsilon > 0$)**.** *Let $w_h \in \boldsymbol{V}_h$ be a solution of (4.6) with fixed $\epsilon > 0$, satisfying the entropy estimate (4.7). Then there exists $w \in H^1(Q_T)^N$ with $\rho := u(w) \in L^2(0,T,H^1(\Omega)^N)$, and a sequence $h_\ell \to 0$ such that*

$$\rho_{h_\ell} := u(w_{h_\ell}) \to \rho \text{ strongly in } L^r(Q_T) \text{ for all } r \in [1,\infty).$$

Moreover, w, ρ solve (4.5) and satisfy the entropy estimate

$$\epsilon \|w\|^2_{H^1(Q_T)^N} + \int_\Omega s(\rho(T)) dx + \gamma \int_{Q_T} |\nabla \rho|^2 dx dt \le \int_\Omega s(\rho_0) dx + C_f |\Omega| T. \tag{4.9}$$

Proof. The first part of the assertion follows from the fact that w_h is uniformly bounded in $H^1(Q_T)^N$, which yields that there exists $w \in H^1(Q_T)^N$ and subsequence $h_\ell \to 0$ such that $w_{h_\ell} \rightharpoonup w$ in $H^1(Q_T)^N$, due to the Banach-Alaoglu theorem, and $w_{h_\ell} \to w$ in $L^2(Q_T)^N$, due to Rellich's theorem. In particular, we can choose this subsequence in such a way that w_{h_ℓ} converges a.e. to w. As u is bounded (see Assumption (H2)), the dominated convergence theorem entails the strong convergence of $\rho_{h_\ell} \equiv u(w_{h_\ell}) \to u(w) =: \rho$ in $L^r(Q_T)^N$ for all $r \in [1, \infty)$. Combining this with the entropy estimate (4.7), there exists another subsequence (which we do not relabel) such that $\rho_\ell \rightharpoonup \rho$ weakly in $L^2(0, T, H^1(\Omega)^N)$.

Finally, owing to assumption (H3), for every $\phi \in H^1(Q_T)^N$, there exists $\phi_{h_\ell} \in \boldsymbol{V}_{h_\ell}$ such that $\phi_{h_\ell} \to \phi$ in $H^1(Q_T)^N$. Using ϕ_{h_ℓ} as a test function in (4.6), we obtain (4.5) in the limit $h_i \to 0$, as each integral in (4.6) converges separately. The entropy inequality (4.9) is a consequence of Fatou's lemma and the weak lower semicontinuity of the norm. □

The following corollary will be used in the analysis of the limit for $\epsilon \to 0$ (see proof of Proposition 4.2.6 below).

Corollary 4.2.5. *Let $\tau, \delta \geq 0$ be such that $\tau + \delta \leq T$. Let $w \in H^1(Q_T)^N$ together with $\rho := u(w) \in L^2(0, T, H^1(\Omega)^N)$ be a solution of (4.5). It holds true that*

$$\epsilon \|w\|^2_{H^1(Q_\tau)^N} + \frac{1}{\delta} \int_\tau^{\tau+\delta} \int_\Omega s(\rho) dx dt + \gamma \int_0^\tau \int_\Omega |\nabla \rho|^2 dx dt \\ \leq (1 + \delta) \int_\Omega s(\rho_0) dx + C_f |\Omega| (\tau + \delta(1/2 + T)), \quad (4.10)$$

where $Q_\tau := (0, \tau) \times \Omega$.

Proof. Set

$$\psi(t) := \begin{cases} 1 & \text{if } t < \tau, \\ 1 - \frac{t-\tau}{\delta} & \text{if } \tau \leq t \leq \tau + \delta, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, $w\psi \in H^1(Q_T)^N$. Similarly as in the proof of Proposition 4.2.2, we use $\rho = u(w)$ and

$$\partial_t(\psi w) \cdot \rho = \partial_t(\psi w \cdot \rho) - \psi w \cdot \partial_t \rho = \partial_t(\psi w \cdot \rho - \psi s(\rho)) + \partial_t \psi s(\rho)$$

and, since $\psi(T) = 0$ and $\psi(0) = 1$,

$$\int_{Q_T} \partial_t(w\psi) \cdot \rho dx dt + \int_\Omega w(0) \cdot \rho_0 dx = \int_{Q_T} \partial_t \psi s(\rho) dx dt + \int_\Omega \big(s(\rho(0)) + w(0) \cdot (\rho_0 - \rho(0))\big) dx.$$

Thus, using the definition of ψ, and treating the last term of the previous equation as in the proof of Proposition 4.2.2, we get

$$\int_{Q_T} \partial_t(w\psi) \cdot \rho dx dt + \int_\Omega \psi(0) w(0) \cdot \rho_0 dx + \frac{1}{\delta} \int_\tau^{\tau+\delta} \int_\Omega s(\rho) dx dt \leq \int_\Omega s(\rho_0) dx.$$

From (4.5) tested with $\phi = w\psi$ and the previous inequality, we get

$$\epsilon(\psi w, w)_{H^1(Q_T)^N} + \sum_{i,j=1}^N \int_0^T \int_\Omega \nabla(\psi w)_i \cdot A_{ij}(\rho) \nabla \rho_j dx dt + \frac{1}{\delta} \int_\tau^{\tau+\delta} \int_\Omega s(\rho) dx dt \\ \leq \int_\Omega s(\rho_0) dx + \int_0^T \int_\Omega \psi w \cdot f(\rho) dx dt$$

which, due to the properties of ψ and the assumption (H2), entails

$$\epsilon(\psi w, w)_{H^1(Q_T)^N} + \frac{1}{\delta}\int_\tau^{\tau+\delta}\int_\Omega s(\rho)dxdt + \gamma\int_0^\tau\int_\Omega |\nabla\rho|^2 dxdt \le \int_\Omega s(\rho_0)dx + C_f|\Omega|(\tau+\delta/2).$$

Finally, we can estimate the first term as

$$\begin{aligned}\epsilon(\psi w, w)_{H^1(Q_T)^N} &= \epsilon\int_0^T\int_\Omega \left(\psi|w|^2 + \psi|\nabla w|^2 + \partial_t(\psi w)\cdot\partial_t w\right)dxdt \\ &= \epsilon\int_0^T\int_\Omega \left(\psi|w|^2 + \psi|\nabla w|^2 + \psi|\partial_t w|^2 + \partial_t\psi w\cdot\partial_t w\right)dxdt \\ &\ge \epsilon\|w\|^2_{H^1(Q_\tau)^N} - \delta\epsilon\int_\tau^{\tau+\delta}\int_\Omega w\cdot\partial_t w dxdt.\end{aligned}$$

Using the Cauchy-Schwarz inequality and the definition of the H^1 norm yields

$$\delta\epsilon\int_\tau^{\tau+\delta}\int_\Omega w\cdot\partial_t w dxdt \le \delta\epsilon\|w\|^2_{H^1((\tau,\tau+\delta)\times\Omega)^N},$$

and therefore

$$\begin{aligned}\epsilon\|w\|^2_{H^1(Q_\tau)^N} + \frac{1}{\delta}\int_\tau^{\tau+\delta}\int_\Omega s(\rho)dxdt + \gamma\int_0^\tau\int_\Omega |\nabla\rho|^2 dxdt& \\ \le \delta\epsilon\|w\|^2_{H^1((\tau,\tau+\delta)\times\Omega)^N} + \int_\Omega s(\rho_0)dx + C_f|\Omega|(\tau+\delta/2).&\end{aligned}$$

Note that we cannot estimate the first term on the right-hand side by the first term on the left-hand side, because the domain of the norms are disjoint. Fortunately, we have the entropy estimate (4.9), which we add δ times to this inequality to get

$$\begin{aligned}\epsilon(1+\delta)\|w\|^2_{H^1(Q_\tau)^N} + \frac{1}{\delta}\int_\tau^{\tau+\delta}\int_\Omega s(\rho)dxdt + \delta\int_\Omega s(\rho(T))dx + \gamma(1+\delta)\int_0^\tau\int_\Omega |\nabla\rho|^2 dxdt& \\ \le (1+\delta)\int_\Omega s(\rho_0)dx + C_f|\Omega|(\tau+\delta(1/2+T)),&\end{aligned}$$

which, since $s(\rho(T)) \ge 0$, implies the assertion. □

4.2.3 Limit of $\epsilon \to 0$

We consider the limiting problem

$$\begin{aligned}-\int_\Omega \phi(0)\cdot\rho_0 dx - \int_0^T\int_\Omega \partial_t\phi\cdot\rho dxdt + \sum_{i,j=1}^N\int_0^T\int_\Omega \nabla\phi_i\cdot A_{ij}(\rho)\nabla\rho_j dxdt& \\ = \int_0^T\int_\Omega \phi\cdot f(\rho)dxdt \quad (4.11)&\end{aligned}$$

and all $\phi \in (H^1(Q_T))^N$ with $\phi(T) = 0$. As in the statement of Lemma 4.2.1, we use the notation $\phi(t) := \mathrm{tr}(\phi)(t,\cdot)$, where tr denotes the trace operator $\mathrm{tr} : H^1(Q_T)^N \to L^2(\{0,T\}\times\Omega)^N$.

Proposition 4.2.6. *Let $\tau, \delta \ge 0$ such that $\tau + \delta \le T$. Let $w \in H^1(Q_T)^N$ together with $\rho := u(w) \in L^2(0,T,H^1(\Omega)^N)$ be a solution of (4.5). Then there exist $\rho \in L^2(0,T;H^1(\Omega)^N)$ with $\rho(t,x) \in \overline{\mathcal{D}}$ for a.e. $(t,x) \in Q_T$ being a solution of (4.11) and a subsequence $\epsilon_j \to 0$ such that*

$$\rho^{\epsilon_j} \to \rho \qquad \text{in every } L^r(Q_T)^N, r \in [1,\infty), \text{ as } \epsilon_j \to 0.$$

Moreover, ρ satisfies the entropy inequality

$$\frac{1}{\delta}\int_\tau^{\tau+\delta}\int_\Omega s(\rho)dxdt + \gamma\int_0^\tau\int_\Omega |\nabla\rho|^2 dxdt \le (1+\delta)\int_\Omega s(\rho_0)dx + C_f|\Omega|(\tau+\delta(1/2+T)). \quad (4.12)$$

In the proof of Proposition 4.2.6, the key ingredient to prove strong convergence of (at least a subsequence of) ρ^ϵ will be the idea of compensated compactness, which is a special technique applying the classical div-curl lemma; see, e.g. [107].

Lemma 4.2.7 (div-curl lemma)**.** *Let $\alpha, \alpha^\ell \in L^2(Q_T)^{1+n}$ and $\beta, \beta^\ell \in L^2(Q_T)^{1+n}$. Then*

$$\alpha^\ell \rightharpoonup \alpha \quad \text{in } L^2(Q_T)^{1+n} \text{ as } \ell \to +\infty, \text{ and } \quad (\mathrm{div}_{(t,x)}\alpha^\ell)_{\ell\in\mathbb{N}} \text{ is bounded in } L^2(Q_T),$$
$$\beta^\ell \rightharpoonup \beta \quad \text{in } L^2(Q_T)^{1+n} \text{ as } \ell \to +\infty, \text{ and } \quad (\mathrm{curl}_{(t,x)}\beta^\ell)_{\ell\in\mathbb{N}} \text{ is bounded in } L^2(Q_T)^{(1+n)\times(1+n)}$$

implies that

$$\alpha^\ell \cdot \beta^\ell \rightharpoonup \alpha \cdot \beta \qquad \text{in } \mathcal{D}'(Q_T) \qquad \text{as } \ell \to +\infty,$$

where $\mathcal{D}'(Q_T)$ denotes the dual space of $\mathcal{D}(Q_T) := C_c^\infty(Q_T)$.

Proof of Proposition 4.2.6. Let $w^\epsilon, \rho^\epsilon := u(w^\epsilon)$ denote the solution of (4.5) satisfying the entropy inequality (4.9). For any fixed i, $i = 1, \dots, N$, we define the vector-valued functions with $(1+n)$ components

$$\alpha^\epsilon = \begin{pmatrix} \rho_i^\epsilon - \epsilon\partial_t w_i^\epsilon \\ J_i^\epsilon - \epsilon\nabla w_i^\epsilon \end{pmatrix} \quad \text{and} \quad \beta^\epsilon := \begin{pmatrix} \rho_i^\epsilon \\ 0 \end{pmatrix}, \qquad \text{where } J_i^\epsilon = -\sum_{j=1}^N A(\rho^\epsilon)_{ij}\nabla\rho_j^\epsilon.$$

Note that, by assumption, $\mathcal{D}$ is bounded and so is $\rho^\epsilon = u(w^\epsilon)$. Thus, thanks to the entropy estimate (4.9), $\alpha^\epsilon, \beta^\epsilon$ are bounded uniformly in $L^2(Q_T)^{1+n}$ w.r.t. $\epsilon \in (0,1)$. By the Banach-Alaoglu theorem, there exist $\alpha, \beta \in L^2(Q_T)^{1+n}$ and a subsequence $\epsilon_\ell \to 0$ such that

$$\alpha^{\epsilon_\ell} \rightharpoonup \alpha, \beta^{\epsilon_\ell} \rightharpoonup \beta \quad \text{in } L^2(Q_T)^{1+n} \qquad \text{as } \epsilon_\ell \to 0.$$

Clearly, β has the form $(\rho_i, 0)$ for some $\rho_i \in L^2(Q_T)$. Due to the entropy estimate (4.9), $\sqrt{\epsilon}w_i^\epsilon$ is bounded in $H^1(Q_T)$. Hence, $\beta_0^\epsilon - \alpha_0^\epsilon = \epsilon\partial_t w_i^\epsilon \to 0$ in $L^2(Q_T)$ as $\epsilon \to 0$, implying that $\rho_i := \beta_0 = \alpha_0$ and $\alpha \cdot \beta = \rho_i^2$, where in this context the index 0 denotes the first component of the $(1+n)$-dimensional vector. Moreover, one can easily show that

$$\|\mathrm{curl}_{(t,x)}\beta^\epsilon\|_{L^2(Q_T)^{(1+n)\times(1+n)}} \le C\|\nabla\rho_i^\epsilon\|_{L^2(Q_T)^n}$$

for some $C > 0$. Again by the entropy estimate (4.9), this implies that $\mathrm{curl}_{(t,x)}\beta^\epsilon$ is uniformly bounded in $L^2(Q_T)^{(1+n)\times(1+n)}$ w.r.t. $\epsilon \in (0,1)$. In order to apply the div-curl lemma, it remains to prove that the space–time divergence of α^ϵ is bounded. For this, we require the equation for ρ_i^ϵ in the interior of Q_T, i.e., from equation (4.5),

$$\epsilon\int_{Q_T}\psi w_i^\epsilon dxdt + \epsilon\int_{Q_T}\partial_t\psi\partial_t w_i^\epsilon dxdt + \epsilon\int_{Q_T}\nabla\psi\cdot\nabla w_i^\epsilon dxdt - \int_{Q_T}\partial_t\psi\rho_i^\epsilon dxdt$$
$$+\sum_{j=1}^N\int_{Q_T}\nabla\psi\cdot A_{ij}(\rho^\epsilon)\nabla\rho_j^\epsilon dxdt = \int_{Q_T}\psi f_i(\rho^\epsilon)dxdt$$

for all $\psi \in H_0^1(Q_T)$. We can rewrite this by using the weak space–time divergence of α^ϵ as

$$-\int_{Q_T}\nabla_{(t,x)}\psi\cdot\alpha^\epsilon dxdt = \int_{Q_T}\partial_t\psi(\epsilon\partial_t w_i^\epsilon - \rho_i^\epsilon)dxdt$$
$$+\int_{Q_T}\nabla\psi\cdot\left(\epsilon\nabla w_i^\epsilon + \sum_{j=1}^N A_{ij}(\rho^\epsilon)\nabla\rho_j^\epsilon\right)dxdt$$
$$= \int_{Q_T}\psi f_i(\rho^\epsilon)dxdt - \epsilon\int_{Q_T}\psi w_i^\epsilon dxdt$$

for all $\psi \in H_0^1(Q_T)$. We observe that the right-hand side defines a bounded operator on $L^2(Q_T)$ due to the entropy estimate (4.9) and the fact that f_i is uniformly bounded as a continuous function defined on a compact set (see (H2)). This yields that $\mathrm{div}_{(t,x)}\alpha^\epsilon$ is uniformly bounded in $L^2(Q_T)$. Therefore, we can apply the div-curl lemma and obtain that

$$(\rho_i^{\epsilon_\ell} - \epsilon_\ell \partial_t w_i^{\epsilon_\ell})\rho_i^{\epsilon_\ell} = \alpha^{\epsilon_\ell} \cdot \beta^{\epsilon_\ell} \rightharpoonup \alpha \cdot \beta = \rho_i^2 \quad \text{in } \mathcal{D}'(Q_T) \qquad \text{as } \epsilon_\ell \to 0.$$

Using that $\rho_i^{\epsilon_\ell} \rightharpoonup \rho_i$ and $\epsilon_\ell \partial_t w_i^{\epsilon_\ell} \to 0$ in $L^2(Q_T)$, we obtain that

$$\int_{Q_T} (\rho_i^{\epsilon_\ell})^2 \phi^2 dxdt \to \int_{Q_T} \rho_i^2 \phi^2 dxdt \qquad \text{as } \epsilon_\ell \to 0$$

for all $\phi \in C_c^\infty(Q_T)$. Hence, $\phi\rho_i^{\epsilon_\ell} \to \phi\rho_i$ in $L^2(Q_T)$ for all $\phi \in C_c^\infty(Q_T)$. In particular, there exists a subsequence not being relabeled such that $\rho_i^{\epsilon_\ell} \to \rho_i$ a.e. in Q_T. For almost every $(t,x) \in Q_T$, we know that $\rho^{\epsilon_\ell}(t,x) \in \mathcal{D}$ and that $\mathcal{D}$ is bounded. Thus, we can apply the dominated convergence theorem, which yields that

$$\rho_i{}^{\epsilon_\ell} \to \rho_i \qquad \text{in every } L^r(Q_T), r \in [1,\infty), \text{ as } \epsilon_\ell \to 0,$$

and that $\rho(t,x) \in \mathcal{D}$ for almost every $(t,x) \in Q_T$.

Moreover, the entropy inequality (4.9) also states that $\nabla\rho_i^\epsilon$ is bounded in $L^2(Q_T)^n$ independently of ϵ. Since $|\rho^\epsilon| = |u(w^\epsilon)| = |(s')^{-1}(w^\epsilon)| \le \sup_{v\in\mathcal{D}} |v|^2$, according to (H2), then, using again (4.9), we obtain

$$\begin{aligned}\|\rho_i^\epsilon\|_{L^2(0,T;H^1(\Omega))}^2 &= \int_{Q_T} (\rho_i^\epsilon)^2 dxdt + \int_{Q_T} |\nabla\rho_i^\epsilon|^2 dxdt \\ &\le |\Omega|T\|\rho_i^\epsilon\|_{L^\infty(Q_T)}^2 + \frac{1}{\gamma}\left(\int_\Omega s(\rho_0)dx + C_f|\Omega|T\right) \\ &\le \frac{1}{\gamma}\int_\Omega s(\rho_0)dx + \left(\sup_{v\in\mathcal{D}}|v|^2 + \frac{C_f}{\gamma}\right)|\Omega|T,\end{aligned}$$

namely, ρ_i^ϵ is bounded in $L^2(0,T;H^1(\Omega))$ independent on ϵ. Taking yet another subsequence, which we do not relabel, we can see that there exists $\rho_i \in L^2(0,T;H^1(\Omega))$ such that $\rho_i^{\epsilon_\ell} \rightharpoonup \rho_i$ in $L^2(0,T;H^1(\Omega))$. In particular, $\nabla\rho_i^{\epsilon_\ell} \rightharpoonup \nabla\rho_i$ in $L^2(Q_T)^n$. We already have seen that $\sqrt{\epsilon}w^\epsilon$ is bounded in $H^1(Q_T)^N$, then $\epsilon w^\epsilon \to 0$ in $H^1(Q_T)^N$.

Now, we prove that ρ is solution to the limiting problem (4.11). Let $\phi \in H^1(Q_T)$ with trace $\phi(T) = 0$. Using that A is bounded, according to (H1), and the dominated convergence theorem yields

$$\int_{Q_T} |\nabla\phi|^2 |A_{ij}(\rho^{\epsilon_\ell})|^2 dxdt \to \int_{Q_T} |\nabla\phi|^2 |A_{ij}(\rho)|^2 dxdt \qquad \text{as } \epsilon_\ell \to 0.$$

In particular, $\nabla\phi A_{ij}(\rho^{\epsilon_\ell})$ converges strongly in $L^2(Q_T)^n$. For each $i = 1,\dots,N$, we test the equation for ρ_i^ϵ (see (4.5)) with functions $\phi \in H^1(Q_T)$ with trace $\phi(T) = 0$, take the limit for $\epsilon = \epsilon_\ell \to 0$, and obtain

$$-\int_\Omega \phi(0)\rho_i^0 dx - \int_0^T\int_\Omega \partial_t\phi\rho_i dxdt + \sum_{j=1}^N \int_0^T\int_\Omega \nabla\phi \cdot A_{ij}(\rho)\nabla\rho_j dxdt = \int_0^T\int_\Omega \phi f_i(\rho)dxdt$$

for all $i = 1,\dots,N$.

Finally, recall that ρ^ϵ satisfies the entropy estimate (4.10) from Corollary 4.2.5. Thus, we obtain the entropy inequality (4.12) as a direct consequence of the lower weak continuity of the L^2 norm and the Fatou lemma. □

4.2.4 Existence of a weak solution

In this section, we prove that problem (4.1) possesses a weak solution ρ in the sense of Definition 1. Moreover, we prove the equivalence stated in Lemma 4.2.1 beetween the weak formulation (4.2) in Definition 1 and the weak formulation (4.4).

Proposition 4.2.8. *Let ρ be given by Proposition 4.2.6. Then $\rho \in H^1(0,T;(H^1(\Omega)')^N)$ and $\rho \in C^0([0,T];L^2(\Omega))$ with $\rho(0)=\rho_0$. Moreover, it satisfies the entropy inequality*

$$\int_\Omega s(\rho(\tau))dx + \gamma \int_0^\tau \int_\Omega |\nabla \rho|^2 \, dxdt \le \int_\Omega s(\rho_0)dx + C_f|\Omega|\tau. \tag{4.13}$$

for almost all $\tau \in (0,T)$.

Proof. Using the equation (4.11), we obtain that

$$\begin{aligned}\left|\int_{Q_T} \partial_t \phi \rho_i dxdt\right| &\le \sum_{j=1}^N \int_{Q_T} |\nabla\phi||A_{ij}(\rho)\nabla\rho_j|dxdt + \int_{Q_T} |\phi||f_i(\rho)|dxdt + \int_\Omega |\phi(0)||\rho_{0,i}|dx \\ &\le C_\rho \|\phi\|_{L^2(0,T;H^1(\Omega))}\end{aligned}$$

using that $\rho \in L^\infty(Q_T) \cap L^2(0,T;H^1(\Omega))$. This implies that, for each $i=1,\dots,N$, ρ_i has a weak time derivative satisfying $\partial_t \rho_i \in L^2(0,T;H^1(\Omega)')$. Then the embedding $H^1(0,T;H^1(\Omega)') \cap L^2(0,T;H^1(\Omega)) \subset C^0([0,T];L^2(\Omega))$, entails that every ρ_i is continuous in time, and so is ρ. We obtain the desired entropy estimate as a limit $\delta \to 0$ of (4.12).

It remains to show that $\rho(0)=\rho_0$ in $L^2(\Omega)^N$. For this, let $\psi \in H^1(\Omega)^N$ and, for $\tau \in (0,T)$, define

$$\phi_\tau(t,\cdot) := \begin{cases} \left(1-\frac{t}{\tau}\right)\psi(\cdot) & \text{in } \Omega\times[0,\tau], \\ 0 & \text{in } \Omega\times(\tau,+\infty). \end{cases}$$

We easily see that $\phi_\tau \to 0$ in $L^2(0,T;H^1(\Omega)^N)$ as $\tau \to 0$. Then, from equation (4.11) tested with ϕ_τ, we get that, for all $\psi \in H^1(\Omega)^N$,

$$\int_\Omega \left(\frac{1}{\tau}\int_0^\tau \rho dt - \rho_0\right)\psi dx \to 0 \qquad \text{as } \tau \to 0.$$

Finally, the continuity implies that $\lim_{\tau\to 0} \frac{1}{\tau}\int_0^\tau \rho dt = \rho(0)$, which entails $\rho(0)=\rho_0$. □

Remark 4.2.9. *Using the last part of the proof of Proposition 4.2.8, we can easily show that any solution ρ of (4.4) satisfies $\rho(0)=\rho_0$. Therefore, the proof of Lemma 1 is a straightforward application of the integration by parts formula and of the embedding (4.3).*

Corollary 4.2.10. *Let ρ be given by Proposition 4.2.6. Then ρ is a solution of (4.2).*

Proof. Thanks to Proposition 4.2.8, we know that ρ possesses enough regularity such that we can integrate in (4.11) w.r.t. t, which yields (4.2) for all $\phi \in H^1(Q_T)^N$ with $\phi(T)=0$. Using a density argument yields the assertion. □

The proof of Propositon 4.2.3 is now straightforward.

Proof of Propositon 4.2.3. We only have to collect the previous results to obtain the proposition using a diagonal sequence argument. □

4.3 Applications and numerical tests

In this section, we apply the general setting of section 4.1 and numerically test the space–time Galerkin method of section 4.2 by considering four problems: the (linear) heat equation (section 4.3.1), the porous medium equation (section 4.3.2), the Fisher-KPP equation (section 4.3.3), and the Maxwell-Stefan system (section 4.3.4); in the latter case, the discussion on the general setting is postponed to section 4.4. We remark that we apply this nonlinear setting to the linear heat equation for validation purposes and, in particular, in order to stress its unconditional stability on a simple test problem.

In all cases, we consider the entropy density $s : \mathcal{D} \to [0, +\infty)$ defined by

$$s(\rho) = \sum_{j=1}^{N} \rho_j \log \rho_j + \left(1 - \sum_{j=1}^{N} \rho_j\right) \log \left(1 - \sum_{j=1}^{N} \rho_j\right) + \log(N+1), \tag{4.14}$$

where $\mathcal{D} := \left\{\rho \in (0,1)^N : \sum_{i=1}^N \rho_i < 1\right\}$. We have

$$(s'(\rho))_\ell = \log \frac{\rho_\ell}{1 - \sum_{j=1}^N \rho_j} \quad \text{and} \quad (s''(\rho))_{k\ell} = \frac{\delta_{k\ell}}{\rho_\ell} + \frac{1}{1 - \sum_{j=1}^N \rho_j}.$$

Then $s \in C^2(\mathcal{D}, [0,\infty)) \cap C^0(\overline{\mathcal{D}})$ and is convex. Moreover, $u : \mathbb{R}^N \to \mathcal{D}$ defined as

$$u_\ell(w) = \frac{e^{w_\ell}}{1 + \sum_{i=1}^N e^{w_i}} \quad \text{for } \ell = 1, \dots, N$$

is in $C^1(\mathbb{R}^N, \mathcal{D})$, and is the inverse of s'. Thus, the preamble of assumption (H2) is satisfied.

In the numerical experiments below, we use continuous space–time finite element discretization spaces. On the space–time cylinder $Q_T = \Omega \times (0,T)$, with Ω bounded interval ($n = 1$) or Lipschitz polytope ($n > 1$), we consider families of shape-regular simplicial or Cartesian meshes $\{\mathcal{T}_h\}_{h>0}$. The parameter h denotes the mesh granularity, namely $\mathcal{T}_h = \{K_i, i = 1, \dots, N_h\}$, $h_K := \mathrm{diam}(K)$, and $h := \max_{K \in \mathcal{T}_h} h_K$.

As discretization spaces, we choose $\{\boldsymbol{V}_h\}_{h>0} = \{\boldsymbol{V}_h^p,\ p \in \mathbb{N}\}_{h>0}$, with

$$\boldsymbol{V}_h^p = \left\{v \in C^0(\overline{Q_T})^N :\ v_{|_K} \in \mathcal{P}_p(K)^N \quad \forall K \in \mathcal{T}_h\right\}, \tag{4.15}$$

where $\mathcal{P}_p(K)$ denotes the space of polynomial functions on K of degree at most p, if K is a simplex, or of degree at most p in each variable, if K is a cuboid. Therefore, the approximability assumption (H3) in the first part of section 4.2 is satisfied.

Defining $B : \mathbb{R}^N \to \mathbb{R}^{N\times N}$ as

$$B(w) = A(u(w))u'(w),$$

the space–time Galerkin method (4.6) can be rewritten more explicitly in terms of the entropy variable unknown as follows:

Find $w_h^\epsilon \in \boldsymbol{V}_h^p$ such that

$$\epsilon(\phi, w_h^\epsilon)_{H^1(Q_T)} + \int_\Omega \phi(T) \cdot u(w_h^\epsilon(T))dx - \int_\Omega \phi(0) \cdot \rho_0 dx - \int_{Q_T} \partial_t \phi \cdot u(w_h^\epsilon)dxdt$$
$$+ \sum_{i,j=1}^N \int_{Q_T} \nabla\phi_i \cdot B_{ij}(w_h^\epsilon)\nabla(w_h^\epsilon)_j dxdt = \int_{Q_T} \phi \cdot f(u(w_h^\epsilon))dxdt$$
$$\text{for all } \phi \in \boldsymbol{V}_h^p. \tag{4.16}$$

Throughout this section, we measure the absolute numerical error defined by $\|\rho - u(w_h^\epsilon)\|_{L^2(Q_T)}$.

4.3.1 Heat equation

We apply our general approach to the linear heat equation:

$$\begin{cases} \partial_t \rho = \Delta \rho & \text{in } \Omega,\ t > 0, \\ \partial_\nu \rho = 0 & \text{on } \partial\Omega,\ t > 0, \\ \rho(0) = \rho_0 & \text{in } \Omega. \end{cases}$$

This corresponds to problem (4.1) with $N = 1$, $A \equiv 1$, and $f \equiv 0$. Furthermore, $\mathcal{D} = (0, 1)$ and the entropy density $s : \mathcal{D} \to [0, +\infty)$ is given by

$$s(\rho) = \rho \log \rho + (1 - \rho) \log(1 - \rho) + \log(2),$$

and thus $s'(\rho) = \log \frac{\rho}{1-\rho}$, and $s''(\rho) = \frac{1}{\rho(1-\rho)}$.

For this choice of $A(\rho)$ and $f(\rho)$, assumption (H1) is obviously satisfied, and assumptions (H2a) and (H2b) are fulfilled with $\gamma = 4$ and $C_f = 0$.

For the numerical tests, we take $\Omega = (0, 1)^2$ and $\rho_0(x) = 0.5 \cos(\pi x_1) \cos(\pi x_2) + 0.5$, so that the problem has the analytical solution given by

$$\rho(t, x) = 0.5 \exp(-2\pi^2 t/\tau) \cos(\pi x_1) \cos(\pi x_2) + 0.5,$$

where we use $\tau = 7$ to rescale the time. The solution is shifted and scaled in order to avoid the singularities of s' at 0 and 1. Without this rescaling, the system matrix is highly ill-conditioned, which prohibits optimal convergence rates. We solve (4.16), setting $\epsilon = 0$ and solving the nonlinearity by Newton's method. We use unstructured space–time simplicial meshes, an example is shown in Figure 4.1. The Newton method converges in 6 steps, for all considered values of h and p. We measure the L^2 error on the whole space–time domain. In Figure 4.2, the convergence rates of the h- and the p-version of the method are shown. We observe optimal rates, exponential in p and of order $p + 1$ in h. In the case of $p = 4$, we observe a preasymptotic region for very large mesh sizes; the exact rates are shown in Table 4.1.

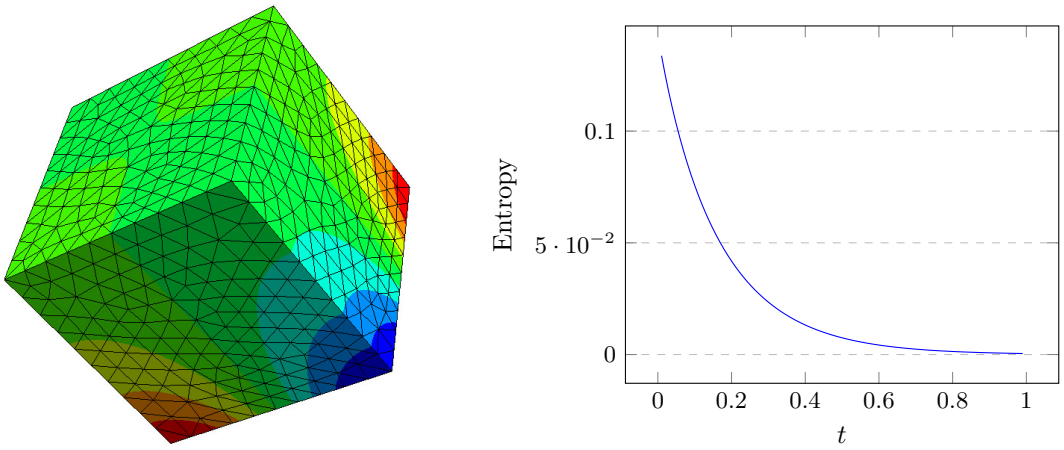

Figure 4.1: Plot of the analytic solution on an unstructured space–time mesh with $h = 0.1$ (left) and its entropy (right).

4.3.2 The porous medium equation

Let $m > 1$. The porous medium equation is given by

$$\begin{cases} \partial_t \rho = \Delta \rho^m & \text{in } \Omega\ t > 0, \\ \partial_\nu (\rho^m) = 0 & \text{on } \partial\Omega,\ t > 0, \\ \rho(0) = \rho_0 & \text{in } \Omega. \end{cases}$$

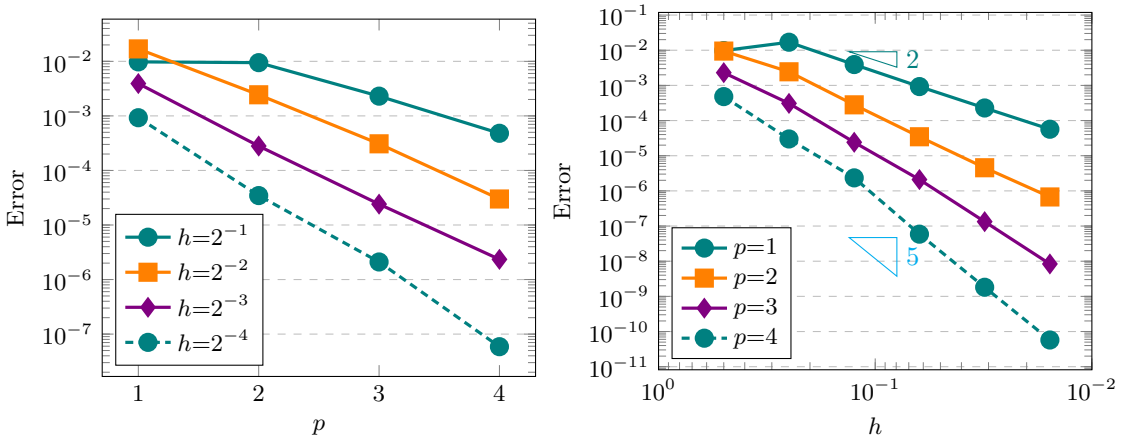

Figure 4.2: Convergence rates for the space–time Galerkin approximation towards the exact solution of the heat equation, in polynomial degree p (left), and mesh size h (right).

$p=3$			$p=4$		
h	error	rate	h	error	rate
2^{-1}	2.3×10^{-3}	0	2^{-1}	4.8×10^{-4}	0
2^{-2}	3.1×10^{-4}	2.9	2^{-2}	3.0×10^{-5}	4.0
2^{-3}	2.4×10^{-5}	3.7	2^{-3}	2.3×10^{-6}	3.7
2^{-4}	2.1×10^{-6}	3.5	2^{-4}	5.9×10^{-8}	5.3
2^{-5}	1.3×10^{-7}	4.0	2^{-5}	1.8×10^{-9}	5.0
2^{-6}	8.4×10^{-9}	4.0	2^{-6}	5.7×10^{-11}	5.0

Table 4.1: Numerical results for the heat equation.

We can write the porous medium equation in the form of (4.1) for N=1, $A(\rho) = m\rho^{m-1}$, and $f \equiv 0$. The entropy density is the same as for the heat equation.

Proposition 4.3.1. *Assumptions (H1) and (H2) are satisfied for $m \in (1, 2]$.*

Proof. For $\mathcal{D} = (0, 1)$ and $m > 1$, $A(\rho) = m\rho^{m-1}$ is in $C^0(\overline{\mathcal{D}})$, thus (H1) is stisfied. As (H2b) is obvious, we only neen to prove that (H2a) is satisfied, namely that $s''(\rho)A(\rho) \geq \gamma$ for some $\gamma > 0$ and all $\rho \in \mathcal{D}$. Thus let $\rho \in (0, 1) = \mathcal{D}$. Then, whenever $m \in (1, 2]$,

$$s''(\rho)A(\rho) = \frac{m\rho^{m-1}}{\rho(1-\rho)} = \frac{m}{\rho^{2-m}(1-\rho)} \geq m =: \gamma.$$

□

We test the space–time Galerkin method for this problem with initial conditions and Neumann boundary conditions chosen such that

$$\rho(x, t) = \left[\frac{(m-1)(x-\alpha)^2}{2m(m+1)(\beta-t)}\right]^{\frac{1}{m-1}}$$

is the exact solution, with α and β real parameters, on $\Omega = (0, 1)$. We consider the case $m = 2$, $\alpha = 5$, $\beta = 5$, $\epsilon = 0$ on unstructured simplicial space–time meshes.

In Figure 4.3, we show the convergence rates of the scheme. Regardless of the nonlinearity, we match the convergence rates of the heat equation, i.e. exponential in p and of order $p + 1$ in h.

In contrast to the heat equation, the power law in the porous medium equation introduces a finite propagation speed of the solution. This is best observed by the interesting behavior of

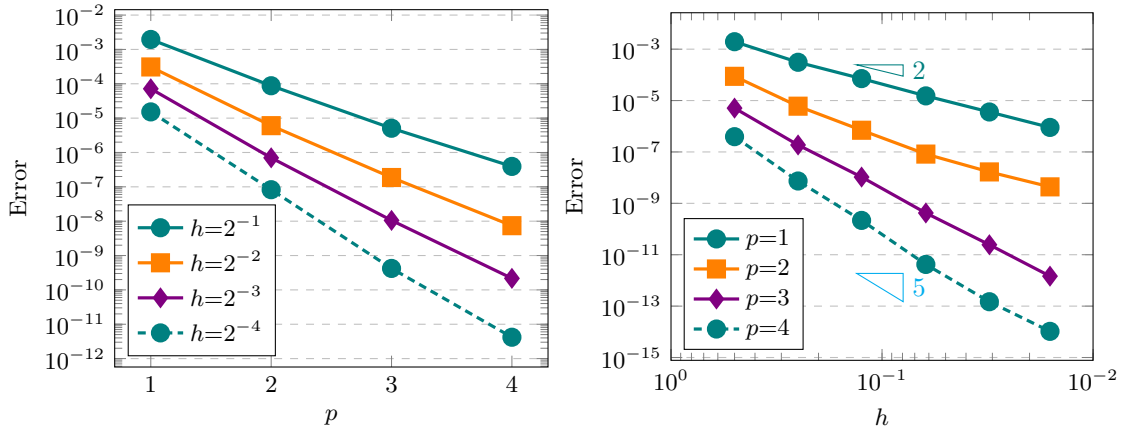

Figure 4.3: Convergence rates towards the exact solution of the porous medium equation, in polynomial degree p (left), and mesh size h (right).

certain initial conditions that induce a waiting time. That is, the solution keeps a fixed support until the waiting time is reached. On $\Omega = (0, \pi)$, the initial condition given by

$$\rho_0(x) = \begin{cases} \sin^{2/(m-1)}(x) & \text{if } 0 \leq x \leq \pi, \\ 0 & \text{otherwise,} \end{cases}$$

produces this behavior. It is shown in [87] that the corresponding solution has a waiting time of $t^* = \frac{m-1}{2m(m+1)}$. As we choose $m = 2$, here $t^* = 0.08\dot{3}$. We choose $u_0 = 10^{-16}$ for $0 \notin [0, \pi]$ to avoid ill-conditioning. Furthermore, to ensure convergence of the Newton method used as a solver, we had to choose $\epsilon = 10^{-8}$, making use of the regularization term. We solve on a Cartesian space–time mesh until final time $T = 0.2$, with spatial mesh size $h_s = 0.05$, and temporal mesh size $h_t = h_s/2$, and fix $p = 5$. The results are shown in Figure 4.4. Looking at snapshots of the numerical solution we can observe that it keeps a compact support set. In Figure 4.4, on the right, we plot the value of the solution on the left interface against time, marking the expected waiting time t^* with the vertical line.

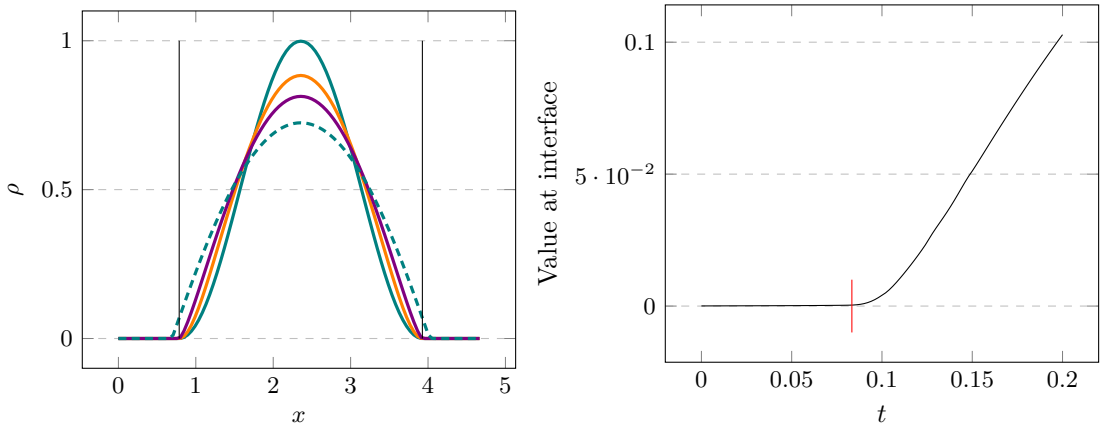

Figure 4.4: Snapshots of the solution of the porous medium equation emitting a waiting time, at different times (left) and the value at the left interface (right).

4.3.3 The Fisher-KPP equation

We consider the Fisher-KPP equation

$$\begin{cases} \partial_t \rho = A\Delta\rho + \rho(1-\rho) & \text{in } \Omega,\ t>0, \\ A\partial_\nu \rho = 0 & \text{on } \partial\Omega,\ t>0, \\ \rho(0) = \rho_0 & \text{in } \Omega, \end{cases}$$

with $A > 0$ now constant. This agrees with formulation (4.1), with $N = 1$, $A(\rho) = A$, and $f(\rho) = \rho(1-\rho)$. We set again $\mathcal{D} := (0,1)$. Assumptions (H1) and (H2a) are clearly satisfied. Choosing an entropy density such that assumption (H2b) is satisfied with $C_f = 0$ allows for the right-hand side of the entropy estimate (4.7) to be independent of time. Motivated by this, we now investigate the rescaled entropy density $s : \mathcal{D} \to (0, +\infty)$ given by

$$s(\rho) = \rho \log \rho + (m-\rho)\log(m-\rho), \tag{4.17}$$

with m to be chosen. Note that $f(\rho) > 0$ for $\rho \in (0,1)$, and $n/\rho - 1 > 1$ if and only if $\rho < m/2$. Thus,

$$f(\rho)s'(\rho) = \rho(1-\rho)\log\frac{\rho}{m-\rho} = -\rho(1-\rho)\log\left(\frac{m}{\rho}-1\right) \le 0$$

for all for $\rho \in (0,1)$ if and only if $m \ge 2$. We choose $m = 2$ so that the hypothesis (H2b) is fulfilled with $C_f = 0$.

We start again by investigate convergence towards a smooth solution. We choose $\Omega = (0,1)$, and initial conditions and Neumann boundary conditions such that

$$\rho(x,t) = \frac{1}{\left[1 + \exp(-\frac{5}{6}t + \frac{1}{\sqrt{6}}x)\right]^2}$$

is the exact solution for $A = 1$. We set $\epsilon = 10^{-16}$ and solve on unstructured simplicial space–time meshes. The results are presented in Figure 4.5. We observe again optimal convergence rates in both p and h, namely exponential in p and of order $p+1$ in h.

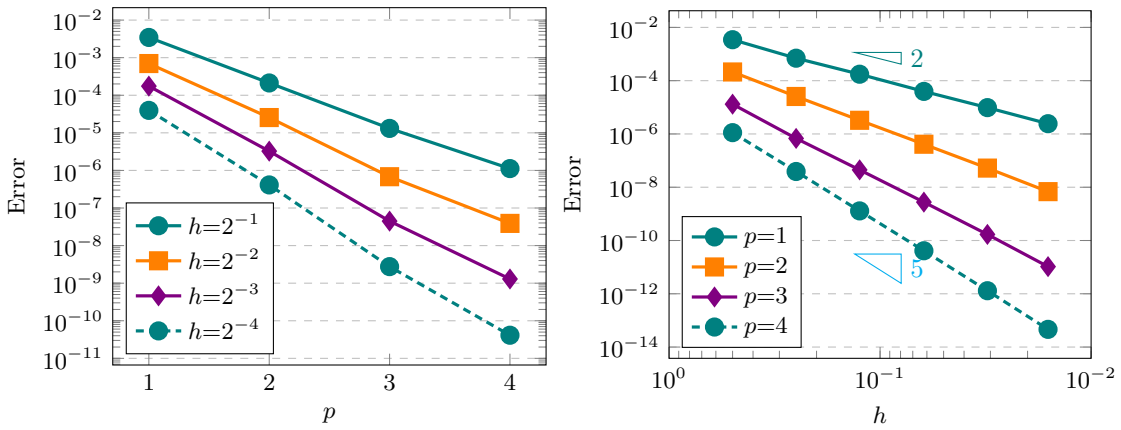

Figure 4.5: Convergence rates in polynomial degree p (left) and mesh size h for the exact solution of the Fisher-KPP equation.

Next, we aim to reproduce the experiments presented in [11], considering an initial condition with a jump, given by $\rho_0(x) = 1$ if $0 < x < 1/2$ and 0 elsewhere, with diffusion coefficient $A = 10^{-4}$. We solve using $p = 3$ on a Cartesian mesh with $h_s = 0.025$, $h_t = 0.4$ up to $T = 8$. Once again, we choose $\epsilon = 10^{-8}$ to avoid ill-conditioning in the solver. Snapshots of the numerical

solution are taken every $1.\dot{3}$ seconds, the results are shown in Figure 4.6 on the left. In Figure 4.6 on the right, we consider different choices for the entropy up to $T = 15$. Note that at the point in time the solution has already converged to $\rho \equiv 1$. The choice for the entropy density in [11] was $\rho \log(\rho) - \rho + 1$. We compare this choice to the entropy in (4.17) for different values of m in Figure 4.6. For the choice of $m = 2$, we recover a similar behavior of the entropy, namely, a region with slow decay followed by an exponential decay. As the solution converges to 1 it can easily be seen that for $m > 2$ the entropy does not convergence to zero exponentially, as exemplified by the choice of $m = 2.1$ in the figure.

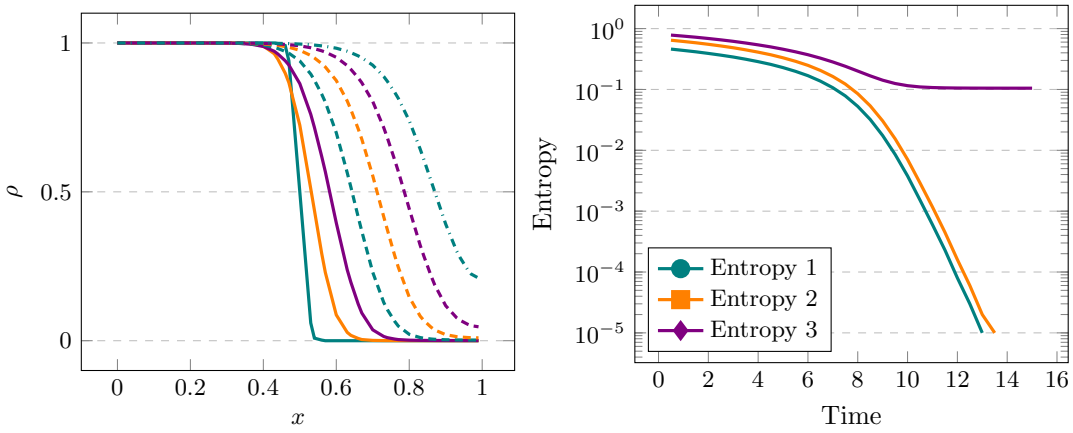

Figure 4.6: Snapshots of the numerical solution for the Fisher-KPP (left) and different choices of the entropy (right). The choices are as follows: Entropy 1 is the one used in [11], Entropy 2 is given by (4.17) with $m = 2$, and Entropy 3 is (4.17) with $m = 2.1$.

4.3.4 The Maxwell-Stefan system

The Maxwell-Stefan system for three-component gas diffusion ($N = 2$) can be written as

$$\begin{cases} \partial_t \rho_i = \nabla \cdot \left(\sum_{j=1}^{2} A_{ij}(\rho_1, \rho_2) \nabla \rho_j \right) & \text{in } \Omega,\ t > 0, \\ \sum_{j=1}^{2} A_{ij}(\rho_1, \rho_2) \partial_\nu \rho_j = 0 & \text{on } \partial\Omega,\ t > 0, \\ \rho_i(0) = (\rho_0)_i & \text{in } \Omega \end{cases}$$

for $i = 1, 2$, with

$$A(\rho_1, \rho_2) = \frac{1}{\delta(\rho_1, \rho_2)} \begin{pmatrix} d_1 + (d_3 - d_1)\rho_1 & (d_3 - d_2)\rho_1 \\ (d_3 - d_1)\rho_2 & d_2 + (d_3 - d_2)\rho_2 \end{pmatrix} \tag{4.18}$$

and

$$\delta(\rho_1, \rho_2) = d_1 d_2 (1 - \rho_1 - \rho_2) + d_2 d_3 \rho_1 + d_3 d_1 \rho_2.$$

The unknowns ρ_1 and ρ_2 represent the concentrations of the first two gases ($\rho_3 = 1 - (\rho_1 + \rho_2)$); the parameters d_1, d_2, and d_3 are the diffusion coefficients of the three gases.

In section 4.4, we derive this form of the Maxwell-Stefan system, prove that it fits our framework, and discuss the case $N > 2$.

In [13, Sec. 2] numerical results were presented for the three component gas diffusion experiment originally performed by Duncan and Toor in [29]. The setting is the following. Consider two bulbs of size 77.99 cm^3 and 78.63 cm^3, respectively, which are connected by a capillary tube of length 85.9 mm and diameter 2.08 mm, with a valve in the middle. We consider the Maxwell-Stefan equations with $N = 2$, corresponding to the gas mixture composed of hydrogen (ρ_1), nitrogen (ρ_2),

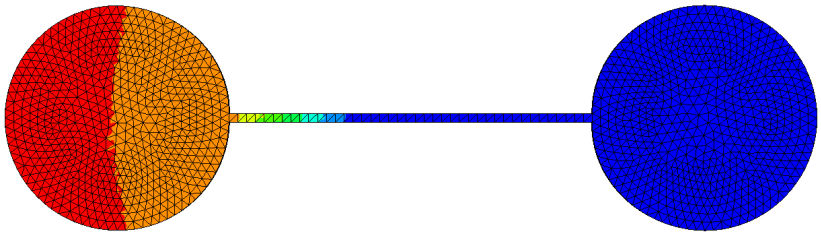

Figure 4.7: The mesh used for the Duncan-Toor example, depicting the Nitrogen content after about one hour.

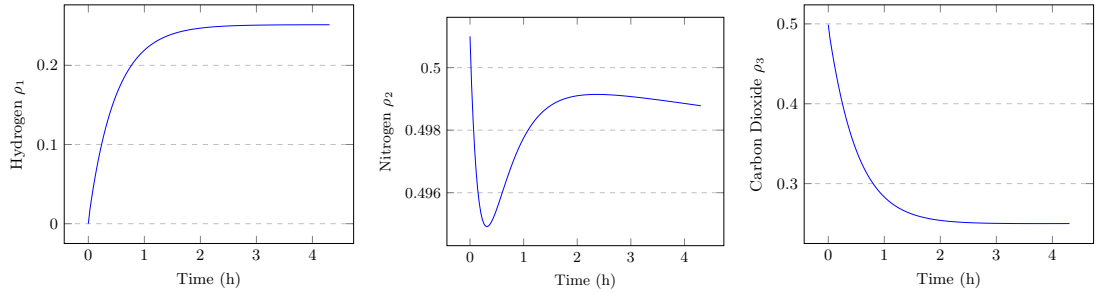

Figure 4.8: Comparison of the mole fractions in the left side of the device.

and carbon dioxide (ρ_3). We consider the following initial gas mixture in the left- and right-hand side of the device.

$$\begin{aligned} \text{Left:} \quad & (\rho_0)_1 = 0.000, \quad (\rho_0)_2 = 0.501, \quad (\rho_0)_3 = 0.499, \\ \text{Right:} \quad & (\rho_0)_1 = 0.501, \quad (\rho_0)_2 = 0.499, \quad (\rho_0)_3 = 0.000. \end{aligned}$$

For these gases, the diffusion coefficients are

$$d_1 = 83.3^{-1}, \quad d_2 = 68.0^{-1}, \quad d_3 = 16.8^{-1}.$$

In Figure 4.7, the computational domain is shown. We choose the spatial mesh size $h_s = 2.08$, equal to the diameter of the tube. The size of the Cartesian product mesh in time is chosen as $h_s/2$. We solve iteratively on these slabs, restarting the computations with the previous solution as initial condition. We fix $p = 1$.

The results are shown in Figure 4.8. We recover the same behavior shown in [13]. Both hydrogen and carbon dioxide converge monotonically to the expected equilibrium. Nitrogen shows the peculiar behavior known from the experiment. Note that the values in [13] differ from the ones found in our experiment, this is most likely due their simplification of the computational domain, using a symmetry argument.

4.4 The Maxwell-Stefan system revisited

In this section, we derive the formulation of the Maxwell-Stefan system as that used in section 4.3.4, and show that it fits into the general framework of section 4.1 (section 4.4.1). For the case $N > 2$,

in which an explicit representation of the currents may not be easily derived, we introduce and analyze an alternative space–time Galerkin method, which is based on a formulation that is implicit for the currents (section 4.4.2).

Let $\rho_0 \in L^\infty(\Omega)^{N+1}$ such that $\rho_0 \geq 0$ and $\sum_{i=1}^{N+1}(\rho_0)_i = 1$. The Maxwell-Stefan equations are given by the continuity equations

$$\begin{cases} \partial_t \rho_i + \nabla \cdot J_i = 0 & \text{in } (0,T)\times\Omega, \\ \nu \cdot J_i = 0 & \text{on } (0,T)\times\partial\Omega, \\ \rho_i(0) = (\rho_0)_i & \text{in } \Omega \end{cases} \tag{4.19}$$

for $i = 1, \ldots, N+1$, where the currents J_i are implicitly given by

$$\nabla \rho_i = \sum_{j=1}^{N+1} \frac{\rho_i J_j - \rho_j J_i}{D_{ij}} \tag{4.20}$$

for some $D_{ij} = D_{ji} > 0$.

4.4.1 Explicit formula for the currents

In this section, we establish an explicit representation of the currents, which allows us to derive the formulation of the Maxwell-Stefan system in the concentration variable unknowns. We follow [12] (see also [62]).

Let $M_{ij}(\rho) := D_{ij}^{-1}\rho_i - \delta_{ij}\sum_{k=1}^{N+1} D_{ik}^{-1}\rho_k$, $i,j = 1,\ldots,N+1$. Thus,

$$\nabla \rho_i = \sum_{j=1}^{N+1} M_{ij}(\rho) J_j.$$

Using $\rho_i \geq 0$ and $D_{ij} = D_{ji} > 0$, it is easy to see that $M(\rho)$ is quasi-positive ($M_{ij}(\rho) \geq 0$ for $i \neq j$). Moreover, provided that $\rho_i > 0$ for all $1 \leq i \leq N+1$, $M(\rho)$ is irreducible. Direct calculations show that

$$\operatorname{Ker} M(\rho) \supseteq \operatorname{span}\{\rho\} \qquad \text{and} \qquad \operatorname{Im} M(\rho) \subseteq \left\{ v : \sum_{i=1}^{N+1} v_i = 0 \right\}.$$

Moreover, $R^{-1}M(\rho)R$, with $R = \operatorname{diag}(\rho_1^{1/2}, \ldots, \rho_{N+1}^{1/2})$, is symmetric, thus all the eigenvalues of $M(\rho)$ are real. By the Perron-Frobenius theory for quasi-positive, irreducible matrices, one deduces that the eigenvalue zero has multiplicity one (we refer to [12] or [62] for details). We deduce

$$\operatorname{Ker} M(\rho) = \operatorname{span}\{\rho\} \qquad \text{and} \qquad \operatorname{Im} M(\rho) = \left\{ v : \sum_{i=1}^{N+1} v_i = 0 \right\}. \tag{4.21}$$

As $M(\rho)$ is not invertible, we have to restrict ourselves to a subspace of all possible currents J in order to obtain an explicit formula for J. For this, we make the assumption that the total current

$$J_{\text{tot}} := \sum_{i=1}^{N+1} J_i$$

vanishes. Then by summing in (4.19) over all $i = 1, \ldots, N+1$, we see that

$$\rho_{\text{tot}} = \sum_{i=1}^{N+1} \rho_i$$

is constant in time, and hence $\rho_{\mathrm{tot}} = \sum_{i=1}^{N+1}(\rho_0)_i = 1$. Using this, we can rewrite the implicit formulation of the currents as

$$\nabla\rho_i = \frac{\rho_i\left(-\sum_{j=1}^N J_j\right) - \left(1-\sum_{j=1}^N \rho_j\right)J_i}{D_{i(N+1)}} + \sum_{j=1}^N \frac{\rho_i J_j - \rho_j J_i}{D_{ij}} \tag{4.22}$$

As before, we can define a matrix

$$\mathcal{M}_{ij}(\rho) := \frac{\rho_i}{D_{ij}} - \frac{\rho_i}{D_{i(N+1)}} - \delta_{ij}\left(\sum_{k=1}^N \frac{\rho_k}{D_{ik}} + \frac{1-\sum_{l=1}^N \rho_l}{D_{i(N+1)}}\right), \qquad i,j = 1,\dots,N. \tag{4.23}$$

From (4.21), the matrix $\mathcal{M}(\rho)$ has full rank, and hence it is invertible. We have

$$J_i = -\sum_{j=1}^N A_{ij}(\rho)\nabla\rho_j \qquad \text{with } A(\rho) := -\mathcal{M}(\rho)^{-1}.$$

Remark 4.4.1. *The matrix $\mathcal{M}(\rho)$ is actually independent from the diagonal elements D_{ii}.*

Proposition 4.4.2. *Let s be as in (4.14), and let $\mathcal{M}$ be given by (4.23). Then, the matrix-valued function $A(\rho) := -\mathcal{M}(\rho)^{-1}$ fulfills (H1) and (H2a).*

Proof. Let $A(\rho) = -\mathcal{M}^{-1}(\rho)$. The fact that $\mathcal{M}$ is smooth directly implies that A is smooth. Similarly as in the proof of [62, Lemma 3.2], one can show that

$$\sum_{i=1}^n \partial_i w \cdot A(u(w))s''(u(w))^{-1}\partial_i w \geq \gamma|\nabla u(w)|^2 \tag{4.24}$$

for some $\gamma > 0$ and all smooth w.

In order to prove (H2a), we have to show that

$$z \cdot s''(\rho)A(\rho)z \geq \gamma|z|^2 \qquad \text{for all } z \in \mathbb{R}^N,\ \rho \in \mathcal{D}.$$

Let $\rho \in \mathcal{D}$, $\boldsymbol{x}_0 \in \Omega$, and $z \in \mathbb{R}^N$. We define the following vector-valued function of $\boldsymbol{x}$:

$$w(\boldsymbol{x}) := s'(\rho) + s''(\rho)z(\boldsymbol{x}-\boldsymbol{x}_0)\cdot\hat{e}_1,$$

where $\hat{e}_1$ denotes the unit vector $(1,0,\dots,0) \in \mathbb{R}^n$. We have

$$\partial_i w(\boldsymbol{x}_0) = \delta_{i1}s''(\rho)z$$

and, for $u = (s')^{-1}$,

$$\partial_i u(w(\boldsymbol{x}_0)) = u'(w(\boldsymbol{x}_0))\partial_i w(\boldsymbol{x}_0) = u'(w(\boldsymbol{x}_0))\delta_{i1}s''(\rho)z = u'(w(\boldsymbol{x}_0))\delta_{i1}s''(u(w(\boldsymbol{x}_0)))z = \delta_{i1}z.$$

This, together with (4.24), implies that

$$\begin{aligned} z\cdot s''(\rho)A(\rho)z &= (s''(\rho)z)\cdot A(\rho)s''(\rho)^{-1}(s''(\rho)z) \\ &= \sum_{i=1}^n \partial_i w(\boldsymbol{x}_0)\cdot A(u(w(\boldsymbol{x}_0)))s''(u(w(\boldsymbol{x}_0)))^{-1}\partial_i w(\boldsymbol{x}_0) \\ &\geq \gamma|\nabla u(w(\boldsymbol{x}_0))|^2 = \gamma|z|^2, \end{aligned}$$

which proves the assertion. □

For $N = 1$, the matrix $\mathcal{M}(\rho)$ is actually a scalar, which is given by

$$\mathcal{M}(\rho) = -\frac{\rho_1}{D_{12}} - \frac{1-\rho_1}{D_{12}} = -\frac{1}{D_{12}}.$$

Hence, $J_1 = D_{12}\nabla\rho_1$. Therefore, in this case the Maxwell-Stefan system reduces to the heat equation.

For three species/gases ($N = 2$), we have

$$\begin{aligned}\mathcal{M}(\rho_1,\rho_2) &= \begin{pmatrix} \frac{\rho_1}{D_{11}} - \frac{\rho_1}{D_{13}} - \frac{\rho_1}{D_{11}} - \frac{\rho_2}{D_{12}} - \frac{1-\rho_1-\rho_2}{D_{13}} & \frac{\rho_1}{D_{12}} - \frac{\rho_1}{D_{13}} \\ \frac{\rho_2}{D_{21}} - \frac{\rho_2}{D_{23}} & \frac{\rho_2}{D_{22}} - \frac{\rho_2}{D_{23}} - \frac{\rho_1}{D_{21}} - \frac{\rho_2}{D_{22}} + \frac{1-\rho_1-\rho_2}{D_{23}} \end{pmatrix} \\ &= -\begin{pmatrix} \frac{1}{D_{13}} + (\frac{1}{D_{12}} - \frac{1}{D_{13}})\rho_2 & (\frac{1}{D_{13}} - \frac{1}{D_{12}})\rho_1 \\ (\frac{1}{D_{23}} - \frac{1}{D_{21}})\rho_2 & \frac{1}{D_{23}} + (\frac{1}{D_{21}} - \frac{1}{D_{23}})\rho_1 \end{pmatrix}.\end{aligned}$$

Let

$$d_1 := \frac{1}{D_{13}}, \qquad d_2 := \frac{1}{D_{23}}, \qquad d_3 := \frac{1}{D_{12}},$$

and recall that $D_{21} = D_{12}$. One can verify that

$$\delta(\rho_1,\rho_2) := \det \mathcal{M}(\rho_1,\rho_2) = d_1d_2(1-\rho_1-\rho_2) + d_2d_3\rho_1 + d_3d_1\rho_2 \neq 0.$$

Let $A(\rho)$ denote the inverse of $-\mathcal{M}(\rho)$. We can rewrite the Maxwell-Stefan equations as the system in section 4.3.4.

4.4.2 Implicit formulation for the currents

In subsection 4.4.1, we have seen that the Maxwell-Stefan system (4.19)-(4.20), can be written in the form (4.1), with $f = 0$ and $A(\rho)$ being given by the inverse of $-\mathcal{M}(\rho)$ for

$$\mathcal{M}_{ij}(\rho) := \frac{\rho_i}{D_{ij}} - \frac{\rho_i}{D_{i(N+1)}} - \delta_{ij}\left(\sum_{k=1}^{N}\frac{\rho_k}{D_{ik}} + \frac{1-\sum_{l=1}^{N}\rho_l}{D_{i(N+1)}}\right), \qquad i,j = 1,\dots,N.$$

Moreover, we have computed $A(\rho)$ explicitly for $N = 1$ and $N = 2$. However, for large N, it is more complicated to find the explicit formulation for $A(\rho)$. In any case we do not expect a simple formulation in these cases. Therefore, this section provides a space–time Galerkin scheme, which avoids the explicit computation of the inverse of $\mathcal{M}$.

Let $q, p \in \mathbb{N}$. We consider the following problem:

Find $w_h^\epsilon \in \boldsymbol{V}_h^p, J^\mu \in \boldsymbol{V}_h^q,\ \mu = 1,\dots,n,$ such that

$$\begin{aligned}0 = \epsilon(\phi^0, w_h^\epsilon)_{H^1(Q_T)} &+ \int_\Omega \phi^0(T)\cdot u(w_h^\epsilon(T))dx - \int_\Omega \phi^0(0)\cdot\rho_0 dx - \int_{Q_T}\partial_t\phi^0\cdot u(w_h^\epsilon)dxdt \\ &- \sum_{\mu=1}^{n}\left(\int_{Q_T}\partial_{x_\mu}\phi^0\cdot J^\mu dxdt + \int_{Q_T}\phi^\mu\cdot\big(\partial_{x_\mu}w_h^\epsilon - s''(u(w_h))\mathcal{M}(u(w_h^\epsilon))J^\mu\big)dxdt\right) \\ &\qquad \forall\phi^0 \in \boldsymbol{V}_h^p,\ \phi^\mu \in \boldsymbol{V}_h^q,\ \mu = 1,\dots,n. \quad (4.25)\end{aligned}$$

Proposition 4.4.3. *Assume that $\rho_0 : \Omega \to \overline{\mathcal{D}}$ is measurable. Then there exists a solution $w_h^\epsilon \in \boldsymbol{V}_h^p, J^\mu \in \boldsymbol{V}_h^q,\ \mu = 1,\dots,n$ of the method (4.25).*

For the proof of Proposition 4.4.3, we need the following lemma.

Lemma 4.4.4. *If $w_h^\epsilon \in \boldsymbol{V}_h^p, J^\mu \in \boldsymbol{V}_h^q,\ \mu = 1,\dots,n,$ solves (4.25), then*

$$\epsilon\|w_h^\epsilon\|^2_{H^1(Q_T)} + \int_\Omega s(u(w_h^\epsilon(T)))dx + \gamma\sum_{\mu=1}^{n}\int_{Q_T}|\mathcal{M}(u(w_h^\epsilon))J^\mu|^2dxdt \le \int_\Omega s(\rho_0)dx.$$

Proof. We can use $\phi^0 = w_h^\epsilon$ and $\phi^\mu = 0$ for $\mu = 1, \dots, n$ as test functions and, similarly to the proof of Proposition 4.2.2, we obtain that

$$\epsilon \|w_h^\epsilon\|^2_{H^1(Q_T)} + \int_\Omega s(u(w_h^\epsilon(T)))dx - \sum_{\mu=1}^n \int_{Q_T} J^\mu \cdot \partial_{x_\mu} w_h^\epsilon dxdt \le \int_\Omega s(\rho_0)dx.$$

The next step is to use the test functions $\phi^0 = 0$ and $\phi^\mu = J^\mu$ for $\mu = 1, \dots, n$ to obtain

$$\sum_{\mu=1}^n \int_{Q_T} J^\mu \cdot \partial_{x_\mu} w_h^\epsilon dxdt = \sum_{\mu=1}^n \int_{Q_T} J^\mu \cdot s''(u(w_h^\epsilon))\mathcal{M}(u(w_h^\epsilon))J^\mu dxdt.$$

According to assumption (H2a), we know that $s''(v)A(v)$ is positive semi-definite and satisfies

$$z \cdot s''(v)A(v)z \ge \gamma |z|^2 \quad \text{for all } z \in \mathbb{R}^N,\ v \in \mathcal{D}.$$

Choosing $v = u(w_h^\epsilon)$, $z := \mathcal{M}(u(w_h^\epsilon))J^\mu$, we see that

$$\gamma |\mathcal{M}(u(w_h^\epsilon))J^\mu|^2 \le J^\mu \cdot \mathcal{M}(v)s''(v)A(v)\mathcal{M}(v)J^\mu = -J^\mu \cdot \mathcal{M}(v)s''(v)J^\mu,$$

where in the last step we have used that $A(v)$ is the inverse of $-\mathcal{M}(v)$. Thus, we conclude that

$$\epsilon \|w_h^\epsilon\|^2_{H^1(Q_T)} + \int_\Omega s(u(w_h^\epsilon(T)))dx + \gamma \sum_{\mu=1}^n \int_{Q_T} |\mathcal{M}(u(w_h^\epsilon))J^\mu|^2 dxdt \le \int_\Omega s(\rho_0)dx.$$

□

Proof of Proposition 4.4.3. The idea of the proof is to proceed similarly to the proof of Proposition 4.2.2. We define the mapping

$$\Phi : \boldsymbol{V}_h^p \times (\boldsymbol{V}_h^q)^n \to \boldsymbol{V}_h^p \times (\boldsymbol{V}_h^q)^n,\ (v, I^1, \dots, I^n) \mapsto (w, J^1, \dots, J^n),$$

where w is (uniquely) defined via the equation

$$0 = \epsilon(\phi^0, w)_{H^1(Q_T)} + \int_\Omega \phi^0(T) \cdot u(v(T))dx - \int_\Omega \phi^0(0) \cdot \rho^0 dx - \int_{Q_T} \partial_t \phi^0 \cdot u(v)dxdt$$
$$- \sum_{\mu=1}^n \int_{Q_T} \partial_{x_\mu} \phi^0 \cdot I^\mu dxdt \quad \text{for all } \phi^0 \in \boldsymbol{V}_h^p,$$

and J^μ denotes the unique solution (see below for a justification) of

$$\int_{Q_T} \phi^\mu \cdot \partial_{x_\mu} v dxdt = \int_{Q_T} \phi^\mu \cdot s''(u(v))\mathcal{M}(u(v))J^\mu dxdt \qquad \text{for all } \phi^\mu \in \boldsymbol{V}_h^q. \tag{4.26}$$

Note that the mapping Φ is well-defined, as (4.26) admits a unique solution for given $v \in \boldsymbol{V}_h^p$ according to the Lemma of Lax-Milgram: we see that $\partial_{x_\mu} v \in L^2(Q_T)^N$ and the matrix $-s''(u(v))\mathcal{M}(u(v)) \in L^\infty(Q_T)^{N\times N}$ is positive definite, because for all $z \in \mathbb{R}^N$

$$\begin{aligned} z \cdot \big(- s''(u(v))\mathcal{M}(u(v))\big)z &= A(u(v))y \cdot s''(u(v))y \\ &= y \cdot s''(u(v))A(u(v))y \\ &\overset{\text{(H2a)}}{\ge} \gamma |y|^2 = \frac{\gamma}{\|A(u(v))\|^2} \|A(u(v))\|^2 |y|^2 \\ &\ge \frac{\gamma}{\|A(u(v))\|^2} |A(u(v))y|^2 = \frac{\gamma}{\|A(u(v))\|^2} |z|^2 \end{aligned}$$

for $y := A(u(v))^{-1}z = -\mathcal{M}(u(v))z$. Moreover, the mapping Φ is continuous since A and u are continuous. Then by the Leray-Schauder fixed-point theorem, we obtain that Φ admits a fixed-point if we can show that the set

$$\{(w, J^1, \dots J^n) \in \boldsymbol{V}_h \times (\boldsymbol{V}_h^q)^n : (w, J^1, \dots J^n) = \sigma\Phi(w, J^1, \dots J^n), \sigma \in [0,1]\}$$

is bounded. Let $(w, J^1, \dots, J^n) = \sigma\Phi(w, J^1, \dots, J^n)$ for $\sigma \in (0,1]$. Similarly to Lemma 4.4.4, we can prove the entropy estimate

$$\frac{\epsilon}{\sigma}\|w\|^2_{H^1(Q_T)} + \int_\Omega s(u(w(T)))dx + \frac{\gamma}{\sigma}\sum_{\mu=1}^n \int_{Q_T} |\mathcal{M}(u(w))J^\mu|^2 dxdt \le \int_\Omega s(\rho_0)dx.$$

Using that $\sigma \in (0,1]$ is bounded from above yields a uniform bound on w in $\boldsymbol{V}_h^q$ and on $\mathcal{M}(u(w))J^\mu$ in $L^2(Q_T)^N$. As $\boldsymbol{V}_h^q$ is finite dimensional, we directly obtain that $\|w\|_{L^\infty(Q_T)^N}$ is uniformly bounded. Thus,

$$\|J^\mu\|_{L^2(Q_T)^N} \le \|A(u(w))\|_{L^\infty(Q_T)^{N\times N}} \|\mathcal{M}(u(w))J^\mu\|_{L^2(Q_T)^N}$$

is also uniformly bounded. As all norms are equivalent on $\boldsymbol{V}_h^q$, this directly implies that J^μ is uniformly bounded in $\boldsymbol{V}_h^q$. Thus, the Leray-Schauder theorem is applicable and yields that Φ has a fixed-point, and therefore the scheme (4.25) admits a solution. □

Proposition 4.4.5. *Let $\rho_0 : \Omega \to \overline{\mathcal{D}}$ be measurable and $w_h^\epsilon \in \boldsymbol{V}_h^p, J_h^{\epsilon,\mu} \in \boldsymbol{V}_h^q,\ \mu = 1, \dots, n$, be a solution for of (4.25) for $\epsilon, h > 0$. Then there exist a solution ρ of (4.4) and sequences $h_i, \epsilon_i \to 0$, as $i \to \infty$, such that*

$$u(w_{h_i}^{\epsilon_i}) \to \rho \qquad \text{in } L^r(Q_T), \text{ as } i \to \infty$$

for all $r \in [1, \infty)$. Moreover, ρ satisfies the entropy estimate

$$\int_\Omega s(\rho(\tau))dx + \gamma \int_0^\tau \int_\Omega |\nabla\rho|^2 \, dxdt \le \int_\Omega s(\rho_0)dx \tag{4.27}$$

for all $\tau \in (0, T]$, where $|\Omega|$ is the volume of Ω.

Proof. The proof is analogue to the proof of Proposition 4.2.3. We only need to replace Proposition 4.2.4 by Lemma 4.4.6 below. □

Lemma 4.4.6 (Convergence of the scheme for fixed $\epsilon > 0$)**.** *Let $w_h \in \boldsymbol{V}_h^p, J_h^\mu \in \boldsymbol{V}_h^q,\ \mu = 1, \dots, n$ be a solution of (4.25), with fixed $\epsilon > 0$. Then there exists $\rho \in H^1(Q_T)^N$ with $\rho(t,x) \in \overline{\mathcal{D}}$ for a.e. $(t,x) \in Q_T$ and $s'(\rho) \in H^1(Q_T)^N$, and a sequence $h_\ell \to 0$ such that*

$$\rho_{h_\ell} := u(w_{h_\ell}) \to \rho \qquad \text{and} \qquad w_{h_\ell} \to s'(\rho)$$

strongly in $L^2(Q_T)$ and weakly in $H^1(Q_T)$. Moreover, ρ solves (4.5) and satisfies the entropy estimate (4.9) for $w = s'(\rho)$.

Proof. The fact that w_h is uniformly bounded in $H^1(Q_T)^N$ yields that there exists $w \in H^1(Q_T)^N$ and subsequence $h_\ell \to 0$ such that $w_{h_\ell} \rightharpoonup w$ in $H^1(Q_T)^N$, due to the Banach-Alaoglu theorem, and $w_{h_\ell} \to w$ in $L^2(Q_T)^N$ due to Rellich's theorem. As u is bounded, the dominated convergence theorem entails the convergence for $\rho_{h_\ell} \equiv u(w_{h_\ell})$ to $\rho := u(w)$ along another subsequence (which we do not relabel).

For the second part, we note that, due to the Banach-Alaoglu theorem and the boundedness of $\mathcal{M}(u(w_h))J_h^\mu$ in $L^2(Q_T)^N$, we know that there exist $\xi^\mu \in L^2(Q_T)^N$ such that, for a subsequence (not being relabeled),

$$\mathcal{M}(u(w_h))J_h^\mu \rightharpoonup \xi^\mu \qquad \text{weakly in } L^2(Q_T)^N.$$

In particular,

$$J_h^\mu = -A(u(w_h))\mathcal{M}(u(w_h))J^\mu \rightharpoonup -A(\rho)\xi^\mu =: J^\mu \qquad \text{weakly in } L^r(Q_T)^N$$

for every $r \in [1,2)$. Finally, for every $\phi^\mu \in H^1(Q_T)^N$, $j = 0,\dots,n$, there exist $\phi^\mu_{h_\ell} \in \boldsymbol{V}^p_{h_\ell} \cap \boldsymbol{V}^q_{h_\ell}$ such that $\phi^\mu_{h_\ell} \to \phi^\mu$ in $H^1(Q_T)^N$. Using $\phi^\mu_{h_\ell}$ as a test function in (4.25), in the limit $h_i \to 0$, we obtain

$$0 = \epsilon(\phi^0, w_h)_{H^1(Q_T)} + \int_\Omega \phi^0(T)\cdot u(w_h(T))dx - \int_\Omega \phi^0(0)\cdot\rho_0 dx - \int_{Q_T} \partial_t\phi^0 \cdot u(w_h)dxdt$$
$$- \sum_{\mu=1}^n \left(\int_{Q_T} \partial_{x_\mu}\phi^0 \cdot J^\mu dxdt + \int_{Q_T} \phi^\mu \cdot \left(\partial_{x_\mu} w_h - s''(u(w_h))\mathcal{M}(u(w_h))J^\mu\right)dxdt \right),$$

as each integral in (4.25) converges separately. In particular, by the fundamental lemma of calculus of variations, we see that $\partial_{x_\mu} w_h = s''(u(w_h))\mathcal{M}(u(w_h))J^\mu$ and equivalently

$$J^\mu = \mathcal{M}(u(w_h))^{-1} s''(u(w_h))^{-1}\partial_{x_\mu} w_h = -A(u(w_h))u'(w_h)\partial_{x_\mu} w_h = -A(u(w_h))\partial_{x_\mu} u(w_h),$$

which implies that ρ solves (4.5). Finally, the entropy inequality is a consequence of Fatou's lemma. □

4.4.3 Numerical Tests

We again turn to [13, Sec. 3] for numerical results we can compare our method to. This time, we consider a model for the lung. The computational domain resembles on branch of the tree structure found in the bottom of the lung. The domain, depicted in Figure 4.9, consists of the inflow, Γ_1, on top, the outflow, Γ_2, located on the bottom of the two branches, and the alveoli, Γ_3, located in the middle of each of the branches. The remaining boundary Γ_4 is a wall where nothing goes in or out. Opposed to the domain presented in the reference, we consider the branches of the lung to be symmetrical and perpendicular to each other. The paper does not mention the angle between the branches used there. Also the size of the alveoli is left unspecified in the paper. Here, we split the boundary of the branches into three equal parts, with the alveoli (Γ_3) in the middle. On $\Gamma_1, \Gamma_2, \Gamma_3$ we impose Dirichlet boundary conditions to model the gas exchange with the other parts of the lung. On the wall, Γ_4, we take homogeneous Neumann boundary conditions.

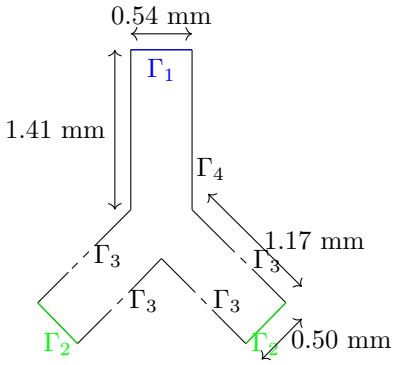

Figure 4.9: Computational domain for the lung model.

We make use of the implicit formulation (4.25) to find the numerical solution. To incorporate the Dirichlet boundary condition, we use Nitsche's method and add to (4.25) the following terms:

$$\sum_{\mu=1}^n \int_{(0,T)\times\Gamma_D} J^\mu \nu^\mu \cdot \phi^0 + \int_{(0,T)\times\Gamma_D} (u(w) - \rho_D)\cdot\phi^\mu\nu^\mu + \int_{(0,T)\times\Gamma_D} \eta {h_s}^{-1}(u(w)-\rho_D)\cdot\phi^0$$

for a parameter $\eta > 0$, h_s being the spatial mesh size, on the Dirichlet boundary Γ_D. In the examples below, we use $\eta = 1$. The first term comes from the integration by parts. The second and third terms are productive zeros that weakly enforce the Dirichlet boundary condition, and are chosen such that they agree with Nitsche's method for the heat equation in the degenerative case.

Diffusion of air

In the following example, compare [13, Sec. 3.4], we choose alveolar air as initial condition and as the Dirichlet data on the outflow and alveoli. On the inflow boundary we choose humidified air as Dirichlet data. See Table 4.2 for the gas components of the different types of air, and Table 4.3 for the diffusion coefficients.

	Humidified air	Alveolar air	Alveolar heliox
Nitrogen	0.7409	0.7490	0.0000
Oxygen	0.1967	0.1360	0.1360
Carbon dioxide	0.0004	0.0530	0.0530
Water	0.0620	0.0620	0.0620
Helium	0.0000	0.0000	0.7490

Table 4.2: Components of the different gas mixtures.

	Oxygen	Carbon dioxide	Water	Helium
Nitrogen	21.87	16.63	23.15	74.07
Oxygen		16.40	22.85	79.07
Carbon dioxide			16.02	63.45
Water				90.59

Table 4.3: Diffusion coefficients of the different gases.

Since there is no helium present we can reduce the number of species involved, setting $N = 3$. For the numerical calculations we choose spatial mesh size $h_s = 0.3$ and measure the value of the gas every 0.001 seconds. The discrete system is not ill-conditioned and we are able to choose $\epsilon = 0$. In Figure 4.10 we show the numerical results for Oxygen and Carbon dioxide as the other gases stay (almost) constant. Both converge to their equilibrium value. Comparing the results to [13], we can see that the equilibrium value slightly differs, which is likely due to the symmetry of the domain and size of the alveoli.

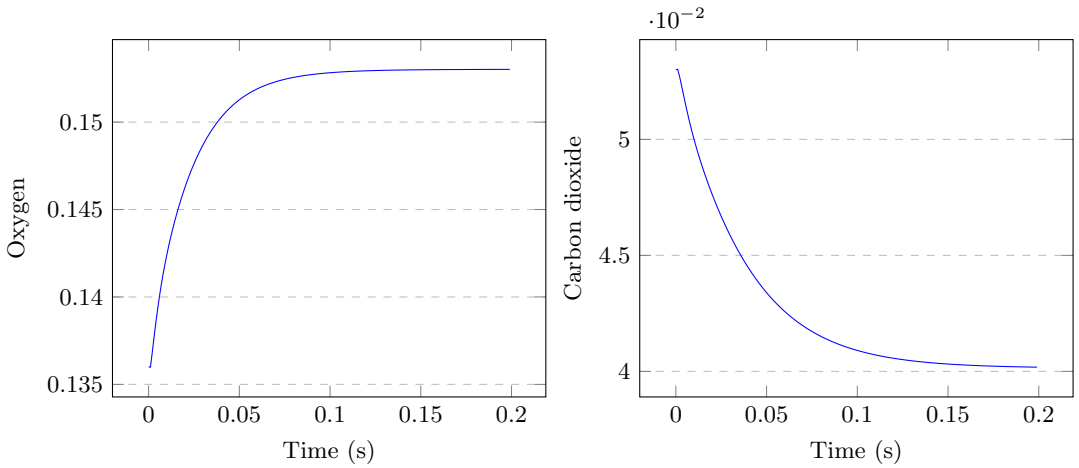

Figure 4.10: Numerical results of the mole fractions Oxygen and Carbon dioxide inside the lung for air mixture.

Diffusion of air/heliox

Next, we try to reproduce the results form [13, Sec. 3.5]. We consider alveolar heliox as initial condition. As the Dirichlet data on the outflow and alveoli, we also choose alveolar heliox, whereas we put humidified air on the inflow. The discrete system is very ill-conditioned due to the gas components taking zero values. In order for the solver to converge, we had to choose $\epsilon = 10^{-4}$. Furthermore, to avoid the singularity of the entropy density, we adjust the helium content in air and the nitrogen content in heliox to be 10^{-6}, subtracting the same amount of water, in order to keep them summing to one. Note that this is not unreasonable, for example, the correct amount of helium in air is about $5.3 \cdot 10^{-7}$. With these adjustments, the solver converges. The numerical results are shown in Figure 4.11. Both oxygen and carbon dioxide levels rise above the values in provided gas mixtures, before they start to decrease towards the equilibrium value. This is the expected behavior. However, the maximum values reached here are slightly lower than the ones found in [13]. This can be attributed to the perturbations of the zero concentrations and, as already seen, to the approximation of the geometry.

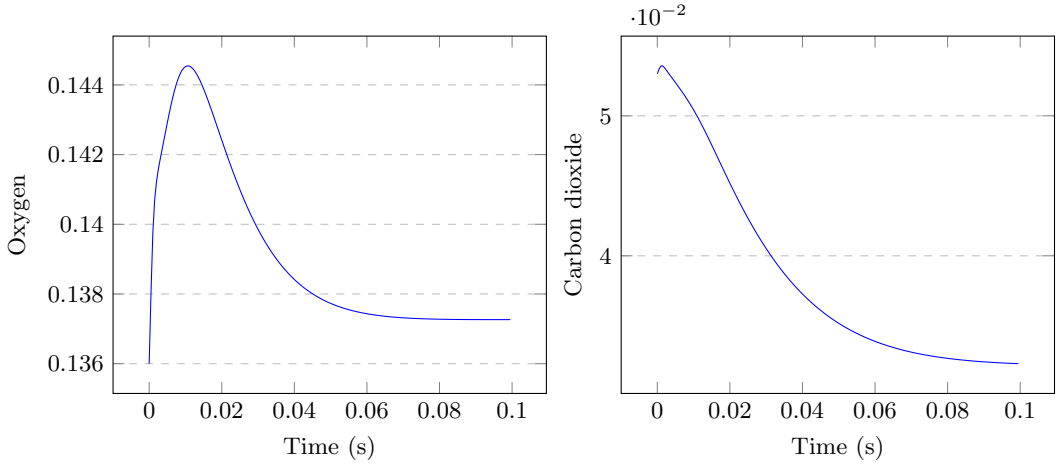

Figure 4.11: Numerical results of the mole fractions Oxygen and Carbon dioxide inside the lung for air/heliox mixture.

Chapter 5

Outlook and open questions

We have considered some applications of space–time approaches to the numerical approximation of evolutionary partial differential equations.

In Chapter 2, we have presented implementational aspects and numerical results for the Trefftz-DG method for the acoustic wave equation, originally presented in [81]. In particular, we have seen that in combination with the tent pitched meshes strategy it gives rise to a very effective algorithm. As we have seen in the numerical examples, for analytic solutions we get exponential convergence rates in the polynomial degree p. A significant and challenging extension, from a theoretical point of view, is the analysis of the approximation properties for increasing polynomial degrees, i.e. the p-convergence. Further possible developments include the extension to the case of electromagnetic waves (Maxwell's equations).

We extended the Trefftz-DG method to a variable coefficient in Chapter 3. We envisage that the quasi-Trefftz method can be extended to elastic and electromagnetic wave propagation in heterogeneous materials, and more generally to a wider class of hyperbolic or Friedrichs systems. For constant-coefficient examples of space–time Trefftz DG schemes for elastodynamics, electromagnetics and kinetic equations/transport models see [10], [30, 66] and [16], respectively. The proposed approach might be effective also for the approximation of PDEs whose nature changes in the computational domain, exemplified by the Euler–Tricomi equation ($\partial_x^2 u + x\partial_y^2 u = 0$), used for applications in transonic flows and plasma physics. The numerical analysis performed here is only a first step towards the establishment of a more comprehensive theory of quasi-Trefftz polynomial schemes. More work is needed to address refined approximation estimates in Sobolev norms (see Remark 3.3.7), the treatment of less regular solutions (e.g. with corner singularities), the proof of error bounds in mesh-independent norms, the analysis of dispersion and dissipation properties. The construction of non-polynomial quasi-Trefftz spaces could be relevant, for example, in order to efficiently approximate solutions that are localised in frequency.

In Chapter 4 we have presented a novel space–time method for cross-diffusion systems with several numerical examples for different PDEs. The method has high potential, allowing for space–time hp-refinement, ready to be extended to other equations and combined with more efficient solvers, with an open source code ready available. Physical consistent schemes for nonlinear PDEs often involve a large number Newton iterations. Our new space–time finite element method allows straightforward high order approximation while using only a minimal number of Newton iterations. The main trick is to solve the equation directly on a space–time cylinder in one step. Furthermore, contrary to the existing space–time FEMs for the heat equation, it does not require a correction term to achieve optimal convergence rates. Therefore, we hope to see more applications of the space–time approach for parabolic equations in the near future. In the numerical examples, we have observed optimal convergence rates, given that the solution stays away from the singularities of the entropy. Lifting this restriction could be the topic of future works. Also, more efficient numerical treatment of the space–time system is of interest.

References

[1] J. M. Alam, N. K.-R. Kevlahan, and O. V. Vasilyev, *Simultaneous space-time adaptive wavelet solution of nonlinear parabolic differential equations*, J. Comput. Phys., 214 (2006), pp. 829–857.

[2] R. Andreev, *Stability of sparse space-time finite element discretizations of linear parabolic evolution equations*, IMA J. Numer. Anal., 33 (2013), pp. 242–260.

[3] P. F. Antonietti, C. Marcati, I. Mazzieri, and A. Quarteroni, *High order discontinuous Galerkin methods on simplicial elements for the elastodynamics equation*, Numer. Algorithms, 71 (2016), pp. 181–206.

[4] T. Apel, *Anisotropic finite elements: local estimates and applications*, Advances in Numerical Mathematics, B. G. Teubner, Stuttgart, 1999.

[5] J. Argyris and D. Scharpf, *Finite elements in time and space*, Nuclear Engineering and Design, 10 (1969), pp. 456–464.

[6] A. K. Aziz and P. Monk, *Continuous finite elements in space and time for the heat equation*, Math. Comp., 52 (1989), pp. 255–274.

[7] W. Bangerth, M. Geiger, and R. Rannacher, *Adaptive Galerkin finite element methods for the wave equation*, Comput. Methods Appl. Math., 10 (2010), pp. 3–48.

[8] L. Banjai, E. H. Georgoulis, and O. Lijoka, *A Trefftz polynomial space-time discontinuous Galerkin method for the second order wave equation*, SIAM J. Numer. Anal., 55 (2017), pp. 63–86.

[9] P. Bansal, A. Moiola, I. Perugia, and C. Schwab, *Space–time discontinuous Galerkin approximation of acoustic waves with point singularities*, IMA J. Numer. Anal., (2020). draa088.

[10] H. Barucq, H. Calandra, J. Diaz, and E. Shishenina, *Space–time Trefftz-DG approximation for elasto-acoustics*, Appl. Anal., 99 (2020), pp. 747–760.

[11] F. Bonizzoni, M. Braukhoff, A. Jüngel, and I. Perugia, *A structure-preserving discontinuous Galerkin scheme for the Fisher-KPP equation*, Numer. Math., 146 (2020), pp. 119–157.

[12] D. Bothe, *On the Maxwell-Stefan approach to multicomponent diffusion*, in Parabolic problems, vol. 80 of Progr. Nonlinear Differential Equations Appl., Birkhäuser/Springer Basel AG, Basel, 2011, pp. 81–93.

[13] L. Boudin, D. Götz, and B. Grec, *Diffusion models of multicomponent mixtures in the lung*, ESAIM: Proceedings, 30 (2010), pp. 91–104.

[14] L. Boudin, B. Grec, and F. Salvarani, *A mathematical and numerical analysis of the Maxwell-Stefan diffusion equations*, Discrete Contin. Dyn. Syst. Ser. B, 17 (2012), pp. 1427–1440.

[15] M. Braukhoff, I. Perugia, and P. Stocker, *An entropy structure preserving space-time Galerkin method for cross-diffusion systems.* arXiv:2006.13069, 2020.

[16] C. Buet, B. Despres, and G. Morel, *Trefftz discontinuous Galerkin basis functions for a class of Friedrichs systems coming from linear transport*, Adv. Comput. Math., 46 (2020), pp. Paper No. 41, 27.

[17] J. J. Callahan, *Advanced Calculus: A Geometric View*, Springer-Verlag New York, 2010.

[18] C. Cancès, V. Ehrlacher, and L. Monasse, *Finite volumes for the Stefan-Maxwell cross-diffusion system.* arXiv:2007.09951, 2020.

[19] B. Carnes and G. F. Carey, *Local boundary value problems for the error in FE approximation of non-linear diffusion systems*, Internat. J. Numer. Methods Engrg., 73 (2008), pp. 665–684.

[20] X. Chen and A. Jüngel, *Analysis of an incompressible Navier-Stokes-Maxwell-Stefan system*, Comm. Math. Phys., 340 (2015), pp. 471–497.

[21] X. Chen and A. Jüngel, *When do cross-diffusion systems have an entropy structure?*, J. Diff. Eq., 278 (2021), pp. 60–72.

[22] R. Courant, *Variational methods for the solution of problems of equilibrium and vibrations*, Bull. Amer. Math. Soc., 49 (1943), pp. 1–23.

[23] E. S. Daus, A. Jüngel, and B. Q. Tang, *Exponential Time Decay of Solutions to Reaction-Cross-Diffusion Systems of Maxwell–Stefan Type*, Arch. Ration. Mech. Anal., 235 (2020), pp. 1059–1104.

[24] D. Devaud, *hp-approximation of linear parabolic evolution problems in $H^{1/2}$*, PhD thesis, ETH Zurich, Zurich, 2017.

[25] D. Devaud and C. Schwab, *Space-time hp-approximation of parabolic equations*, Calcolo, 55 (2018), pp. Paper No. 35, 23.

[26] K. Dieter-Kissling, H. Marschall, and D. Bothe, *Numerical method for coupled interfacial surfactant transport on dynamic surface meshes of general topology*, Comput. & Fluids, 109 (2015), pp. 168–184.

[27] W. Dörfler, S. Findeisen, and C. Wieners, *Space-time discontinuous Galerkin discretizations for linear first-order hyperbolic evolution systems*, Comput. Methods Appl. Math., 16 (2016), pp. 409–428.

[28] D. Drake, J. Gopalakrishnan, J. Schöberl, and C. Wintersteiger, *Convergence analysis of some tent-based schemes for linear hyperbolic systems*, 2021.

[29] J. B. Duncan and H. L. Toor, *An experimental study of three component gas diffusion*, AIChE Journal, 8 (1962), pp. 38–41.

[30] H. Egger, F. Kretzschmar, S. M. Schnepp, and T. Weiland, *A Space-Time Discontinuous Galerkin Trefftz Method for Time Dependent Maxwell's Equations*, SIAM J. Sci. Comput., 37 (2015), pp. B689–B711.

[31] J. Erickson, D. Guoy, J. Sullivan, and A. Üngör, *Building space-time meshes over arbitrary spatial domains*, Eng. Comput., 20 (2005), pp. 342–353.

[32] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems. I. A linear model problem*, SIAM J. Numer. Anal., 28 (1991), pp. 43–77.

[33] ——, *Adaptive finite element methods for parabolic problems. IV. Nonlinear problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1729–1749.

[34] J. Ernesti and C. Wieners, *Space-time discontinuous Petrov-Galerkin methods for linear wave equations in heterogeneous media*, Comput. Methods Appl. Math., 19 (2019), pp. 465–481.

[35] L. C. Evans, *Partial differential equations*, vol. 19 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, second ed., 2010.

[36] R. S. Falk and G. R. Richter, *Explicit finite element methods for symmetric hyperbolic equations*, SIAM J. Numer. Anal., 36 (1999), pp. 935–952.

[37] B. G. Galerkin, *Rods and Plates: Series in Some Questions of Elastic Equilibrium of Rods and Plates*, Engineers Bulletin (Vestnik Inzhenerov), 1915.

[38] J. Geiser, *Iterative solvers for the Maxwell-Stefan diffusion equations: methods and applications in plasma and particle transport*, Cogent Math., 2 (2015), pp. Art. ID 1092913, 16.

[39] V. Giovangigli, *Multicomponent flow modeling*, Sci. China Math., 55 (2012), pp. 285–308.

[40] V. Giovangigli and M. Massot, *Asymptotic stability of equilibrium states for multicomponent reactive flows*, Math. Models Methods Appl. Sci., 8 (1998), pp. 251–297.

[41] ——, *The local Cauchy problem for multicomponent reactive flows in full vibrational non-equilibrium*, Math. Methods Appl. Sci., 21 (1998), pp. 1415–1439.

[42] C. J. Gittelson, R. Hiptmair, and I. Perugia, *Plane wave discontinuous Galerkin methods: analysis of the h-version*, M2AN Math. Model. Numer. Anal., 43 (2009), pp. 297–332.

[43] J. Gopalakrishnan, M. Hochsteger, J. Schöberl, and C. Wintersteiger, *An explicit mapped tent pitching scheme for Maxwell equations*, in Spectral and High Order Methods for Partial Differential Equations ICOSAHOM 2018, Cham, 2020, Springer International Publishing, pp. 359–369.

[44] J. Gopalakrishnan, P. Monk, and P. Sepúlveda, *A tent pitching scheme motivated by Friedrichs theory*, Comput. Math. Appl., 70 (2015), pp. 1114–1135.

[45] J. Gopalakrishnan, J. Schöberl, and C. Wintersteiger, *Mapped tent pitching schemes for hyperbolic systems*, SIAM J. Sci. Comput., 39 (2017), pp. B1043–B1063.

[46] J. Gopalakrishnan, J. Schöberl, and C. Wintersteiger, *Structure aware Runge–Kutta time stepping for spacetime tents*, SN Partial Differential Equations and Applications, 1 (2020), pp. 1–24.

[47] M. J. Grote, M. Mehlin, and S. A. Sauter, *Convergence analysis of energy conserving explicit local time-stepping methods for the wave equation*, SIAM J. Numer. Anal., 56 (2018), pp. 994–1021.

[48] M. J. Grote and T. Mitkova, *High-order explicit local time-stepping methods for damped wave equations*, J. Comput. Appl. Math., 239 (2013), pp. 270–289.

[49] M. Herberg, M. Meyries, J. Prüss, and M. Wilke, *Reaction-diffusion systems of Maxwell-Stefan type with reversible mass-action kinetics*, Nonlinear Anal., 159 (2017), pp. 264–284.

[50] R. Hiptmair, A. Moiola, and I. Perugia, *A survey of Trefftz methods for the Helmholtz equation*, in Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations, G. R. Barrenechea, F. Brezzi, A. Cangiani, and E. H. Georgoulis, eds., Lect. Notes Comput. Sci. Eng., Springer, 2016. pp. 237–278.

[51] T. J. R. Hughes and G. M. Hulbert, *Space-time finite element methods for elastodynamics: formulations and error estimates*, Comput. Methods Appl. Mech. Engrg., 66 (1988), pp. 339–363.

[52] L.-M. Imbert-Gérard, *Interpolation properties of generalized plane waves*, Numer. Math., (2015), pp. 1–29.

[53] L.-M. Imbert-Gérard and B. Després, *A generalized plane-wave numerical method for smooth nonconstant coefficients*, IMA J. Numer. Anal., 34 (2014), pp. 1072–1103.

[54] L.-M. Imbert-Gérard, A. Moiola, and P. Stocker, *A space–time quasi-Trefftz DG method for the wave equation with piecewise-smooth coefficients*. arXiv:2011.04617, 2020.

[55] L.-M. Imbert-Gérard and P. Monk, *Numerical simulation of wave propagation in inhomogeneous media using generalized plane waves*, ESAIM Math. Model. Numer. Anal., 51 (2017), pp. 1387–1406.

[56] C. Johnson, *Discontinuous Galerkin finite element methods for second order hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 107 (1993), pp. 117–129.

[57] O. Junge, D. Matthes, and H. Osberger, *A fully discrete variational scheme for solving nonlinear Fokker-Planck equations in multiple space dimensions*, SIAM J. Numer. Anal., 55 (2017), pp. 419–443.

[58] A. Jüngel, *The boundedness-by-entropy method for cross-diffusion systems*, Nonlinearity, 28 (2015), pp. 1963–2001.

[59] ——, *Entropy methods for diffusive partial differential equations*, SpringerBriefs in Mathematics, Springer, [Cham], 2016.

[60] A. Jüngel and O. Leingang, *Convergence of an implicit Euler Galerkin scheme for Poisson-Maxwell-Stefan systems*, Adv. Comput. Math., 45 (2019), pp. 1469–1498.

[61] A. Jüngel and S. Schuchnigg, *Entropy-dissipating semi-discrete Runge-Kutta schemes for nonlinear diffusion equations*, Commun. Math. Sci., 15 (2017), pp. 27–53.

[62] A. Jüngel and I. V. Stelzer, *Existence analysis of Maxwell-Stefan systems for multicomponent mixtures*, SIAM J. Math. Anal., 45 (2013), pp. 2421–2440.

[63] A. Jüngel, *Cross-diffusion systems with entropy structure*, in Proceedings of Equadiff 2017 Conference, 2017, pp. 181–190.

[64] U. Köcher and M. Bause, *Variational space-time methods for the wave equation*, J. Sci. Comput., 61 (2014), pp. 424–453.

[65] A. Y. Kokotov and B. A. Plamenevskiĭ, *On the asymptotic behavior of solutions of the Neumann problem for hyperbolic systems in domains with conical points*, Algebra i Analiz, 16 (2004), pp. 56–98.

[66] F. Kretzschmar, *The discontinuous Galerkin Trefftz method*, PhD thesis, Technische Universität Darmstadt, 2015. Available at http://tuprints.ulb.tu-darmstadt.de/5166/.

[67] F. Kretzschmar, A. Moiola, I. Perugia, and S. M. Schnepp, *A priori error analysis of space-time Trefftz discontinuous Galerkin methods for wave problems*, IMA J. Numer. Anal., 36 (2016), pp. 1599–1635.

[68] F. Kretzschmar, S. M. Schnepp, I. Tsukerman, and T. Weiland, *Discontinuous Galerkin methods with Trefftz approximations*, J. Comput. Appl. Math., 270 (2014), pp. 211–222.

[69] P. C. Kunstmann, B. Li, and C. Lubich, *Runge-Kutta time discretization of nonlinear parabolic equations studied via discrete maximal parabolic regularity*, Found. Comput. Math., 18 (2018), pp. 1109–1130.

[70] U. Langer, S. E. Moore, and M. Neumüller, *Space-time isogeometric analysis of parabolic evolution problems*, Comput. Methods Appl. Mech. Engrg., 306 (2016), pp. 342–363.

[71] S. Larsson and C. Schwab, *Compressive space-time Galerkin discretizations of parabolic partial differential equations*. arXiv:1501.04514, 2015.

[72] E. Leonardi and C. Angeli, *On the Maxwell–Stefan approach to diffusion: A general resolution in the transient regime for one-dimensional systems*, J. Phys. Chem. B, 114 (2009), pp. 151–164.

[73] M. Lilienthal, S. M. Schnepp, and T. Weiland, *Non-dissipative space-time hp-discontinuous Galerkin method for the time-dependent Maxwell equations*, J. Comput. Phys., 275 (2014), pp. 589–607.

[74] J.-B. W. P. Loos, P. J. T. Verheijen, and J. A. Moulijn, *Numerical simulation of the generalized Maxwell–Stefan model for multicomponent diffusion in microporous sorbents*, Collection of Czechoslovak Chemical Communications, 57 (1992), pp. 687–697.

[75] R. B. Lowrie, P. L. Roe, and B. van Leer, *Space-time methods for hyperbolic conservation laws*, in Barriers and challenges in computational fluid dynamics (Hampton, VA, 1996), vol. 6 of ICASE/LaRC Interdiscip. Ser. Sci. Eng., Kluwer Acad. Publ., Dordrecht, 1998, pp. 79–98.

[76] V. T. Luong and N. T. Tung, *The Dirichlet–Cauchy problem for nonlinear hyperbolic equations in a domain with edges*, Nonlinear Analysis, 125 (2015), pp. 457–467.

[77] M. Marion and R. Temam, *Global existence for fully nonlinear reaction-diffusion systems describing multicomponent reactive flows*, J. Math. Pures Appl. (9), 104 (2015), pp. 102–138.

[78] J. C. Maxwell, *On the dynamical theory of gases*, Philosophical Transactions of the Royal Society of London, 157 (1867), pp. 49–88.

[79] M. McLeod and Y. Bourgault, *Mixed finite element methods for addressing multi-species diffusion using the Maxwell–Stefan equations*, Computer Methods in Applied Mechanics and Engineering, 279 (2014), pp. 515–535.

[80] A. Moiola, *Trefftz-discontinuous Galerkin methods for time-harmonic wave problems*, PhD thesis, Seminar for applied mathematics, ETH Zürich, 2011.
Available at http://e-collection.library.ethz.ch/view/eth:4515.

[81] A. Moiola and I. Perugia, *A space–time Trefftz discontinuous Galerkin method for the acoustic wave equation in first-order formulation*, Numer. Math., 138 (2018), pp. 389–435.

[82] P. Monk and G. R. Richter, *A discontinuous Galerkin method for linear symmetric hyperbolic systems in inhomogeneous media*, J. Sci. Comput., 22/23 (2005), pp. 443–477.

[83] S. E. Moore, *A stable space-time finite element method for parabolic evolution problems*, Calcolo, 55 (2018), pp. Paper No. 18, 19.

[84] F. Müller, *Numerical analysis of finite element methods for second order wave equations in polygons*, PhD thesis, ETH Zurich, 2017. Available at https://www.research-collection.ethz.ch/handle/20.500.11850/167502.

[85] F. Müller, D. Schötzau, and C. Schwab, *Discontinuous Galerkin methods for acoustic wave propagation in polygons*, J. Sci. Comput., (2018), pp. 1–27.

[86] F. Müller and C. Schwab, *Finite elements with mesh refinement for wave equations in polygons*, J. Comput. Appl. Math., 283 (2015), pp. 163–181.

[87] T. Nakaki and K. Tomoeda, *Numerical approach to the waiting time for the one-dimensional porous medium equation*, Quart. Appl. Math., 61 (2003), pp. 601–612.

[88] M. Neumüller, *Space-Time Methods: Fast Solvers and Applications*, PhD thesis, Graz University of Technology, 2013.

[89] K. S. C. Peerenboom, J. van Dijk, J. H. M. Ten Thije Boonkkamp, L. Liu, W. J. Goedheer, and J. J. A. M. van der Mullen, *Mass conservative finite volume discretization of the continuity equations in multi-component mixtures*, J. Comput. Phys., 230 (2011), pp. 3525–3537.

[90] I. Perugia, J. Schöberl, P. Stocker, and C. Wintersteiger, *Tent pitching and Trefftz-DG method for the acoustic wave equation*, Comput. Math. Appl., 79 (2020), pp. 2987–3000.

[91] S. Petersen, C. Farhat, and R. Tezaur, *A space-time discontinuous Galerkin method for the solution of the wave equation in the time domain*, Internat. J. Numer. Methods Engrg., 78 (2009), pp. 275–295.

[92] D. Portillo, J. C. García Orden, and I. Romero, *Energy–entropy–momentum integration schemes for general discrete non-smooth dissipative problems in thermomechanics*, Internat. J. Numer. Methods Engrg., 112 (2017), pp. 776–802.

[93] J. Prüss and G. Simonett, *Moving Interfaces and Quasilinear Parabolic Evolution Equations*, vol. 105, Springer, 2016.

[94] G. R. Richter, *An explicit finite element method for the wave equation*, Appl. Numer. Math., 16 (1994), pp. 65–80.

[95] T. Richter, A. Springer, and B. Vexler, *Efficient numerical realization of discontinuous Galerkin methods for temporal discretization of parabolic problems*, Numer. Math., 124 (2013), pp. 151–182.

[96] W. Ritz, *Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik.*, J. für die reine und angew. Math., 135 (1909), pp. 1–61.

[97] F. Salvarani and A. J. Soares, *On the relaxation of the Maxwell–Stefan system to linear diffusion*, Applied Mathematics Letters, 85 (2018), pp. 15–21.

[98] A. Schafelner, *Space-time finite element methods for parabolic initial-boundary problems*, Master's thesis, Johannes Kepler University Linz, 2017.

[99] J. Schöberl, *Netgen/NGSolve*. https://ngsolve.org. Accessed: 2021-03-25.

[100] J. Schöberl, *C++11 implementation of finite elements in NGSolve*, ASC Report 30/2014, Institute for Analysis and Scientific Computing, Vienna University of Technology, (2014).

[101] C. Schwab and R. Stevenson, *Space-time adaptive wavelet methods for parabolic evolution problems*, Math. Comp., 78 (2009), pp. 1293–1318.

[102] J. Stefan, *Über das Gleichgewicht und die Bewegung, insbesondere die Diffusion von Gasgemengen*, Akad. Wiss. Wien, 63 (1871).

[103] O. Steinbach, *Space-time finite element methods for parabolic problems*, Comput. Methods Appl. Math., 15 (2015), pp. 551–566.

[104] O. Steinbach and H. Yang, *An algebraic multigrid method for an adaptive space-time finite element discretization*, in Large-scale scientific computing, vol. 10665 of Lecture Notes in Comput. Sci., Springer, Cham, 2018, pp. 66–73.

[105] ——, *Comparison of algebraic multigrid methods for an adaptive space-time finite-element discretization of the heat equation in 3D and 4D*, Numer. Linear Algebra Appl., 25 (2018), pp. e2143, 17.

[106] O. Steinbach and M. Zank, *A stabilized space–time finite element method for the wave equation*, Springer International Publishing, Cham, 2019, pp. 341–370.

[107] L. Tartar, *The general theory of homogenization - A personalized introduction*, vol. 7 of Lecture Notes of the Unione Matematica Italiana, Springer-Verlag, Berlin; UMI, Bologna, 2009.

[108] I. Toulopoulos, *Space-time finite element methods stabilized using bubble function spaces*, Applicable Analysis, 0 (2018), pp. 1–18.

[109] E. Trefftz, *Ein Gegenstuck zum Ritzschen Verfahren*, Proc. 2nd Int. Cong. Appl. Mech., Zurich, 1926, (1926), pp. 131–137.

[110] A. Üngör and A. Sheffer, *Pitching tents in space-time: mesh generation for discontinuous Galerkin method*, Internat. J. Found. Comput. Sci., 13 (2002), pp. 201–221. Volume and surface triangulations.

[111] J. Česenek and M. Feistauer, *Theory of the space-time discontinuous Galerkin method for nonstationary parabolic problems with nonlinear convection and diffusion*, SIAM J. Numer. Anal., 50 (2012), pp. 1181–1206.

[112] D. Wang, R. Tezaur, and C. Farhat, *A hybrid discontinuous in space and time Galerkin method for wave propagation problems*, Internat. J. Numer. Methods Engrg., 99 (2014), pp. 263–289.

[113] C. Wintersteiger, *Mapped tent pitching method for hyperbolic conservation laws*, Diplomarbeit, (2015).

[114] L. Yin, A. Acharya, N. Sobh, R. B. Haber, and D. A. Tortorelli, *A space-time discontinuous Galerkin method for elastodynamic analysis*, in Discontinuous Galerkin methods (Newport, RI, 1999), vol. 11 of Lect. Notes Comput. Sci. Eng., Springer, Berlin, 2000, pp. 459–464.

[115] M. Zank, *Inf-Sup Stable Space-Time Methods for Time-Dependent Partial Differential Equations*, vol. 36 of Monographic Series TU Graz, Computation in Engineering and Science, TU Graz, 2020.