



# **On local and structure-adapted time integration for wave-type problems**

Zur Erlangung des akademischen Grades eines  
**DOKTORS DER NATURWISSENSCHAFTEN**

von der KIT-Fakultät für Mathematik des  
Karlsruher Instituts für Technologie  
genehmigte

**DISSERTATION**

von

**Malik Scheifinger**

Tag der mündlichen Prüfung: 21. September 2026

Referentin: Prof. Dr. Marlis Hochbruck

Korreferent: Prof. Dr. Tobias Jahnke



# Acknowledgement

I gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 258734477 – CRC 1173.

*In German.*

Zuerst möchte ich mich bei allen bedanken, die mich während meiner Promotion wissenschaftlich, beruflich oder persönlich unterstützt und begleitet haben. Mein besonderer Dank gilt meiner Betreuerin, Prof. Dr. Marlis Hochbruck. Ohne sie wäre diese Arbeit nicht möglich gewesen. Schon während meines Studiums gab sie mir die Möglichkeit, als wissenschaftliche Hilfskraft in ihrer Arbeitsgruppe mitzuarbeiten und durch die Time-Integration-Workshops diese kennenzulernen. Auch während der vergangenen drei Jahre konnte ich stets Fragen stellen und auf ihre Unterstützung zählen. Ich danke ebenso meinem Zweitbetreuer, Prof. Dr. Tobias Jahnke. Trotz seines vollen Terminkalenders hat er sich die Zeit genommen, die Arbeit zu lesen und sie mit vielen hilfreichen Anmerkungen zu bereichern.

Außerdem danke ich Jessy, Lisa und Lea sowie dem Rest der variierenden 6:00er Stammlaufgruppe, die die teils anstrengenden Arbeitsphasen durch mehr oder weniger entspannte Läufe und schöne Abende aufgelockert haben. Auch Eliane, Till und Lucas danke ich für ihre Freundschaft, ebenso Carl, Daniel und Tim, die im Laufe der Promotion für mich zu mehr als Kollegen geworden sind.

Benjamin und Eliane haben sich die Zeit genommen, frühe Fassungen von Kapitel 2 und der Einleitung zu lesen, wofür ich ihnen besonders dankbar bin. Außerdem danke ich Benjamin dafür, dass ich mich mit Fragen jederzeit an ihn wenden durfte, sowie für die witzigen Abende auf verschiedenen Konferenzen und Workshops. Laurette, Barbara, Mathias und Christian danke ich für ihre stets hilfreiche Unterstützung bei administrativen und technischen Fragen. Markus danke ich für die angenehme Zusammenarbeit in den HM-Vorlesungen.

Schließlich danke ich meinen Eltern, die mir mein Studium ermöglicht haben und auf deren Unterstützung ich mich jederzeit verlassen kann.



# Abstract

This thesis studies time integration for wave-type problems whose numerical approximation is constrained by fast oscillations in space and time, small discretization parameters, or materials with high wave speeds. Two settings are considered. In the first part, the constraint is spatially localized. Two-field Friedrichs' systems, such as Maxwell's equations and the acoustic wave equation, are discretized in space by discontinuous Galerkin (dG) methods on locally refined meshes. The smallest cells then impose a severe CFL restriction on explicit time integrators, although they represent only a small part of the mesh.

For central-flux discretizations, we construct and analyze a second-order, leapfrog-based local time-integration framework. It combines the leapfrog method on the coarse part of the mesh with a filtered treatment of the fine part, and it contains fully explicit local time-stepping schemes as well as locally implicit variants. Under suitable assumptions on the filter, the resulting stability and error estimates show convergence with a CFL restriction governed by the coarse part of the mesh. This framework is then extended to stabilized dG discretizations, like upwind fluxes, where dissipative face penalties are applied on the coarse part of the mesh. The analysis yields stability and error estimates under a CFL restriction which again is governed by the coarse part of the mesh.

To obtain higher-order time integration, we use implicit Gauß collocation methods and solve the arising linear systems by preconditioned Krylov subspace methods. The preconditioner uses only the fine part of the mesh. It is shown that the number of preconditioned QMR iterations needed to reach a fixed tolerance can be bounded independently of the fine mesh. Thus the same local fine and coarse split is used at a different algorithmic level, namely inside the linear solver for an implicit method.

The second part treats the semiclassical magnetic Schrödinger equation. Here, a small semiclassical parameter appears which leads to rapid oscillations of the solution in space and time making a mesh-based approach very expensive. A mesh-free variational Gaussian wave packet ansatz [BDH<sup>+</sup>24] captures the rapid oscillations and reduces the evolution to a finite-dimensional system of parameter ODEs. For this system we extend the Boris algorithm [BL18] from classical charged-particle dynamics and propose a modified fourth-order Runge–Kutta method. Both methods preserve the  $L^2$ -norm of the Gaussian wave packet by construction. Numerical experiments, including the Penning trap, investigate convergence, norm preservation and energy behavior of the proposed schemes.



# Contents

|                                                                                             |           |
|---------------------------------------------------------------------------------------------|-----------|
| <b>Introduction</b>                                                                         | <b>1</b>  |
| <b>I. Local time integration for Friedrichs' systems</b>                                    | <b>17</b> |
| <b>1. Local time integration for Friedrichs' systems</b>                                    | <b>19</b> |
| 1.1. Introduction . . . . .                                                                 | 19        |
| 1.2. Friedrichs' systems . . . . .                                                          | 20        |
| 1.2.1. Notation . . . . .                                                                   | 21        |
| 1.3. Local time-integration . . . . .                                                       | 22        |
| 1.3.1. Stability . . . . .                                                                  | 23        |
| 1.3.2. Error analysis of the local time-integration scheme . . . . .                        | 28        |
| 1.3.3. Stability and convergence of local time-integration schemes . . . . .                | 32        |
| 1.4. Numerical examples . . . . .                                                           | 34        |
| 1.4.1. Linear Maxwell's equations . . . . .                                                 | 34        |
| 1.4.2. Order of convergence . . . . .                                                       | 35        |
| 1.4.3. Necessity of stabilization . . . . .                                                 | 36        |
| 1.4.4. Runtime comparison . . . . .                                                         | 36        |
| 1.5. Concluding remarks . . . . .                                                           | 38        |
| <b>2. Local time integration for Friedrichs' systems<br/>with upwind-type stabilization</b> | <b>41</b> |
| 2.1. Space discretization with stabilization . . . . .                                      | 41        |
| 2.1.1. Notation . . . . .                                                                   | 41        |
| 2.1.2. Stabilization operators . . . . .                                                    | 42        |
| 2.1.3. Properties of the stabilization operators . . . . .                                  | 44        |
| 2.2. Full discretization . . . . .                                                          | 48        |
| 2.3. Stability . . . . .                                                                    | 49        |
| 2.4. Error analysis . . . . .                                                               | 53        |
| 2.5. Numerical examples . . . . .                                                           | 62        |
| 2.5.1. Spatial convergence . . . . .                                                        | 63        |
| 2.5.2. Temporal convergence . . . . .                                                       | 64        |
| <b>A. Weak and strong forms of the discrete Friedrichs' operators</b>                       | <b>65</b> |
| A.1. Weak and strong forms of the discrete Friedrichs' operators . . . . .                  | 65        |

|                                                                                                             |                |
|-------------------------------------------------------------------------------------------------------------|----------------|
| <b>3. Higher order local time integration for Friedrichs' systems</b>                                       | <b>67</b>      |
| 3.1. Introduction . . . . .                                                                                 | 68             |
| 3.2. Problem setting . . . . .                                                                              | 69             |
| 3.3. Higher-order implicit Runge–Kutta methods . . . . .                                                    | 71             |
| 3.3.1. Gauß collocation methods . . . . .                                                                   | 72             |
| 3.4. Preconditioned Krylov subspace methods . . . . .                                                       | 74             |
| 3.4.1. Krylov subspace methods for complex symmetric matrices . . . . .                                     | 74             |
| 3.4.2. Preconditioning for locally refined grids . . . . .                                                  | 76             |
| 3.5. Numerical experiments . . . . .                                                                        | 79             |
| 3.5.1. Illustration of Theorems 3.4 and 3.6 . . . . .                                                       | 80             |
| 3.5.2. Runtime comparison . . . . .                                                                         | 81             |
| 3.6. Conclusion . . . . .                                                                                   | 83             |
| <br><b>II. Time-integration of Gaussian variational approximation for the magnetic Schrödinger equation</b> | <br><b>85</b>  |
| <b>4. Time-integration of Gaussian variational approximation for the magnetic Schrödinger equation</b>      | <b>87</b>      |
| 4.1. Introduction . . . . .                                                                                 | 87             |
| 4.1.1. Contributions of the paper . . . . .                                                                 | 88             |
| 4.1.2. Outline of the paper . . . . .                                                                       | 88             |
| 4.1.3. Notation . . . . .                                                                                   | 89             |
| 4.2. Penning trap . . . . .                                                                                 | 89             |
| 4.3. Variational approximation . . . . .                                                                    | 91             |
| 4.3.1. Variational equations of motion . . . . .                                                            | 91             |
| 4.3.2. Asymptotic accuracy . . . . .                                                                        | 92             |
| 4.3.3. Classical versus variational approximation . . . . .                                                 | 92             |
| 4.3.4. Hagedorn parametrization . . . . .                                                                   | 93             |
| 4.4. Transformation of equations of motion . . . . .                                                        | 93             |
| 4.4.1. Averaged magnetic momenta . . . . .                                                                  | 93             |
| 4.4.2. Equations for the center and the width . . . . .                                                     | 94             |
| 4.4.3. Equations for the phase . . . . .                                                                    | 98             |
| 4.5. Time integration for the equations of motion . . . . .                                                 | 99             |
| 4.5.1. Boris algorithm for classical equations of motion . . . . .                                          | 99             |
| 4.5.2. Boris-type algorithm . . . . .                                                                       | 99             |
| 4.5.3. Discretization of the phase . . . . .                                                                | 100            |
| 4.5.4. Complete Boris-type algorithm . . . . .                                                              | 100            |
| 4.5.5. Modified classical Runge–Kutta method . . . . .                                                      | 101            |
| 4.6. Numerical experiments . . . . .                                                                        | 101            |
| 4.6.1. Sublinear magnetic potential in two dimensions . . . . .                                             | 101            |
| 4.6.2. Penning trap . . . . .                                                                               | 103            |
| <br><b>B. Gaussian calculus and energy formula</b>                                                          | <br><b>107</b> |
| <br><b>Bibliography</b>                                                                                     | <br><b>109</b> |

# Introduction

Various physical phenomena are described by time-dependent partial differential equations. Important examples are electromagnetic, acoustic, and quantum waves. Even though the existence and uniqueness of a solution can be guaranteed under certain assumptions on the data and the domain, the solution typically cannot be stated in a closed expression. This necessitates the use of numerical methods to approximate the solution. Computing the numerical approximation accurately and at a cost that remains affordable as the problem size grows is a central task of numerical analysis.

The difficulty addressed throughout this thesis is a disparity of scales that appears in two distinct forms. In Part I, for example, a few tiny mesh elements in a spatial discretization impose a severe time step size restriction on an explicit time stepping method. In Part II, a small parameter produces rapid oscillations of the wave function in time and space which have to be resolved by a numerical method. In both parts, the aim is to prevent the restrictive feature from dictating the cost of the entire computation by using a numerical method adapted to the problem structure.

This thesis pursues two answers. The first is local time-integration, where extra computational work is only employed where it is needed (Chapters 1 to 3). The second builds upon a structure-exploiting reduced model, in which a single Gaussian wave packet captures the rapid oscillations. This ansatz yields a small system of nonoscillatory ordinary differential equations. Our contribution is to propose two efficient numerical schemes to advance these equations in time (Chapter 4).

The results in Chapter 1 and Chapter 4 have been published in [HS26] and [SBH<sup>+</sup>25], respectively. The results in Chapter 3 have been published as a preprint in [HKS26].

## Friedrichs' systems and local stiffness

Friedrichs introduced a functional-analytic framework for certain first-order partial differential equations in 1958, now called Friedrichs' systems [Fri58]. Modern frameworks are presented by Ern and Guermond [EG06a, EG06b], by Di Pietro and Ern [DPE12], and, more recently, in [HK19, DHK<sup>+</sup>23].

Throughout Part I, we work with the two-field structure [DHK<sup>+</sup>23, §11.2] of these systems. The domain  $\Omega \subset \mathbb{R}^d$  is assumed to be open, bounded, and polygonal or polyhedral. The two-field Friedrichs' system couples the fields  $u$  and  $v$  through

$$\partial_t u = \mathcal{L}_v v + g_u, \quad \partial_t v = \mathcal{L}_u u + g_v, \quad \text{on } \Omega \times \mathbb{R}_+, \quad (\text{I.1})$$

with given source terms  $g_u$  and  $g_v$ . The Friedrichs' operators  $\mathcal{L}_u$  and  $\mathcal{L}_v$  are the first-order differential operators

$$\mathcal{M}_v \mathcal{L}_u u = \sum_{i=1}^d L_i \partial_i u, \quad \mathcal{M}_u \mathcal{L}_v v = \sum_{i=1}^d L_i^T \partial_i v,$$

with constant coefficient matrices  $L_i \in \mathbb{R}^{m_v \times m_u}$ , where  $m_u, m_v \in \mathbb{N}$  are the numbers of components of  $u$  and  $v$ , respectively. The material tensors  $\mathcal{M}_u \in L^\infty(\Omega)^{m_u \times m_u}$  and  $\mathcal{M}_v \in L^\infty(\Omega)^{m_v \times m_v}$  are symmetric and uniformly positive definite.

We write

$$(u, \psi)_\Omega := (\mathcal{M}_u u, \psi)_{L^2(\Omega)}, \quad u, \psi \in L^2(\Omega)^{m_u}$$

for the  $L^2$ -inner product weighted by the material tensor  $\mathcal{M}_u$  and  $\|\cdot\|_\Omega$  for the induced norm. The weighted inner product and the induced norm on  $L^2(\Omega)^{m_v}$  are defined analogously. The arguments determine which inner product or norm is considered. Since the material tensors are uniformly positive definite, the weighted inner products are equivalent to the standard  $L^2$ -inner products. Therefore, equipped with the weighted inner products, the spaces  $L^2(\Omega)^{m_u}$  and  $L^2(\Omega)^{m_v}$  are Hilbert spaces.

Stacking the two fields into  $x = \begin{pmatrix} u \\ v \end{pmatrix}$  and denoting the source term by  $g$ , the system (I.1) reads

$$\partial_t x = \mathcal{L}x + g, \quad \mathcal{L} = \begin{pmatrix} 0 & \mathcal{L}_v \\ \mathcal{L}_u & 0 \end{pmatrix}. \quad (\text{I.2})$$

The boundary conditions are encoded in the operator domains  $\mathcal{D}(\mathcal{L}_u)$  and  $\mathcal{D}(\mathcal{L}_v)$ . If the boundary conditions do not introduce damping [DHK<sup>+</sup>23, Assump. 11.8], the two Friedrichs' operators satisfy the adjointness relation

$$(\mathcal{L}_u u, v)_\Omega = -(\mathcal{L}_v v, u)_\Omega \quad \text{for } u \in \mathcal{D}(\mathcal{L}_u), v \in \mathcal{D}(\mathcal{L}_v), \quad (\text{I.3})$$

see [DHK<sup>+</sup>23, Lem. 13.1], leading to the skew-adjointness of  $\mathcal{L}$  on  $\mathcal{D}(\mathcal{L}) = \mathcal{D}(\mathcal{L}_u) \times \mathcal{D}(\mathcal{L}_v)$ . Equipped with an appropriate initial condition  $x^0$ , the system (I.2) is well-posed.

**Theorem I.1** ([DHK<sup>+</sup>23, Cor. 9.23]). *Let  $x^0 \in \mathcal{D}(\mathcal{L})$  and  $g \in C^1(\mathbb{R}_+, L^2(\Omega)^m) + C(\mathbb{R}_+, \mathcal{D}(\mathcal{L}))$ . Then the system (I.2) has a unique solution  $x \in C^1(\mathbb{R}_+, L^2(\Omega)^m) \cap C(\mathbb{R}_+, \mathcal{D}(\mathcal{L}))$ , given by the variation-of-constants formula*

$$x(t) = e^{t\mathcal{L}} x^0 + \int_0^t e^{(t-s)\mathcal{L}} g(s) \, ds.$$

It satisfies

$$\|x(t)\|_\Omega \leq \|x^0\|_\Omega + \int_0^t \|g(s)\|_\Omega \, ds,$$

and  $\|x(t)\|_\Omega = \|x^0\|_\Omega$  if  $g = 0$ .

The variation-of-constants formula represents the solution through the unitary evolution operator  $e^{t\mathcal{L}}$  acting on the initial data and the accumulated source. The norm equality for  $g = 0$  corresponds to the energy conservation of the wave-type system. For the sake of presentation, in the following part we neglect forcing terms.

Many wave-type systems fall into the Friedrichs' class. Two of them are the acoustic wave equation and Maxwell's equations in first-order formulation.

a) Maxwell's equations with the perfectly conducting boundary condition read

$$\varepsilon \partial_t E = \nabla \times H, \quad \mu \partial_t H = -\nabla \times E \quad \text{in } \Omega, \quad E \times n = 0 \quad \text{on } \partial\Omega, \quad (\text{I.4})$$

with outward unit normal  $n$ . The first component is the electric field  $u = E$  with permittivity  $\mathcal{M}_u = \varepsilon$ , and the second component is the magnetic field  $v = H$  with permeability  $\mathcal{M}_v = \mu$ .

Writing  $[a]_\times \in \mathbb{R}^{3 \times 3}$  for the skew-symmetric matrix with  $[a]_\times b = a \times b$ , this fits the abstract form with coefficient matrices  $L_i = -[e_i]_\times$ , built from the standard basis vectors  $e_1, e_2, e_3$ . The perfectly conducting boundary condition is contained in the operator domains  $\mathcal{D}(\mathcal{L}_u) = \{E \in H(\text{curl}, \Omega) \mid E \times n = 0 \text{ on } \partial\Omega\}$  and  $\mathcal{D}(\mathcal{L}_v) = H(\text{curl}, \Omega)$ , see [DHK<sup>+</sup>23, Example 2.1 and §9.3], [Köh19, §2.5.3] for details.

b) The acoustic wave equation in mixed div-grad form with homogeneous Dirichlet boundary condition reads

$$\rho \partial_t v = \nabla p, \quad \kappa^{-1} \partial_t p = \nabla \cdot v \quad \text{in } \Omega, \quad p = 0 \quad \text{on } \partial\Omega. \quad (\text{I.5})$$

We choose the pressure  $u = p$  as the first component, with  $\mathcal{M}_u = \kappa^{-1}$ , where  $\kappa$  is the bulk modulus. We choose the velocity as the second component, with  $\mathcal{M}_v = \rho$ , where  $\rho$  is the density. This fits the abstract form with coefficient matrices given by the rows  $L_i = e_i \in \mathbb{R}^{d \times 1}$ . The boundary condition is imposed through the operator domains  $\mathcal{D}(\mathcal{L}_u) = H_0^1(\Omega)$  and  $\mathcal{D}(\mathcal{L}_v) = H(\text{div}, \Omega)$ , see [DHK<sup>+</sup>23, Example 2.1], [Zie20, §2.4] for details.

**Finite element and discontinuous Galerkin space discretization.** Discretizing the wave-type problems in Chapters 1 to 3, we follow the method-of-lines approach, which first discretizes in space and then integrates the resulting semidiscrete system in time. The space discretization is carried out by a finite element method, turning the partial differential equation (PDE) into a large system of ordinary differential equations (ODEs).

A finite element method partitions the domain  $\Omega$  into a mesh  $\mathcal{T}_h$  of cells  $K$ , e.g., triangles or rectangles in  $\mathbb{R}^2$ , and approximates each field in (I.1) by functions that are polynomial in each component on each cell. The approximation is determined by the weak form of the PDE reduced to a discrete ansatz space. In Part I, we use a discontinuous Galerkin (dG) finite element method, where the piecewise polynomials do not need to be continuous across interfaces between cells.

Historically, dG methods entered numerical analysis through first-order transport problems. A compact overview of this development is given in [HW08, §1.1]. Reed and Hill [RH73] proposed a triangular mesh method for the neutron transport equation in 1973, and Lesaint and Raviart [LR74] analyzed a closely related finite element method shortly after. These early works already contain the structural ingredients used below: elementwise polynomial approximation and weak coupling by numerical fluxes. The same ingredients also occur in modern wave computations. For example, high-order dG methods are used for time-dependent Maxwell's equations and related wave propagation problems, including unstructured meshes, piecewise polynomial spaces, and various numerical fluxes [HW08, Nie09, HP15].

This history also points to the mixed character of the method. On the one hand, dG methods are finite element methods, since they use local polynomial trial and test spaces in a weak formulation. On the other hand, their coupling between cells resembles finite volume methods.

Neighboring cells communicate only through numerical fluxes on their common interfaces. These fluxes encode, for example, conservation or dissipative stabilization.

Two properties make the dG method a natural fit for the aims of this thesis. First, dG methods are flexible, permitting unstructured meshes, varying dG polynomial degrees on different cells, and hanging nodes without constraints. Second, the cells couple only weakly through numerical fluxes built from traces, jumps, and averages of the fields on the faces. The resulting mass matrix contains only volume terms, which completely decouple. It therefore has a block-diagonal structure, where the size of each block depends only on the local degrees of freedom of the corresponding cell. The inverse is therefore local and cheap to apply. Explicit time stepping is then truly explicit and needs no global solve per step, which is the main reason why the dG method pairs well with explicit and local time integration. The dG method delivers this block-diagonal, positive definite mass matrix without a mass-lumping construction. In continuous finite element methods, the continuity constraint across faces makes neighboring elements share unknowns, and the mass matrix is sparse but globally coupled [EG21]. Thus, making the mass inverse local usually requires an additional mass-lumping step, whose construction can be restrictive at high polynomial degrees.

A few objects recur throughout Part I, with their precise definitions given in the respective chapters. The mesh  $\mathcal{T}_h$  partitions the domain, and  $h_K$  denotes the diameter of a cell  $K \in \mathcal{T}_h$ . Moreover,  $h$  denotes a piecewise defined function taking the value  $h_K$  on a cell  $K$ . The faces split into interior and boundary faces,  $\mathcal{F}_h = \mathcal{F}_h^{\text{int}} \cup \mathcal{F}_h^{\text{bnd}}$ . The finite-dimensional dG ansatz spaces  $\mathbb{V}_h^u$  and  $\mathbb{V}_h^v$  contain piecewise polynomials of total degree at most  $k$  in  $d$  variables, which are globally in  $L^2(\Omega)^{m_u}$  and  $L^2(\Omega)^{m_v}$ , respectively.

The dG method replaces the Friedrichs' operator  $\mathcal{L}_u$  by a discrete operator  $\mathcal{L}_{u,h}$  on the dG space  $\mathbb{V}_h^u$ , and  $\mathcal{L}_{v,h}$  analogously, see [DPE12, §7.1.2] and [DHK<sup>+</sup>23, §10.3], or Lemma A.3. On each cell, the discrete operator is built from the local action of the Friedrichs' operator  $\mathcal{L}_u$  locally. The cells couple only through face terms built from jumps and averages on the interior faces, and boundary terms weakly enforce the boundary conditions. The flux is called central because it uses the average of the two traces and adds no penalization.

The central-flux discretizations  $\mathcal{L}_{u,h}$  and  $\mathcal{L}_{v,h}$  inherit the adjointness property (I.3), i.e.,

$$(\mathcal{L}_{u,h}u_h, v_h)_\Omega = -(u_h, \mathcal{L}_{v,h}v_h)_\Omega, \quad u_h \in \mathbb{V}_h^u, v_h \in \mathbb{V}_h^v, \quad (\text{I.6})$$

so the stacked operator  $\mathcal{L}_h$  is skew-adjoint. Moreover, they satisfy the inverse inequality

$$\|\mathcal{L}_{u,h}u_h\|_\Omega \lesssim \|h^{-1}u_h\|_\Omega, \quad u_h \in \mathbb{V}_h^u. \quad (\text{I.7})$$

Here,  $\lesssim$  hides a positive constant independent of the mesh size  $h$ , which may depend on the polynomial degree, the mesh regularity, and the operator and material coefficients, see [DHK<sup>+</sup>23, §10.3].

The choice of the numerical flux is the main difference between Chapter 1 and Chapter 2. Central fluxes, used in Chapter 1, are conservative, but spurious modes can remain undamped. Chapter 2 keeps the central-flux framework but adds a dissipative jump penalty on selected faces. This damps spurious modes, at the price of replacing exact energy conservation by controlled dissipation [CW17].

After a central-flux dG discretization of the system (I.1), we obtain the semidiscrete two-field system

$$\partial_t u_h = \mathcal{L}_{v,h}v_h, \quad \partial_t v_h = \mathcal{L}_{u,h}u_h, \quad \text{on } \mathbb{R}_+, \quad (\text{I.8})$$

where we have neglected forcing terms for the sake of presentation. Stacking the two fields into a single vector  $x_h$  yields the system of ordinary differential equations

$$\partial_t x_h = \mathcal{L}_h x_h, \quad \text{on } \mathbb{R}_+, \quad (\text{I.9})$$

with projected initial datum  $x_h(0) = \pi_h x^0$ . In our setting,  $\pi_h$  is a material-weighted orthogonal projection  $\pi_h: L^2(\Omega)^m \rightarrow \mathbb{V}_h$ , defined in [DHK<sup>+</sup>23, §10.1].

The semidiscrete system (I.9) is well-posed [DHK<sup>+</sup>23, Cor. 10.11], and assuming a certain regularity of the exact solution  $x$  of (I.2), the semi-discrete solution  $x_h$  of (I.9) satisfies the error bound

$$\sup_{0 \leq t \leq T} \|x(t) - x_h(t)\|_{\Omega} \leq C_T h^k,$$

i.e., it converges to the exact solution  $x$  as the mesh is refined, see [DHK<sup>+</sup>23, Thm. 10.12]. For every fixed  $T > 0$ , the constant  $C_T$  is independent of the mesh size but may depend on  $T$  and the exact solution. Here  $k$  denotes the dG polynomial degree.

The semidiscrete system (I.9) now has to be integrated in time.

**Explicit time integration.** A widely used explicit time integrator for wave-type problems is the leapfrog method, also called the Störmer–Verlet method. It is symmetric and conserves a modified energy, as discussed in [HLW06, §I.5.2] and [HS16]. However, the leapfrog method is only stable under a Courant–Friedrichs–Lewy (CFL) condition, a restriction on the time step size  $\tau$ , making the leapfrog method conditionally stable. Using the inverse inequality (I.7), the restriction on the time step size can be stated as

$$\tau \lesssim h_{\min},$$

with  $h_{\min}$  being the smallest diameter over all cells, see [DHK<sup>+</sup>23, §11.2]. Thus, the smallest cell can determine the step for the whole mesh.

This is illustrated in Figure 1 by the one-dimensional wave equation in mixed form (I.5), discretized with the dG method and the leapfrog scheme on a mesh with a few artificially refined cells. With a time step size just above the CFL limit, high-frequency modes associated with the refined cells leave the stability interval of the leapfrog method, and the numerical solution grows rapidly.

**Implicit time integration.** The natural alternative to the conditionally stable leapfrog method is an unconditionally stable one, the Crank–Nicolson method, analyzed for Maxwell’s equations in [Stu17, §4.2] and for Friedrichs’ systems in [DHK<sup>+</sup>23, §11.1]. It carries no CFL restriction, and for the skew-adjoint discrete operator  $\mathcal{L}_h$  it conserves the semi-discrete energy exactly, as long as no external forcing is applied. One might therefore ask why not to integrate every problem with this method.

The reason is the cost of a single step. The Crank–Nicolson method is implicit and couples all unknowns into one linear system per step. For large three-dimensional problems, a direct factorization of the system matrix is unfeasible because of its high memory consumption due to the size and dense structure of the factorization matrices. An alternative is to use an iterative solver in each step. After applying a Schur complement, one has to solve a symmetric positive definite system for only one of the two possibly vector-valued components. This can be done using the conjugate gradient method [Stu17, §4.5]. However, the conditioning of the linear

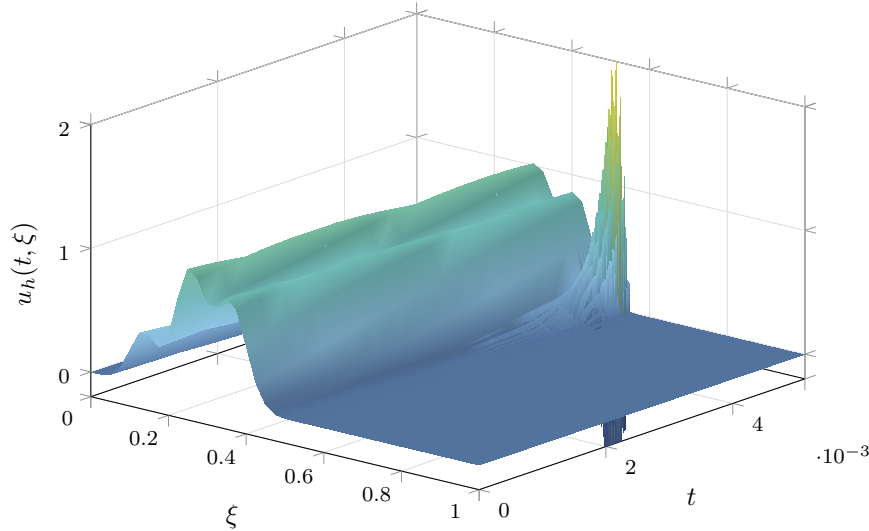


Figure 1.: Space-time surface illustrating the CFL instability of the leapfrog method for a central-flux dG discretization of the one-dimensional wave equation (I.5). A smooth wave is evolved on a mesh of 40 coarse cells, with three cells around  $\xi = 0.5$  refined four times. The time step size is 0.1% above the CFL limit, causing high-frequency oscillations localized near the refined cells.

system strongly influences the number of iterations needed to achieve a given accuracy [Hac16], making a single step very expensive. In our situation, the condition of the linear system scales like  $h_{\min}^{-2}$ . Thus, again, the smallest cell dictates the global cost of the time integration method.

The two options therefore sit at opposite extremes. An explicit method is cheap per step but constrained by the CFL condition, which leads to many steps. In addition to the high cost, a large number of steps can increase the contribution of accumulated roundoff errors to the global error [HLW06, §VIII.5]. An unconditionally stable method has no CFL restriction but is typically implicit and therefore expensive per step.

## Local time integration for Friedrichs' systems

In Part I, we are interested in meshes which are locally refined. Local refinement may be used to resolve geometric features, such as re-entrant corners, cracks, thin gaps, small inclusions, or curved boundaries, or localized solution features, such as singularities or wavefronts. Also, adaptive mesh refinement places small cells around singularities, wavefronts, or regions of high error. As we have already seen, the CFL condition is dictated by the smallest cell, so a few small cells can force a very small time step size, which is wasteful if the rest of the mesh is coarse. However, local mesh refinement is only one way to create such a local CFL bottleneck. A high local wave speed raises the local inverse inequality constant exactly as a small cell does [DHK<sup>+</sup>23, Rem. 11.12]. Moreover, a cell with a high dG polynomial degree raises the same constant [DPE12, Rem. 1.48].

In the terminology of ODE solvers, stiffness describes the numerical situation in which an explicit method is forced to use a very small time step size by stability rather than by the desired accuracy [HW96, §IV.1]. For the present linear ODE system (I.9), whose discrete operator

$\mathcal{L}_h$  is skew-adjoint by (I.6), this stability restriction can also be read from the eigenvalues of  $\mathcal{L}_h$  [HW96, §IV.2]. The inverse inequality (I.7) provides a mesh-dependent upper bound on the spectral radius of the discrete operator  $\mathcal{L}_h$  and hence on the magnitudes of its eigenvalues. The global inverse inequality follows from elementwise bounds involving each cell and its face neighbors [Köh19, Lem. 3.40 and Prop. 3.41]. Writing  $\Omega_K$  for the union of a cell  $K$  and its face neighbors, these bounds take the form

$$\|\mathcal{L}_h x_h\|_K \lesssim \|h^{-1} x_h\|_{\Omega_K}.$$

These local bounds therefore depend on the cell diameters on the patch, the polynomial degree, the mesh regularity, and the operator and material coefficients. We use these local bounds to identify the part of the mesh that causes stiffness.

From now on, we call the part of the mesh where the local inverse bound is large relative to the rest of the mesh the stiff part of the mesh. The stiffness may be caused by small cells, a high local wave speed, or a high local dG polynomial degree. We also include one layer of adjacent cells in the stiff part. The nonstiff part is defined as the remaining larger part of the

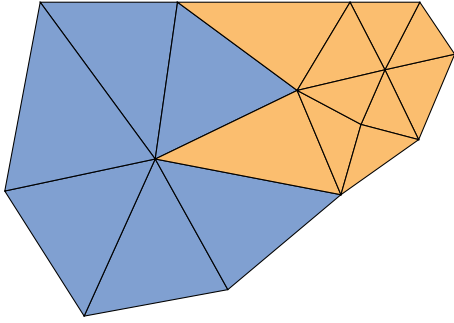


Figure 2: Excerpt of a mesh partitioned into the nonstiff region (blue) and the stiff region (orange). The stiff region includes the fine cells and one adjacent layer of coarse cells.

mesh. Note that the part called nonstiff here is not necessarily nonstiff in the classical sense of ODEs but only relative to the part called stiff in this context. We use a similar distinction for the degrees of freedom (dofs), with the stiff dofs containing the dofs associated with the stiff part of the mesh and the nonstiff dofs containing the remaining dofs. The mesh separation is illustrated in Figure 2. In our setting, the stiff part is spatially localized. Paying the price for an implicit solver on the whole domain is therefore wasteful. We want to apply the expensive, implicit treatment only where it is needed and a cheap explicit method everywhere else.

This is the idea of the locally implicit (LI) method proposed for linear Maxwell’s equations in [Ver11] and analyzed in [HS16]. The extension to Friedrichs’ systems was analyzed in [DHK<sup>+</sup>23]. The LI method treats only the stiff part implicitly, through a Crank-Nicolson-type solve on the small system coming from those dofs. It is unconditionally stable on the stiff part, making the nonstiff CFL condition the only stability factor. Furthermore, it requires solving only a small linear system rather than one global system.

A second approach is local time-stepping (LTS), which is fully explicit and does extra explicit work on the stiff part to relax its CFL condition so that the nonstiff CFL condition becomes the dominating one. Classically, local time-stepping comes from a multirate approach, where the stiff part is advanced by several small substeps for each coarse step on the nonstiff part, see, for example, [DG09] for the second-order wave equation and [GM10] for second-order Maxwell’s equations. The approach proposed in [CH24] for the second-order wave equation is following a different ansatz. It uses a filter function to modify the leapfrog method on the stiff part, relaxing its CFL condition. A special case therein is the stabilized Chebyshev filter, which leads to an explicit LTS method.

Motivated by the filtered leapfrog method of [CH24] for the second-order wave equation, in Chapter 1 we develop a family of local time-integration (LTI) methods for the general class of Friedrichs’ systems (1.1) satisfying a two-field structure discretized in space by a dG method with central fluxes. Our LTI method advances the nonstiff part by the cheap explicit leapfrog method. The stiff part receives a tailored treatment that removes its CFL restriction or weakens it until the nonstiff CFL condition dominates. The goal is that the effective time step size restriction is set by the nonstiff part. The trade-off for a removed or weakened CFL condition of the tailored method is a higher computational cost per step than the plain leapfrog method, as we already observed with the Crank-Nicolson method. However, the extra cost is only paid on the few stiff dofs and is therefore small as long as the stiff part is small. Overall, this results in a problem-adapted and efficient time integrator.

A uniform analysis, based on sufficient conditions on the filter function, then yields stability for the whole family under a CFL condition set by the leapfrog method on the nonstiff part of the mesh and a CFL condition set by the stiff part depending on the chosen filter function. Our main result is to prove that the leapfrog-based local time-integration method is second-order in time and of full order  $k$  in space. Here,  $k$  denotes the dG polynomial degree of the spatial discretization. A rational filter mimicking the Crank–Nicolson method locally leads to the LI method analyzed in [DHK<sup>+</sup>23], which removes the CFL condition of the stiff part completely. A filter based on properly scaled and shifted Chebyshev polynomials of the first kind (1.45) results in an explicit LTS method. Using a degree  $p$  Chebyshev polynomial, the maximal stable time step size on the stiff part scales like  $p$  times the maximum stable time step size of the leapfrog method on the stiff part. Therefore, a moderate degree already lets the nonstiff CFL condition dominate. The filter is never formed explicitly. Its action on a vector can be computed efficiently via a three-term recurrence [Car21, §4.4], resulting in  $p$  matrix-vector products with the matrix associated with the stiff part. This is the only extra cost of the LTS method over the plain leapfrog method. However, one caveat remains. Without stabilization, the explicit filter may become unstable for certain choices of the time step size, as illustrated in Section 1.4.3. Adding stabilization avoids these instabilities at the price of a slightly weaker CFL condition, see [CH22] for details.

The goal of Chapter 2 is to extend the analysis of the local time-integration method developed in Chapter 1 to a setting where we use upwind fluxes instead of central fluxes in the dG discretization. Following the approach of [HS19], we apply the upwind stabilization only on the nonstiff part of the mesh. The resulting method retains the family of filter-based local time-integration schemes from Chapter 1 and remains fully explicit when a polynomial filter is used. We prove stability under the CFL conditions from the central-flux case together with an additional CFL condition caused by the explicit upwind fluxes. Under suitable regularity assumptions and given a dG polynomial degree  $k$ , the method converges with order two in time, with spatial order  $k$  on the stiff part and with spatial order  $k + 1/2$  on the nonstiff part.

## Higher-order local time integration.

Under suitable regularity assumptions, the spatial error estimate in Chapter 1 is of order  $h_{\max}^k$  for dG polynomial degree  $k$ . Increasing the polynomial degree can therefore improve the spatial convergence rate without refining the mesh. Once the spatial error is sufficiently small, the second-order temporal error can dominate the total error and limit the benefit of the higher spatial degree. As discussed earlier, reducing the time step size requires more steps and may

therefore increase the contribution of accumulated roundoff errors to the global error.

Chapter 3 introduces a family of time integrators that reach a higher temporal order and are tailored to the same local efficiency principle as the local time-integration methods of Chapters 1 and 2. The whole system is integrated with a higher-order, unconditionally stable implicit method, a Gauß collocation method [HW96, §IV.5]. A Gauß collocation method is a Runge–Kutta method with nodes coming from the Gauß–Legendre quadrature points. It is unconditionally stable and therefore carries no CFL restriction. With  $s$  stages, it has classical order  $2s$ , the highest possible order per stage for Runge–Kutta methods. Gauß collocation methods are well understood and have been analyzed for linear Maxwell’s equations in [HP15] and for quasilinear Maxwell’s equations in [HPS18]. Moreover, Gauß collocation methods are symplectic [HLW06, §VI.4]. The stability function of a Gauß collocation method maps the imaginary axis to the unit circle (see [HW96, §IV.3] and [Gau97, §5.9.2]) and is therefore conservative on purely oscillatory modes. This matches the energy-conserving, purely imaginary spectrum of the discrete operator  $\mathcal{L}_h$ .

However, applying an implicit time integrator reintroduces the implicit price, i.e., solving large, globally coupled linear systems at every time step, which we aimed to avoid. Even though an efficient implementation of a Gauß collocation method reduces the large coupled linear system to  $s$  decoupled smaller but still large linear systems, one for each stage, a direct factorization may still be infeasible for large-scale problems because of the high memory consumption of the factorization matrices. Therefore, it is natural to employ a Krylov subspace method [Saa03] to solve the linear system iteratively at each stage. After decoupling the stages, the linear system at each stage has a coefficient matrix that is either symmetric positive definite or complex symmetric. Thus, we apply either the conjugate gradient (CG) method or the complex symmetric quasi-minimal residual (QMR) method [FN91, FN94]. The relevant cost is then proportional to the number of Krylov iterations. However, the number of iterations needed to reach a given accuracy depends on the spectrum of the linear system matrix [HL98]. In our scenario, the mesh contains a small stiff part that raises the spectral radius of the linear system matrices, increasing the number of iterations required. To avoid this, we construct a symmetric positive definite preconditioner for the Krylov solvers that is motivated by the same local efficiency principle as the local time-integration methods of Chapters 1 and 2. The preconditioner is based on splitting the system matrix into a stiff and a nonstiff part, where we use the stiff contribution to precondition the whole system. As long as the stiff part is small, the preconditioner is cheap to apply. The analysis proves that the field of values of the preconditioned matrix is bounded independently of the stiff part of the mesh. Consequently, the number of iterations needed to reach a fixed tolerance is also bounded independently of the stiff part of the mesh.

Even though we only consider linear problems in this thesis, the preconditioning technique carries over to nonlinear problems, where linear systems of the same type arise within a Newton-type iteration.

The stiff and nonstiff split of Chapter 1 thus reappears in a different place. In Chapter 1, it defines a time integrator which itself acts locally, whereas in Chapter 3 the integrator is globally implicit and the locality re-enters through the preconditioner. The philosophy is the same, namely to spend extra work only on a few stiff cells, achieving an overall efficiency boost.

**Contributions of Part I.** Part I develops two different local time-integration approaches for two-field Friedrichs’ systems. The first contribution is a general leapfrog-based local time-integration

framework for central-flux dG discretizations, which contains explicit local time-stepping and locally implicit variants in one unified formulation. We perform a stability and error analysis in which the admissible time step size is governed by the nonstiff part of the mesh, provided the filter satisfies its stiff part stability condition, which is weaker than the global leapfrog stability condition. The local time-integration method is then extended to an upwind-flux dG discretization where dissipative interface penalizations are applied on the nonstiff part of the mesh. The second contribution is a higher-order variant based on Gauß collocation methods in time. It removes the explicit stability restriction from the time integrator, while accuracy and the cost of the implicit solves still constrain the usable time step size. Its local character does not lie in the time step itself but in the preconditioner used for the arising linear systems. Our analysis shows that the preconditioned Krylov iteration count needed to reach a fixed tolerance is controlled independently of the stiff part of the mesh. Taken together, these results show that the same stiff and nonstiff split can be used at different algorithmic layers: inside leapfrog-based time integrators and inside linear solvers used in implicit time integration.

## Gaussian wave packets for semiclassical dynamics

Part II changes the source of the limiting factor. In Part I, localized stiffness caused by the spatial discretization or by materials with high wave speeds limit the efficiency of standard numerical methods. In Part II, a small parameter makes the wave function highly oscillatory in space and time on an unbounded spatial domain  $\Omega = \mathbb{R}^d$ . Thus, there is no nonstiff region to isolate. The guiding principle remains the same, namely to identify the problematic part and treat it by a method adapted to its structure. The mechanism, however, is different. Instead of a local split in space, the next part uses a mesh-free reduced description from [BDH<sup>+</sup>24] in which a Gaussian wave packet ansatz captures the rapid oscillations. This leaves a small system of parameter ODEs which are independent of oscillations induced by the small parameter. The aim of Part II is to approximate the solution of these ODEs efficiently while (nearly) preserving important physical quantities. As a physically relevant benchmark, we compare our methods against the analytical solution of the Penning trap.

**The magnetic Schrödinger equation.** The model problem of Part II is the semiclassical, time-dependent magnetic Schrödinger equation

$$i\varepsilon \partial_t \psi(t) = H(t)\psi(t), \quad \psi(0) = \psi_0, \quad t \in \mathbb{R}, \quad (\text{I.10a})$$

on  $\mathbb{R}^d$  with initial value  $\psi_0$  and magnetic Hamiltonian

$$H(t) = \frac{1}{2}(-i\varepsilon \nabla_x - A(t, \cdot))^2 + \phi(t, \cdot). \quad (\text{I.10b})$$

Here, the square in (I.10b) denotes the dot product of the vector-valued differential operator with itself. Moreover,  $\psi$  denotes the complex-valued wave function,  $A$  the magnetic vector potential chosen in Coulomb gauge,  $\varphi$  the electric scalar potential, and  $0 < \varepsilon \ll 1$  the small semiclassical parameter, often obtained by rescaling Planck's constant  $\hbar$  to a dimensionless quantity. The magnetic Schrödinger equation (I.10) is a fundamental model in quantum mechanics, describing the evolution of charged particles in the presence of external electromagnetic fields. One physically relevant example is the Penning trap considered in Section 4.2 (see, e.g., [BG84]). There, a charged particle, such as an electron or a proton, is confined by magnetic coils and

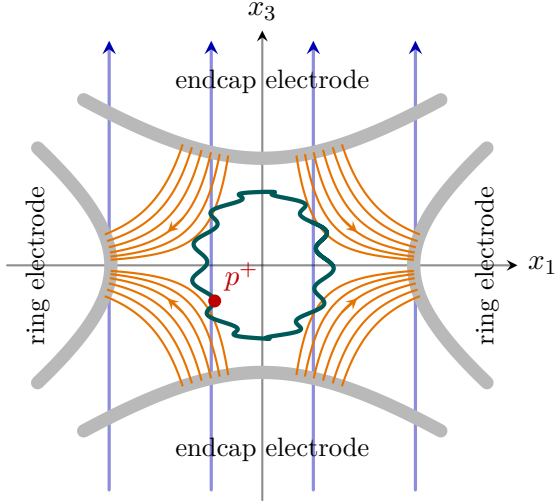


Figure 3: Gray curves denote the endcap and ring electrodes of the Penning trap in the  $(x_1, x_3)$ -plane. The orange curves indicate electric field lines, and the blue arrows represent the homogeneous magnetic field. The green curve shows an exemplary trajectory of a proton, marked by the red dot labeled  $p^+$ . The figure was created with the assistance of OpenAI's ChatGPT 5.5.

electrodes. The trap features a static magnetic field along the  $x_3$ -direction and a quadrupole-like static electric field that is rotationally symmetric about the  $x_3$ -axis. A classical point particle moving in such an electromagnetic field configuration executes an oscillatory motion along the  $x_3$ -axis while simultaneously executing an epicyclic motion in the  $x_1x_2$ -plane, where the low-frequency magnetron orbit is combined with high-frequency cyclotron orbits. For the proton trap parameters listed in Table 4.1 and the nondimensionalization used in Section 4.2, this gives  $\varepsilon \approx 1.2 \times 10^{-8}$ .

From a numerical point of view, approximating (I.10) raises three major problems. First, it is a high-dimensional problem. For an  $N$ -particle system in three spatial dimensions, the configuration space has dimension  $d = 3N$ . Further, the computational domain  $\mathbb{R}^d$  is naturally unbounded, and thus most numerical methods require truncation before discretization. For the method-of-lines approach (first discretize space, then time), high dimension combined with an unbounded domain leads to very large semidiscrete systems that then have to be integrated in time. Another challenge is given by the rapid oscillations induced by the small semiclassical parameter  $\varepsilon$ . For a direct method-of-lines discretization, these oscillations require fine resolution in space and time, so standard time integration of the resulting large system quickly becomes impracticable in the semiclassical regime [LL20].

**Gaussian wave packets.** A localized initial state provides a reduced description. For finite times, it remains concentrated near the corresponding classical trajectory [Lub08]. The relevant information is then carried by one moving center, an oscillatory phase, and the local shape of the packet.

A single Gaussian wave packet uses exactly this structure. It is of the form

$$u(t, x) = \exp\left(\frac{i}{\varepsilon} \left( \frac{1}{2} (x - q)^\top \mathcal{C} (x - q) + (x - q)^\top p + \zeta \right)\right),$$

with time-dependent position and momentum centers  $q, p \in \mathbb{R}^d$ , a complex-symmetric width matrix  $\mathcal{C}$  having a positive definite imaginary part, and a complex phase  $\zeta \in \mathbb{C}$ .

The position  $q$  locates the packet. Splitting  $\mathcal{C} = \mathcal{C}_R + i\mathcal{C}_I$  and  $\zeta = \zeta_R + i\zeta_I$  into their real and imaginary parts separates the Gaussian into its spatial envelope and its carrier phase. The

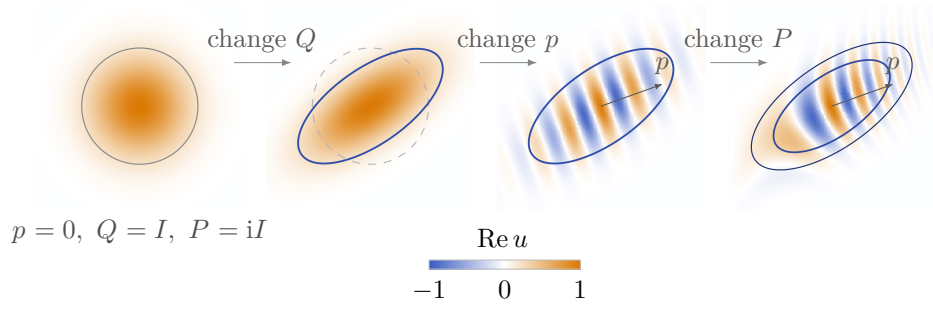


Figure 4.: Successive construction of a wave packet  $u$ . The first panel shows the initial packet with  $p = 0$ ,  $Q = I$ , and  $P = iI$ . Each arrow changes only the indicated parameter. Changing  $Q$  deforms the Gaussian envelope, changing  $p$  adds linear phase oscillations, and changing  $P$  adds a quadratic phase contribution. Colors show the signed real part  $\text{Re } u$  on the scale from  $-1$  to  $1$ , while the contours indicate levels of the Gaussian envelope. The dashed contour retains the preceding envelope for comparison. The figure was created with the assistance of OpenAI’s ChatGPT 5.5.

matrix  $\mathcal{C}_I$  determines the shape, orientation, and width of the spatial envelope,

$$|u(t, x)|^2 = \exp\left(-\frac{1}{\varepsilon}(x - q)^\top \mathcal{C}_I(x - q) - \frac{2}{\varepsilon}\zeta_I\right),$$

whereas  $\mathcal{C}_R$  determines the quadratic curvature of the oscillatory phase. More precisely, the phase function is given by

$$S(t, x) = \frac{1}{2}(x - q)^\top \mathcal{C}_R(x - q) + (x - q)^\top p + \zeta_R.$$

Thus,  $\nabla_x S(t, x) = \mathcal{C}_R(x - q) + p$ , so  $p$  is the constant part of the phase gradient, while  $\mathcal{C}_R$  determines how this phase gradient varies across the packet.

Hagedorn’s parametrization writes the width matrix as  $\mathcal{C} = PQ^{-1}$ , with matrices  $P$  and  $Q$  satisfying the structural conditions recalled in Section 4.3.4, see also [LL20, §3.5]. This parametrization separates the effects of the envelope and the phase gradient. The spatial envelope is determined only by  $QQ^*$ , while the curvature of the carrier phase is determined by

$$\mathcal{C}_R = \text{Re}(PQ^{-1}).$$

Therefore,  $P$  affects the position-dependent phase gradient only together with  $Q$ .

The complex phase  $\zeta$  contains the global phase and the normalization. Its real part  $\zeta_R$  shifts the phase of the whole packet by a constant. The imaginary part  $\zeta_I$  fixes the overall normalization of the Gaussian, i.e.,

$$\|u\|_{L^2(\mathbb{R}^d)}^2 = \exp\left(-\frac{2}{\varepsilon}\zeta_I\right) (\varepsilon\pi)^{d/2} |\det Q|. \quad (\text{I.11})$$

Thus, normalization ties  $\zeta_I$  to  $Q$ . The roles of the parameters  $p$ ,  $Q$ , and  $P$  are illustrated in Figure 4, which shows the successive construction of a wave packet. Gaussian wave packets are natural states, not only convenient trial functions. They include coherent and squeezed states, which are standard localized states in quantum mechanics [RC21]. They also have a practical advantage because their moments and many expectation values can be evaluated in closed form.

**Classical, frozen, and variational Gaussian dynamics.** Gaussian wave packets have a long history in semiclassical analysis and chemical quantum dynamics. A broad overview is given in [LL20, §3.8], which discusses the time-dependent variational principle from the work of Dirac [Dir30] and Frenkel [Fre34] to Heller’s variational Gaussian wave packets [Hel76], as well as later error analyses [Lub08].

These variants of Gaussian methods make different compromises. Frozen Gaussian methods keep the width matrix fixed [Hel81]. A single frozen packet cannot describe spreading, so frozen Gaussians are typically used in superpositions. The Herman–Kluk propagator is a prominent superposition method [HK84, LS17]. Hagedorn wave packets play a related but different role. They form orthonormal families generated from a ground state Gaussian by raising and lowering operators [Hag98]. This framework is central in rigorous semiclassical analysis and has also been applied to magnetic Schrödinger equations [Zho14, GR21].

The variational Gaussian dynamics utilized in Chapter 4 is the branch closest to Heller’s time-dependent Gaussian approximation [Hel76]. It does not require a mesh and remains low dimensional. Furthermore, it determines the parameter evolution by projecting the full Schrödinger dynamics onto the tangent space of the Gaussian manifold. In the nonmagnetic setting discussed in [LL20, Lub08], these variational dynamics conserve norm and energy and have error bounds that are not valid for diagonal or fixed-width restrictions, as emphasized in [LL20, §3.8]. The magnetic extension used in Chapter 4 was developed in [BDH<sup>+</sup>24], where the Hamiltonian contains a time-dependent magnetic vector potential.

**Variational equations of motion.** The time evolution of the parameters is determined by the Dirac–Frenkel variational principle. At each time, the Schrödinger velocity is projected onto the tangent space of the Gaussian manifold. Equivalently, one chooses the tangent vector for which the residual in the Schrödinger equation is minimal.

For the magnetic Schrödinger equation, the variational principle was applied in [BDH<sup>+</sup>24] to derive the equations of motion

$$\dot{q} = \langle \partial_p h \rangle_u, \quad (\text{I.12a})$$

$$\dot{p} = -\langle \partial_q h \rangle_u, \quad (\text{I.12b})$$

$$\dot{\mathcal{C}} = -\mathcal{B}(\mathcal{C}), \quad (\text{I.12c})$$

$$\dot{\zeta} = -\langle h \rangle_u + \frac{\varepsilon}{4} \text{tr} \left( \mathcal{B}(\mathcal{C}) \mathcal{C}_I^{-1} \right) + p^\top \langle \partial_p h \rangle_u, \quad (\text{I.12d})$$

with

$$\mathcal{B}(\mathcal{C}) = \begin{pmatrix} \text{Id} & \mathcal{C} \end{pmatrix} \langle \nabla^2 h \rangle_u \begin{pmatrix} \text{Id} \\ \mathcal{C} \end{pmatrix}, \quad (\text{I.12e})$$

and the classical Hamiltonian for a charged particle in a magnetic field is

$$h(t, q, p) = \frac{1}{2} (p - A(t, q))^2 + \phi(t, q). \quad (\text{I.12f})$$

Here, for a suitable observable  $A$ , we write

$$\langle A \rangle_u = \langle u, Au \rangle = \int_{\mathbb{R}^d} \overline{u(x)} \cdot (Au)(x) \, dx$$

for its expectation value in  $u$ , details are in Section 4.1. For phase-space functions such as  $h$  and its derivatives, the brackets denote expectation values of their Weyl quantizations, applied

componentwise to vector- and matrix-valued functions [LL20, §6.1]. Thus, the forces in (I.12) are not evaluated only at the center  $q$ , but averaged over the current packet.

This reduced model retains several structural properties of the Schrödinger equation. It preserves the  $L^2$ -norm of the wave function and, for time-independent Hamiltonians, its energy [BDH<sup>+</sup>24, Lem. 3.6]. For Gaussian initial data, the variational dynamics are exact if the magnetic potential is linear and the electric potential is quadratic in position [BDH<sup>+</sup>24, Prop. 2.5]. This includes the Penning trap model used in Chapter 4. Under suitable smoothness and growth assumptions on the electromagnetic potentials, the wave-function error of the variational approximation is of order  $\sqrt{\varepsilon}$  on fixed finite time intervals. For suitable observables, the corresponding expectation-value error is of order  $\varepsilon^2$  [BDH<sup>+</sup>24, Theorems 3.8 and 3.10].

The parameter equations also explain the relation to classical mechanics. If  $\varepsilon$  tends to zero, the Gaussian averages collapse to point evaluations at the center  $q$ , see [LL20, Lem. 3.15]. Then the equations for  $q$  and  $p$  tend to the classical Hamiltonian equations generated by  $h$ . The variational packet may therefore be viewed as a quantum correction of the classical trajectory dynamics, where the width keeps track of the evolving uncertainty.

In Hagedorn's parametrization  $\mathcal{C} = PQ^{-1}$ , the width equation (I.12c) is equivalent to the pair of coupled linear equations

$$\dot{Q} = \langle \partial_{pq} h \rangle_u Q + \langle \partial_{pp} h \rangle_u P, \quad (\text{I.13a})$$

$$\dot{P} = -\langle \partial_{qq} h \rangle_u Q - \langle \partial_{qp} h \rangle_u P, \quad (\text{I.13b})$$

see [BDH<sup>+</sup>24, Theorem 4.2].

## Time integration of Gaussian parameters

The variational Gaussian approximation has replaced the magnetic Schrödinger equation by a small system of ODEs for the parameters of one packet. The next step is to evolve these parameters in time.

The starting point is the Boris algorithm [Bor70], originally introduced for the classical equations of motion of a charged particle in an electromagnetic field. For a magnetic vector potential  $A$  in spatial dimension  $d$ , let

$$\mathcal{B}_A = J_A^\top - J_A$$

denote the skew-symmetric magnetic matrix, where  $J_A$  is the Jacobian of  $A$ . The classical equations of motion take the form

$$\dot{q} = v, \quad \dot{v} = \mathcal{B}_A v + E,$$

where  $E$  is the electric field. In three dimensions,  $\mathcal{B}_A v = v \times B$ , which recovers the usual formulation in terms of the magnetic field  $B = \nabla \times A$ . The Boris method is one of the standard particle pushers in plasma-physics particle-in-cell codes [BL18]. It updates the velocity by a kick-rotate-kick procedure. First, the electric field is applied over a half step. Second, a rotation with the magnetic part is applied. Third, the electric field is applied over another half step. The position is then advanced by a drift with the updated velocity. The Boris method is of second-order [QZX<sup>+</sup>13]. In position and canonical momentum, the classical charged-particle equations have a Hamiltonian formulation. However, the Boris method preserves phase-space volume while

generally failing to preserve the associated symplectic structure [QZX<sup>+</sup>13]. Related long-time analyses for charged-particle dynamics in strong magnetic fields are given in [HL20, HLW20].

Chapter 4 extends this idea from classical charged-particle dynamics to the parameter equations (I.12) of the variational Gaussian wave packet ansatz. It first introduces the averaged magnetic momenta

$$v = p - \langle A \rangle_u, \quad \Upsilon = P - \langle J_A \rangle_u Q.$$

In these variables, the center and width equations take the Euler–Lagrange-type form

$$\begin{aligned} \dot{q} &= v, & \dot{v} &= \langle \mathcal{B}_A \rangle_u v + E_{\text{eff}}, \\ \dot{Q} &= \Upsilon, & \dot{\Upsilon} &= \langle \mathcal{B}_A \rangle_u \Upsilon + S_{\text{eff}} Q, \end{aligned}$$

see Theorem 4.3. Here,  $E_{\text{eff}}$  and  $S_{\text{eff}}$  collect additional averaged force terms.

The phase, however, is updated separately. Its imaginary part is computed from the discrete normalization relation

$$\zeta_{\text{I}}^{n+1} = \zeta_{\text{I}}^n + \frac{\varepsilon}{2} \left( \ln |\det Q^{n+1}| - \ln |\det Q^n| \right),$$

which enforces preservation of the  $L^2$ -norm at every step. The real part of the phase is advanced by a midpoint rule.

Chapter 4 also introduces a modified classical fourth-order Runge–Kutta (RK4) method. All parameters except the imaginary part of the phase are treated by the standard RK4 stages. At each stage, the imaginary phase is instead updated from the same normalization relation as in the Boris-type method. Consequently, the modified RK4 method preserves the  $L^2$ -norm of the Gaussian packet by construction, in contrast to the unmodified RK4 method.

The numerical experiments in Chapter 4 use the norm and the energy as diagnostics for this structure-preserving aim. For the tested examples, the Boris-type method and the modified RK4 method preserve the  $L^2$ -norm up to machine precision. The computed energy error of the Boris-type method remains bounded and oscillatory in the reported long-time runs, whereas the standard RK4 method and the modified RK4 method show an energy drift, which could be reduced significantly for the modified method. Thus, the variational Gaussian and Boris-type discretizations close the time-integration discussion for Part II in the same spirit as the local methods of Part I: they isolate the difficult component and treat it by a method adapted to the structure of the problem.

Beyond the present thesis, recent work by Funk et al. compares a mesh-free wave-packet-based solver with a benchmark grid-based solver for the time-dependent Schrödinger equation and demonstrates accurate mesh-free simulations of physically relevant examples with significantly shorter simulation times than those obtained with the grid-based solver [FML<sup>+</sup>26]. Moreover, recent work by Merk and Lasser develops structure-preserving time integrators for variational Gaussian wave packet dynamics that exhibit near-conservation of the averaged Hamiltonian over long time intervals [ML26].

## Outline of the thesis

The two parts of the thesis address different computational bottlenecks by related principles. In Part I, a spatially localized stiff component would force the whole semidiscrete Friedrichs’ system to use the small time step required only near a refined part of the mesh. Local time integration avoids this global restriction by separating stiff and nonstiff degrees of freedom. In

Part II, the computational bottleneck is not localized in space. It is the oscillation caused by the small semiclassical parameter, which is captured by the mesh-free variational Gaussian wave packet ansatz. In both cases, the goal is to keep the low-cost part of the computation inexpensive, while treating the restrictive component by a method adapted to its structure.

Within Part I, the same fine/coarse split is used at two different algorithmic levels. Chapter 1 and Chapter 2 place it directly inside a leapfrog-based time integrator. Chapter 3 instead uses implicit Gauß collocation for higher-order time integration and moves the local construction into a preconditioner for the arising Krylov iteration. Thus, the local idea is not tied to one particular time step formula. It can also enter through the linear algebra used to realize an implicit method.

Chapter 1 constructs and analyzes a second-order, leapfrog-based local time-integration class for two-field Friedrichs' systems with central-flux dG discretization. The filter formulation contains fully explicit local time-stepping variants as well as a locally implicit variant, and the stability condition is governed by the nonstiff part of the mesh, once the filter's relaxed stiff part CFL condition is satisfied. Chapter 2 extends the second-order local time-integration framework to a stabilized dG discretization. The stabilization terms add interface dissipation to the central-flux discretization, at the price of replacing exact energy conservation by a discrete energy identity with controlled dissipation. Chapter 3 develops a higher-order variant based on Gauß collocation. The resulting complex-symmetric linear systems are solved by preconditioned QMR, and the analysis shows that the number of iterations needed to reach a fixed tolerance is bounded independently of the fine mesh. Chapter 4 then turns to the semiclassical magnetic Schrödinger equation. There, a variational Gaussian wave packet reduces the evolution to parameter ODEs, and a Boris-type integrator or a modified fourth-order Runge–Kutta method advance these equations while preserving the  $L^2$ -norm of the packet by construction.

**Part I.**

**Local time integration for Friedrichs'  
systems**



# CHAPTER 1

---

## Local time integration for Friedrichs' systems

The content of this chapter is published in [HS26].

In this paper, we address the full discretization of Friedrichs' systems with a two-field structure, such as Maxwell's equations or the acoustic wave equation in div-grad form, cf. [DHK<sup>+</sup>23]. We focus on a discontinuous Galerkin space discretization applied to a locally refined mesh or a small region with high wave speed. This results in a stiff system of ordinary differential equations, where the stiffness is mainly caused by a small region of the spatial mesh. When using explicit time-integration schemes, the time stepsize is severely restricted by a few spatial elements, leading to a loss of efficiency. As a remedy, we propose and analyze a general leapfrog-based scheme which is motivated by [CH22]. The new, fully explicit, local time-integration method filters the stiff part of the system in such a way that its CFL condition is significantly weaker than that of the leapfrog scheme while its computational cost is only slightly larger. For this scheme, the filter function is a suitably scaled and shifted Chebyshev polynomial. While our main interest is in explicit local time-stepping schemes, the filter functions can be much more general, for instance, a certain rational function leads to the locally implicit method, proposed and analyzed in [HS16]. Our analysis provides sufficient conditions on the filter function to ensure full order of convergence in space and second order in time for the whole class of local time-integration schemes.

### 1.1. Introduction

The aim of this paper is to construct and analyze a new local time-integration (LTI) method for the discontinuous Galerkin (dG) space discretization of Friedrichs' systems in a two-field structure. Using an explicit time-integrator like the leapfrog method on the full spatial domain necessitates satisfying a Courant–Friedrichs–Lewy (CFL) condition, meaning that the time stepsize  $\tau$  is proportional to the inverse of the smallest diameter within the spatial mesh. On the other hand, implicit methods like the Crank–Nicolson method are unconditionally stable, but require the solution of large linear systems of equations. We are interested in problems, where the CFL condition is dominated by a small number of mesh elements, i.e., tiny elements or elements with a high wave speed. In this situation, standard explicit or implicit methods are

inefficient. The basic idea of LTI methods is to use an explicit time-integrator, for example, the leapfrog method, on the large, nonstiff part of the system and to couple it to a tailored method on the small, stiff part. The modified method can either be an explicit method with a weaker CFL condition (resulting in a local time-stepping (LTS) method) or an unconditionally stable implicit scheme (leading to a locally implicit (LI) method). Overall one aims at using a method with a CFL condition which only depends on the nonstiff part of the system. Since the higher computational cost of the modified method arises only on a small part of the degrees of freedom (dofs), the construction leads to a problem-adapted and efficient LTI scheme.

In recent years, many LTI schemes have been introduced for different wave-type equations. LTS methods for the acoustic wave equation in second-order formulation were constructed in [AM17, CH24, DG09, DG15, GMM15, GMS21, GM13a, GM13b] and for Maxwell's equations in second-order formulation in [GM10, KFL20, MPF<sup>+</sup>08, Pip06], and LI methods for Maxwell's equations in [DLM13, DLM14, DLM17, DFF<sup>+</sup>10, HS16, HS19, RB02, Ver11], for instance. Moreover, in [DHK<sup>+</sup>23] the LI method from [HS16] was generalized to Friedrichs' systems.

However, to the best of our knowledge, fully explicit LTS methods for Friedrichs' systems have not been considered so far. Here, we construct and analyze a class of leapfrog-based LTI schemes motivated by [CH24]. We will show that the method is stable under a CFL condition which is independent of the stiff part of the problem and is of second order in time and optimal order in space. As a special case, one variant of the LTI scheme corresponds to the LI method from [DHK<sup>+</sup>23] and the theoretical results reduce to the ones given there.

**Outline** The paper is structured as follows. First, in Section 1.2, we briefly review Friedrichs' systems and introduce the notation which we will use throughout the paper.

Our main results are contained in Section 1.3. Here, we start with an introduction of the splitting of the discrete operators into a stiff and a nonstiff part. Then, we present the new LTI scheme based on the leapfrog method with a general filter function and study its stability. Important examples leading to LTS and LI methods are discussed in detail. For the whole class of methods, we prove stability if a time stepsize restriction based on properties of the filter function is satisfied. Moreover, we prove that LTI methods satisfy error bounds with optimal order in space and second order in time.

At last, in Section 1.4, we present numerical experiments which substantiate our theoretical results for Maxwell's equations. Moreover, we show that the LTS method outperforms the LI and the leapfrog method on these test problems.

## 1.2. Friedrichs' systems

Let  $\Omega \subset \mathbb{R}^d$  be an open, polygonal domain and denote its boundary by  $\Gamma = \partial\Omega$ .

Following [DHK<sup>+</sup>23, §11.2] we consider the two-field structured Friedrichs' system

$$\partial_t u = \mathcal{L}_v v + g_u, \quad \text{on } \Omega \times \mathbb{R}_+, \quad (1.1a)$$

$$\partial_t v = \mathcal{L}_u u + g_v, \quad \text{on } \Omega \times \mathbb{R}_+, \quad (1.1b)$$

$$u(0) = u^0, \quad v(0) = v^0, \quad \text{on } \Omega$$

with initial values  $u^0, v^0$  and right-hand-side  $g$ . The Friedrichs' operators  $\mathcal{L}_u$  and  $\mathcal{L}_v$  are given as

$$\mathcal{M}_v \mathcal{L}_u u = \sum_{i=1}^d L_i \partial_i u, \quad \mathcal{M}_u \mathcal{L}_v v = \sum_{i=1}^d L_i^T \partial_i v, \quad (1.2)$$

with  $L_i \in \mathbb{R}^{m_v \times m_u}$  for  $i = 0, \dots, d$ . The material tensors  $\mathcal{M}_u \in L^\infty(\Omega)^{m_u \times m_u}$ ,  $\mathcal{M}_v \in L^\infty(\Omega)^{m_v \times m_v}$  are assumed to be symmetric and positive definite, and  $m_u, m_v \in \mathbb{N}$  denote the number of components of  $u$  and  $v$ . The boundary conditions of (1.1) are embodied into the domains  $\mathcal{D}(\mathcal{L}_u)$  and  $\mathcal{D}(\mathcal{L}_v)$ , see [DPE12, DHK<sup>+</sup>23] for details. We further assume that the boundary conditions do not introduce damping, cf. [DHK<sup>+</sup>23, Assump. 11.8]. This leads to the important adjointness property

$$(\mathcal{L}_u u, v) = -(u, \mathcal{L}_v v) \quad \text{for all } u \in \mathcal{D}(\mathcal{L}_u), v \in \mathcal{D}(\mathcal{L}_v). \quad (1.3)$$

Special cases are linear Maxwell's equations or the acoustic wave equation in a div-grad formulation. For details see [DPE12, §7.1] or [DHK<sup>+</sup>23, §9.3].

### 1.2.1. Notation

Let  $K \subseteq \mathbb{R}^d$  be open. For two vector-valued functions  $u, \tilde{u} \in L^2(K)^{m_u}$  we denote the  $L^2$ -inner product by

$$(u, \tilde{u})_{L^2(K)} = \int_K u \cdot \tilde{u} \, dx$$

and the  $L^2$ -inner product weighted by  $\mathcal{M}_u$  as

$$(u, \tilde{u})_K = (u, \tilde{u})_{\mathcal{M}_u, K} = (\mathcal{M}_u u, \tilde{u})_{L^2(K)},$$

where we suppress the indices  $\mathcal{M}_u$  whenever possible. The inner product weighted by  $\mathcal{M}_v$  is defined analogously. From the arguments of the inner product it is clear which weight is used. The induced norms are denoted accordingly.

Furthermore, we denote by  $\partial_i$  the  $i$ th weak derivative,  $i = 1, \dots, d$ . Higher derivatives are denoted by  $\partial^\alpha = \partial_1^{\alpha_1} \dots \partial_d^{\alpha_d}$  for a multi index  $\alpha \in \mathbb{N}_0^d$ , and we write  $|\alpha| = \alpha_1 + \dots + \alpha_d$ .

The Sobolev spaces in  $L^2(K)$  are denoted as  $H^\ell(K)$ ,  $\ell \in \mathbb{N}_0$ , and are equipped with the weighted norms

$$\|u\|_{\ell, K}^2 = \|u\|_{\mathcal{M}_u, \ell, K}^2 = \sum_{j=0}^{\ell} |u|_{\mathcal{M}_u, j, K}^2, \quad |u|_{j, K}^2 = |u|_{\mathcal{M}_u, j, K}^2 = \sum_{|\alpha|=j} \|\partial^\alpha u\|_{\mathcal{M}_u, K}^2.$$

Again we suppress the index  $\mathcal{M}_u$  whenever possible.

In what follows, we recall some notation and results from [DPE12, DHK<sup>+</sup>23]. We denote by  $\mathcal{T}_h$  meshes of  $\Omega$  which we assume to be admissible [DPE12, Def. 1.57]. By  $h_K$  we denote the diameter of a cell  $K \in \mathcal{T}_h$ . We further define  $h \in L^\infty(\Omega)$  by  $h|_K = h_K$  and  $h_{\max} = \max_K h_K$  as well as  $h_{\min} = \min_K h_K$ .

For each component of (1.1) we denote by

$$\begin{aligned} \mathbb{V}_h^u &= \{u \in L^2(\Omega) \mid u|_K \in \mathbb{P}_d^k(K) \text{ for all } K \in \mathcal{T}_h\}^{m_u}, \\ \mathbb{V}_h^v &= \{v \in L^2(\Omega) \mid v|_K \in \mathbb{P}_d^k(K) \text{ for all } K \in \mathcal{T}_h\}^{m_v}, \end{aligned}$$

the spaces of piecewise polynomials of total degree at most  $k$  and the product space by  $\mathbb{V}_h = \mathbb{V}_h^u \times \mathbb{V}_h^v$ .

Furthermore, we symbolize the  $L^2$ -projection w.r.t. the weighted inner product  $(\cdot, \cdot)_\Omega$  by  $\pi_h: L^2(\Omega) \rightarrow \mathbb{V}_h$ , see [DHK<sup>+</sup>23] for details.

At last, we recall the broken Sobolev spaces

$$H^\ell(\mathcal{T}_h) = \{u \in L^2(\Omega) \mid u|_K \in H^\ell(K) \text{ for all } K \in \mathcal{T}_h\},$$

consisting of functions which are piecewise in the Sobolev space of order  $\ell$ . The space  $H^\ell(\mathcal{T}_h)$  is equipped with the norm

$$\|u\|_{\ell, \mathcal{T}_h}^2 = \sum_{j=0}^{\ell} |u|_{j, \mathcal{T}_h}^2, \quad |u|_{j, \mathcal{T}_h}^2 = \sum_{K \in \mathcal{T}_h} |u|_{j, K}^2.$$

### 1.3. Local time-integration

After a central-fluxes dG discretization [DHK<sup>+</sup>23, EG21, HK20], the spatially discrete wave-type problem (1.1) takes the form

$$\partial_t u_h = \mathcal{L}_{v,h} v_h + g_{u,h}, \quad \text{on } \mathbb{R}_+, \quad (1.4a)$$

$$\partial_t v_h = \mathcal{L}_{u,h} u_h + g_{v,h}, \quad \text{on } \mathbb{R}_+, \quad (1.4b)$$

$$u_h(0) = \pi_h u^0, \quad v_h(0) = \pi_h v^0, \quad g_h = \pi_h g, \quad (1.4c)$$

where the discrete operators  $\mathcal{L}_{u,h}, \mathcal{L}_{v,h}$  inherit the adjointness property (1.3), i.e.,

$$(\mathcal{L}_{u,h} u_h, v_h)_\Omega = -(u_h, \mathcal{L}_{v,h} v_h)_\Omega \quad \text{for all } u_h \in \mathbb{V}_h^u, v_h \in \mathbb{V}_h^v. \quad (1.5)$$

Before we present our local time-integration scheme, we follow [HS16] and split the mesh into two parts

$$\mathcal{T}_h = \mathcal{T}_{h,f} \dot{\cup} \mathcal{T}_{h,c}. \quad (1.6)$$

The set  $\mathcal{T}_{h,f}$  contains all elements with a small diameter or a high wave speed and the set  $\mathcal{T}_{h,c}$  all others. We further define

$$h_{c,\min} = \min_{K \in \mathcal{T}_{h,c}} h_K, \quad h_{f,\min} = \min_{K \in \mathcal{T}_{h,f}} h_K. \quad (1.7)$$

Our approach necessitates an additional partition of the mesh such that the set  $\mathcal{T}_{h,m}$  contains the set  $\mathcal{T}_{h,f}$  and the neighbors of elements in  $\mathcal{T}_{h,f}$  and  $\mathcal{T}_{h,lf}$  only contains coarse elements with coarse neighbors, i.e.,

$$\begin{aligned} \mathcal{T}_{h,m} &= \{K \in \mathcal{T}_h : \exists K_f \in \mathcal{T}_{h,f} \text{ s.t. } K, K_f \text{ share a face}\}, \\ \mathcal{T}_{h,lf} &= \mathcal{T}_h \setminus \mathcal{T}_{h,m}. \end{aligned}$$

These extra coarse elements in  $\mathcal{T}_{h,m}$  are essential to the independence of small mesh diameters or high wave speeds in the CFL condition, since mesh diameters and material properties enter through the flux terms in the definition [HK20, Def. 4.8] of the discrete operators  $\mathcal{L}_{u,h}$  and  $\mathcal{L}_{v,h}$  from the element itself and from the neighboring element sharing the same face by the discrete trace inequality, cf. [HK20, Proof of Prop. 4.13].

We further define the indicator functions

$$\chi_m u = \begin{cases} u, & \text{on } K \in \mathcal{T}_{h,m}, \\ 0, & \text{on } K \in \mathcal{T}_{h,lf}, \end{cases} \quad \text{and} \quad \chi_{lf} u = \begin{cases} 0, & \text{on } K \in \mathcal{T}_{h,m}, \\ u, & \text{on } K \in \mathcal{T}_{h,lf}. \end{cases} \quad (1.8)$$

Now we propose our local time-integration method, motivated by the leapfrog-based scheme in [CH24]. With a smooth function  $\Psi: [0, \infty) \rightarrow \mathbb{R}$  satisfying  $\Psi(0) = 1$ , the fully discrete local time-integration scheme is defined as

$$v_h^{n+1/2} - v_h^n = \frac{\tau}{2} \mathcal{L}_{u,h} u_h^n + \frac{\tau}{2} g_{v,h}^{n+1/2}, \quad (1.9a)$$

$$u_h^{n+1} - u_h^n = \tau \Psi \mathcal{L}_{v,h} v_h^{n+1/2} + \tau \Psi g_{u,h}^{n+1/2}, \quad (1.9b)$$

$$v_h^{n+1} - v_h^{n+1/2} = \frac{\tau}{2} \mathcal{L}_{u,h} u_h^{n+1} + \frac{\tau}{2} g_{v,h}^{n+1/2}, \quad (1.9c)$$

$$\Psi = \Psi(-\tau^2 \mathcal{L}_{v,h} \chi_m \mathcal{L}_{u,h}), \quad (1.9d)$$

$$g_{w,h}^{n+1/2} = \frac{1}{2} (g_{w,h}(t^{n+1}) + g_{w,h}(t^n)), \quad w \in \{u, v\}. \quad (1.9e)$$

In (1.9b) and in the following we write  $\Psi w_h$  for the application of the discrete operator  $\Psi$  on a discrete function  $w_h \in \mathbb{V}_h^u$ . The scheme consists of a combination of the leapfrog method on  $\mathcal{T}_{h,\text{lf}}$  and a second scheme, which is defined by the filter function  $\Psi$ , on  $\mathcal{T}_{h,m}$ . For  $\Psi \equiv 1$ , the scheme reduces to the leapfrog scheme on the whole domain. Choosing  $\Psi$  as a polynomial of a certain degree  $p$  yields a fully explicit scheme (with  $p$  matrix vector multiplications in each time step), while choosing a rational function leads to a scheme, which is implicit in the dofs on  $\mathcal{T}_{h,m}$ . In Section 1.3.3 we consider polynomials  $\Psi$ , which are chosen in such a way that the CFL condition is optimized, and the rational filter corresponding to the unconditionally stable Crank-Nicolson method. The main contribution of our paper is to prove stability and full order of convergence of the combined scheme.

Moreover, we denote by

$$\widehat{w}^{n+1/2} = \frac{1}{2} (w(t^{n+1}) + w(t^n)), \quad w \in \{u, v\}, \quad (1.10)$$

the mean of the components of the exact solution at two consecutive time steps.

The remaining part of this paper is devoted to the proof of the following full-discretization error estimate.

**Theorem 1.1.** *For  $x_h^n = (u_h^n \ v_h^n)$  given by (1.9) with  $\tau$  satisfying a CFL condition independent of  $h_{f,\min}$  and a sufficiently regular solution  $x = (u \ v)$  of (1.1) we obtain*

$$\|x(t^n) - x_h^n\|_\Omega \leq C(\tau^2 + h_{\max}^k) \quad (1.11)$$

with a constant  $C$  independent of  $\tau$  and  $h_{\max}$ .

Details on the CFL condition will be given later.

### 1.3.1. Stability

In the following, we assume that  $\Psi$  defined in (1.9d) is invertible, which is indeed the case under a certain restriction on the time stepsize as we will later see. We start by rewriting (1.9) into a system inspired by the analysis of the Crank-Nicolson method [HS16] imitating a discrete semi-group technique. To do so, we define

$$\mathcal{R}_\pm = \begin{pmatrix} \Theta & \pm \frac{\tau}{2} \mathcal{L}_{v,h} \\ \pm \frac{\tau}{2} \mathcal{L}_{u,h} & \mathcal{I}_v \end{pmatrix} - \frac{\tau^2}{4} \begin{pmatrix} -\mathcal{L}_{v,h} \chi_{\text{lf}} \mathcal{L}_{u,h} & 0 \\ 0 & 0 \end{pmatrix} \quad (1.12a)$$

with

$$\Theta = \Theta(-\tau^2 \mathcal{L}_{v,h} \chi_m \mathcal{L}_{u,h}), \quad (1.12b)$$

$$\Theta(z) = \Psi(z)^{-1} \left(1 - \frac{z}{4} \Psi(z)\right) = \frac{1}{\Psi(z)} - \frac{z}{4}, \quad \Theta(0) = 1. \quad (1.12c)$$

In Table 1.1, we collect special choices of  $\Psi$  and  $\Theta$  together with certain parameters, which we will introduce later for the analysis. More details will be given in Section 1.3.3 below.

**Lemma 1.2.** *The approximations  $x_h^n = \begin{pmatrix} u_h^n & v_h^n \end{pmatrix}$  of the local time-integration scheme (1.9) satisfy*

$$\mathcal{R}_- x_h^{n+1} = \mathcal{R}_+ x_h^n + \tau g_h^{n+1/2}, \quad g_h^{n+1/2} = \begin{pmatrix} g_{u,h}^{n+1/2} \\ g_{v,h}^{n+1/2} \end{pmatrix}. \quad (1.13)$$

*Proof.* The recursion directly follows from (1.9) and the definition of  $\mathcal{R}_\pm$ .  $\square$

A key idea for stability analysis is to split  $\mathcal{R}_\pm$  into a positive definite operator  $\Theta_{\mathcal{I}}$ , a skew-adjoint operator  $\mathcal{L}_h$ , and a self-adjoint, positive semidefinite perturbation operator  $\mathcal{D}_{\text{lf}}$  via

$$\mathcal{R}_\pm = \Theta_{\mathcal{I}} \pm \frac{\tau}{2} \mathcal{L}_h - \frac{\tau^2}{4} \mathcal{D}_{\text{lf}}, \quad (1.14a)$$

where

$$\Theta_{\mathcal{I}} = \begin{pmatrix} \Theta & 0 \\ 0 & \mathcal{I}_v \end{pmatrix}, \quad \mathcal{L}_h = \begin{pmatrix} 0 & \mathcal{L}_{v,h} \\ \mathcal{L}_{u,h} & 0 \end{pmatrix}, \quad \mathcal{D}_{\text{lf}} = \begin{pmatrix} -\mathcal{L}_{v,h} \chi_{\text{lf}} \mathcal{L}_{u,h} & 0 \\ 0 & 0 \end{pmatrix}. \quad (1.14b)$$

The adjointness condition (1.5) directly implies that  $\mathcal{L}_h$  is skew-adjoint,  $\mathcal{D}_{\text{lf}}$  and  $\Theta$  are self-adjoint, and that  $\mathcal{D}_{\text{lf}}$  is positive semidefinite. However, the definiteness of  $\Theta_{\mathcal{I}}$  is only guaranteed for special choices of  $\Theta$  (see Table 1.1) or, if the modified scheme is not unconditionally stable, under the weaker CFL condition resulting from the argument  $-\tau^2 \mathcal{L}_{v,h} \chi_m \mathcal{L}_{u,h}$  of  $\Theta$  in (1.12b). To ensure this, we introduce additional notation.

**Definition 1.3.** *Let  $\Psi: [0, \infty) \rightarrow \mathbb{R}$  be a smooth function satisfying  $\Psi(0) = 1$  and consider  $\Theta$  defined in (1.12c). For a constant  $c_\Theta \in (0, 1]$  we define  $\beta_\Psi = \beta_\Psi(c_\Theta) \in (0, \infty]$  as the maximum value such that*

$$0 < \Psi(z) \leq 1 \quad \text{and} \quad \Theta(z) \geq c_\Theta \quad \text{for all } z \in [0, \beta_\Psi^2] \cap \mathbb{R}, \quad (1.15)$$

and  $\beta_\Psi = \infty$ , if (1.15) holds for all  $z \geq 0$ .

The existence of  $c_\Theta$  and  $\beta_\Psi$  fulfilling Definition 1.3 is guaranteed since  $\Psi$  and  $\Theta$  are continuous and we assume  $\Psi(0) = 1$  which implies  $\Theta(0) = 1$ . For a given filter function  $\Psi$ , one can first maximize  $\beta_\Psi$  such that  $\Theta(z) > 0$  on  $[0, \beta_\Psi^2] \cap \mathbb{R}$  and then set  $c_\Theta = \min_{z \in [0, \beta_\Psi^2]} \Theta(z)$ . For certain local time-stepping and locally implicit schemes, these constants are given in Table 1.1.

The parameters  $\beta_\Psi^2$  and  $c_\Theta$  directly enter the CFL conditions to ensure stability of the scheme.

**Definition 1.4.** *For given  $\vartheta, \vartheta_c \in (0, 1]$  we define*

$$\tau_{\text{CFL}, \Psi}^2 = \frac{\beta_\Psi^2}{\|\mathcal{L}_{v,h} \chi_m \mathcal{L}_{u,h}\|}, \quad (1.16a)$$

|                    | $\Psi$                   | $\Theta$          | $\varphi$      | $\beta_\Psi^2$    | $c_\Theta$        | $C_\varphi$   |
|--------------------|--------------------------|-------------------|----------------|-------------------|-------------------|---------------|
| Leapfrog           | 1                        | $1 - \frac{z}{4}$ | $-\frac{1}{4}$ | $4(1 - c_\Theta)$ | $1 - \vartheta^2$ | $\frac{1}{4}$ |
| Crank-Nicolson     | $(1 + \frac{z}{4})^{-1}$ | 1                 | 0              | $\infty$          | 1                 | 0             |
| Leapfrog-Chebyshev | (1.45)                   | (1.12c)           | (1.36a)        | (1.46)            |                   |               |

Table 1.1.: Examples for filter functions with corresponding constants.

$$\tau_{\text{CFL,lf,c}}^2 = \tau_{\text{CFL,lf,c}}^2(\vartheta_c) = \frac{4c_\Theta \vartheta_c^2}{\|\mathcal{L}_{v,h} \chi_{\text{lf}} \mathcal{L}_{u,h}\|}, \quad (1.16b)$$

and

$$\tau_{\text{CFL,lf}}^2 = \tau_{\text{CFL,lf}}^2(\vartheta) = \frac{4\vartheta^2}{\|\mathcal{L}_{v,h} \mathcal{L}_{u,h}\|}. \quad (1.17)$$

Recall that the leapfrog scheme on the whole spatial domain is stable if  $\tau \leq \tau_{\text{CFL,lf}}$ . We aim at choosing  $\Psi$  in such a way that the CFL condition of the local time-integration method is significantly relaxed compared to the leapfrog scheme. For the dependence of the CFL condition on the submeshes and the material parameters, note that

$$\|\mathcal{L}_{v,h} \chi_{\text{lf}} \mathcal{L}_{u,h}\| \lesssim h_{c,\min}^{-2}, \quad \|\mathcal{L}_{v,h} \chi_m \mathcal{L}_{u,h}\| \lesssim h_{f,\min}^{-2},$$

with constants that, roughly speaking, scale like the square of the degree  $k^2$  of the dG polynomials (see, e.g., [DPE12, Rem. 1.48]) and the inverse of the product of the smallest eigenvalues of  $\mathcal{M}_u$  and  $\mathcal{M}_v$  on the respective submeshes  $\mathcal{T}_{h,c}$  and  $\mathcal{T}_{h,f}$ , cf. [HK22, Eq. 6] or [DPE12, Lem. 7.32].

It is important to note that the eigenvalues of  $\mathcal{L}_{v,h} \chi_{\text{lf}} \mathcal{L}_{u,h}$  and hence  $\tau_{\text{CFL,lf,c}}$  are independent of the material parameters and the diameters within the fine mesh  $\mathcal{T}_{h,f}$ , cf. [DHK<sup>+</sup>23, HS16] for details.

**Corollary 1.5.** *Let  $\Psi: [0, \infty) \rightarrow \mathbb{R}$  be a smooth function satisfying  $\Psi(0) = 1$  and assume that the CFL condition  $\tau \leq \tau_{\text{CFL},\Psi}$  with  $\tau_{\text{CFL},\Psi}$  defined in (1.16a) is fulfilled. Then, the filter operator  $\Psi$  defined in (1.9d) is invertible and satisfies  $\|\Psi\|_{\mathbb{V}_h^u \leftarrow \mathbb{V}_h^u} \leq 1$ .*

*Proof.* The statement follows immediately from  $-\mathcal{L}_{v,h} \chi_m \mathcal{L}_{u,h}$  being symmetric, and positive semidefinite together with Definitions 1.3 and 1.4.  $\square$

**Lemma 1.6.** *For  $\tau \leq \tau_{\text{CFL},\Psi}$  defined (1.16a), the operator  $\Theta$  induces a norm  $\|\cdot\|_\Theta$  such that for all  $u_h \in \mathbb{V}_h^u$  we have*

$$c_\Theta \|u_h\|_\Omega^2 \leq \|u_h\|_\Theta^2 = (\Theta u_h, u_h)_\Omega. \quad (1.18)$$

*Proof.* The statement follows immediately from (1.15) since we can bound the eigenvalues  $z$  of  $-\tau^2 \mathcal{L}_{v,h} \chi_m \mathcal{L}_{u,h}$  by  $\beta_\Psi^2$  if  $\tau \leq \tau_{\text{CFL},\Psi}$ .  $\square$

Lemma 1.6 implies that  $\Theta_{\mathcal{I}}$  defined in (1.12b) is also self-adjoint and positive definite for  $\tau \leq \tau_{\text{CFL},\Psi}$  and hence it induces a norm  $\|\cdot\|_{\Theta_{\mathcal{I}}}$  on  $\mathbb{V}_h$ .

Using both CFL conditions in (1.16) we can bound the action of the operators  $\mathcal{R}_\pm$  defined in (1.12a) from below.

**Lemma 1.7.** *Let  $\vartheta_c \in (0, 1)$  and  $\tau \leq \min\{\tau_{\text{CFL,lf,c}}(\vartheta_c), \tau_{\text{CFL},\Psi}\}$  satisfy the CFL conditions (1.16). Then we have*

$$c_\Theta(1 - \vartheta_c^2)\|u_h\|_\Omega^2 + \|v_h\|_\Omega^2 \leq (\mathcal{R}_\pm x_h, x_h)_\Omega. \quad (1.19)$$

for all  $x_h = \begin{pmatrix} u_h & v_h \end{pmatrix} \in \mathbb{V}_h$ .

*Proof.* Let  $x_h = \begin{pmatrix} u_h & v_h \end{pmatrix} \in \mathbb{V}_h$ . Using (1.14b) and (1.16b) we obtain

$$\frac{\tau^2}{4}(\mathcal{D}_{\text{lf}}x_h, x_h)_\Omega = -\frac{\tau^2}{4}(\mathcal{L}_{v,h}\chi_{\text{lf}}\mathcal{L}_{u,h}u_h, u_h)_\Omega \leq c_\Theta\vartheta_c^2\|u_h\|_\Omega^2,$$

and thus with (1.14a) and (1.18)

$$\begin{aligned} (\mathcal{R}_\pm x_h, x_h)_\Omega &= (\Theta u_h, u_h)_\Omega + \|v_h\|_\Omega^2 - \frac{\tau^2}{4}(\mathcal{D}_{\text{lf}}x_h, x_h)_\Omega \\ &\geq c_\Theta(1 - \vartheta_c^2)\|u_h\|_\Omega^2 + \|v_h\|_\Omega^2. \end{aligned}$$

This proves the statement.  $\square$

In the next Lemma 1.8 we show that the CFL condition for the local time-integration scheme (1.9) remains less restrictive than that of the leapfrog scheme used on the whole grid.

**Lemma 1.8.** *Let  $\vartheta \in (0, 1)$ ,  $x_h = \begin{pmatrix} u_h & v_h \end{pmatrix} \in \mathbb{V}_h$ , and  $\tau \leq \tau_{\text{CFL,lf}}(\vartheta)$  defined in (1.17). Moreover, assume that  $0 < \Psi(z) \leq 1$  holds for all  $z \in [0, 4]$ . Then we have*

$$(\mathcal{R}_\pm x_h, x_h)_\Omega \geq (1 - \vartheta^2)\|u_h\|_\Omega^2 + \|v_h\|_\Omega^2. \quad (1.20)$$

*Proof.* By assumption,  $\Psi$  is invertible since the leapfrog CFL condition (1.17) implies  $\| -\tau^2\mathcal{L}_{v,h}\mathcal{L}_{u,h}\|_\Omega \leq 4\vartheta^2 \leq 4$  and thus  $(\Psi^{-1}u_h, u_h)_\Omega \geq \|u_h\|_\Omega^2$ . Using (1.12c) we can write

$$\begin{aligned} \Theta + \frac{\tau^2}{4}\mathcal{L}_{v,h}\chi_{\text{lf}}\mathcal{L}_{u,h} &= \Psi^{-1} + \frac{\tau^2}{4}(\mathcal{L}_{v,h}\chi_m\mathcal{L}_{u,h} + \mathcal{L}_{v,h}\chi_{\text{lf}}\mathcal{L}_{u,h}) \\ &= \Psi^{-1} + \frac{\tau^2}{4}\mathcal{L}_{v,h}\mathcal{L}_{u,h}. \end{aligned}$$

From (1.12a) and the CFL condition (1.17) we conclude

$$\begin{aligned} (\mathcal{R}_\pm x_h, x_h)_\Omega &= ((\Theta + \frac{\tau^2}{4}\mathcal{L}_{v,h}\chi_{\text{lf}}\mathcal{L}_{u,h})u_h, u_h)_\Omega + \|v_h\|_\Omega^2 \\ &\geq (1 - \vartheta^2)\|u_h\|_\Omega^2 + \|v_h\|_\Omega^2. \end{aligned}$$

This completes the proof.  $\square$

Lemma 1.7 and the splitting of  $\mathcal{R}_\pm$  in (1.12a) lead to the following properties.

**Lemma 1.9.** *Let  $\tau \leq \min\{\tau_{\text{CFL},\Psi}, \tau_{\text{CFL,lf,c}}\}$  defined in (1.16) for some  $\vartheta_c \in (0, 1)$ . Then, for  $x_h, y_h \in \mathbb{V}_h$  and  $\mathcal{R} = \mathcal{R}_-^{-1}\mathcal{R}_+$  we have for  $n = 0, 1, 2, \dots$  the identities*

$$(\mathcal{R}_-x_h, y_h)_\Omega = (x_h, \mathcal{R}_+y_h)_\Omega, \quad (1.21a)$$

$$(\mathcal{R}_-\mathcal{R}^n x_h, \mathcal{R}^n x_h)_\Omega = (\mathcal{R}_-x_h, x_h)_\Omega, \quad (1.21b)$$

$$(\mathcal{R}_-\mathcal{R}^n \mathcal{R}_-^{-1} x_h, \mathcal{R}^n \mathcal{R}_-^{-1} x_h)_\Omega = (\mathcal{R}_-^{-1} x_h, x_h)_\Omega, \quad (1.21c)$$

and in addition the bounds<sup>1</sup>

$$\|\mathcal{R}^n x_h\|_\Omega \leq C_{stb,c} \|x_h\|_{\Theta_I}, \quad (1.22a)$$

$$\|\mathcal{R}^n \mathcal{R}_-^{-1} x_h\|_\Omega \leq C_{stb,c} \|x_h\|_\Omega, \quad (1.22b)$$

with  $C_{stb,c} = (c_\Theta(1 - \vartheta_c^2))^{-1/2}$ .

*Proof.* The adjointness property (1.21a) follows from the skew-adjointness of  $\mathcal{L}_h$  and the self-adjointness of  $\mathcal{L}_{v,h} \chi_{\text{lf}} \mathcal{L}_{u,h}$ , while (1.21b) can be seen by induction with

$$\begin{aligned} (\mathcal{R}_- \mathcal{R} x_h, \mathcal{R} x_h)_\Omega &= (\mathcal{R}_+ x_h, \mathcal{R} x_h)_\Omega \\ &= (x_h, \mathcal{R}_- \mathcal{R} x_h)_\Omega \\ &= (x_h, \mathcal{R}_+ x_h)_\Omega \\ &= (\mathcal{R}_- x_h, x_h)_\Omega. \end{aligned}$$

One can easily verify that the inverses are given by

$$\begin{aligned} \mathcal{R}_\pm^{-1} &= \begin{pmatrix} \Psi & \mp \frac{\tau}{2} \Psi \mathcal{L}_{v,h} \\ \mp \frac{\tau}{2} \mathcal{L}_{u,h} \Psi & \mathcal{I}_v + \frac{\tau^2}{4} \mathcal{L}_{u,h} \Psi \mathcal{L}_{v,h} \end{pmatrix} \\ &= \begin{pmatrix} \Psi & \\ & \mathcal{I}_v + \frac{\tau^2}{4} \mathcal{L}_{u,h} \Psi \mathcal{L}_{v,h} \end{pmatrix} \mp \frac{\tau}{2} \begin{pmatrix} & \Psi \mathcal{L}_{v,h} \\ \mathcal{L}_{u,h} \Psi & \end{pmatrix}. \end{aligned} \quad (1.23)$$

These are decompositions of  $\mathcal{R}_\pm^{-1}$  into their symmetric and skew-symmetric parts.

Replacing  $x_h$  by  $\mathcal{R}_-^{-1} x_h$  in (1.21b) proves (1.21c). Furthermore, (1.19) and (1.21b) imply

$$\begin{aligned} c_\Theta(1 - \vartheta_c^2) \|\mathcal{R}^n x_h\|_\Omega^2 &\leq (\mathcal{R}_- \mathcal{R}^n x_h, \mathcal{R}^n x_h)_\Omega \\ &= (\mathcal{R}_- x_h, x_h)_\Omega \\ &= \|x_h\|_{\Theta_I}^2 - (\mathcal{D}_{\text{lf}} x_h, x_h)_\Omega \leq \|x_h\|_{\Theta_I}^2, \end{aligned}$$

where we have used the positive semi-definiteness of  $\mathcal{D}_{\text{lf}}$  in the last estimate. This proves (1.22a).

By (1.15),  $\Psi$  is positive definite if the CFL condition (1.16a) is satisfied. Thus we have with the adjointness property (1.5)

$$(\mathcal{L}_{u,h} \Psi \mathcal{L}_{v,h} v_h, v_h)_\Omega = -(\Psi \mathcal{L}_{v,h} v_h, \mathcal{L}_{v,h} v_h)_\Omega \leq 0$$

which shows

$$(\mathcal{R}_\pm^{-1} x_h, x_h)_\Omega = (\Psi u_h, u_h)_\Omega + (v_h, v_h)_\Omega + \frac{\tau^2}{4} (\mathcal{L}_{u,h} \Psi \mathcal{L}_{v,h} v_h, v_h)_\Omega \leq \|x_h\|_\Omega^2. \quad (1.24)$$

Using (1.19), (1.21c) and (1.24) shows

$$\begin{aligned} c_\Theta(1 - \vartheta_c^2) \|\mathcal{R}^n \mathcal{R}_-^{-1} x_h\|_\Omega^2 &\leq (\mathcal{R}_- \mathcal{R}^n \mathcal{R}_-^{-1} x_h, \mathcal{R}^n \mathcal{R}_-^{-1} x_h)_\Omega \\ &= (\mathcal{R}_-^{-1} x_h, x_h)_\Omega \\ &\leq \|x_h\|_\Omega^2 \end{aligned}$$

and hence (1.22b). □

---

<sup>1</sup>In [HS26, Lem. 3.9] the estimate (3.19c) is wrong. Since it is not used anywhere else in the manuscript it could be removed without any trouble.

With the previous Lemma 1.9 we conclude the following stability estimate.

**Lemma 1.10** (Stability). *Let  $x_h^{n+1}$  be defined in (1.13). Under the same assumptions as in Lemma 1.9 the numerical solution is bounded by*

$$\|x_h^{n+1}\|_\Omega \leq C_{stb,c} \left( \|x_h^0\|_{\Theta_I} + \tau \sum_{j=0}^n \|g_h^{j+1/2}\|_\Omega \right). \quad (1.25)$$

*Proof.* Since  $\mathcal{R}_-^{-1}$  is invertible, the scheme (1.13) is equivalent to

$$x_h^{n+1} = \mathcal{R}x_h^n + \tau \mathcal{R}_-^{-1} g_h^{n+1/2}, \quad \mathcal{R} = \mathcal{R}_-^{-1} \mathcal{R}_+.$$

Then, the discrete variation-of-constants formula yields

$$x_h^{n+1} = \mathcal{R}^{n+1} x_h^0 + \tau \sum_{j=0}^n \mathcal{R}^{n-j} \mathcal{R}_-^{-1} g_h^{j+1/2}. \quad (1.26)$$

The claim now follows from (1.22a) and (1.22b) and the triangle inequality.  $\square$

### 1.3.2. Error analysis of the local time-integration scheme

In the following we denote the exact solution of (1.1) evaluated at time  $t^n$  by

$$\hat{x}^n = x(t^n), \quad x = \begin{pmatrix} u & v \end{pmatrix}. \quad (1.27)$$

The error of the full discretization is given by

$$e^n = \hat{x}^n - x_h^n = e_\pi^n + e_h^n, \quad e_\pi^n = \hat{x}^n - \pi_h \hat{x}^n, \quad e_h^n = \pi_h \hat{x}^n - x_h^n, \quad (1.28)$$

where  $e_\pi^n$  denotes the  $L^2$ -projection error and  $e_h^n$  the discretization error respectively.

To define defects  $\delta_h^{n+1}$  we insert the  $L^2$ -projected exact solution into the numerical scheme (1.13). This yields

$$\mathcal{R}_- \pi_h \hat{x}^{n+1} = \mathcal{R}_+ \pi_h \hat{x}^n + \tau g_h^{n+1/2} + \tau \delta_h^{n+1}. \quad (1.29)$$

Subtracting (1.13) from (1.29) yields the error recursion

$$\mathcal{R}_- e_h^{n+1} = \mathcal{R}_+ e_h^n + \tau \delta_h^{n+1}. \quad (1.30)$$

This recursion is of the same form as (1.13). Hence, we can apply Lemma 1.10 to bound the error. Unfortunately, it will turn out later, that we need a more careful inspection of the defects to deal with the cutoff functions within the Friedrichs' operators. Otherwise, this would lead to suboptimal error bounds.

**Theorem 1.11.** *Let  $\tau \leq \min\{\tau_{\text{CFL},\Psi}, \tau_{\text{CFL},\text{lf},c}\}$  defined in (1.16) for some  $\vartheta_c \in (0, 1)$ . If one can decompose the defect into*

$$\delta_h^j = \delta_{h,s}^j + \mathcal{L}_h \delta_{h,\mathcal{L}_h}^j \quad \text{with} \quad \delta_{h,\mathcal{L}_h}^j = \begin{pmatrix} 0 \\ \delta_{h,v,\mathcal{L}_h}^j \end{pmatrix}, \quad (1.31a)$$

then it holds

$$\begin{aligned} \|e_h^{n+1}\|_\Omega &\leq C_{stb,c} \left( \|e_h^0\|_{\Theta_{\mathcal{I}}} + \tau \sum_{j=1}^{n+1} \|\delta_{h,s}^j\|_\Omega + \|\delta_{h,v,\mathcal{L}_h}^1\|_\Omega + \tau \sum_{j=2}^{n+1} \|\partial_\tau \delta_{h,v,\mathcal{L}_h}^j\|_\Omega \right) \\ &\quad + \|\delta_{h,v,\mathcal{L}_h}^{n+1}\|_\Omega. \end{aligned} \quad (1.31b)$$

Here,

$$\partial_\tau \hat{u}^{n+1} = \frac{1}{\tau} (\hat{u}^{n+1} - \hat{u}^n) \quad (1.31c)$$

denotes the discrete time derivative of a function  $\hat{u}$ .

*Proof.* Solving the error recursion (1.30) with the discrete variation-of-constants formula yields

$$\begin{aligned} e_h^{n+1} &= \mathcal{R}^{n+1} e_h^0 + \tau \sum_{j=0}^n \mathcal{R}^{n-j} \mathcal{R}_-^{-1} \delta_h^{j+1} \\ &= \mathcal{R}^{n+1} e_h^0 + \tau \sum_{j=0}^n \mathcal{R}^{n-j} \mathcal{R}_-^{-1} \delta_{h,s}^{j+1} + \sum_{j=0}^n \mathcal{R}^{n-j} (\mathcal{R} - \mathcal{I}) \delta_{h,\mathcal{L}_h}^{j+1}, \end{aligned} \quad (1.32)$$

since

$$\tau \mathcal{R}_-^{-1} \mathcal{L}_h \delta_{h,\mathcal{L}_h}^{j+1} = \mathcal{R}_-^{-1} (\mathcal{R}_+ - \mathcal{R}_-) \delta_{h,\mathcal{L}_h}^{j+1} = (\mathcal{R} - \mathcal{I}) \delta_{h,\mathcal{L}_h}^{j+1}.$$

Now we use the well-known summation by parts formula: for suitable sequences  $\{\rho_j\}_j$  and  $\{\delta_j\}_j$  we have

$$\sum_{j=0}^n \rho_{n-j} \delta_{j+1} = r_n \delta_1 + \sum_{j=1}^n r_{n-j} (\delta_{j+1} - \delta_j), \quad r_k = \sum_{j=0}^k \rho_j. \quad (1.33)$$

For  $\rho_k = \mathcal{R}^k (\mathcal{R} - \mathcal{I})$ , this yields  $r_k = \mathcal{R}^{k+1} - \mathcal{I}$  and

$$\sum_{j=0}^n \mathcal{R}^{n-j} (\mathcal{R} - \mathcal{I}) \delta_{h,\mathcal{L}_h}^{j+1} = \mathcal{R}^{n+1} \delta_{h,\mathcal{L}_h}^1 - \delta_{h,\mathcal{L}_h}^{n+1} + \tau \sum_{j=1}^n \mathcal{R}^{n-j+1} \partial_\tau \delta_{h,\mathcal{L}_h}^{j+1}.$$

The bounds (1.22a) and (1.22b) imply (1.31b), since the first component of the defect  $\delta_{h,\mathcal{L}_h}$  vanishes.  $\square$

For the discrete time-derivative (1.31c) of a sufficiently smooth function  $u$ , a simple calculation shows the representations

$$\partial_\tau \hat{u}^{n+1} = \int_0^1 \partial_t u(t^n + \tau s) \, ds, \quad \partial_\tau^2 \hat{u}^{n+1} = \int_{-1}^1 (1 - |s|) \partial_t^2 u(t^n + \tau s) \, ds. \quad (1.34)$$

Note that the average  $\partial_t \hat{u}^{n+1/2}$  corresponds to the trapezoidal rule applied to the first integral in (1.34). It is well-known that the error

$$\delta_{\text{tr},u}^{n+1} = \partial_\tau \hat{u}^{n+1} - \partial_t \hat{u}^{n+1/2} \quad (1.35a)$$

satisfies

$$\|\delta_{\text{tr},u}^{n+1}\|_\Omega \leq \frac{\tau^2}{8} \int_0^1 \|\partial_t^3 u(t^n + \tau s)\|_\Omega \, ds. \quad (1.35b)$$

Moreover, we define

$$\varphi = \varphi(-\tau^2 \mathcal{L}_{v,h} \chi_m \mathcal{L}_{u,h}), \quad \text{with} \quad \varphi(z) = \frac{\Theta(z) - 1}{z}, \quad z > 0, \quad (1.36a)$$

and set  $\varphi(0) = \Theta'(0)$ . In addition, we will use

$$\tilde{\varphi} = \varphi(-\tau^2 \chi_m \mathcal{L}_{u,h} \mathcal{L}_{v,h} \chi_m). \quad (1.36b)$$

Note that the nonzero eigenvalues of  $\chi_m \mathcal{L}_{u,h} \mathcal{L}_{v,h} \chi_m$  and  $\mathcal{L}_{v,h} \chi_m \mathcal{L}_{u,h}$  coincide.

Furthermore, we recall the consistency and approximation properties of the dG discretized Friedrichs' operators  $\mathcal{L}_{w,h}$ , with  $w \in \{u, v\}$ ,

$$\mathcal{L}_{w,h} w = \pi_h \mathcal{L}_w w \quad \text{for all } w \in \mathcal{D}(\mathcal{L}_w) \cap H^1(\mathcal{T}_h)^{m_w}, \quad (1.37a)$$

$$\|\mathcal{L}_{w,h} e_{\pi,w}\|_{\Omega} \leq C_{\pi,\mathcal{L},w} |h^k w|_{k+1,\mathcal{T}_h} \quad \text{for all } w \in \mathcal{D}(\mathcal{L}_w) \cap H^{k+1}(\mathcal{T}_h)^{m_w}, \quad (1.37b)$$

cf. [DPE12, DHK<sup>+</sup>23].

To derive a representation of the defect  $\delta_h^j$  defined in (1.29), we write (1.14) as a perturbation of the Crank-Nicolson scheme.

**Lemma 1.12.** *The operators  $\mathcal{R}_{\pm}$  defined in (1.14) satisfy*

$$\mathcal{R}_{\pm} = \mathcal{R}_{\pm,\text{CN}} + \mathcal{L}_h \mathcal{D}_{\text{LTI}}, \quad (1.38a)$$

where we have

$$\mathcal{R}_{\pm,\text{CN}} = \mathcal{I} \pm \frac{\tau}{2} \mathcal{L}_h \quad \text{and} \quad \mathcal{D}_{\text{LTI}} = \tau^2 \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{4} \chi_{\text{If}} \mathcal{I}_v - \tilde{\varphi} \chi_m \end{pmatrix} \mathcal{L}_h. \quad (1.38b)$$

*Proof.* The representation follows from the definition of  $\varphi$  in (1.36a) and  $\chi_m = \chi_m^2$  since

$$\Theta = \mathcal{I}_u - \tau^2 \mathcal{L}_{v,h} \chi_m \mathcal{L}_{u,h} \varphi = \mathcal{I}_u - \tau^2 \mathcal{L}_{v,h} \chi_m \tilde{\varphi} \chi_m \mathcal{L}_{u,h},$$

where the second equality follows from [Hig08, Cor. 1.34].  $\square$

With Lemma 1.12, we can split the defect  $\delta_h^j$  defined in (1.29) as

$$\delta_h^j = \delta_{h,\text{CN}}^j + \mathcal{L}_h \delta_{h,\text{LTI}}^j \quad (1.39a)$$

with

$$\delta_{h,\text{CN}}^j = \pi_h \partial_{\tau} \hat{x}^j - \mathcal{L}_h \pi_h \hat{x}^{j-1/2} - g_h^{j-1/2}, \quad (1.39b)$$

$$\begin{aligned} \delta_{h,\text{LTI}}^j &= \mathcal{D}_{\text{LTI}} \pi_h \partial_{\tau} \hat{x}^j \\ &= \mathcal{D}_{\text{LTI}} \partial_{\tau} \hat{x}^j - \mathcal{D}_{\text{LTI}} \partial_{\tau} e_{\pi}^j. \end{aligned} \quad (1.39c)$$

The first component of  $\delta_{h,\text{LTI}}^j$  vanishes such that we can apply Theorem 1.11 later.

*Remark 1.13.* We could analyze the defect  $\delta_h^j$  directly (without splitting). One would proceed by using the discrete variations-of-constants formula in Lemma 1.10. However, this requires a bound on  $\mathcal{L}_{v,h} \chi_m \phi_h$  for some  $\phi_h \in \mathbb{V}_h^v$  resulting in a suboptimal error estimate.  $\diamond$

We now have to bound the defect (1.39a). Following [DHK<sup>+</sup>23, Lemma 12.2] the defect (1.39b) stemming from the Crank-Nicolson scheme can be bounded under appropriate regularity assumptions on the exact solution of (1.1) as

$$\|\delta_{h,\text{CN}}^j\|_\Omega \leq C_{\pi,\mathcal{L}} |h^k \hat{u}^{j-1/2}|_{k+1,\mathcal{T}_h} + \frac{\tau^2}{8} \int_0^1 \|\partial_t^3 v(t^{j-1} + \tau s)\|_\Omega ds. \quad (1.40)$$

To bound the defect (1.39c) we need a bound on  $\varphi$  introduced in (1.36a).

**Definition 1.14.** With  $\beta_\Psi$  from Definition 1.3, we define  $C_\varphi$  as the smallest constant such that for  $\varphi$  defined in (1.36a) it holds

$$|\varphi(z)| \leq C_\varphi \quad \text{for all } z \in [0, \beta_\Psi^2] \cap \mathbb{R}. \quad (1.41)$$

Note that such a bound exists since  $\varphi$  is continuous on  $[0, \beta_\Psi^2]$  and hence bounded.

Now we are able to state a bound for the defect (1.39c) under appropriate regularity assumptions on the exact solution of (1.1).

**Lemma 1.15.** Let  $x = \begin{pmatrix} u & v \end{pmatrix}$  with

$$x \in C(t^0, T; D(\mathcal{L}) \cap H^{k+1}(\mathcal{T}_h)^m) \cap C^3(t^0, T; L^2(\Omega)^m)$$

be the solution of (1.1) and  $\tau \leq \tau_{\text{CFL},\Psi}$  defined in (1.4). We further assume

$$g_v \in C^2(t^0, T; L^2(\Omega)^{m_v}).$$

Then the defect (1.39c) is bounded by

$$\begin{aligned} \|\delta_{h,\text{LTI}}^{n+1}\|_\Omega &\leq \tau^2 \hat{C}_\varphi \left( C_{\pi,\mathcal{L}} |h^k \partial_\tau \hat{u}^{n+1}|_{k+1,\mathcal{T}_h} + \int_0^1 \|\partial_t^2 v(t^n + \tau s)\|_\Omega ds \right. \\ &\quad \left. + \int_0^1 \|\partial_t g_v(t^n + \tau s)\|_\Omega ds \right) \end{aligned} \quad (1.42a)$$

$$\begin{aligned} \|\partial_\tau \delta_{h,\text{LTI}}^{n+1}\|_\Omega &\leq \tau^2 \hat{C}_\varphi \left( C_{\pi,\mathcal{L}} |h^k \partial_\tau^2 \hat{u}^{n+1}|_{k+1,\mathcal{T}_h} + \int_{-1}^1 \|\partial_t^3 v(t^n + \tau s)\|_\Omega ds \right. \\ &\quad \left. + \int_{-1}^1 \|\partial_t^2 g_v(t^n + \tau s)\|_\Omega ds \right) \end{aligned} \quad (1.42b)$$

with  $\hat{C}_\varphi = \sqrt{2} \max\{\frac{1}{4}, C_\varphi\}$ .

*Proof.* With (1.37a), (1.37b), (1.41), and (1.1) we get

$$\begin{aligned} \|\delta_{h,\text{LTI}}^{n+1}\|_\Omega &\leq \frac{\tau^2}{4} \|\chi_{\text{If}} \pi_h \partial_\tau \partial_t \hat{v}^{n+1}\|_\Omega + \tau^2 C_\varphi \|\chi_m \pi_h \partial_\tau \partial_t \hat{v}^{n+1}\|_\Omega \\ &\quad + \frac{\tau^2}{4} \|\chi_{\text{If}} \pi_h \partial_\tau g_v^{n+1}\|_\Omega + \tau^2 C_\varphi \|\chi_m \pi_h \partial_\tau g_v^{n+1}\|_\Omega \\ &\quad + \tau^2 C_\varphi C_{\pi,\mathcal{L}} |\chi_m h^k \partial_\tau \hat{u}^{n+1}|_{k+1,\mathcal{T}_h} + \frac{\tau^2}{4} C_{\pi,\mathcal{L}} |\chi_{\text{If}} h^k \partial_\tau \hat{u}^{n+1}|_{k+1,\mathcal{T}_h} \end{aligned}$$

and in the same manner

$$\begin{aligned} \|\partial_\tau \delta_{h,\text{LTI}}^{n+1}\|_\Omega &\leq \frac{\tau^2}{4} \|\chi_{\text{If}} \pi_h \partial_\tau^2 \partial_t \hat{v}^{n+1}\|_\Omega + \tau^2 C_\varphi \|\chi_m \pi_h \partial_\tau^2 \partial_t \hat{v}^{n+1}\|_\Omega \\ &\quad + \frac{\tau^2}{4} \|\chi_{\text{If}} \pi_h \partial_\tau^2 g_v^{n+1}\|_\Omega + \tau^2 C_\varphi \|\chi_m \pi_h \partial_\tau^2 g_v^{n+1}\|_\Omega \\ &\quad + \tau^2 C_\varphi C_{\pi,\mathcal{L}} |\chi_m h^k \partial_\tau^2 \hat{u}^{n+1}|_{k+1,\mathcal{T}_h} + \frac{\tau^2}{4} C_{\pi,\mathcal{L}} |\chi_{\text{If}} h^k \partial_\tau^2 \hat{u}^{n+1}|_{k+1,\mathcal{T}_h}. \end{aligned}$$

Using (1.34) completes the proof.  $\square$

Now we state our main theorem which yields under certain regularity assumptions convergence of order two in time and order  $k$  in space if we choose as dG polynomial degree  $k$ .

**Theorem 1.16.** *By the same regularity assumptions as in Lemma 1.15 together with  $\tau \leq \min\{\tau_{\text{CFL,lf,c}}, \tau_{\text{CFL},\Psi}\}$  and  $x_h^0 = \pi_h x(t^0)$ , the error of the full discretization satisfies*

$$\|x(t^{n+1}) - x_h^{n+1}\|_{\Omega} \leq C(\tau^2 + h_{\max}^k) \quad (1.43)$$

with a constant  $C$  independent of  $\tau$  and  $h$ .

*Proof.* The projection error in (1.28) is bounded by

$$\|e_{\pi}^{n+1}\|_{\Omega} \leq C_{\pi} |h^{k+1} \hat{x}^{n+1}|_{k+1, \mathcal{T}_h},$$

cf. [DHK<sup>+</sup>23, Lemma 3.2]. With Theorem 1.11, Lemma 1.15, and  $x_h^0 = \pi_h \hat{x}^0$  we get for the discretization error

$$\begin{aligned} \|e_h^{n+1}\|_{\Omega} &\leq C_{stb,c} \left( \tau \sum_{j=1}^{n+1} \|\delta_{h,s}^j\|_{\Omega} + \|\delta_{h,v,\mathcal{L}_h}^1\|_{\Omega} + \tau \sum_{j=2}^{n+1} \|\partial_{\tau} \delta_{h,v,\mathcal{L}_h}^j\|_{\Omega} \right) + \|\delta_{h,v,\mathcal{L}_h}^{n+1}\|_{\Omega} \\ &\leq C_{stb,c} \left( \frac{\tau^2}{8} \|\partial_t^3 x\|_{L^1([t^0, T], L^2(\Omega))} \right. \\ &\quad + 2\tau^2 \widehat{C}_{\varphi} \left( \|\partial_t^3 v\|_{L^1([t^0, T], L^2(\Omega))} + \|\partial_t^2 g_v\|_{L^1([t^0, T], L^2(\Omega))} \right) \\ &\quad + \widehat{C}_{\varphi} C_{\pi, \mathcal{L}} \max_{j=0, \dots, n} |h^k \tau^2 \partial_{\tau} \hat{u}^{j+1}|_{k+1, \mathcal{T}_h} \\ &\quad + \widehat{C}_{\varphi} \tau^2 \left( \max_{s \in [t^0, T]} \|\partial_t^2 v(s)\|_{\Omega} + \max_{s \in [t^0, T]} \|\partial_t^1 g_v(s)\|_{\Omega} \right) \\ &\quad \left. + C_{\pi, \mathcal{L}} \tau \sum_{j=0}^n |h^k \hat{x}^{j+1/2}|_{k+1, \mathcal{T}_h} + \widehat{C}_{\varphi} C_{\pi, \mathcal{L}} \tau \sum_{j=1}^n |h^k \tau^2 \partial_{\tau}^2 \hat{u}^{j+1}|_{k+1, \mathcal{T}_h} \right). \end{aligned}$$

This proves the statement.  $\square$

*Remark 1.17.* If we use the midpoint evaluation  $g_h^{n+1/2} = g_h(t^{n+1/2})$  in (1.9) instead of the average (1.10), we get an additional term  $\tau^2 \|\partial_t^2 g\|_{L^1([t^0, T], L^2(\Omega))}$  in the bound of Theorem 1.16. Nevertheless, this modification leads to the same order of convergence.  $\diamond$

### 1.3.3. Stability and convergence of local time-integration schemes

So far, we proved our theoretical results for general filter functions  $\Psi$ . In this section, we investigate the constants from Definition 1.3 for the special case of using leapfrog-Chebyshev polynomials and for a rational function from our previous work [CH22, CH24].

We start with considering the LI method from [HS16], where, for all  $z \in [0, \infty)$ , it holds

$$\Psi(z) = (1 + \frac{z}{4})^{-1} > 0, \quad \Theta(z) = 1, \quad \varphi(z) = 0. \quad (1.44)$$

Thus we obtain  $\beta_{\Psi} = \infty$  with the constants  $c_{\Theta} = 1$  and  $C_{\varphi} = 0$ . Inserting these constants into Definition 1.4 we get exactly the CFL conditions from the literature, i.e., [DHK<sup>+</sup>23, Assumption 11.26] for Friedrichs' systems or from [HS16] for the special case of Maxwell's equations. This indicates that our new general theory does not require stronger assumptions than for the known results.

Now we turn to the new local time-stepping methods. Here, motivated by [CH22], we choose the filter function as the following polynomial of degree  $p$

$$\Psi(z)z = P_p(z)z = 2 - \frac{2}{T_p(\nu_p)}T_p\left(\nu_p - \frac{z}{\alpha_p}\right), \quad \alpha_p = 2\frac{T'_p(\nu_p)}{T_p(\nu_p)}, \quad (1.45)$$

with a stabilization parameter  $\nu_p > 1$ . With  $T_p$ , we denote the  $p$ th Chebychev polynomial of first kind.

**Theorem 1.18.** *Let  $\Psi$  be given by (1.45) for some  $p \in \mathbb{N}$  and  $\nu_p > 1$ . Then, (1.15) and (1.41) are satisfied for*

$$c_\Theta = \frac{1}{2}\left(1 - \frac{1}{T_p(\nu_p)}\right) \in (0, \frac{1}{2}), \quad \beta_\Psi^2 = \alpha_p(\nu_p + 1), \quad C_\varphi = \frac{1}{4}\left(\frac{1}{c_\Theta} - 1\right). \quad (1.46)$$

*Proof.* Following [CH22, Lemmas 5.1, 5.4] we have

$$\Psi(z)z \leq 4(1 - c_\Theta), \quad 0 < \Psi(z) \leq 1, \quad 0 \leq z \leq \beta_\Psi^2$$

and hence

$$\Theta(z) = \Psi(z)^{-1}(1 - \frac{z}{4}\Psi(z)) \geq 1 - \frac{z}{4}\Psi(z) \geq c_\Theta.$$

The formula for  $C_\varphi$  can be shown analogously to the proof of [CH22, Lemma 5.4].  $\square$

Note that for  $p = 1$  we have  $\Psi \equiv 1$  and the scheme (1.9) is just the leapfrog method (which is independent of the stabilization parameter  $\nu_p$ ). If we choose  $c_\Theta = 1 - \vartheta^2$ , for  $\vartheta^2 \in (0, 1)$ , we get  $\beta_\Psi^2 = 4(1 - c_\Theta) = 4\vartheta^2$ .

If, for  $p \geq 2$ , we choose

$$\nu_p = 1 + \frac{\eta^2}{2p^2}, \quad \eta > 0, \quad (1.47)$$

then  $c_\Theta$  defined in (1.46) can be bounded independently of  $p$ , see [CH22, Lemma 5.5] for a more in-depth view. In this case it holds

$$\beta_\Psi^2 = \beta_p^2 \geq 2p \geq 4.$$

Thus, the assumption of Lemma 1.8 on  $\Psi$  is satisfied for the leapfrog-Chebychev polynomials (1.45).

*Remark 1.19 (Implementation).* (i) It is important to note that one does not have to evaluate the polynomials (1.45) of degree  $p - 1$  in each step. Instead one calculates the action of a vector  $\mathbf{u}$  on  $\Psi(-\tau^2 \mathbf{A}_m)$  in an efficient way by a three-term recurrence relation, see [Car21, Algo. 4.2] for details. Here  $\mathbf{A}_m$  denotes the system matrix of  $\mathcal{L}_{v,h} \chi_m \mathcal{L}_{u,h}$  one obtains after choosing an appropriate basis of the dG space  $\mathbb{V}_h$ , see [HS19].  
 (ii) By sorting the degrees-of-freedom in an appropriate way [HS19, Stu17] one can see that  $\Psi(-\tau^2 \mathbf{A}_m)$  only acts on the few fine elements in  $\mathcal{T}_{h,f}$  plus two additional layers.

$\diamond$

## 1.4. Numerical examples

We support our theoretical findings numerically with three examples. First we substantiate the error bounds of Theorem 1.16 and we study the influence of stabilization within the leapfrog-Chebyshev local time-stepping (LFC-LTS) method. Afterward, we investigate the efficiency of the LFC-LTS method compared to the locally implicit (LI) and the original leapfrog scheme. The linear systems in the LI method are solved with the conjugate gradient method without preconditioning since it required only a few iterations (not more than four in our examples). We would like to stress that it is essential to utilize the correct inner product induced by the mass matrix within the conjugate gradient method.

The codes to reproduce our results are available at

<https://github.com/MalikScheifinger/DG-LTS-Maxwell>

The software is based on the FEM library `deal.II` [ABB<sup>+</sup>23] version 9.5 and the Maxwell toolbox `TiMaxdG` [CDK<sup>+</sup>]. Since we use `deal.II`, all our examples are done with quadrilateral mesh elements.

### 1.4.1. Linear Maxwell's equations

Linear Maxwell's equations in transverse-electric (TE) mode, see [Nie09, §2.3], are given by

$$\varepsilon \partial_t E_x = \partial_y H_z - J_x, \quad \Omega \times (0, T), \quad (1.48a)$$

$$\varepsilon \partial_t E_y = -\partial_x H_z - J_y, \quad \Omega \times (0, T), \quad (1.48b)$$

$$\mu \partial_t H_z = \partial_y E_x - \partial_x E_y, \quad \Omega \times (0, T), \quad (1.48c)$$

$$E(0) = E^0, \quad H(0) = H^0, \quad \Omega, \quad (1.48d)$$

$$E \times \mathbf{n} = 0, \quad \partial\Omega \times (0, T). \quad (1.48e)$$

For the spatial domain we choose  $\Omega = (0, 1)^2$  and final time  $T = 1$ . We set  $\varepsilon = \mu = 1$  and use initial values

$$E_x^0(x, y, t) = \cos(2\pi x) \sin(2\pi y), \quad (1.49a)$$

$$E_y^0(x, y, t) = -\sin(2\pi x) \cos(2\pi y), \quad (1.49b)$$

$$H_z^0(x, y, t) = 4\pi \cos(2\pi x) \cos(2\pi y), \quad (1.49c)$$

as well as right-hand sides

$$J_x(x, y, t) = -(1 + 8\pi^2) \cos(2\pi x) \sin(2\pi y) e^t, \quad (1.50a)$$

$$J_y(x, y, t) = (1 + 8\pi^2) \sin(2\pi x) \cos(2\pi y) e^t. \quad (1.50b)$$

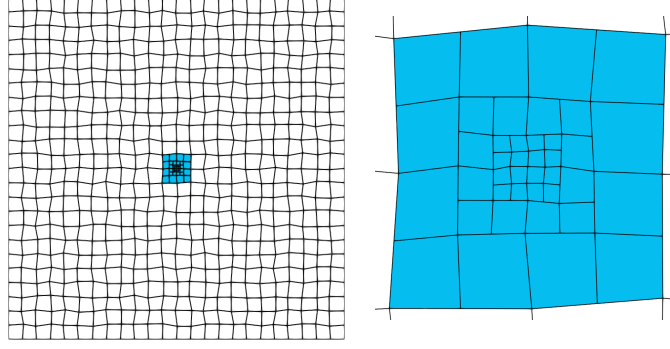
The exact solution to (1.48) with (1.49) and (1.50) is given by a variant of the cavity solution [HS16]

$$E_x(x, y, t) = \cos(2\pi x) \sin(2\pi y) e^t, \quad (1.51a)$$

$$E_y(x, y, t) = -\sin(2\pi x) \cos(2\pi y) e^t, \quad (1.51b)$$

$$H_z(x, y, t) = 4\pi \cos(2\pi x) \cos(2\pi y) e^t. \quad (1.51c)$$

Thereby we can compute the exact  $L^2$ -error of our scheme.

Figure 1.1.: Locally refined mesh of the domain  $(0, 1)^2$ .

We use the spatial mesh illustrated in Figure 1.1 in which we apply threefold refinement in the blue central box. The mesh has been randomly perturbed to circumvent superconvergence effects of the space discretization. The mesh elements are quadrilaterals. Moreover, the dG degree is chosen as  $k = 5$ .

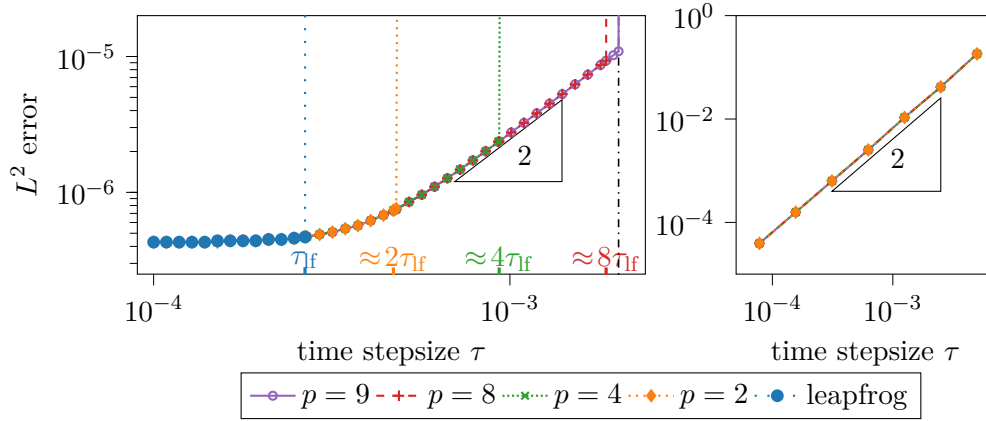


Figure 1.2.: Left: Error of the numerical solution of (1.48) with initial data given by (1.51) obtained by the leapfrog method and the local time-stepping method (1.9) using the filter (1.45) constructed with polynomial degrees  $p = 2, 4, 8, 9$  and stabilization (1.47) with  $\eta = 1$ . The space discretization is done with dG degree  $k = 5$  and a locally refined mesh at the center. The dash-dotted line depicts the maximal stable time step size of leapfrog method used on the coarse mesh  $\mathcal{T}_{h,lf}$ . The vertical dashed or dotted lines indicate the largest time step size that can be used by the corresponding modified method also indicated by the annotations at the x-axis. Right: Error of the LFC-LTS method with simultaneous refinement in space and time by a factor  $1/2$ . The space discretization is done with dG degree  $k = 2$ . Regardless of the filter used, second-order convergence is observed.

### 1.4.2. Order of convergence

We apply the LTS method (1.9) with the LFC filter (1.45) and the stabilization parameter (1.47) for  $\eta = 1$  and various values of  $p$ .

In Figure 1.2 on the left, we illustrate the stability and convergence behavior with LFC polynomials of degrees  $p = 2, 4, 8, 9$ . We observe second-order convergence in time until the error reaches a plateau stemming from the spatial discretization. We denote the maximal stable time stepsize of each method by either dashed or dotted vertical lines. Annotations at the bottom of the picture illustrate that by increasing the LFC polynomial degree  $p$ , the maximal stable time stepsize becomes larger and scales roughly as  $p\tau_{\text{CFL,lf}}$  where  $\tau_{\text{CFL,lf}}$  denotes the maximal stable time stepsize of the leapfrog method (blue) on the whole mesh. Obviously one cannot exceed the maximal stable time stepsize of the leapfrog method on the coarse part  $\mathcal{T}_{h,\text{lf}}$  illustrated by the vertical dash-dotted line. On the right side of Figure 1.2 we illustrate the convergence in space and time of the LFC-LTS method with  $p = 2, 4, 8, 9$  and a mesh which is refined two times at the center. We start with a maximal mesh diameter  $h_{\text{max}} = 0.1345$  and a time stepsize  $\tau = 0.005$  which is necessary to achieve stability for the LFC-LTS method with  $p = 2$ . We halve each value simultaneously and measure the  $L^2$  error at the endtime  $T = 1$ . We observe second-order convergence in space and time for all polynomial degrees  $p$ .

### 1.4.3. Necessity of stabilization

In this example we show that stabilization, i.e., choosing  $\eta > 0$ , is indispensable in the LFC-LTS method, see also [CH22] for second-order differential equations. For this we consider the following one-dimensional example

$$\partial_t u = -\partial_x v, \quad \Omega \times (0, T), \quad (1.52a)$$

$$\partial_t v = -\partial_x u, \quad \Omega \times (0, T), \quad (1.52b)$$

$$u(0) = u^0, \quad v(0) = v^0, \quad \Omega, \quad (1.52c)$$

$$u = 0, \quad \partial\Omega \times (0, T), \quad (1.52d)$$

on  $\Omega = (0, 1)$ . We choose the initial values such that

$$\begin{aligned} u(x, t) &= \sin(2\pi x) \cos(2\pi t) \\ v(x, t) &= -\cos(2\pi x) \sin(2\pi t). \end{aligned}$$

is the exact solution of (1.52).

We use a spatial grid where all cells have diameter  $h_{\text{max}} = 0.009975$  except for one cell in the middle of  $\Omega$  with diameter  $h_{\text{min}} = 0.0025 \approx h_{\text{max}}/4$ . In Figure 1.3 we show the error of the LFC-LTS method with polynomial degrees  $p = 3, 4, 5$  and stabilization (1.47) with  $\eta = 0$  and  $\eta = 0.1$ , on the left and the right picture, respectively. As we can see, for  $\eta = 0$ , the method becomes unstable for certain time stepsizes whereas the slightly stabilized method with  $\eta = 0.1$  has no deviations. This is caused by double roots of  $\Theta$  in the interval  $[0, \pi^2]$ , which is independent of  $p$ , cf. [CHS20, Theorem 5.2(a)]. The stabilization removes these double roots, as shown in Figure 1.4 for  $p = 4$  and  $\eta \in \{0, 0.1\}$ , where the first double root is marked in red.

### 1.4.4. Runtime comparison

Next, we compare the runtimes of the LFC-LTS method, the LI method, and the leapfrog method at two different examples. We revisit (1.48) on  $\Omega = (0, 4)^2$  and use an equidistant spatial grid with  $h \approx 0.0221$ . For the dG space we use the polynomial degree  $k = 2$ . Cells with center in the ball  $\|x\| \leq r$  with  $r \in \{0.1, 0.5\}$  are refined 2, 3, and 4 times. The minimal mesh

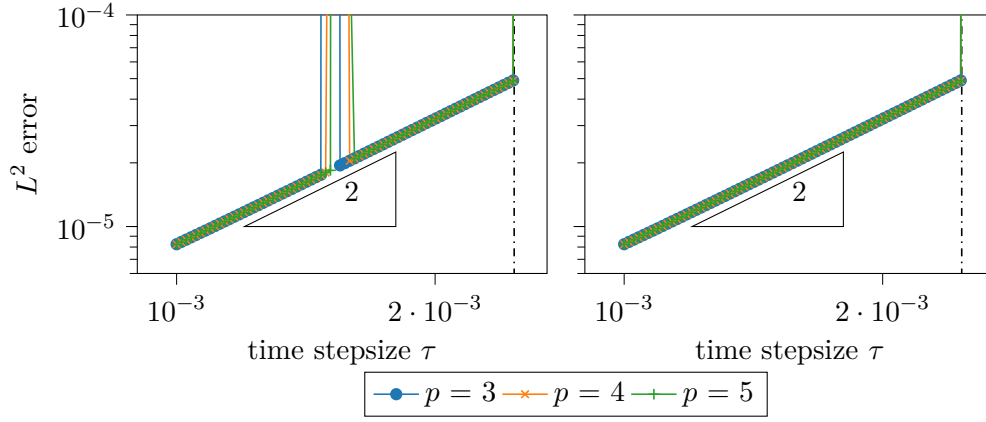


Figure 1.3.: Error of the LFC-LTS method for the example in Section 1.4.3 with polynomial degrees  $p = 3, 4, 5$ . Left: without stabilization ( $\eta = 0$ ), right: with stabilization ( $\eta = 0.1$ ). The dash-dotted line depicts the maximal stable time-step size of the leapfrog method on the coarse mesh  $\mathcal{T}_{h,\text{lf}}$ .

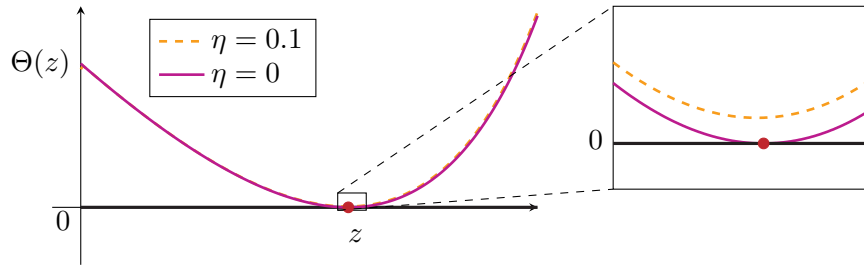


Figure 1.4.: Values  $\Theta(z)$  using the filter function (1.45) with  $p = 4$  and  $\eta \in \{0, 0.1\}$ . The thick red point marks the first root of  $\Theta$ .

| local refinements | 2         | 3         | 4         |
|-------------------|-----------|-----------|-----------|
| total dofs        | 1 782 756 | 1 823 823 | 1 984 932 |
| refined dofs      | 15 012    | 56 133    | 217 296   |
| runtime leapfrog  | 64.1sec   | 133.0sec  | 291.0sec  |
| runtime LI        | 50.6sec   | 57.3sec   | 125.0sec  |
| runtime LFC-LTS   | 31.4sec   | 37.2sec   | 74.3sec   |

Table 1.2.: Runtime comparison between the leapfrog, LI and LFC-LTS methods where the region  $r = 0.1$  was refined 2, 3 and 4 times.

| local refinements | 2         | 3         | 4         |
|-------------------|-----------|-----------|-----------|
| total dofs        | 2 097 522 | 3 141 369 | 7 298 127 |
| refined dofs      | 353 430   | 1 397 655 | 5 554 656 |
| runtime leapfrog  | 76.8sec   | 230sec    | 1 050sec  |
| runtime LI        | 83.5sec   | 334sec    | 2 140sec  |
| runtime LFC-LTS   | 45.8sec   | 156sec    | 1 160sec  |

Table 1.3.: Runtime comparison between the leapfrog, LI and LFC-LTS methods where the region  $r = 0.5$  was refined 2, 3 and 4 times.

diameter on the whole mesh is then respectively  $h_{\min} = 0.005524$ ,  $0.002762$ ,  $0.001381$  which corresponds to a ratio of  $h_{\max}/h_{\min} = 4, 8, 16$  respectively. For the leapfrog method we choose the maximal stable stepsizes  $\tau = 6.8 \cdot 10^{-4}$ ,  $3.4 \cdot 10^{-4}$ ,  $1.7 \cdot 10^{-4}$  respectively. For the local time-integration schemes we aim at choosing  $\tau = 2.2 \cdot 10^{-3}$  which is small enough to balance time and space discretization errors in each scenario. This was achieved by choosing the LFC polynomial degrees  $p = 4, 7, 14$  respectively.

In Table 1.2, we summarize the runtimes of the leapfrog, LI, and LFC-LTS method for  $r = 0.1$ . In this example, the LFC-LTS method outperforms the leapfrog method where the difference gets larger with an increasing number of local refinements. The runtimes of the LI method are shorter than the runtimes of the leapfrog method but longer than the runtimes of the LFC-LTS method. This can be explained by the fact that the LI method has to solve a linear system in each time step in the refined region, while the LFC-LTS method is fully explicit.

In Table 1.3, we show the data for  $r = 0.5$ . Due to the larger radius, the dofs in the refined section are much higher than in the situation of Table 1.2. This leads to a different behavior. In this example, the LFC-LTS method is at first faster than the leapfrog method but with more refinements the leapfrog method becomes faster. The LI method is outperformed by both the leapfrog and the LFC-LTS method.

Overall we see that the LFC-LTS method performs much better than the leapfrog and the LI method in regimes where the mesh is locally refined in a small region.

## 1.5. Concluding remarks

In this paper we have constructed a family of second-order in time local time-integration methods for central-fluxes dG discretizations of Friedrichs' systems by combining the leapfrog

scheme on the coarse part of the mesh with a modified explicit or implicit scheme on the fine part. We proved error bounds of the overall scheme under a CFL condition, which is independent of the fine part of the mesh, if the filter is chosen appropriately. Numerical experiments showed the efficiency of the new scheme compared to the standard leapfrog method.

For linear Maxwell equations, higher-order locally implicit methods have been investigated in [HKK22], and for discretizations with upwind fluxes we refer to [HS19].

## Acknowledgments

We thank Constantin Carle and Benjamin Dörich for inspiring discussions on local time integration and its error analysis and helpful comments on the manuscript.



## CHAPTER 2

---

### Local time integration for Friedrichs' systems with upwind-type stabilization

This chapter extends the central-flux local time-integration scheme from Chapter 1 by adding explicit stabilization on faces between elements in the leapfrog part  $\mathcal{T}_{h,\text{lf}}$ . The additional stabilization terms provide control of selected interior jumps and boundary traces. Since the stabilization terms are only applied on faces between elements in  $\mathcal{T}_{h,\text{lf}}$ , the fully explicit structure of the local time-integration method is preserved if a polynomial filter is used. The upwind fluxes considered in [HS19] are a special case of these interface penalties.

The main purpose of this chapter is to show that the stability and error analysis of the central-flux method carries over to this setting, but with a further CFL restriction due to the explicit stabilization terms. Under the combined CFL condition, the method satisfies a stability estimate and an error bound of the same temporal order as in the central-flux case, and an enhanced spatial convergence order on  $\mathcal{T}_{h,\text{lf}}$ .

#### 2.1. Space discretization with stabilization

This section collects the ingredients needed for the extension of the central-flux discretization. We first refine the face partition according to the leapfrog and modified mesh regions and introduce the enlarged modified region used in the estimates. We then define the stabilization operators on the leapfrog faces, record their structural properties, and derive the inverse, projection, and discrete-operator estimates needed for the stability and error analysis.

##### 2.1.1. Notation

The stabilization terms and the subsequent error analysis require a partition of the face set  $\mathcal{F}_h = \mathcal{F}_h^{\text{int}} \dot{\cup} \mathcal{F}_h^{\text{bnd}}$ . We decompose the interior faces  $\mathcal{F}_h^{\text{int}}$  into three disjoint subsets according to the mesh regions they connect. By  $\mathcal{F}_{h,\text{lf}}^{\text{int}}$  we denote the set of interior faces between two elements in  $\mathcal{T}_{h,\text{lf}}$ , by  $\mathcal{F}_{h,m}^{\text{int}}$  the set of interior faces between two elements in  $\mathcal{T}_{h,m}$ , and by  $\mathcal{F}_{h,\text{lf}|m}^{\text{int}}$  the set of interior faces connecting an element in  $\mathcal{T}_{h,\text{lf}}$  with one in  $\mathcal{T}_{h,m}$ . The sets  $\mathcal{T}_{h,\text{lf}}$  and  $\mathcal{T}_{h,m}$  are defined in Section 1.3. Analogously, the boundary faces are split as  $\mathcal{F}_h^{\text{bnd}} = \mathcal{F}_{h,\text{lf}}^{\text{bnd}} \dot{\cup} \mathcal{F}_{h,m}^{\text{bnd}}$ . For a compact notation in the error analysis, we further collect all non-leapfrog interior faces as  $\mathcal{F}_{h,m+}^{\text{int}} = \mathcal{F}_{h,\text{lf}|m}^{\text{int}} \dot{\cup} \mathcal{F}_{h,m}^{\text{int}}$ .

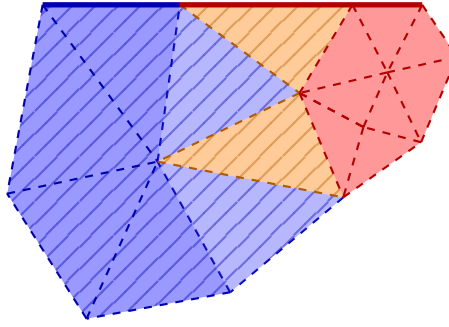


Figure 2.1.: Excerpt of the mesh partitioning described in Section 2.1.1. The leapfrog region  $\mathcal{T}_{h,\text{lf}}$  (blue shades) and the modified region  $\mathcal{T}_{h,m}$  (red–orange) tile the domain.  $\mathcal{T}_{h,m}$  consists of the fine cells  $\mathcal{T}_{h,f}$  (red) together with their coarse neighbors (orange). The extended modified region  $\mathcal{T}_{h,m+}$  contains  $\mathcal{T}_{h,m}$  (red–orange) and another layer of coarse neighbors (light blue). The blue shaded and orange cells form the coarse mesh. Interfaces are drawn with dashed lines and boundary faces with solid lines. The interior faces split into leapfrog faces  $\mathcal{F}_{h,\text{lf}}^{\text{int}}$  (dashed blue), mixed faces  $\mathcal{F}_{h,\text{lf}|m}^{\text{int}}$  (dashed orange) connecting the two regions, and modified faces  $\mathcal{F}_{h,m}^{\text{int}}$  (dashed red). The boundary faces split into  $\mathcal{F}_{h,\text{lf}}^{\text{bnd}}$  (solid blue) and  $\mathcal{F}_{h,m}^{\text{bnd}}$  (solid red).

In addition to the mesh regions from Section 1.3, we define the extended modified region

$$\mathcal{T}_{h,m+} = \{K \in \mathcal{T}_h \mid \exists \widehat{K} \in \mathcal{T}_{h,m} \text{ s.t. } K, \widehat{K} \text{ share a face}\},$$

which consists of  $\mathcal{T}_{h,m}$  together with all elements sharing a face with elements in  $\mathcal{T}_{h,m}$ . By construction, the mesh regions are nested, i.e.,  $\mathcal{T}_{h,f} \subseteq \mathcal{T}_{h,m} \subseteq \mathcal{T}_{h,m+}$ . Figure 2.1 illustrates the resulting partition of the mesh and its faces.

### 2.1.2. Stabilization operators

We now define the stabilization operators motivated by [DPE12, §7.3.2]. They will be added to the semi-discrete system (1.4). The key idea is to penalize jumps of the discrete solution across interior faces and traces on boundary faces in the leapfrog region  $\mathcal{T}_{h,\text{lf}}$ , weighted by appropriate fields.

**Definition 2.1.** *Given the fields*

$$S_{i,u}^F, S_{b,u}^F \in \mathbb{R}^{m_u \times m_u}, \quad S_{i,v}^F, S_{b,v}^F \in \mathbb{R}^{m_v \times m_v}, \quad \text{for all } F \in \mathcal{F}_h, \quad (2.1)$$

we define  $\mathcal{S}_{u,h}^{\text{lf}} : V_h^{\mathcal{L}_u} \rightarrow \mathbb{V}_h^u$  and  $\mathcal{S}_{v,h}^{\text{lf}} : V_h^{\mathcal{L}_v} \rightarrow \mathbb{V}_h^v$  by

$$(\mathcal{S}_{u,h}^{\text{lf}} u, \psi_h)_\Omega := \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{int}}} (S_{i,u}^F \llbracket u \rrbracket_F, \llbracket \psi_h \rrbracket_F)_{L^2(F)} + \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{bnd}}} (S_{b,u}^F u, \psi_h)_{L^2(F)}, \quad \psi_h \in \mathbb{V}_h^u, \quad (2.2a)$$

$$(\mathcal{S}_{v,h}^{\text{lf}} v, \phi_h)_\Omega := \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{int}}} (S_{i,v}^F \llbracket v \rrbracket_F, \llbracket \phi_h \rrbracket_F)_{L^2(F)} + \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{bnd}}} (S_{b,v}^F v, \phi_h)_{L^2(F)}, \quad \phi_h \in \mathbb{V}_h^v. \quad (2.2b)$$

The fields (2.1) are not uniquely determined by the problem, with possible choices given in Example 2.3. Following [DPE12, §7.3.2], we impose the following conditions on their

choice to obtain the properties needed for our analysis. In words, the conditions are threefold. Symmetry and positive semi-definiteness ensure that the operators (2.2) induce the seminorms of Definition 2.5. The vanishing weighted jump and boundary traces are a consistency requirement, i.e., the stabilization is invisible to functions in  $\mathcal{D}(\mathcal{L}) \cap H^1(\mathcal{T}_h)^m$ .

**Assumption 2.2.** *The fields defined in (2.1) are symmetric and positive semi-definite. Moreover,*

$$S_{i,u}^F \llbracket u \rrbracket_F = S_{i,v}^F \llbracket v \rrbracket_F = 0, \quad \text{for all } F \in \mathcal{F}_{h,\text{lf}}^{\text{int}}, \quad (2.3)$$

$$S_{b,u}^F u = S_{b,v}^F v = 0, \quad \text{for all } F \in \mathcal{F}_{h,\text{lf}}^{\text{bnd}}, \quad (2.4)$$

for  $u \in \mathcal{D}(\mathcal{L}_u) \cap H^1(\mathcal{T}_h)$  and  $v \in \mathcal{D}(\mathcal{L}_v) \cap H^1(\mathcal{T}_h)$ . Additionally, for  $w \in \{u, v\}$

$$\eta_{w,i} (L_{\partial,w}^F L_{\partial,w}^F)^\top)^{1/2} \leq S_{i,w}^F \leq \hat{\eta}_{w,i} \mathbf{I}_{m_w}, \quad (2.5)$$

$$\eta_{w,b} ((L_{\partial,w}^F - L_{\Gamma,w})^\top (L_{\partial,w}^F - L_{\Gamma,w}))^{1/2} \leq S_{b,w}^F \leq \hat{\eta}_{w,b} \mathbf{I}_{m_w}, \quad (2.6)$$

uniformly on each face  $F \in \mathcal{F}_h$  with positive constants  $\eta_{w,i}, \hat{\eta}_{w,i}, \eta_{w,b}, \hat{\eta}_{w,b}$ . The fields  $L_{\partial,w}^F, L_{\Gamma,w}$  are stated in [DHK<sup>+</sup>23, §10, §13]. For symmetric matrices  $A, B \in \mathbb{R}^{k \times k}$ , the matrix inequality  $A \leq B$  means that  $x^\top A x \leq x^\top B x$  for all  $x \in \mathbb{R}^k$ , or equivalently, that  $B - A$  is positive semi-definite.

*Example 2.3.* We give examples of stabilization fields for the acoustic wave equation in div-grad form and for Maxwell's equations, following [DPE12, §7.3.4–7.3.5].

- (a) For the acoustic wave equation the pressure  $u = p$  is scalar,  $m_u = 1$ , and the velocity is vector valued,  $m_v = d$ , with face fields  $L_{\partial,u}^F = \mathbf{n}^F$  and  $L_{\partial,v}^F = (\mathbf{n}^F)^\top$ . A possible choice of stabilization fields is

$$S_{i,u}^F = \eta_1, \quad S_{i,v}^F = \eta_2 \mathbf{n}^F (\mathbf{n}^F)^\top, \quad S_{b,u}^F = \eta_3, \quad S_{b,v}^F = 0,$$

with positive user-dependent constants  $\eta_1, \eta_2$ , and  $\eta_3$ , penalizing the jump of  $u$  and the normal jump of  $v$ . On the boundary, only the first component  $u$  is penalized.

- (b) For the three-dimensional Maxwell's equations  $u = E$  and  $v = H$  are vector-valued with  $m_u = m_v = 3$ , and face fields acting as  $L_{\partial,u}^F u = -\mathbf{n}^F \times E$  and  $L_{\partial,v}^F v = \mathbf{n}^F \times H$ . A possible choice of stabilization fields is the tangential projection

$$S_{i,u}^F = S_{i,v}^F = \eta_{1/2} (\mathbf{I}_3 - \mathbf{n}^F (\mathbf{n}^F)^\top), \quad S_{b,u}^F = \eta_3 (\mathbf{I}_3 - \mathbf{n}^F (\mathbf{n}^F)^\top), \quad S_{b,v}^F = 0,$$

with positive user-dependent constants  $\eta_1, \eta_2$ , and  $\eta_3$ , penalizing the tangential jumps of  $u$  and  $v$ . On the boundary, only the first component  $u$  is penalized.

◇

With the stabilization operators at hand, given a constant  $\alpha \in [0, 1]$ , the stabilized semi-discrete system reads

$$\partial_t u_h = \mathcal{L}_{v,h} v_h - \alpha S_{u,h}^{\text{lf}} u_h + g_{u,h}, \quad \text{on } \mathbb{R}_+, \quad (2.7a)$$

$$\partial_t v_h = \mathcal{L}_{u,h} u_h - \alpha S_{v,h}^{\text{lf}} v_h + g_{v,h}, \quad \text{on } \mathbb{R}_+, \quad (2.7b)$$

or compactly written

$$\partial_t x_h = (\mathcal{L}_h - \alpha S_h^{\text{lf}}) x_h + g_h, \quad \text{on } \mathbb{R}_+, \quad (2.8a)$$

with

$$x_h = \begin{pmatrix} u_h \\ v_h \end{pmatrix}, \quad \mathcal{L}_h = \begin{pmatrix} 0 & \mathcal{L}_{v,h} \\ \mathcal{L}_{u,h} & 0 \end{pmatrix}, \quad \mathcal{S}_h^{\text{lf}} = \begin{pmatrix} \mathcal{S}_{u,h}^{\text{lf}} & 0 \\ 0 & \mathcal{S}_{v,h}^{\text{lf}} \end{pmatrix}, \quad g_h = \begin{pmatrix} g_{u,h} \\ g_{v,h} \end{pmatrix}. \quad (2.8b)$$

The system (2.8a) is equipped with the  $L^2$ -projected initial condition  $x_h(0) = \pi_h x^0$  and  $L^2$ -projected forcing  $g_h = \pi_h g$ . The parameter  $\alpha \in [0, 1]$  controls the strength of the stabilization: for  $\alpha = 0$  we recover the central-flux formulation (1.4), while an increasing  $\alpha$  strengthens the penalty.

### 2.1.3. Properties of the stabilization operators

The stabilization operators  $\mathcal{S}_{u,h}^{\text{lf}}$  and  $\mathcal{S}_{v,h}^{\text{lf}}$  share the same structural properties as their counterparts in [HS19], which we collect in the following Lemma 2.4 for later reference.

**Lemma 2.4.** *Let  $w \in \{u, v\}$ . The stabilization operator  $\mathcal{S}_{w,h}^{\text{lf}}$  from Definition 2.1 is consistent, symmetric, and positive semidefinite, i.e.,*

$$\mathcal{S}_{w,h}^{\text{lf}} w = 0, \quad w \in \mathcal{D}(\mathcal{L}_w) \cap H^1(\mathcal{T}_h)^{m_w}, \quad (2.9a)$$

$$(\mathcal{S}_{w,h}^{\text{lf}} w_h, \tilde{w}_h)_\Omega = (w_h, \mathcal{S}_{w,h}^{\text{lf}} \tilde{w}_h)_\Omega, \quad w_h, \tilde{w}_h \in \mathbb{V}_h^w, \quad (2.9b)$$

$$(\mathcal{S}_{w,h}^{\text{lf}} w_h, w_h)_\Omega \geq 0, \quad w_h \in \mathbb{V}_h^w. \quad (2.9c)$$

*Proof.* All three properties follow directly from Definition 2.1 and Assumption 2.2.  $\square$

The stabilization operators thus penalize jumps across the interior leapfrog faces  $\mathcal{F}_{h,\text{lf}}^{\text{int}}$  and traces on the boundary faces  $\mathcal{F}_{h,\text{lf}}^{\text{bnd}}$ , weighted by the respective stabilization fields, while remaining invisible to functions in  $\mathcal{D}(\mathcal{L}_w) \cap H^1(\mathcal{T}_h)^{m_w}$ . The symmetry (2.9b) and positive semi-definiteness (2.9c) allow us to define a seminorm on  $\mathbb{V}_h^w$ , which plays a central role in the stability and error analysis. By (2.9a), these seminorms extend to functions in  $V_h^{\mathcal{L}_w} = \mathbb{V}_h^w + \mathcal{D}(\mathcal{L}_w) \cap H^1(\mathcal{T}_h)^{m_w}$ .

**Definition 2.5.** *For  $u \in V_h^{\mathcal{L}_u}$ ,  $v \in V_h^{\mathcal{L}_v}$ , we define the stabilization seminorms*

$$|u|_{\mathcal{S}_{u,h}^{\text{lf}}}^2 := (\mathcal{S}_{u,h}^{\text{lf}} u, u)_\Omega, \quad (2.10a)$$

$$|v|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 := (\mathcal{S}_{v,h}^{\text{lf}} v, v)_\Omega. \quad (2.10b)$$

Moreover, for  $x = (u \ v) \in V_h^{\mathcal{L}}$ , we define

$$|x|_{\mathcal{S}_h^{\text{lf}}}^2 = |u|_{\mathcal{S}_{u,h}^{\text{lf}}}^2 + |v|_{\mathcal{S}_{v,h}^{\text{lf}}}^2. \quad (2.10c)$$

The remainder of this section collects several auxiliary estimates that quantify the interaction of the stabilization seminorms with inverse inequalities, projection errors, and the discrete Friedrichs' operators. The estimates are performed largely analogously to those in [HS19] for the Maxwell setting, adapted here to the two-field Friedrichs' framework. They will be needed in the stability and error analysis of Sections 2.3 and 2.4.

We begin by establishing an inverse inequality, relating the stabilization seminorm to a weighted  $L^2$ -norm on the leapfrog region  $\mathcal{T}_{h,\text{lf}}$ . This is the analogue of [HS19, Lem. 3.5] in our Friedrichs' setting. While the CFL conditions in Section 2.3 are formulated in terms of operator norms, the following estimate makes the dependence on the time-step size  $\tau$ , the mesh diameter  $h$ , and the material tensor  $\mathcal{M}$ , defined in Section 1.2, explicit.

**Lemma 2.6.** For  $u_h \in \mathbb{V}_h^u$ ,  $v_h \in \mathbb{V}_h^v$ , the following estimates hold:

$$|u_h|_{\mathcal{S}_{u,h}^{\text{lf}}}^2 \leq C_{\text{inv},\mathcal{S}_u^{\text{lf}}} \|h^{-1/2} u_h\|_{\mathcal{T}_{h,\text{lf}}}^2, \quad (2.11a)$$

$$|v_h|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 \leq C_{\text{inv},\mathcal{S}_v^{\text{lf}}} \|h^{-1/2} v_h\|_{\mathcal{T}_{h,\text{lf}}}^2, \quad (2.11b)$$

where  $C_{\text{inv},\mathcal{S}_u^{\text{lf}}}$ ,  $C_{\text{inv},\mathcal{S}_v^{\text{lf}}}$  are independent of  $h$ .

*Proof.* To prove (2.11a), we use Assumption 2.2 and the discrete trace inequality (A.1) for an interface  $F \in \mathcal{F}_{h,\text{lf}}^{\text{int}}$  between two elements  $K_1^F, K_2^F \in \mathcal{T}_{h,\text{lf}}$  and  $u_h \in \mathbb{V}_h^u$

$$\begin{aligned} (S_{i,u}^F \llbracket u_h \rrbracket_F, \llbracket u_h \rrbracket_F)_{L^2(F)} &\leq \hat{\eta}_{u,i} \|\llbracket u_h \rrbracket_F\|_{L^2(F)}^2 \\ &\leq 2\hat{\eta}_{u,i} (\|u_h|_{K_1^F}\|_{L^2(F)}^2 + \|u_h|_{K_2^F}\|_{L^2(F)}^2) \\ &\leq 2\hat{\eta}_{u,i} C_{\text{tr}}^2 (\|h^{-1/2} u_h\|_{L^2(K_1^F)}^2 + \|h^{-1/2} u_h\|_{L^2(K_2^F)}^2) \\ &\leq 2\hat{\eta}_{u,i} C_{\text{tr}}^2 (\|\mathcal{M}_u^{-1}\|_{L^\infty(K_1^F)}^2 \|h^{-1/2} u_h\|_{K_1^F}^2 + \|\mathcal{M}_u^{-1}\|_{L^\infty(K_2^F)}^2 \|h^{-1/2} u_h\|_{K_2^F}^2), \end{aligned}$$

where the material tensor  $\mathcal{M}_u$  is defined in Section 1.2. For a boundary face  $F \in \mathcal{F}_h^{\text{bnd}}$  and the corresponding cell  $K$  touching that boundary face, we estimate

$$\begin{aligned} (S_{b,u}^F u_h, u_h)_{L^2(F)} &\leq \hat{\eta}_{u,b} \|u_h\|_{L^2(F)}^2 \\ &\leq \hat{\eta}_{u,b} C_{\text{tr}}^2 \|h^{-1/2} u_h\|_{L^2(K)}^2 \\ &\leq \hat{\eta}_{u,b} C_{\text{tr}}^2 \|\mathcal{M}_u^{-1}\|_{L^\infty(K)}^2 \|h^{-1/2} u_h\|_K^2. \end{aligned}$$

Inserting the above estimates into (2.10) and using that each cell has at most  $N_\partial$  faces yields

$$|u_h|_{\mathcal{S}_{u,h}^{\text{lf}}}^2 \leq C_{\text{tr}}^2 (2\hat{\eta}_{u,i} + \hat{\eta}_{u,b}) N_\partial \|\mathcal{M}_u^{-1}\|_{L^\infty(\mathcal{T}_{h,\text{lf}})}^2 \|h^{-1/2} u_h\|_{\mathcal{T}_{h,\text{lf}}}^2.$$

Note that the constant  $C_{\text{tr}}$  only comes from interface to cell estimates on  $\mathcal{T}_{h,\text{lf}}$ . The second inequality (2.11b) follows analogously.  $\square$

The following Lemma 2.7 bounds the interaction of the stabilization operator with the projection error. It shows that the stabilization applied to the projection error can be controlled by the stabilization seminorm of the test function and a Sobolev seminorm of a sufficiently smooth function on the leapfrog region  $\mathcal{T}_{h,\text{lf}}$ . Compared to the central-flux estimate (1.37b), the bound gains an additional half power of  $h$  but is restricted to the leapfrog region  $\mathcal{T}_{h,\text{lf}}$ . Lemma 2.7 is the analogue of [HS19, Thm. 3.8] in our Friedrichs' setting.

**Lemma 2.7.** Let  $w \in \{u, v\}$ . For  $w \in \mathcal{D}(\mathcal{L}_w) \cap H^{k+1}(\mathcal{T}_h)^{m_w}$  and  $\tilde{w}_h \in \mathbb{V}_h^w$ , the following estimate holds:

$$(\mathcal{S}_{w,h}^{\text{lf}} e_{\pi,w}, \tilde{w}_h)_\Omega \leq C_{\pi,\mathcal{S}_w^{\text{lf}}} |\tilde{w}_h|_{\mathcal{S}_{w,h}^{\text{lf}}} |h^{k+1/2} w|_{k+1,\mathcal{T}_{h,\text{lf}}}, \quad (2.12)$$

where  $C_{\pi,\mathcal{S}_w^{\text{lf}}}$  is independent of  $h$ .

*Proof.* Using the Cauchy–Schwarz inequality in  $L^2(F)^{m_w}$  and afterwards in  $\mathbb{R}^{|\mathcal{F}_{h,\text{lf}}^{\text{int}}| + |\mathcal{F}_{h,\text{lf}}^{\text{bnd}}|}$ , we estimate

$$(\mathcal{S}_{w,h}^{\text{lf}} e_{\pi,w}, \tilde{w}_h)_\Omega \leq \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{int}}} \|(S_{i,w}^F)^{1/2} \llbracket e_{\pi,w} \rrbracket_F\|_{L^2(F)} \|(S_{i,w}^F)^{1/2} \llbracket \tilde{w}_h \rrbracket_F\|_{L^2(F)}$$

$$\begin{aligned}
 & + \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{bnd}}} \|(S_{b,w}^F)^{1/2} e_{\pi,w}\|_{L^2(F)} \|(S_{b,w}^F)^{1/2} \tilde{w}_h\|_{L^2(F)} \\
 & \leq \left( \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{int}}} \|(S_{i,w}^F)^{1/2} \llbracket e_{\pi,w} \rrbracket_F\|_{L^2(F)}^2 + \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{bnd}}} \|(S_{b,w}^F)^{1/2} e_{\pi,w}\|_{L^2(F)}^2 \right)^{1/2} \\
 & \quad \cdot \left( \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{int}}} \|(S_{i,w}^F)^{1/2} \llbracket \tilde{w}_h \rrbracket_F\|_{L^2(F)}^2 + \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{bnd}}} \|(S_{b,w}^F)^{1/2} \tilde{w}_h\|_{L^2(F)}^2 \right)^{1/2}.
 \end{aligned}$$

Applying the projection error estimate (A.3), we get for an interface  $F \in \mathcal{F}_{h,\text{lf}}^{\text{int}}$  between two elements  $K_1^F, K_2^F \in \mathcal{T}_{h,\text{lf}}$  and  $e_{\pi,w} \in V_h^{\mathcal{L}_w}$

$$(S_{i,w}^F \llbracket e_{\pi,w} \rrbracket_F, \llbracket e_{\pi,w} \rrbracket_F)_{L^2(F)} \leq 2\hat{\eta}_{w,i} C_{\pi,\partial}^2 (|h^{k+1/2} w|_{k+1,K_1^F}^2 + |h^{k+1/2} w|_{k+1,K_2^F}^2).$$

For a boundary face  $F \in \mathcal{F}_{h,\text{lf}}^{\text{bnd}}$  of a cell  $K$ , we get

$$(S_{i,w}^F e_{\pi,w}, e_{\pi,w})_{L^2(F)} \leq \hat{\eta}_{w,b} C_{\pi,\partial}^2 |h^{k+1/2} w|_{k+1,K}^2.$$

Altogether, using that each cell has at most  $N_\partial$  faces yields

$$(\mathcal{S}_{w,h}^{\text{lf}} e_{\pi,w}, \tilde{w}_h) \leq (2\hat{\eta}_{w,i} + \hat{\eta}_{w,b})^{1/2} C_{\pi,\partial} N_\partial^{1/2} |h^{k+1/2} w|_{k+1,\mathcal{T}_{h,\text{lf}}} |\tilde{w}_h|_{\mathcal{S}_{w,h}^{\text{lf}}},$$

proving the statement.  $\square$

The last auxiliary estimate of this section splits the action of the discrete Friedrichs' operators on the projection error into two contributions. One is controlled by the stabilization seminorm on  $\mathcal{T}_{h,\text{lf}}$  and the other by the  $L^2$ -norm on the extended modified region  $\mathcal{T}_{h,m+}$ . Lemma 2.8 is the analogue of [HS19, Thm. 3.8] in our Friedrichs' setting.

**Lemma 2.8.** *For  $u \in \mathcal{D}(\mathcal{L}_u) \cap H^{k+1}(\mathcal{T}_h)^{m_u}$ ,  $v \in \mathcal{D}(\mathcal{L}_v) \cap H^{k+1}(\mathcal{T}_h)^{m_v}$ , the following estimates hold:*

$$\begin{aligned}
 (\mathcal{L}_{u,h} e_{\pi,u}, \phi_h)_\Omega & \leq C_{\pi,\mathcal{L}_u,c}^{\text{up}} |\phi_h|_{\mathcal{S}_{v,h}^{\text{lf}}} |h^{k+1/2} u|_{k+1,\mathcal{T}_{h,\text{lf}}} \\
 & \quad + C_{\pi,\mathcal{L}_u,f}^{\text{up}} \|\phi_h\|_{\mathcal{T}_{h,m+}} |h^k u|_{k+1,\mathcal{T}_{h,m+}}, \quad \text{for all } \phi_h \in \mathbb{V}_h^v,
 \end{aligned} \tag{2.13a}$$

$$\begin{aligned}
 (\mathcal{L}_{v,h} e_{\pi,v}, \psi_h)_\Omega & \leq C_{\pi,\mathcal{L}_v,c}^{\text{up}} |\psi_h|_{\mathcal{S}_{u,h}^{\text{lf}}} |h^{k+1/2} v|_{k+1,\mathcal{T}_{h,\text{lf}}} \\
 & \quad + C_{\pi,\mathcal{L}_v,f}^{\text{up}} \|\psi_h\|_{\mathcal{T}_{h,m+}} |h^k v|_{k+1,\mathcal{T}_{h,m+}}, \quad \text{for all } \psi_h \in \mathbb{V}_h^u,
 \end{aligned} \tag{2.13b}$$

where  $C_{\pi,\mathcal{L}_u,c}^{\text{up}}$ ,  $C_{\pi,\mathcal{L}_v,c}^{\text{up}}$  and  $C_{\pi,\mathcal{L}_u,f}^{\text{up}}$ ,  $C_{\pi,\mathcal{L}_v,f}^{\text{up}}$  are independent of  $h$ .

*Proof.* To prove (2.13a), we use the weak formulation (A.6a) and  $\mathcal{L}_v \phi_h|_K \in \mathbb{V}_h^u$ . Since  $e_{\pi,u}$  is  $L^2$ -orthogonal to  $\mathbb{V}_h^u$ , the volume term drops. We then proceed as in the proof of Lemma 2.7 using the Cauchy–Schwarz inequality

$$(\mathcal{L}_{u,h} e_{\pi,u}, \phi_h)_\Omega = - \sum_{K \in \mathcal{T}_h} (e_{\pi,u}, \mathcal{L}_v \phi_h)_K + \sum_{F \in \mathcal{F}_h^{\text{int}}} (\llbracket e_{\pi,u} \rrbracket_F, L_{\partial,v}^F \llbracket \phi_h \rrbracket_F)_{L^2(F)}$$

$$\begin{aligned}
 & + \sum_{F \in \mathcal{F}_h^{\text{bnd}}} (e_{\pi,u}, (L_{\partial,v}^F - L_{\Gamma,v})\phi_h)_{L^2(F)} \\
 = & \sum_{F \in \mathcal{F}_h^{\text{int}}} (\llbracket e_{\pi,u} \rrbracket_F, L_{\partial,v}^F \llbracket \phi_h \rrbracket_F)_{L^2(F)} + \sum_{F \in \mathcal{F}_h^{\text{bnd}}} (e_{\pi,u}, (L_{\partial,v}^F - L_{\Gamma,v})\phi_h)_{L^2(F)} \\
 = & \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{int}}} (\llbracket e_{\pi,u} \rrbracket_F, L_{\partial,v}^F \llbracket \phi_h \rrbracket_F)_{L^2(F)} + \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{bnd}}} (e_{\pi,u}, (L_{\partial,v}^F - L_{\Gamma,v})\phi_h)_{L^2(F)} \\
 & + \sum_{F \in \mathcal{F}_{h,\text{m}+}^{\text{int}}} (\llbracket e_{\pi,u} \rrbracket_F, L_{\partial,v}^F \llbracket \phi_h \rrbracket_F)_{L^2(F)} + \sum_{F \in \mathcal{F}_{h,\text{m}}^{\text{bnd}}} (e_{\pi,u}, (L_{\partial,v}^F - L_{\Gamma,v})\phi_h)_{L^2(F)} \\
 \leq & \left( \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{int}}} \|\llbracket e_{\pi,u} \rrbracket_F\|_{L^2(F)}^2 + \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{bnd}}} \|e_{\pi,u}\|_{L^2(F)}^2 \right)^{1/2} \\
 & \cdot \left( \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{int}}} \|L_{\partial,v}^F \llbracket \phi_h \rrbracket_F\|_{L^2(F)}^2 + \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{bnd}}} \|(L_{\partial,v}^F - L_{\Gamma,v})\phi_h\|_{L^2(F)}^2 \right)^{1/2} \\
 & + \left( \sum_{F \in \mathcal{F}_{h,\text{m}+}^{\text{int}}} \frac{1}{\gamma} \|\llbracket e_{\pi,u} \rrbracket_F\|_{L^2(F)}^2 + \sum_{F \in \mathcal{F}_{h,\text{m}}^{\text{bnd}}} \frac{1}{\gamma} \|e_{\pi,u}\|_{L^2(F)}^2 \right)^{1/2} \\
 & \cdot \left( \sum_{F \in \mathcal{F}_{h,\text{m}+}^{\text{int}}} \gamma \|L_{\partial,v}^F \llbracket \phi_h \rrbracket_F\|_{L^2(F)}^2 + \sum_{F \in \mathcal{F}_{h,\text{m}}^{\text{bnd}}} \gamma \|(L_{\partial,v}^F - L_{\Gamma,v})\phi_h\|_{L^2(F)}^2 \right)^{1/2}
 \end{aligned}$$

For faces  $F \in \mathcal{F}_{h,\text{lf}}^{\text{int}}$ , we use the lower bounds from Assumption 2.2 to estimate

$$\begin{aligned}
 \|L_{\partial,v}^F \llbracket \phi_h \rrbracket_F\|_{L^2(F)}^2 & = (L_{\partial,u}^F L_{\partial,v}^F \llbracket \phi_h \rrbracket_F, \llbracket \phi_h \rrbracket_F)_{L^2(F)} \\
 & \leq C_{\partial,v}^F ((L_{\partial,u}^F L_{\partial,v}^F)^{1/2} \llbracket \phi_h \rrbracket_F, \llbracket \phi_h \rrbracket_F)_{L^2(F)} \\
 & \leq C_{\partial,v}^F \eta_{v,i}^{-1} (S_{i,v}^F \llbracket \phi_h \rrbracket_F, \llbracket \phi_h \rrbracket_F)_{L^2(F)},
 \end{aligned}$$

and similarly for faces  $F \in \mathcal{F}_{h,\text{lf}}^{\text{bnd}}$

$$\|(L_{\partial,v}^F - L_{\Gamma,v})\phi_h\|_{L^2(F)}^2 \leq C_{\Gamma,v}^F \eta_{v,b}^{-1} (S_{b,v}^F \llbracket \phi_h \rrbracket_F, \llbracket \phi_h \rrbracket_F)_{L^2(F)},$$

with constants

$$\begin{aligned}
 C_{\partial,v}^F & = \|(L_{\partial,u}^F L_{\partial,v}^F)^{1/2}\|_{L^\infty(F)}, \\
 C_{\Gamma,v}^F & = \|((L_{\partial,v}^F - L_{\Gamma,v})^\top (L_{\partial,v}^F - L_{\Gamma,v}))^{1/2}\|_{L^\infty(F)}.
 \end{aligned}$$

We use the discrete trace inequality (A.1) for interfaces  $F \in \mathcal{F}_{h,\text{m}+}^{\text{int}}$  to estimate

$$\|L_{\partial,v}^F \llbracket \phi_h \rrbracket_F\|_{L^2(F)}^2 \leq 2\tilde{C}_{\partial,v}^F C_{\text{tr}}^2 (\|\mathcal{M}_v^{-1}\|_{L^\infty(K_1^F)}^2 \|h^{-1/2}\phi_h\|_{K_1^F}^2 + \|\mathcal{M}_v^{-1}\|_{L^\infty(K_2^F)}^2 \|h^{-1/2}\phi_h\|_{K_2^F}^2),$$

and for boundary faces  $F \in \mathcal{F}_{h,\text{m}}^{\text{bnd}}$

$$\|(L_{\partial,v}^F - L_{\Gamma,v})\phi_h\|_F^2 \leq \tilde{C}_{\Gamma,v}^F C_{\text{tr}}^2 \|\mathcal{M}_v^{-1}\|_{L^\infty(K)}^2 \|h^{-1/2}\phi_h\|_K^2,$$

with constants

$$\begin{aligned}\tilde{C}_{\partial,v}^F &= \|L_{\partial,u}^F L_{\partial,v}^F\|_{L^\infty(F)}^2, \\ \tilde{C}_{\Gamma,v}^F &= \|(L_{\partial,v}^F - L_{\Gamma,v})^\top (L_{\partial,v}^F - L_{\Gamma,v})\|_{L^\infty(F)}^2.\end{aligned}$$

Using the projection error estimate (A.3), we get for an interface  $F \in \mathcal{F}_h^{\text{int}}$  between two elements  $K_1^F, K_2^F \in \mathcal{T}_h$

$$\|\llbracket e_{\pi,w} \rrbracket_F\|_{L^2(F)}^2 \leq 2C_{\pi,\partial}^2 \left( |h^{k+1/2}w|_{k+1,K_1^F}^2 + |h^{k+1/2}w|_{k+1,K_2^F}^2 \right),$$

and for a boundary face  $F \in \mathcal{F}_h^{\text{bnd}}$

$$\|e_{\pi,w}\|_F^2 \leq C_{\pi,\partial}^2 |h^{k+1/2}w|_{k+1,K}^2.$$

Choosing  $\gamma = \frac{1}{2}(h|_{K_1^F} + h|_{K_2^F})$ , employing the shape- and contact-regularity estimate [DPE12, Lem. 1.43] with constant  $\rho$ , and using that each cell has at most  $N_\partial$  faces proves (2.13a). The estimate (2.13b) follows analogously.  $\square$

*Remark 2.9.* The second term in (2.13) only yields spatial convergence order  $k$  on  $\mathcal{T}_{h,m+}$ , compared to order  $k + 1/2$  on  $\mathcal{T}_{h,\text{lf}}$ . This reduced order stems from the interfaces in  $\mathcal{F}_{h,m+}^{\text{int}}$ . One can recover the full order  $k + 1/2$  on the whole mesh  $\mathcal{T}_h$  by increasing the polynomial degree on the coarse elements contained in  $\mathcal{T}_{h,m+}$  and exploiting the smaller diameter of the fine elements in  $\mathcal{T}_{h,f}$ , see [Stu17, Remark 5.20b]. This is demonstrated in the numerical experiment in Section 2.5.1.  $\diamond$

## 2.2. Full discretization

We now combine the spatial discretization from Section 2.1 with the leapfrog-based local time-integration method from Section 1.3. The resulting fully discrete scheme extends (1.9) by incorporating the stabilization operators  $\mathcal{S}_{u,h}^{\text{lf}}, \mathcal{S}_{v,h}^{\text{lf}}$  from Definition 2.1. As in [HS19], the stabilization enters explicitly in each sub-step. Compared to [HS19], we treat a more general filter function  $\Psi$ . The locally implicit method [HS19] is a special case, see Section 1.3.3. Moreover, we allow a more general forcing in the  $v$ -component. Both generalizations lead to additional terms and problems in the stability and error analysis.

With a smooth function  $\Psi: [0, \infty) \rightarrow \mathbb{R}$  satisfying  $\Psi(0) = 1$ , the fully discrete local time-integration scheme is defined as

$$v_h^{n+1/2} - v_h^n = \frac{\tau}{2} (\mathcal{L}_{u,h} u_h^n - \alpha \mathcal{S}_{v,h}^{\text{lf}} v_h^n + g_{v,h}^n), \quad (2.14a)$$

$$u_h^{n+1} - u_h^n = \tau \Psi (\mathcal{L}_{v,h} v_h^{n+1/2} - \alpha \mathcal{S}_{u,h}^{\text{lf}} u_h^n + g_{u,h}^{n+1/2}), \quad (2.14b)$$

$$v_h^{n+1} - v_h^{n+1/2} = \frac{\tau}{2} (\mathcal{L}_{u,h} u_h^{n+1} - \alpha \mathcal{S}_{v,h}^{\text{lf}} v_h^n + g_{v,h}^{n+1}), \quad (2.14c)$$

$$\Psi = \Psi(-\tau^2 \mathcal{L}_{v,h} \chi_m \mathcal{L}_{u,h}), \quad (2.14d)$$

$$g_{u,h}^{n+1/2} = \frac{1}{2} (g_{u,h}(t^{n+1}) + g_{u,h}(t^n)). \quad (2.14e)$$

The scheme proceeds in three stages per time step: a half-step for the  $v$ -component (2.14a), a full filtered step for the  $u$ -component (2.14b), and a second half-step for the  $v$ -component (2.14c). The explicit stabilization in each sub-step keeps the method fully explicit if a polynomial filter  $\Psi$  is applied. As in (1.9), the action of the filter operator  $\Psi$  localizes to the region  $\mathcal{T}_{h,m}$ .

## 2.3. Stability

While the stability analysis in Section 1.3.1 relies on a discrete semigroup technique, the stabilization terms require a different approach based on discrete energy estimates. The starting point, however, is the same: we first rewrite the scheme (2.14) in the compact operator form already used in the central-flux case. The stability analysis is performed analogously to [HS19], with the exception of the additional forcing term.

**Lemma 2.10.** *The approximations  $x_h^{n+1} = (u_h^{n+1} \ v_h^{n+1})$  defined by (2.14) satisfy*

$$\mathcal{R}_- x_h^{n+1} = \mathcal{R}_+ x_h^n - \tau \alpha \mathcal{S}_h^{\text{lf}} x_h^n + \tau g_h^{n+1/2} - \tau \frac{\tau^2}{4} \mathcal{L}_h \begin{pmatrix} 0 \\ \partial_\tau g_{v,h}^{n+1} \end{pmatrix}. \quad (2.15)$$

Here,  $\mathcal{R}_\pm$  are defined in (1.12a), and

$$\partial_\tau u^{n+1} = \frac{1}{\tau} (u^{n+1} - u^n) \quad (2.16)$$

denotes the discrete time derivative of a sequence  $(u^n)$  for  $\tau > 0$ .

*Proof.* The recursion follows from (2.14) and the definition of  $\mathcal{R}_\pm$  in (1.12a).  $\square$

Compared to the central-flux recursion (1.13), the transformed scheme (2.15) contains the additional dissipative term  $-\tau \alpha \mathcal{S}_h^{\text{lf}} x_h^n$  on the right-hand side, as well as an additional forcing term arising from the different implementation of the forcing  $g_v$ . The next Lemma 2.11 makes use of the summation-by-parts formula

$$\tau \sum_{j=0}^n (a^{j+1}, \partial_\tau b^{j+1})_\Omega = (a^{n+1}, b^{n+1})_\Omega - (a^1, b^0)_\Omega - \tau \sum_{j=1}^n (\partial_\tau a^{j+1}, b^j)_\Omega, \quad (2.17)$$

given two sequences  $(a^j)_{j \geq 0}$ ,  $(b^j)_{j \geq 0}$  in  $L^2(\Omega)$  and  $\partial_\tau$  defined in (2.16).

**Lemma 2.11.** *If  $x_h^{n+1}$  is defined by (2.15), then*

$$\begin{aligned} & ((\mathcal{R}_- - \frac{\tau}{2} \alpha \mathcal{S}_h^{\text{lf}}) x_h^{n+1}, x_h^{n+1})_\Omega - ((\mathcal{R}_- - \frac{\tau}{2} \alpha \mathcal{S}_h^{\text{lf}}) x_h^0, x_h^0)_\Omega + \frac{\tau}{2} \alpha \sum_{j=0}^n |x_h^{j+1} + x_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 \\ &= \frac{\tau^2}{2} ((\partial_\tau g_{v,h}^{n+1}, v_h^{n+1})_\Omega - (\partial_\tau g_{v,h}^1, v_h^0)_\Omega - \frac{1}{2} \|g_{v,h}^{n+1}\|_\Omega^2 + \frac{1}{2} \|g_{v,h}^0\|_\Omega^2) \\ &+ \tau \sum_{j=0}^n (g_h^{j+1/2}, x_h^{j+1} + x_h^j)_\Omega - \frac{\tau}{2} \sum_{j=1}^n (\tau^2 \partial_\tau^2 g_{v,h}^{j+1}, v_h^j)_\Omega + \tau \sum_{j=0}^n (\tau \partial_\tau g_{v,h}^{j+1}, \frac{\tau}{2} \alpha \mathcal{S}_h^{\text{lf}} v_h^j)_\Omega. \end{aligned} \quad (2.18)$$

*Proof.* First of all, using (1.21a) gives

$$(\mathcal{R}_- x_h^{j+1}, x_h^{j+1} + x_h^j)_\Omega - (\mathcal{R}_+ x_h^j, x_h^{j+1} + x_h^j)_\Omega = (\mathcal{R}_- x_h^{j+1}, x_h^{j+1})_\Omega - (\mathcal{R}_+ x_h^j, x_h^j)_\Omega. \quad (2.19)$$

Using the self-adjointness of  $\mathcal{S}_h^{\text{lf}}$ , we also obtain

$$\begin{aligned} (\mathcal{S}_h^{\text{lf}} x_h^j, x_h^{j+1} + x_h^j)_\Omega &= \frac{1}{2} |x_h^{j+1} + x_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 - \frac{1}{2} (\mathcal{S}_h^{\text{lf}} (x_h^{j+1} - x_h^j), x_h^{j+1} + x_h^j)_\Omega \\ &= \frac{1}{2} |x_h^{j+1} + x_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 + \frac{1}{2} |x_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 - \frac{1}{2} |x_h^{j+1}|_{\mathcal{S}_h^{\text{lf}}}^2. \end{aligned} \quad (2.20)$$

Then, testing (2.15) with  $x_h^{j+1} + x_h^j$ , summing over  $j = 0, \dots, n$ , and inserting (2.19) as well as (2.20) yields

$$\begin{aligned} & (\mathcal{R}_- x_h^{n+1}, x_h^{n+1})_\Omega - (\mathcal{R}_+ x_h^0, x_h^0)_\Omega \\ & - \frac{\tau}{2} \alpha (\mathcal{S}_h^{\text{lf}} x_h^{n+1}, x_h^{n+1})_\Omega + \frac{\tau}{2} \alpha (\mathcal{S}_h^{\text{lf}} x_h^0, x_h^0)_\Omega + \frac{\tau}{2} \alpha \sum_{j=0}^n |x_h^{j+1} + x_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 \\ & = \tau \sum_{j=0}^n (g_h^{j+1/2}, x_h^{j+1} + x_h^j)_\Omega - \frac{\tau^2}{2} \sum_{j=0}^n (\frac{\tau}{2} \mathcal{L}_{v,h} \partial_\tau g_{v,h}^{j+1}, u_h^{j+1} + u_h^j)_\Omega. \end{aligned}$$

Using (2.14a) and (2.14c), we get

$$\begin{aligned} & \sum_{j=0}^n (\partial_\tau g_{v,h}^{j+1}, \frac{\tau}{2} \mathcal{L}_{u,h} (u_h^{j+1} + u_h^j))_\Omega \\ & = \tau \sum_{j=0}^n (\partial_\tau g_{v,h}^{j+1}, \partial_\tau v_h^{j+1})_\Omega + \sum_{j=0}^n (\partial_\tau g_{v,h}^{j+1}, \tau \alpha \mathcal{S}_{v,h}^{\text{lf}} v_h^j)_\Omega - \sum_{j=0}^n (\partial_\tau g_{v,h}^{j+1}, \frac{\tau}{2} (g_{v,h}^{j+1} + g_{v,h}^j))_\Omega. \end{aligned} \quad (2.21)$$

Applying the summation-by-parts formula (2.17) to the first part of (2.21), we obtain

$$\tau \sum_{j=0}^n (\partial_\tau g_{v,h}^{j+1}, \partial_\tau v_h^{j+1})_\Omega = (\partial_\tau g_{v,h}^{n+1}, v_h^{n+1})_\Omega - (\partial_\tau g_{v,h}^1, v_h^0)_\Omega - \tau \sum_{j=1}^n (\partial_\tau^2 g_{v,h}^{j+1}, v_h^j)_\Omega.$$

Collapsing the telescopic sum

$$\sum_{j=0}^n (\partial_\tau g_{v,h}^{j+1}, \frac{\tau}{2} (g_{v,h}^{j+1} + g_{v,h}^j))_\Omega = \frac{1}{2} \|g_{v,h}^{n+1}\|_\Omega^2 - \frac{1}{2} \|g_{v,h}^0\|_\Omega^2$$

completes the proof.  $\square$

The stabilization introduces an additional restriction on the time-step size. Together with the CFL condition (1.16) from the central-flux analysis, it ensures that the combined operator  $\mathcal{R}_\pm - \frac{\tau}{2} \alpha \mathcal{S}_h^{\text{lf}}$  remains positive definite.

**Definition 2.12.** For given  $\vartheta_{s,c} \in (0, 1]$ , we define

$$\tau_{\text{CFL},s,c} = \tau_{\text{CFL},s,c}(\vartheta_{s,c}) = \frac{2c_\Theta \vartheta_{s,c}}{\|\mathcal{S}_h^{\text{lf}}\|_{\mathbb{V}_h \leftarrow \mathbb{V}_h}}. \quad (2.22)$$

The following Lemma 2.13 shows that under the combined CFL conditions (1.16) and (2.22), the operator  $\mathcal{R}_\pm - \frac{\tau}{2} \alpha \mathcal{S}_h^{\text{lf}}$  is positive definite. The proof extends Lemma 1.7 by absorbing the stabilization via (2.22). The resulting constant  $C_{\text{stb},\text{upw},c}$  couples the filter and stabilization CFL parameters.

**Lemma 2.13.** Let  $\vartheta_c^2 + \alpha \vartheta_{s,c} \in (0, 1)$  and  $\tau \leq \min\{\tau_{\text{CFL},\text{lf},c}(\vartheta_c), \tau_{\text{CFL},s,c}(\vartheta_{s,c}), \tau_{\text{CFL},\Psi}\}$  satisfy the CFL conditions (1.16) and (2.22). Then

$$C_{\text{stb},\text{upw},c}^{-1} \|x_h\|_\Omega^2 := c_\Theta (1 - \vartheta_c^2 - \alpha \vartheta_{s,c}) \|x_h\|_\Omega^2 \leq ((\mathcal{R}_\pm - \frac{\tau}{2} \alpha \mathcal{S}_h^{\text{lf}}) x_h, x_h)_\Omega, \quad (2.23)$$

for all  $x_h \in \mathbb{V}_h$ .

*Proof.* The proof is analogous to the proof of Lemma 1.7. Additionally, we use (2.22) to bound

$$\frac{\tau}{2}(\mathcal{S}_h^{\text{lf}} x_h, x_h)_\Omega \leq c_\Theta \vartheta_{s,c} \|x_h\|_\Omega^2.$$

Overall, it follows

$$\begin{aligned} ((\mathcal{R}_\pm - \frac{\tau}{2}\alpha\mathcal{S}_h^{\text{lf}})x_h, x_h)_\Omega &= (\Theta u_h, u_h)_\Omega + \|v_h\|_\Omega^2 - \frac{\tau^2}{4}(\mathcal{D}_{\text{lf}} x_h, x_h)_\Omega - \frac{\tau}{2}\alpha(\mathcal{S}_h^{\text{lf}} x_h, x_h)_\Omega \\ &\geq c_\Theta(1 - \vartheta_c^2 - \alpha\vartheta_{s,c})\|u_h\|_\Omega^2 + (1 - \alpha c_\Theta \vartheta_{s,c})\|v_h\|_\Omega^2. \end{aligned}$$

Since  $c_\Theta \leq 1$  per Definition 1.3, we have  $c_\Theta(1 - \vartheta_c^2) \leq 1$ , and hence it follows

$$1 - \alpha c_\Theta \vartheta_{s,c} \geq c_\Theta(1 - \vartheta_c^2 - \alpha\vartheta_{s,c}) = C_{\text{stb,upw,c}}^{-1},$$

proving the statement.  $\square$

With the lower bound from Lemma 2.13 at hand, we combine the energy identity (2.18) with Cauchy–Schwarz and Young’s inequalities to derive the stability estimate. The additional right-hand side terms involving  $\partial_\tau g_{v,h}$  require a careful choice of the Young weights  $\gamma_1, \dots, \gamma_4$  and an application of the discrete Gronwall inequality [HS19, Lemma A.1].

**Lemma 2.14.** *Under the same assumptions as in Lemma 2.13, the numerical solution  $x_h^{n+1} = (u_h^{n+1} \ v_h^{n+1})$  of (2.14) satisfies*

$$\begin{aligned} C_{\text{stb,upw,c}}^{-1} \|x_h^{n+1}\|_\Omega^2 + \frac{\tau}{2}\alpha \sum_{j=0}^n |x_h^{j+1} + x_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 \\ \leq e^{3/2} C \left( \|x_h^0\|_{\Theta_{\mathcal{I}}}^2 + \frac{\tau^2}{4} \|v_h^0\|_\Omega^2 + \tau \sum_{j=0}^{n+1} \|g_h^j\|_\Omega^2, \right) \end{aligned} \quad (2.24)$$

with a constant  $C$  dependent on  $C_{\text{stb,upw,c}}$  and  $T$ .

*Proof.* First, we bound the left-hand side of (2.18) using (2.23) from below

$$\begin{aligned} ((\mathcal{R}_- - \frac{\tau}{2}\alpha\mathcal{S}_h^{\text{lf}})x_h^{n+1}, x_h^{n+1})_\Omega - (\mathcal{R}_+ x_h^0, x_h^0)_\Omega + \frac{\tau}{2}\alpha(\mathcal{S}_h^{\text{lf}} x_h^0, x_h^0)_\Omega + \frac{\tau}{2}\alpha \sum_{j=0}^n |x_h^{j+1} + x_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 \\ \geq C_{\text{stb,upw,c}}^{-1} \|x_h^{n+1}\|_\Omega^2 - \|x_h^0\|_{\Theta_{\mathcal{I}}}^2 + \frac{\tau}{2}\alpha \sum_{j=0}^n |x_h^{j+1} + x_h^j|_{\mathcal{S}_h^{\text{lf}}}^2. \end{aligned}$$

Bounding the right-hand side of (2.18) from above using the Cauchy–Schwarz inequality and Young’s inequality together with the CFL condition (2.22) gives

$$\begin{aligned} \tau \sum_{j=0}^n (g_h^{j+1/2}, x_h^{j+1} + x_h^j)_\Omega - \frac{\tau}{2} \sum_{j=1}^n (\tau^2 \partial_\tau^2 g_{v,h}^{j+1}, v_h^j)_\Omega + \tau \sum_{j=0}^n (\tau \partial_\tau g_{v,h}^{j+1}, \frac{\tau}{2}\alpha \mathcal{S}_{v,h}^{\text{lf}} v_h^j)_\Omega \\ + \frac{\tau^2}{2} ((\partial_\tau g_{v,h}^{n+1}, v_h^{n+1})_\Omega - (\partial_\tau g_{v,h}^1, v_h^0)_\Omega - \frac{1}{2} \|g_{v,h}^{n+1}\|_\Omega^2 + \frac{1}{2} \|g_{v,h}^0\|_\Omega^2) \\ \leq \tau \sum_{j=0}^n \gamma_1 (\|x_h^{j+1}\|_\Omega^2 + \|x_h^j\|_\Omega^2) + \frac{\tau}{2} \sum_{j=0}^n \frac{1}{\gamma_1} \|g_h^{j+1/2}\|_\Omega^2 \\ + \frac{\tau}{4} \sum_{j=1}^n \gamma_2 \|v_h^j\|_\Omega^2 + \frac{\tau}{4} \sum_{j=1}^n \frac{1}{\gamma_2} \|\tau^2 \partial_\tau^2 g_{v,h}^{j+1}\|_\Omega^2 \end{aligned}$$

$$\begin{aligned}
 & + \frac{\tau}{2} \sum_{j=0}^n \gamma_3 \frac{\tau}{2} \alpha |v_h^j|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 + \frac{\tau}{2} \sum_{j=0}^n \frac{1}{\gamma_3} \frac{\tau}{2} \alpha |\tau \partial_\tau g_{v,h}^{j+1}|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 \\
 & + \frac{\tau^2}{4} (\gamma_4 \|v_h^{n+1}\|_\Omega^2 + \|v_h^0\|_\Omega^2 + \frac{1}{\gamma_4} \|\partial_\tau g_{v,h}^{n+1}\|_\Omega^2 + \|\partial_\tau g_{v,h}^1\|_\Omega^2 - \|g_{v,h}^{n+1}\|_\Omega^2 + \|g_{v,h}^0\|_\Omega^2) \\
 \leq & \tau \sum_{j=0}^n \gamma_1 (\|x_h^{j+1}\|_\Omega^2 + \|x_h^j\|_\Omega^2) + \frac{\tau}{2} \sum_{j=0}^n \frac{1}{\gamma_1} \|g_h^{j+1/2}\|_\Omega^2 \\
 & + \frac{\tau}{4} \sum_{j=0}^n \gamma_2 \|x_h^{j+1}\|_\Omega^2 + \frac{\tau}{4} \sum_{j=1}^n \frac{1}{\gamma_2} \|\tau^2 \partial_\tau^2 g_{v,h}^{j+1}\|_\Omega^2 \\
 & + \frac{\tau}{2} \sum_{j=0}^n \gamma_3 \alpha c_\Theta \vartheta_{s,c} \|x_h^j\|_\Omega^2 + \frac{\tau}{2} \sum_{j=0}^n \frac{1}{\gamma_3} \alpha c_\Theta \vartheta_{s,c} \|\tau \partial_\tau g_{v,h}^{j+1}\|_\Omega^2 \\
 & + \frac{\tau^2}{4} (\|v_h^0\|_\Omega^2 + \frac{\tau}{\gamma_2} \|\partial_\tau g_{v,h}^{n+1}\|_\Omega^2 + \|\partial_\tau g_{v,h}^1\|_\Omega^2 + \|g_{v,h}^0\|_\Omega^2),
 \end{aligned}$$

where we have chosen  $\gamma_4 := \gamma_2/\tau$ . Next, we choose the weights  $\gamma_2 := 4\gamma_1$  and  $\gamma_3 := 2\gamma_1/(\alpha c_\Theta \vartheta_{s,c})$ , and we group terms. This gives

$$\begin{aligned}
 & \tau \sum_{j=0}^n \gamma_1 (\|x_h^{j+1}\|_\Omega^2 + \|x_h^j\|_\Omega^2) + \frac{\tau}{2} \sum_{j=0}^n \frac{1}{\gamma_1} \|g_h^{j+1/2}\|_\Omega^2 \\
 & + \frac{\tau}{4} \sum_{j=0}^n \gamma_2 \|x_h^{j+1}\|_\Omega^2 + \frac{\tau}{4} \sum_{j=1}^n \frac{1}{\gamma_2} \|\tau^2 \partial_\tau^2 g_{v,h}^{j+1}\|_\Omega^2 \\
 & + \frac{\tau}{2} \sum_{j=0}^n \gamma_3 \alpha c_\Theta \vartheta_{s,c} \|x_h^j\|_\Omega^2 + \frac{\tau}{2} \sum_{j=0}^n \frac{1}{\gamma_3} \alpha c_\Theta \vartheta_{s,c} \|\tau \partial_\tau g_{v,h}^{j+1}\|_\Omega^2 \\
 & + \frac{\tau^2}{4} (\|v_h^0\|_\Omega^2 + \frac{\tau}{\gamma_2} \|\partial_\tau g_{v,h}^{n+1}\|_\Omega^2 + \|\partial_\tau g_{v,h}^1\|_\Omega^2 + \|g_{v,h}^0\|_\Omega^2) \\
 = & \tau \sum_{j=0}^n (\gamma_1 + \frac{\gamma_2}{4}) \|x_h^{j+1}\|_\Omega^2 + (\gamma_1 + \frac{\gamma_3}{2} \alpha c_\Theta \vartheta_{s,c}) \|x_h^j\|_\Omega^2 \\
 & + \frac{\tau}{2} \sum_{j=0}^n \frac{1}{\gamma_1} \|g_h^{j+1/2}\|_\Omega^2 + \frac{\tau}{4} \sum_{j=1}^n \frac{1}{\gamma_2} \|\tau^2 \partial_\tau^2 g_{v,h}^{j+1}\|_\Omega^2 + \frac{\tau}{2} \sum_{j=0}^n \frac{1}{\gamma_3} \alpha c_\Theta \vartheta_{s,c} \|\tau \partial_\tau g_{v,h}^{j+1}\|_\Omega^2 \\
 & + \frac{\tau^2}{4} (\|v_h^0\|_\Omega^2 + \frac{\tau}{\gamma_2} \|\partial_\tau g_{v,h}^{n+1}\|_\Omega^2 + \|\partial_\tau g_{v,h}^1\|_\Omega^2 + \|g_{v,h}^0\|_\Omega^2) \\
 = & 2\gamma_1 \tau \sum_{j=0}^n \|x_h^{j+1}\|_\Omega^2 + \|x_h^j\|_\Omega^2 \\
 & + \frac{\tau}{2} \sum_{j=0}^n \frac{1}{\gamma_1} \|g_h^{j+1/2}\|_\Omega^2 + \frac{\tau}{4} \sum_{j=1}^n \frac{1}{4\gamma_1} \|\tau^2 \partial_\tau^2 g_{v,h}^{j+1}\|_\Omega^2 + \frac{\tau}{2} \sum_{j=0}^n \frac{\alpha^2 c_\Theta^2 \vartheta_{s,c}^2}{2\gamma_1} \|\tau \partial_\tau g_{v,h}^{j+1}\|_\Omega^2 \\
 & + \frac{\tau^2}{4} (\|v_h^0\|_\Omega^2 + \frac{\tau}{4\gamma_1} \|\partial_\tau g_{v,h}^{n+1}\|_\Omega^2 + \|\partial_\tau g_{v,h}^1\|_\Omega^2 + \|g_{v,h}^0\|_\Omega^2).
 \end{aligned}$$

Together, we obtain

$$\begin{aligned}
 & C_{\text{stb,upw,c}}^{-1} \|x_h^{n+1}\|_\Omega^2 + \frac{\tau}{2} \alpha \sum_{j=0}^n |x_h^{j+1} + x_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 \\
 & \leq \|x_h^0\|_{\Theta_{\mathcal{I}}}^2 + \frac{\tau^2}{4} \|v_h^0\|_\Omega^2 + 2\gamma_1 \tau \sum_{j=0}^n \|x_h^{j+1}\|_\Omega^2 + \|x_h^j\|_\Omega^2
 \end{aligned}$$

$$\begin{aligned}
& + \frac{\tau}{2} \sum_{j=0}^n \frac{1}{\gamma_1} \|g_h^{j+1/2}\|_{\Omega}^2 + \frac{\tau}{4} \sum_{j=1}^n \frac{1}{4\gamma_1} \|\tau^2 \partial_{\tau}^2 g_{v,h}^{j+1}\|_{\Omega}^2 + \frac{\tau}{2} \sum_{j=0}^n \frac{\alpha^2 c_{\Theta}^2 v_{s,c}^2}{2\gamma_1} \|\tau \partial_{\tau} g_{v,h}^{j+1}\|_{\Omega}^2 \\
& + \frac{\tau^2}{4} \left( \frac{\tau}{4\gamma_1} \|\partial_{\tau} g_{v,h}^{n+1}\|_{\Omega}^2 + \|\partial_{\tau} g_{v,h}^1\|_{\Omega}^2 + \|g_{v,h}^0\|_{\Omega}^2 \right).
\end{aligned}$$

The discrete Gronwall-type inequality [HS19, Lemma A.1] together with  $\gamma_1 := \frac{C_{\text{stb,upw,c}}^{-1}}{4(T+1)}$  yields

$$\begin{aligned}
& C_{\text{stb,upw,c}}^{-1} \|x_h^{n+1}\|_{\Omega}^2 + \frac{\tau}{2} \alpha \sum_{j=0}^n |x_h^{j+1} + x_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 \\
& \leq e^{\frac{3}{2}} \left( \|x_h^0\|_{\Theta_x}^2 + \frac{\tau^2}{4} \|v_h^0\|_{\Omega}^2 + \frac{\tau^2}{4} (\tau C_{\text{stb,upw,c}}(T+1) \|\partial_{\tau} g_{v,h}^{n+1}\|_{\Omega}^2 + \|\partial_{\tau} g_{v,h}^1\|_{\Omega}^2 + \|g_{v,h}^0\|_{\Omega}^2) \right. \\
& \quad \left. + C_{\text{stb,upw,c}}(T+1) \left( 2\tau \sum_{j=0}^n \|g_h^{j+1/2}\|_{\Omega}^2 + \frac{\tau}{4} \sum_{j=1}^n \|\tau^2 \partial_{\tau}^2 g_{v,h}^{j+1}\|_{\Omega}^2 + \tau \sum_{j=0}^n \alpha^2 \|\tau \partial_{\tau} g_{v,h}^{j+1}\|_{\Omega}^2 \right) \right).
\end{aligned}$$

For example, the triangle inequality gives

$$\|\tau \partial_{\tau} g_{v,h}^{j+1}\|_{\Omega}^2 \leq 2(\|g_{v,h}^{j+1}\|_{\Omega}^2 + \|g_{v,h}^j\|_{\Omega}^2),$$

Estimating the remaining right-hand side terms analogously finishes the proof.  $\square$

## 2.4. Error analysis

In the following, we briefly recall notation from Section 1.3.2. We denote the exact solution of (1.1) evaluated at time  $t^n$  by

$$\hat{x}^n = x(t^n), \quad x = (u \ v), \quad (2.25)$$

and the error of the full discretization by

$$e^n = \hat{x}^n - x_h^n = e_{\pi}^n + e_h^n, \quad e_{\pi}^n = \hat{x}^n - \pi_h \hat{x}^n, \quad e_h^n = \pi_h \hat{x}^n - x_h^n, \quad (2.26)$$

where  $e_{\pi}^n$  is the  $L^2$ -projection error and  $e_h^n$  is the discretization error.

We define the defect  $\delta_h^{n+1}$  by inserting the  $L^2$ -projected exact solution into the numerical scheme (2.15). This yields

$$\mathcal{R}_{-\pi_h} \hat{x}^{n+1} = \mathcal{R}_{+\pi_h} \hat{x}^n - \tau \alpha \mathcal{S}_h^{\text{lf}} \pi_h \hat{x}^n + \tau g_h^{n+1/2} - \tau \frac{\tau^2}{4} \mathcal{L}_h \begin{pmatrix} 0 \\ \partial_{\tau} g_{v,h}^{n+1} \end{pmatrix} + \tau \delta_h^{n+1}. \quad (2.27)$$

Subtracting (2.15) from (2.27) yields the error recursion

$$\mathcal{R}_{-} e_h^{n+1} = \mathcal{R}_{+} e_h^n - \tau \alpha \mathcal{S}_h^{\text{lf}} e_h^n + \tau \delta_h^{n+1}. \quad (2.28)$$

This recursion has the same form as (2.15). Thus, one could apply Lemma 2.14 to bound the error. However, a sharper estimate requires a more careful inspection of the defects, because the cutoff functions  $\chi_{\text{lf}}$ ,  $\chi_m$ , defined in Section 1.3, occur within the Friedrichs' operators. The authors of [HS19] use the coarse-grid CFL condition (1.16b) to bound terms containing  $\mathcal{L}_{v,h} \chi_{\text{lf}}$ . In addition, we have to handle terms containing  $\mathcal{L}_{v,h} \chi_m$ . These terms could be controlled by the CFL condition (1.16a) of the modified method, but using it would introduce the constant

$\widehat{\beta}_\Psi^2$  into the error estimate. This would result in a suboptimal error bound. If the filter (1.45) is used in (2.14), this constant scales with the polynomial degree of the Chebychev polynomials. Therefore, we take a different approach. We start with the error equality in the following Lemma 2.15 where we apply the summation-by-parts formula (2.17) to exchange a discrete spatial derivative for a discrete temporal derivative.

**Lemma 2.15.** *Assume that the defect can be decomposed as*

$$\delta_h^j = \delta_{h,s}^j + \mathcal{L}_h \delta_{h,\mathcal{L}_h}^j \quad \text{with} \quad \delta_{h,\mathcal{L}_h}^j = \begin{pmatrix} 0 \\ \delta_{h,v,\mathcal{L}_h}^j \end{pmatrix}. \quad (2.29a)$$

Then

$$\begin{aligned} & ((\mathcal{R}_- - \frac{\tau}{2}\alpha \mathcal{S}_h^{\text{lf}})e_h^{n+1}, e_h^{n+1})_\Omega + \frac{\tau}{2}\alpha \sum_{j=0}^n |e_h^{j+1} + e_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 \\ &= \tau \sum_{j=0}^n (\delta_{h,s}^{j+1}, e_h^{j+1} + e_h^j)_\Omega - \sum_{j=0}^n (\delta_{h,v,\mathcal{L}_h}^{j+1}, \tau\alpha \mathcal{S}_{v,h}^{\text{lf}}(e_{h,v}^{j+1} + e_{h,v}^j))_\Omega \\ &\quad + \tau \sum_{j=1}^n 2(\partial_\tau \delta_{h,v,\mathcal{L}_h}^{j+1}, e_{h,v}^j)_\Omega - \tau \sum_{j=1}^n (\partial_\tau \delta_{h,v,\mathcal{L}_h}^{j+1}, \tau\alpha \mathcal{S}_{v,h}^{\text{lf}} e_{h,v}^j)_\Omega \\ &\quad + \tau \sum_{j=0}^n 2(\delta_{h,v,\mathcal{L}_h}^{j+1}, \delta_{h,v,s}^{j+1})_\Omega - 2(\delta_{h,v,\mathcal{L}_h}^{n+1}, (\mathcal{I}_v - \frac{\tau}{2}\alpha \mathcal{S}_{v,h}^{\text{lf}})e_{h,v}^{n+1})_\Omega. \end{aligned} \quad (2.29b)$$

*Proof.* We begin similarly as in the proof of Lemma 2.14 by testing (2.28) with  $e_h^{n+1} + e_h^n$ , which leads to

$$\begin{aligned} & (\mathcal{R}_- e_h^{n+1}, e_h^{n+1})_\Omega - (\mathcal{R}_+ e_h^0, e_h^0)_\Omega \\ & - \frac{\tau}{2}\alpha (\mathcal{S}_h^{\text{lf}} e_h^{n+1}, e_h^{n+1})_\Omega + \frac{\tau}{2}\alpha (\mathcal{S}_h^{\text{lf}} e_h^0, e_h^0)_\Omega + \frac{\tau}{2}\alpha \sum_{j=0}^n |e_h^{j+1} + e_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 \\ &= \tau \sum_{j=0}^n (\delta_{h,s}^{j+1}, e_h^{j+1} + e_h^j)_\Omega + \tau \sum_{j=0}^n (\mathcal{L}_{v,h} \delta_{h,v,\mathcal{L}_h}^{j+1}, e_{h,u}^{j+1} + e_{h,u}^j)_\Omega. \end{aligned}$$

Then, inserting the second component of (2.28) given by

$$e_{h,v}^{j+1} - \frac{\tau}{2}\mathcal{L}_{u,h} e_{h,u}^{j+1} = e_{h,v}^j + \frac{\tau}{2}\mathcal{L}_{u,h} e_{h,u}^j - \tau\alpha \mathcal{S}_{v,h}^{\text{lf}} e_{h,v}^j + \tau\delta_{h,v,s}^{j+1},$$

since by assumption  $\delta_{h,v}^j = \delta_{h,v,s}^j$ , yields

$$\begin{aligned} & -\tau \sum_{j=0}^n (\mathcal{L}_{v,h} \delta_{h,v,\mathcal{L}_h}^{j+1}, e_{h,u}^{j+1} + e_{h,u}^j)_\Omega \\ &= \sum_{j=0}^n 2(\delta_{h,v,\mathcal{L}_h}^{j+1}, e_{h,v}^{j+1} - e_{h,v}^j)_\Omega + \sum_{j=0}^n 2(\delta_{h,v,\mathcal{L}_h}^{j+1}, \tau\alpha \mathcal{S}_{v,h}^{\text{lf}} e_{h,v}^j)_\Omega - \sum_{j=0}^n 2(\delta_{h,v,\mathcal{L}_h}^{j+1}, \tau\delta_{h,v,s}^{j+1})_\Omega \\ &= \sum_{j=0}^n (\delta_{h,v,\mathcal{L}_h}^{j+1}, \tau\alpha \mathcal{S}_{v,h}^{\text{lf}}(e_{h,v}^{j+1} + e_{h,v}^j))_\Omega - \sum_{j=0}^n 2(\delta_{h,v,\mathcal{L}_h}^{j+1}, \tau\delta_{h,v,s}^{j+1})_\Omega \end{aligned}$$

$$\begin{aligned}
& + \sum_{j=0}^n 2(\delta_{h,v,\mathcal{L}_h}^{j+1}, e_{h,v}^{j+1} - e_{h,v}^j)_\Omega - \sum_{j=0}^n (\delta_{h,v,\mathcal{L}_h}^{j+1}, \tau \alpha \mathcal{S}_{v,h}^{\text{lf}}(e_{h,v}^{j+1} - e_{h,v}^j))_\Omega \\
& = \sum_{j=0}^n (\delta_{h,v,\mathcal{L}_h}^{j+1}, \tau \alpha \mathcal{S}_{v,h}^{\text{lf}}(e_{h,v}^{j+1} + e_{h,v}^j))_\Omega - \sum_{j=0}^n 2(\delta_{h,v,\mathcal{L}_h}^{j+1}, \tau \delta_{h,v,s}^{j+1})_\Omega \\
& \quad - \tau \sum_{j=1}^n 2(\partial_\tau \delta_{h,v,\mathcal{L}_h}^{j+1}, e_{h,v}^j)_\Omega + 2(\delta_{h,v,\mathcal{L}_h}^{n+1}, e_{h,v}^{n+1})_\Omega - 2(\delta_{h,v,\mathcal{L}_h}^1, e_{h,v}^0)_\Omega \\
& \quad + \tau \sum_{j=1}^n (\partial_\tau \delta_{h,v,\mathcal{L}_h}^{j+1}, \tau \alpha \mathcal{S}_{v,h}^{\text{lf}} e_{h,v}^j)_\Omega - (\delta_{h,v,\mathcal{L}_h}^{n+1}, \tau \alpha \mathcal{S}_{v,h}^{\text{lf}} e_{h,v}^{n+1})_\Omega + (\delta_{h,v,\mathcal{L}_h}^1, \tau \alpha \mathcal{S}_{v,h}^{\text{lf}} e_{h,v}^0)_\Omega
\end{aligned}$$

where we have used the summation-by-parts formula (2.17) two times. Since  $e_{h,v}^0 = 0$ , we get

$$\begin{aligned}
& ((\mathcal{R}_- - \frac{\tau}{2} \alpha \mathcal{S}_h^{\text{lf}}) e_h^{n+1}, e_h^{n+1})_\Omega + \frac{\tau}{2} \alpha \sum_{j=0}^n |e_h^{j+1} + e_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 \\
& = \tau \sum_{j=0}^n (\delta_{h,s}^{j+1}, e_h^{j+1} + e_h^j)_\Omega - \sum_{j=0}^n (\delta_{h,v,\mathcal{L}_h}^{j+1}, \tau \alpha \mathcal{S}_{v,h}^{\text{lf}}(e_{h,v}^{j+1} + e_{h,v}^j))_\Omega + \sum_{j=0}^n 2(\delta_{h,v,\mathcal{L}_h}^{j+1}, \tau \delta_{h,v,s}^{j+1})_\Omega \\
& \quad + \tau \sum_{j=1}^n 2(\partial_\tau \delta_{h,v,\mathcal{L}_h}^{j+1}, e_{h,v}^j)_\Omega - 2(\delta_{h,v,\mathcal{L}_h}^{n+1}, e_{h,v}^{n+1})_\Omega \\
& \quad - \tau \sum_{j=1}^n (\partial_\tau \delta_{h,v,\mathcal{L}_h}^{j+1}, \tau \alpha \mathcal{S}_{v,h}^{\text{lf}} e_{h,v}^j)_\Omega + (\delta_{h,v,\mathcal{L}_h}^{n+1}, \tau \alpha \mathcal{S}_{v,h}^{\text{lf}} e_{h,v}^{n+1})_\Omega,
\end{aligned}$$

completing the proof.  $\square$

The error equality (2.29b) reduces the convergence analysis to bounding the six right-hand side terms. To this end, we show how we decompose the defect into the two parts needed to apply Lemma 2.15.

With Lemma 1.12, we can split the defect  $\delta_h^{j+1}$  defined in (2.27) into

$$\delta_h^{j+1} = \delta_{h,s}^{j+1} + \mathcal{L}_h \delta_{h,\mathcal{L}_h}^{j+1}, \quad (2.30a)$$

with

$$\delta_{h,s}^{j+1} = \delta_{\text{tr}}^{j+1} + \mathcal{L}_h e_\pi^{j+1} + \alpha \mathcal{S}_h^{\text{lf}} e_\pi^j, \quad (2.30b)$$

$$\delta_{h,\mathcal{L}_h}^{j+1} = \delta_{h,\text{LTI}}^{j+1} + \delta_{h,g}^{j+1}, \quad (2.30c)$$

where the second component of (2.30c) is given by

$$\begin{aligned}
\delta_{h,\text{LTI},v}^{j+1} & = \frac{\tau^2}{4} \chi_{\text{lf}} \mathcal{L}_{u,h} \partial_\tau \hat{u}^{j+1} - \frac{\tau^2}{4} \chi_{\text{lf}} \mathcal{L}_{u,h} \partial_\tau e_{\pi,u}^{j+1} \\
& \quad + \tau^2 \chi_m \tilde{\varphi} \chi_m \mathcal{L}_{u,h} \partial_\tau \hat{u}^{j+1} - \tau^2 \chi_m \tilde{\varphi} \chi_m \mathcal{L}_{u,h} \partial_\tau e_{\pi,u}^{j+1},
\end{aligned} \quad (2.31a)$$

$$\delta_{h,g,v}^{j+1} = \frac{\tau^2}{4} \partial_\tau g_{v,h}^{j+1}. \quad (2.31b)$$

The first component of (2.30c) vanishes.

We now state the main result.

**Theorem 2.16.** *Let  $\alpha \in (0, 1]$ ,  $\vartheta_c^2 + \alpha\vartheta_{s,c} \in (0, 1)$ , and assume that*

$$\tau \leq \min\{\tau_{\text{CFL,lf,c}}(\vartheta_c), \tau_{\text{CFL,s,c}}(\vartheta_{s,c}), \tau_{\text{CFL},\Psi}\},$$

*as specified in (1.16) and (2.22). Assume that the exact solution  $x = (u \ v)$  of (1.1) satisfies*

$$\begin{aligned} x &\in C^3(0, T; L^2(\Omega)^m) \cap C(0, T; H^{k+1}(\mathcal{T}_h)^m \cap \mathcal{D}(\mathcal{L})), \\ u &\in C^2(0, T; H^1(\mathcal{T}_h)^{m_u} \cap \mathcal{D}(\mathcal{L}_u)) \cap C^1(0, T; H^2(\mathcal{T}_h)^{m_u} \cap \mathcal{D}(\mathcal{L}_u)), \\ v &\in C^2(0, T; H^1(\mathcal{T}_h)^{m_v} \cap \mathcal{D}(\mathcal{L}_v)), \end{aligned}$$

*and that the forcing term satisfies*

$$g_v \in C^2(0, T; L^2(\Omega)^{m_v}) \cap C^1(0, T; H^1(\mathcal{T}_h)^{m_v} \cap \mathcal{D}(\mathcal{L}_v)).$$

*If, additionally,  $\|h\|_{L^\infty(\Omega)} \leq 1$ , then for  $0 \leq t^{n+1} \leq T$ ,*

$$\begin{aligned} &\tilde{C}_{\text{stb,upw,c}}^{-1} \|x(t^{n+1}) - x_h^{n+1}\|_\Omega^2 + \frac{\tau}{4} \alpha \sum_{j=0}^n |e_h^{j+1} + e_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 \\ &\leq \tau^4 \hat{C}_1(x, g) + \tilde{C}_{\text{stb,upw,c}}^{-1} C_\pi^2 |h^{k+1} x(t^{n+1})|_{k+1, \mathcal{T}_h}^2 \\ &\quad + \hat{C}_2 \tau \sum_{j=0}^n \left( |h^k x(t^{j+1})|_{k+1, \mathcal{T}_{h,m+}}^2 + \frac{1+\alpha^2}{\alpha} |h^{k+1/2} x(t^{j+1})|_{k+1, \mathcal{T}_{h,\text{lf}}}^2 \right), \end{aligned}$$

*where  $\tilde{C}_{\text{stb,upw,c}}^{-1} = \min\{C_{\text{stb,upw,c}}^{-1}, \frac{1}{2}(1 - \alpha c_\Theta \vartheta_{s,c})\}$ . The data-dependent constant  $\hat{C}_1(x, g)$  and the constant  $\hat{C}_2$  are independent of  $\tau$  and  $h$ .*

For fixed  $T$  and  $\alpha > 0$ , suppressing the factors depending on the solution and forcing, the estimate takes the simplified form

$$\|x(t^{n+1}) - x_h^{n+1}\|_\Omega \lesssim \tau^2 + h_{m+}^k + h_{\text{lf}}^{k+1/2},$$

where  $h_{m+} = \max_{K \in \mathcal{T}_{h,m+}} h_K$  and  $h_{\text{lf}} = \max_{K \in \mathcal{T}_{h,\text{lf}}} h_K$ . Compared with the central-flux bound in Theorem 1.16, stabilization thus gains half an order in the spatial contribution from  $\mathcal{T}_{h,\text{lf}}$ , while the contribution from  $\mathcal{T}_{h,m+}$  remains of order  $k$  because stabilization is absent on the modified region  $\mathcal{T}_{h,m}$ .

We postpone the proof of Theorem 2.16 to the end of this section since we first have to provide estimates for the various defects appearing in the right-hand side of (2.29b).

We start with Lemma 2.17, in which an estimate for the tested defect  $\delta_{h,s}^{j+1}$  is stated.

**Lemma 2.17.** *Assume that the exact solution  $x$  of (1.1) satisfies*

$$x \in C(0, T; H^{k+1}(\mathcal{T}_h)^m \cap \mathcal{D}(\mathcal{L})).$$

*Then*

$$\begin{aligned} (\delta_{h,s}^{j+1}, e_h^{j+1} + e_h^j)_\Omega &\leq \gamma_1 \|e_h^{j+1} + e_h^j\|_\Omega^2 + \frac{1}{2} \gamma_2 (1 + \alpha^2) |e_h^{j+1} + e_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 \\ &\quad + \frac{1}{\gamma_1} C_1 |h^k \hat{x}^{j+1}|_{k+1, \mathcal{T}_{h,m+}}^2 + \frac{1}{2\gamma_1} \|\delta_{\text{tr}}^{j+1}\|_\Omega^2 \\ &\quad + \frac{1}{\gamma_2} C_2 |h^{k+1/2} \hat{x}^{j+1}|_{k+1, \mathcal{T}_{h,\text{lf}}}^2, \end{aligned} \tag{2.32}$$

*with constants  $C_1, C_2$  independent of  $\tau$  and  $h$  and positive scalars  $\gamma_1, \gamma_2$ .*

*Proof.* With the representation of the defect  $\delta_{h,s}^{j+1}$  (2.30b) together with the bounds (2.12) and (2.13) as well as the Cauchy–Schwarz inequality and Young’s inequality with weights  $\gamma_1, \gamma_2 > 0$ , we obtain

$$\begin{aligned}
& (\delta_{h,s}^{j+1}, e_h^{j+1} + e_h^j)_\Omega \\
&= (\delta_{\text{tr}}^{j+1}, e_h^{j+1} + e_h^j)_\Omega + (\mathcal{L}_h e_\pi^{j+1}, e_h^{j+1} + e_h^j)_\Omega + (\alpha \mathcal{S}_h^{\text{lf}} e_\pi^j, e_h^{j+1} + e_h^j)_\Omega \\
&\leq \frac{1}{2} \gamma_1 \|e_h^{j+1} + e_h^j\|_\Omega^2 + \frac{1}{2\gamma_1} \|\delta_{\text{tr}}^{j+1}\|_\Omega^2 \\
&\quad + \frac{1}{2} \gamma_2 |e_h^{j+1} + e_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 + \frac{1}{2\gamma_2} (C_{\pi,\mathcal{L},c}^{\text{up}})^2 |h^{k+1/2} \hat{x}^{j+1}|_{k+1,\mathcal{T}_{h,\text{lf}}}^2 \\
&\quad + \frac{1}{2} \gamma_1 \|e_h^{j+1} + e_h^j\|_{\mathcal{T}_{h,m+}}^2 + \frac{1}{2\gamma_1} (C_{\pi,\mathcal{L},f}^{\text{up}})^2 |h^k \hat{x}^{j+1}|_{k+1,\mathcal{T}_{h,m+}}^2 \\
&\quad + \frac{1}{2} \gamma_2 \alpha^2 |e_h^{j+1} + e_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 + \frac{1}{2\gamma_2} (C_{\pi,\mathcal{S}^{\text{lf}}})^2 |h^{k+1/2} \hat{x}^{j+1}|_{k+1,\mathcal{T}_{h,\text{lf}}}^2.
\end{aligned}$$

□

In the next Lemma 2.18, a bound for the term  $(\delta_{h,v,\mathcal{L}_h}^{j+1}, \delta_{h,v,s}^{j+1})_\Omega$  is given.

**Lemma 2.18.** Assume  $\|h\|_{L^\infty(\Omega)} \leq 1$  and assume that the exact solution  $x = (u \ v)$  of (1.1) satisfies

$$\begin{aligned}
x &\in C(0, T; H^{k+1}(\mathcal{T}_h)^m \cap \mathcal{D}(\mathcal{L})), \\
u &\in C^1(0, T; H^2(\mathcal{T}_h)^{m_u} \cap \mathcal{D}(\mathcal{L}_u)), \\
v &\in C^2(0, T; H^1(\mathcal{T}_h)^{m_v} \cap \mathcal{D}(\mathcal{L}_v)),
\end{aligned}$$

and that the forcing term satisfies

$$g_v \in C^1(0, T; H^1(\mathcal{T}_h)^{m_v} \cap \mathcal{D}(\mathcal{L}_v)).$$

Then

$$\begin{aligned}
2(\delta_{h,v,\mathcal{L}_h}^{j+1}, \delta_{h,v,s}^{j+1})_\Omega &\leq \tilde{C} \left( |h^{k+1/2} x|_{k+1,\mathcal{T}_{h,\text{lf}}}^2 + |h^k u|_{k+1,\mathcal{T}_{h,m+}}^2 \right. \\
&\quad + \tau^4 \left( |\partial_\tau \hat{u}^{j+1}|_1^2 + \|\partial_\tau \partial_t \hat{v}^{j+1}\|_\Omega^2 + \|\partial_\tau g_v^{j+1}\|_\Omega^2 \right) \\
&\quad + \tau^4 \left( |\partial_\tau \hat{u}^{j+1}|_{2,\mathcal{T}_{h,\text{lf}}}^2 + |\partial_\tau \partial_t \hat{v}^{j+1}|_{1,\mathcal{T}_{h,\text{lf}}}^2 + |\partial_\tau g_v^{j+1}|_{1,\mathcal{T}_{h,\text{lf}}}^2 \right) \\
&\quad \left. + \|\delta_{\text{tr},v}^{j+1}\|_\Omega^2 \right)
\end{aligned} \tag{2.33}$$

with a constant  $\tilde{C}$  independent of  $h$  and  $\tau$ .

*Proof.* Similar to the proof of Lemma 2.17 we get

$$\begin{aligned}
(\delta_{h,v,\mathcal{L}_h}^{j+1}, \delta_{h,v,s}^{j+1})_\Omega &= (\delta_{h,v,\mathcal{L}_h}^{j+1}, \delta_{\text{tr},v}^{j+1})_\Omega + (\delta_{h,v,\mathcal{L}_h}^{j+1}, \mathcal{L}_{u,h} e_{\pi,u}^{j+1})_\Omega + (\delta_{h,v,\mathcal{L}_h}^{j+1}, \alpha \mathcal{S}_{v,h}^{\text{lf}} e_{\pi,v}^j)_\Omega \\
&\leq \frac{1}{2} \left( \|\delta_{h,v,\mathcal{L}_h}^{j+1}\|_\Omega^2 + \|\delta_{\text{tr},v}^{j+1}\|_\Omega^2 \right. \\
&\quad + C_{\pi,\mathcal{L}_u,c}^{\text{up}} (|\delta_{h,v,\mathcal{L}_h}^{j+1}|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 + |h^{k+1/2} \hat{u}^{j+1}|_{k+1,\mathcal{T}_{h,\text{lf}}}^2) \\
&\quad + C_{\pi,\mathcal{L}_u,f}^{\text{up}} (\|\delta_{h,v,\mathcal{L}_h}^{j+1}\|_{\mathcal{T}_{h,m+}}^2 + |h^k \hat{u}^{j+1}|_{k+1,\mathcal{T}_{h,m+}}^2) \\
&\quad \left. + \alpha C_{\pi,\mathcal{S}_v^{\text{lf}}} (|\delta_{h,v,\mathcal{L}_h}^{j+1}|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 + |h^{k+1/2} \hat{v}^{j+1}|_{k+1,\mathcal{T}_{h,\text{lf}}}^2) \right)
\end{aligned}$$

Applying Lemma 2.22 to estimate  $|\delta_{h,v,\mathcal{L}_h}^{j+1}|_{\mathcal{S}_{v,h}^{\text{lf}}}$  and (1.42a) with  $k = 0$  for  $\|\delta_{h,v,\mathcal{L}_h}^{j+1}\|_\Omega$  together with  $\frac{\tau^2}{4} \|\partial_\tau g_v^{j+1}\|_\Omega \leq \frac{\tau^2}{4} \|\partial_\tau g_v^{j+1}\|_\Omega$  yields the statement. □

In the next two Lemmas 2.19 and 2.20, bounds for the terms involving the discrete time derivative  $\partial_\tau \delta_{h,v,\mathcal{L}_h}^{j+1}$  in the error equality (2.29b) are stated. These terms arise from the summation-by-parts step in the proof of the error equality. This step contributes an additional factor  $\tau$ , which allows us to use the CFL condition (2.22) to absorb the stabilization operator norm in Lemma 2.19. Lemma 2.20 bounds the  $L^2$ -inner product directly.

**Lemma 2.19.** *Assume that the exact solution  $x = (u \ v)$  of (1.1) satisfies*

$$u \in C^2(0, T; H^1(\mathcal{T}_h)^{m_u} \cap \mathcal{D}(\mathcal{L}_u)), \quad v \in C^3(0, T; L^2(\Omega)^{m_v}),$$

*and that the right-hand side satisfies*

$$g_v \in C^2(0, T; L^2(\Omega)^{m_v}).$$

*Moreover, assume that  $\tau \leq \tau_{\text{CFL},s,c}(\vartheta_{s,c})$ , defined in (2.22). Then*

$$\begin{aligned} & (\partial_\tau \delta_{h,v,\mathcal{L}_h}^{j+1}, \tau \alpha \mathcal{S}_{v,h}^{\text{lf}} e_{h,v}^j)_\Omega \\ & \leq \alpha \gamma_5 c_\Theta \vartheta_{s,c} \|e_{h,v}^j\|_{\mathcal{T}_{h,\text{lf}}}^2 \\ & \quad + \alpha \frac{1}{\gamma_5} c_\Theta \vartheta_{s,c} C_5 \tau^4 \left( \|\partial_\tau^2 \partial_t \widehat{v}^{j+1}\|_{\mathcal{T}_{h,\text{lf}}}^2 + \|\partial_\tau^2 g_v^{j+1}\|_{\mathcal{T}_{h,\text{lf}}}^2 + |\partial_\tau^2 \widehat{u}^{j+1}|_{1,\mathcal{T}_{h,\text{lf}}}^2 \right), \end{aligned}$$

*with a constant  $C_5$  independent of  $h$  and  $\tau$  and a positive scalar  $\gamma_5$ .*

*Proof.* The Cauchy–Schwarz inequality, Young’s inequality with weight  $\gamma_5$ , and the CFL condition (2.22) give

$$\begin{aligned} (\partial_\tau \delta_{h,v,\mathcal{L}_h}^{j+1}, \tau \alpha \mathcal{S}_{v,h}^{\text{lf}} e_{h,v}^j)_\Omega & \leq \frac{\tau}{2} \alpha \frac{1}{\gamma_5} |\partial_\tau \delta_{h,v,\mathcal{L}_h}^{j+1}|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 + \frac{\tau}{2} \alpha \gamma_5 |e_{h,v}^j|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 \\ & \leq \alpha \frac{1}{\gamma_5} c_\Theta \vartheta_{s,c} \|\partial_\tau \delta_{h,v,\mathcal{L}_h}^{j+1}\|_{\mathcal{T}_{h,\text{lf}}}^2 + \alpha \gamma_5 c_\Theta \vartheta_{s,c} \|e_{h,v}^j\|_{\mathcal{T}_{h,\text{lf}}}^2 \end{aligned}$$

Using (1.42b) with  $k = 0$  and  $\frac{\tau^2}{4} \|\partial_\tau^2 g_{v,h}^{j+1}\|_{\mathcal{T}_{h,\text{lf}}} \leq \frac{\tau^2}{4} \|\partial_\tau^2 g_v^{j+1}\|_{\mathcal{T}_{h,\text{lf}}}$  gives the result.  $\square$

**Lemma 2.20.** *Assume that the exact solution  $x = (u \ v)$  of (1.1) satisfies*

$$u \in C^2(0, T; H^1(\mathcal{T}_h)^{m_u} \cap \mathcal{D}(\mathcal{L}_u)), \quad v \in C^3(0, T; L^2(\Omega)^{m_v}),$$

*and that the forcing term satisfies*

$$g_v \in C^2(0, T; L^2(\Omega)^{m_v}).$$

*Then*

$$2(\partial_\tau \delta_{h,v,\mathcal{L}_h}^{j+1}, e_{h,v}^j)_\Omega \leq \frac{1}{\gamma_4} C_4 \tau^4 \left( \|\partial_\tau^2 \partial_t \widehat{v}^{j+1}\|_\Omega^2 + \|\partial_\tau^2 g_v^{j+1}\|_\Omega^2 + |\partial_\tau^2 \widehat{u}^{j+1}|_1^2 \right) + \gamma_4 \|e_{h,v}^j\|_\Omega^2,$$

*with a constant  $C_4$  independent of  $h$  and  $\tau$  and a positive scalar  $\gamma_4$ .*

*Proof.* The Cauchy–Schwarz inequality and Young’s inequality with weight  $\gamma_4$  give

$$(\partial_\tau \delta_{h,v,\mathcal{L}_h}^{j+1}, e_{h,v}^j)_\Omega \leq \frac{1}{2\gamma_4} \|\partial_\tau \delta_{h,v,\mathcal{L}_h}^{j+1}\|_\Omega^2 + \frac{\gamma_4}{2} \|e_{h,v}^j\|_\Omega^2$$

Applying (1.42b) with  $k = 0$  and  $\frac{\tau^2}{4} \|\partial_\tau^2 g_{v,h}^{j+1}\|_\Omega \leq \frac{\tau^2}{4} \|\partial_\tau^2 g_v^{j+1}\|_\Omega$  gives the result.  $\square$

The following Lemma 2.21 deals with the Friedrichs' defect tested in the stabilization form.

**Lemma 2.21.** *Assume  $\|h\|_{L^\infty(\Omega)} \leq 1$  and assume that the exact solution  $x = (u \ v)$  of (1.1) satisfies*

$$u \in C^1(0, T; H^2(\mathcal{T}_h)^{m_u} \cap \mathcal{D}(\mathcal{L}_u)), \quad v \in C^2(0, T; H^1(\mathcal{T}_h)^{m_v} \cap \mathcal{D}(\mathcal{L}_v)),$$

*and that the forcing term satisfies*

$$g_v \in C^1(0, T; H^1(\mathcal{T}_h)^{m_v} \cap \mathcal{D}(\mathcal{L}_v)).$$

*Then*

$$\begin{aligned} & (\delta_{h,v,\mathcal{L}_h}^{j+1}, \alpha \mathcal{S}_{v,h}^{\text{lf}}(e_{h,v}^{j+1} + e_{h,v}^j))_\Omega \\ & \leq \frac{1}{2} \alpha \gamma_3 |e_{h,v}^{j+1} + e_{h,v}^j|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 \\ & \quad + \frac{1}{2} \alpha \frac{1}{\gamma_3} C_3 \tau^4 \left( |\partial_\tau \hat{u}^{j+1}|_{2,\mathcal{T}_{h,\text{lf}}}^2 + |\partial_\tau \partial_t \hat{v}^{j+1}|_{1,\mathcal{T}_{h,\text{lf}}}^2 + |\partial_\tau g_v^{j+1}|_{1,\mathcal{T}_{h,\text{lf}}}^2 \right), \end{aligned}$$

*with a constant  $C_3$  independent of  $h$  and  $\tau$  and a positive scalar  $\gamma_3$ .*

*Proof.* The Cauchy–Schwarz inequality and Young's inequality with weight  $\gamma_3$  give

$$(\delta_{h,v,\mathcal{L}_h}^{j+1}, \alpha \mathcal{S}_{v,h}^{\text{lf}}(e_{h,v}^{j+1} + e_{h,v}^j))_\Omega \leq \frac{1}{2} \alpha \frac{1}{\gamma_3} |\delta_{h,v,\mathcal{L}_h}^{j+1}|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 + \frac{1}{2} \alpha \gamma_3 |e_{h,v}^{j+1} + e_{h,v}^j|_{\mathcal{S}_{v,h}^{\text{lf}}}^2.$$

Applying Lemma 2.22 yields the result.  $\square$

The following Lemma 2.22 is an auxiliary result used in the previous Lemmas 2.18 and 2.21.

**Lemma 2.22.** *Under the assumptions of Lemma 2.21 we have*

$$|\delta_{h,v,\mathcal{L}_h}^{j+1}|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 \leq C \tau^4 \left( |\partial_\tau \hat{u}^{j+1}|_{2,\mathcal{T}_{h,\text{lf}}}^2 + |\partial_\tau \partial_t \hat{v}^{j+1}|_{1,\mathcal{T}_{h,\text{lf}}}^2 + |\partial_\tau g_v^{j+1}|_{1,\mathcal{T}_{h,\text{lf}}}^2 \right), \quad (2.34)$$

*with a constant  $C$  independent of  $h$  and  $\tau$ .*

*Proof.* Recall the decomposition (2.30c)

$$\begin{aligned} \delta_{h,v,\mathcal{L}_h}^{j+1} &= \frac{\tau^2}{4} \chi_{\text{lf}} \mathcal{L}_{u,h} \partial_\tau \hat{u}^{j+1} - \frac{\tau^2}{4} \chi_{\text{lf}} \mathcal{L}_{u,h} \partial_\tau e_{\pi,u}^{j+1} \\ &\quad + \tau^2 \chi_m \tilde{\varphi} \chi_m \mathcal{L}_{u,h} \partial_\tau \hat{u}^{j+1} - \tau^2 \chi_m \tilde{\varphi} \chi_m \mathcal{L}_{u,h} \partial_\tau e_{\pi,u}^{j+1} + \frac{\tau^2}{4} \partial_\tau g_v^{j+1}. \end{aligned}$$

Due to the definition of  $\chi_m$  in (1.8) and  $|\cdot|_{\mathcal{S}_{v,h}^{\text{lf}}}$  in (2.10), we have

$$|\tau^2 \chi_m \tilde{\varphi} \chi_m \mathcal{L}_{u,h} \partial_\tau \hat{u}^{j+1}|_{\mathcal{S}_{v,h}^{\text{lf}}} = 0, \quad \text{and} \quad |\tau^2 \chi_m \tilde{\varphi} \chi_m \mathcal{L}_{u,h} \partial_\tau e_{\pi,u}^{j+1}|_{\mathcal{S}_{v,h}^{\text{lf}}} = 0.$$

Using the consistency property (1.37a) and (1.1) gives

$$|\delta_{h,v,\mathcal{L}_h}^{j+1}|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 \leq 3 \frac{\tau^4}{16} \left( |\pi_h \partial_\tau (\partial_t \hat{v}^{j+1} - g_v^{j+1})|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 + |\mathcal{L}_{u,h} \partial_\tau e_{\pi,u}^{j+1}|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 + |\partial_\tau g_v^{j+1}|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 \right).$$

The consistency property (2.9a) together with Assumption 2.2 gives us

$$|\pi_h \partial_\tau (\partial_t \hat{v}^{j+1} - g_v^{j+1})|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 = |\partial_\tau (\partial_t e_{\pi,v}^{j+1} - e_{\pi,g_v}^{j+1})|_{\mathcal{S}_{v,h}^{\text{lf}}}^2$$

$$\begin{aligned}
 &\leq \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{int}}} \hat{\eta}_{v,i} \|\partial_\tau [\partial_t e_{\pi,v}^{j+1} - e_{\pi,g_v}^{j+1}]_F\|_{L^2(F)}^2 + \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{bnd}}} \hat{\eta}_{v,b} \|\partial_\tau (\partial_t e_{\pi,v}^{j+1} - e_{\pi,g_v}^{j+1})\|_{L^2(F)}^2 \\
 &\leq 2 \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{int}}} \hat{\eta}_{v,i} C_{\pi,\partial}^2 \left( |h^{1/2} \partial_\tau (\partial_t \hat{v}^{j+1} - g_v^{j+1})|_{1,K_1^F}^2 + |h^{1/2} \partial_\tau (\partial_t \hat{v}^{j+1} - g_v^{j+1})|_{1,K_2^F}^2 \right) \\
 &\quad + \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{bnd}}} \hat{\eta}_{v,b} C_{\pi,\partial}^2 |h^{1/2} \partial_\tau (\partial_t \hat{v}^{j+1} - g_v^{j+1})|_{1,K}^2.
 \end{aligned}$$

In the last estimate we have used the projection error estimate (A.3) with  $\ell = 0$  and  $K_1^F, K_2^F \in \mathcal{T}_{h,\text{lf}}$  with common interface  $F \in \mathcal{F}_{h,\text{lf}}^{\text{int}}$ .

We proceed as in the previous estimate, but use the discrete trace inequality (A.1) and the approximation property (1.37b) with  $k = 1$  in the last step:

$$\begin{aligned}
 |\mathcal{L}_{u,h} \partial_\tau e_{\pi,u}^{j+1}|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 &\leq 2 \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{int}}} \hat{\eta}_{v,i} C_{\text{tr}}^2 C_{\pi,\mathcal{L}}^2 \left( |h^{1/2} \partial_\tau \hat{u}^{j+1}|_{2,K_1^F}^2 + |h^{1/2} \partial_\tau \hat{u}^{j+1}|_{2,K_2^F}^2 \right) \\
 &\quad + \frac{1}{2} \sum_{F \in \mathcal{F}_{h,\text{lf}}^{\text{bnd}}} \hat{\eta}_{v,b} C_{\text{tr}}^2 C_{\pi,\mathcal{L}}^2 |h^{1/2} \partial_\tau \hat{u}^{j+1}|_{2,K}^2
 \end{aligned}$$

The consistency property (2.9a) and the projection error estimate (2.12) with  $k = 0$  give

$$|\partial_\tau g_{v,h}^{j+1}|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 = |\partial_\tau e_{\pi,g_v}^{j+1}|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 \leq C_{\pi,\mathcal{S}_v} |h^{1/2} \partial_\tau g_{v,h}^{j+1}|_{1,\mathcal{T}_{h,\text{lf}}}^2.$$

Using that each cell has at most  $N_\partial$  faces together with the assumption  $\|h\|_{L^\infty(\Omega)} \leq 1$  proves the statement.  $\square$

Combining the previous results, we can now prove Theorem 2.16.

*Proof of Theorem 2.16.* Using the Cauchy–Schwarz inequality, Young’s inequality, and the CFL condition (2.22), we estimate

$$\begin{aligned}
 &2(\delta_{h,v,\mathcal{L}_h}^{n+1}, (\mathcal{I}_v - \frac{\tau}{2} \alpha \mathcal{S}_{v,h}^{\text{lf}}) e_{h,v}^{n+1})_\Omega \\
 &\leq 2\|\delta_{h,v,\mathcal{L}_h}^{n+1}\|_\Omega^2 + \frac{1}{2}\|(\mathcal{I}_v - \frac{\tau}{2} \alpha \mathcal{S}_{v,h}^{\text{lf}}) e_{h,v}^{n+1}\|_\Omega^2 \\
 &\leq 2\|\delta_{h,v,\mathcal{L}_h}^{n+1}\|_\Omega^2 + \frac{1}{2}\left(\|e_{h,v}^{n+1}\|_\Omega^2 - \tau \alpha |e_{h,v}^{n+1}|_{\mathcal{S}_{v,h}^{\text{lf}}}^2 + \frac{\tau}{2} \alpha^2 c_\Theta \vartheta_{s,c} |e_{h,v}^{n+1}|_{\mathcal{S}_{v,h}^{\text{lf}}}^2\right) \\
 &\leq 2\|\delta_{h,v,\mathcal{L}_h}^{n+1}\|_\Omega^2 + \frac{1}{2}\left(\|e_{h,v}^{n+1}\|_\Omega^2 - \frac{\tau}{2} \alpha |e_{h,v}^{n+1}|_{\mathcal{S}_{v,h}^{\text{lf}}}^2\right).
 \end{aligned} \tag{2.35}$$

Moreover, applying the estimate (2.23) for the  $v$  component, we get the lower bound

$$((\mathcal{R}_- - \frac{\tau}{2} \alpha \mathcal{S}_h^{\text{lf}}) e_h^{n+1}, e_h^{n+1})_\Omega \geq C_{\text{stb,upw,c}}^{-1} \|e_{h,u}^{n+1}\|_\Omega^2 + \|e_{h,v}^{n+1}\|_\Omega^2 - \frac{\tau}{2} \alpha |e_{h,v}^{n+1}|_{\mathcal{S}_{v,h}^{\text{lf}}}^2. \tag{2.36}$$

Combining (2.35) and (2.36) in (2.29b) and using the CFL condition (2.22), we get

$$\begin{aligned}
& C_{\text{stb,upw,c}}^{-1} \|e_{h,u}^{n+1}\|_{\Omega}^2 + \frac{1}{2}(1 - \alpha c_{\Theta} \vartheta_{s,c}) \|e_{h,v}^{n+1}\|_{\Omega}^2 + \frac{\tau}{2} \alpha \sum_{j=0}^n |e_h^{j+1} + e_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 \\
& \leq \tau \sum_{j=0}^n (\delta_{h,s}^{j+1}, e_h^{j+1} + e_h^j)_{\Omega} + \tau \sum_{j=0}^n (\delta_{h,v,\mathcal{L}_h}^{j+1}, \alpha \mathcal{S}_{v,h}^{\text{lf}}(e_{h,v}^{j+1} + e_{h,v}^j))_{\Omega} \\
& \quad + \tau \sum_{j=1}^n (\partial_{\tau} \delta_{h,v,\mathcal{L}_h}^{j+1}, \tau \alpha \mathcal{S}_{v,h}^{\text{lf}} e_{h,v}^j)_{\Omega} \\
& \quad - \tau \sum_{j=0}^n 2(\delta_{h,v,\mathcal{L}_h}^{j+1}, \delta_{h,v,s}^{j+1})_{\Omega} - \tau \sum_{j=1}^n 2(\partial_{\tau} \delta_{h,v,\mathcal{L}_h}^{j+1}, e_{h,v}^j)_{\Omega} + 2\|\delta_{h,v,\mathcal{L}_h}^{n+1}\|_{\Omega}^2.
\end{aligned} \tag{2.37}$$

The six terms on the right-hand side of (2.37) are bounded, in their order of appearance, by Lemma 2.17, Lemma 2.21, Lemma 2.19, Lemma 2.18, Lemma 2.20, and (1.42a) with  $k = 0$ , respectively. Inserting these estimates and grouping the terms, we get

$$\begin{aligned}
& \tilde{C}_{\text{stb,upw,c}}^{-1} \|e_h^{n+1}\|_{\Omega}^2 + \frac{\tau}{2}(\alpha - \gamma_2(1 + \alpha^2) - \alpha\gamma_3) \sum_{j=0}^n |e_h^{j+1} + e_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 \\
& \leq \tau \sum_{j=0}^n 2\gamma_1 \|e_h^{j+1}\|_{\Omega}^2 + (2\gamma_1 + \gamma_4 + \alpha\gamma_5 c_{\Theta} \vartheta_{s,c}) \|e_h^j\|_{\Omega}^2 + C(x, g)
\end{aligned}$$

with  $\tilde{C}_{\text{stb,upw,c}}^{-1} = \min\{C_{\text{stb,upw,c}}^{-1}, \frac{1}{2}(1 - \alpha c_{\Theta} \vartheta_{s,c})\}$ . The constant  $C(x, g)$  is defined in (2.38) and depends on the exact solution as well as on the data. We first fix the Young weights and apply the discrete Gronwall argument, and return to the estimate of  $C(x, g)$  at the end of the proof. Choosing  $\gamma_5 = \frac{\gamma_4}{\alpha c_{\Theta} \vartheta_{s,c}}$ ,  $\gamma_4 = \frac{1}{2}\gamma_1$ ,  $\gamma_2 = \frac{\alpha}{1+\alpha^2}\gamma_3$ , and  $\gamma_3 = \frac{1}{4}$  leads to

$$\tilde{C}_{\text{stb,upw,c}}^{-1} \|e_h^{n+1}\|_{\Omega}^2 + \frac{\tau}{4} \alpha \sum_{j=0}^n |e_h^{j+1} + e_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 \leq 3\gamma_1 \tau \sum_{j=0}^n \|e_h^{j+1}\|_{\Omega}^2 + \|e_h^j\|_{\Omega}^2 + C(x, g).$$

The discrete Gronwall-type inequality [HS19, Lemma A.1] yields

$$\tilde{C}_{\text{stb,upw,c}}^{-1} \|e_h^{n+1}\|_{\Omega}^2 + \frac{\tau}{4} \alpha \sum_{j=0}^n |e_h^{j+1} + e_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 \leq e^{9\tilde{C}_{\text{stb,upw,c}}\gamma_1(n+1)\tau} C(x, g).$$

Choosing  $\gamma_1 = \frac{1}{6} \frac{1}{T+1} \tilde{C}_{\text{stb,upw,c}}^{-1}$  gives

$$\tilde{C}_{\text{stb,upw,c}}^{-1} \|e_h^{n+1}\|_{\Omega}^2 + \frac{\tau}{4} \alpha \sum_{j=0}^n |e_h^{j+1} + e_h^j|_{\mathcal{S}_h^{\text{lf}}}^2 \leq e^{\frac{3}{2}} C(x, g).$$

The constant  $C(x, g)$  is given as

$$\begin{aligned}
 C(x, g) = & \tau \sum_{j=0}^n 6(T+1) \tilde{C}_{\text{stb,upw,c}} \|\delta_{\text{tr}}^{j+1}\|_{\Omega}^2 \\
 & + \tau^4 (2\alpha C_3 + \tilde{C}) \left( |\partial_{\tau} \hat{u}^{j+1}|_{2, \mathcal{T}_{h,\text{lf}}}^2 + |\partial_{\tau} \partial_t \hat{v}^{j+1}|_{1, \mathcal{T}_{h,\text{lf}}}^2 + |\partial_{\tau} g_v^{j+1}|_{1, \mathcal{T}_{h,\text{lf}}}^2 \right) \\
 & + \tau^4 \tilde{C} \left( |\partial_{\tau} \hat{u}^{j+1}|_1^2 + \|\partial_{\tau} \partial_t \hat{v}^{j+1}\|_{\Omega}^2 + \|\partial_{\tau} g_v^{j+1}\|_{\Omega}^2 \right) \\
 & + \tau^4 12 \tilde{C}_{\text{stb,upw,c}} (T+1) \\
 & \cdot \tau \sum_{j=1}^n C_4 \left( \|\partial_{\tau}^2 \partial_t \hat{v}^{j+1}\|_{\Omega}^2 + \|\partial_{\tau}^2 g_v^{j+1}\|_{\Omega}^2 + |\partial_{\tau}^2 \hat{u}^{j+1}|_1^2 \right) \\
 & + \alpha^2 c_{\Theta}^2 \vartheta_{s,c}^2 C_5 \left( \|\partial_{\tau}^2 \partial_t \hat{v}^{j+1}\|_{\mathcal{T}_{h,\text{lf}}}^2 + \|\partial_{\tau}^2 g_v^{j+1}\|_{\mathcal{T}_{h,\text{lf}}}^2 + |\partial_{\tau}^2 \hat{u}^{j+1}|_{1, \mathcal{T}_{h,\text{lf}}}^2 \right) \\
 & + \tau^4 C_4 \left( \|\partial_{\tau} \partial_t \hat{v}^{n+1}\|_{\Omega}^2 + \|\partial_{\tau} g_v^{n+1}\|_{\Omega}^2 + |\partial_{\tau} \hat{u}^{n+1}|_1^2 \right) \\
 & + \tau \sum_{j=0}^n 6(T+1) \tilde{C}_{\text{stb,upw,c}} C_1 |h^k \hat{x}^{j+1}|_{k+1, \mathcal{T}_{h,m+}}^2 \\
 & + 4 \frac{1+\alpha^2}{\alpha} C_2 |h^{k+1/2} \hat{x}^{j+1}|_{k+1, \mathcal{T}_{h,\text{lf}}}^2 \\
 & + \tilde{C} \left( |h^{k+1/2} \hat{x}^{j+1}|_{k+1, \mathcal{T}_{h,\text{lf}}}^2 + |h^k \hat{u}^{j+1}|_{k+1, \mathcal{T}_{h,m+}}^2 \right).
 \end{aligned} \tag{2.38}$$

We use (1.34) to estimate

$$\tau \sum_{j=0}^n \|\partial_{\tau} \hat{x}^{j+1}\|_{\Omega}^2 \leq \int_{t^0}^{t^{n+1}} \|\partial_t x(s)\|_{\Omega}^2 ds, \quad \text{and} \quad \tau \sum_{j=1}^n \|\partial_{\tau}^2 \hat{x}^{j+1}\|_{\Omega}^2 \leq 4 \int_{t^0}^{t^{n+1}} \|\partial_t^2 x(s)\|_{\Omega}^2 ds.$$

Together with (1.35b), we get the bound

$$\begin{aligned}
 C(x, g) \leq & \tau^4 C(T+1) \int_{t^0}^{t^{n+1}} \left( \|\partial_t^3 x(s)\|_{\Omega}^2 + \|\partial_t^2 v(s)\|_{\Omega}^2 + \|\partial_t g_v(s)\|_{\Omega}^2 + \|\partial_t^2 g_v(s)\|_{\Omega}^2 \right. \\
 & + |\partial_t u(s)|_{2, \mathcal{T}_{h,\text{lf}}}^2 + |\partial_t^2 v(s)|_{1, \mathcal{T}_{h,\text{lf}}}^2 + |\partial_t g_v(s)|_{1, \mathcal{T}_{h,\text{lf}}}^2 \\
 & \left. + |\partial_t u(s)|_1^2 + |\partial_t^2 u(s)|_1^2 \right) ds \\
 & + \tau^4 C \max_{s \in [t^n, t^{n+1}]} \left( \|\partial_t^2 v(s)\|_{\Omega}^2 + \|\partial_t g_v(s)\|_{\Omega}^2 + |\partial_t u(s)|_1^2 \right) \\
 & + C \tau \sum_{j=0}^n (T+1) |h^k \hat{x}^{j+1}|_{k+1, \mathcal{T}_{h,m+}}^2 + \frac{1+\alpha^2}{\alpha} |h^{k+1/2} \hat{x}^{j+1}|_{k+1, \mathcal{T}_{h,\text{lf}}}^2.
 \end{aligned} \tag{2.39}$$

The projection error estimate (see [DHK<sup>+</sup>23, Lem. 10.2])

$$\|e_{\pi}^{n+1}\|_{\Omega} \leq C_{\pi} |h^{k+1} \hat{x}^{n+1}|_{k+1, \mathcal{T}_h},$$

and the identity  $\|e^{n+1}\|_{\Omega}^2 = \|e_{\pi}^{n+1}\|_{\Omega}^2 + \|e_h^{n+1}\|_{\Omega}^2$  conclude the proof.  $\square$

## 2.5. Numerical examples

We illustrate the spatial and temporal convergence as well as stability behavior of the local time-integration method extended to upwind-type stabilization in space. As an example equation we

use the two-dimensional first-order wave system

$$\partial_t u = \nabla \cdot v + g_u, \quad \text{on } \Omega \times (0, T], \quad (2.40a)$$

$$\partial_t v = \nabla u + g_v, \quad \text{on } \Omega \times (0, T], \quad (2.40b)$$

$$u(0) = u_0, \quad v(0) = v_0, \quad \text{on } \Omega, \quad (2.40c)$$

$$u = 0, \quad \text{on } \partial\Omega \times (0, T], \quad (2.40d)$$

on the unit square  $\Omega = (0, 1)^2$ . In each experiment, we choose the initial values and forcing so that the exact solution is  $u(x, y, t) = v_x(x, y, t) = v_y(x, y, t) = w(x, y, t)$ , with  $w$  specified below. Both experiments use center-refined quadrilateral meshes, see Figure 3.1 for an example, and the stabilization fields from Example 2.3 with all three constants equal to one. We choose  $\alpha = 1$  and measure the  $L^2$ -error at the final time  $T = 1$ .

### 2.5.1. Spatial convergence

For the spatial convergence study, presented in Figure 2.2, we choose

$$w(x, y, t) = \sin(\pi x) \sin(\pi y)(1 + t).$$

The mesh has two additional refinements at its center and is randomly perturbed to reduce superconvergence effects caused by its otherwise structured geometry. We choose dG degrees  $k = 1, 2, 3$  as a basis and increased it to  $k + 1$  on the coarse cells contained in  $\mathcal{T}_{h,m+}$ , as explained in Remark 2.9. We integrate in time using the locally implicit filter (1.44) with fixed time step size  $\tau = 10^{-5}$  to prevent instabilities from the underlying leapfrog method. For each base degree  $k$ , the observed convergence rates approach  $k + 1/2$ .

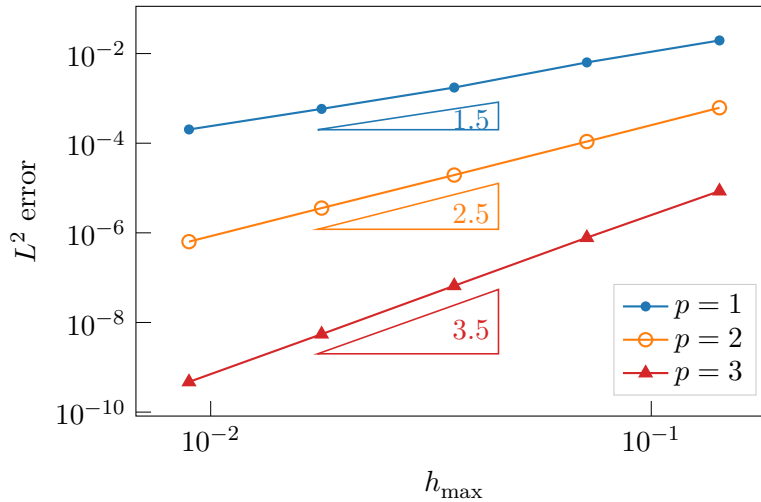


Figure 2.2.: Errors of the local time-integration method with locally implicit filter (1.44) combined with an upwind fluxes spatial discretization with different dG polynomial degrees  $k = 1, 2, 3$  and data from Section 2.5.1. The horizontal axis shows the maximal cell diameter  $h_{\max} = \max_{K \in \mathcal{T}_h} h_K$ . The vertical axis shows the  $L^2$ -error at the final time  $T = 1$ . The triangles indicate the different convergence orders.

### 2.5.2. Temporal convergence

For the temporal convergence study, we choose

$$w(x, y, t) = x(1 - x)y(1 - y) \exp(t) \sin(t).$$

In the applied upwind fluxes spatial discretization we choose fixed dG degree  $k = 2$  on every cell of a mesh with two global refinements and four additional refinements at the center. Since the spatial factor is a tensor-product polynomial of degree two in each coordinate, the exact state is contained in the dG space.

In Figure 2.3, we depict the errors and stability behavior of the local time-integration method with locally implicit filter (1.44) and the leapfrog-Chebyshev filter (1.45) with degrees  $p = 1, 2, 4, 5$  and stabilization parameter (1.47) with  $\eta = 0.1$ . For  $p = 1$ , the resulting method is the leapfrog method. All methods exhibit second-order convergence within their stable regimes. The largest stable time step size approximately doubles from  $p = 1$  to  $p = 2$  and again from  $p = 2$  to  $p = 4$ . It remains unchanged from  $p = 4$  to  $p = 5$  because the leapfrog CFL condition on the coarse grid then becomes the dominant time step size restriction.

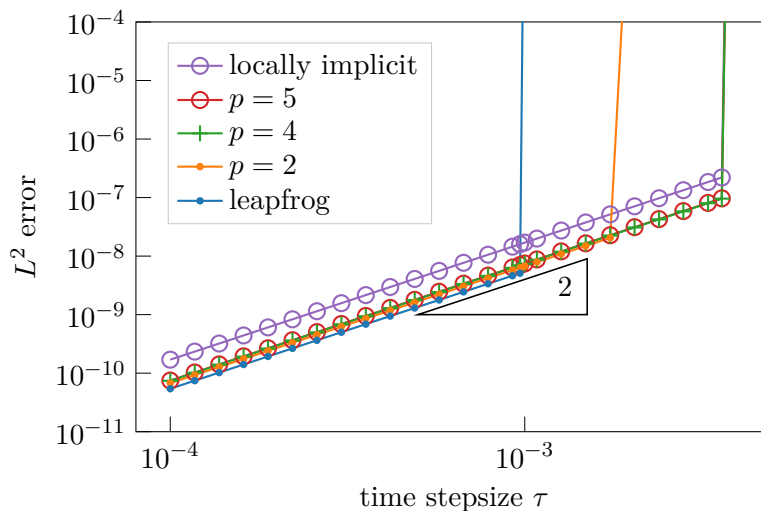


Figure 2.3.: Errors of the leapfrog method, and local time-integration method with locally implicit filter (1.44) and leapfrog-Chebyshev filter (1.45) with  $p = 2, 4, 5$  for different time stepsizes and data from Section 2.5.2. The upwind fluxes spatial discretization uses fixed dG degree  $k = 2$ . The horizontal axis shows the time stepsizes  $\tau$ . The vertical axis shows the  $L^2$ -error at the final time  $T = 1$ . The triangle indicates second-order convergence.

# APPENDIX A

---

## Weak and strong forms of the discrete Friedrichs' operators

### A.1. Weak and strong forms of the discrete Friedrichs' operators

We briefly recall notation and results from [DHK<sup>+</sup>23] needed for Chapter 2.

Whenever  $F \in \mathcal{F}_h^{\text{bnd}}$ , we define  $\mathbf{n}^F$  as the outward unit normal vector to  $\Gamma$ . For  $F \in \mathcal{F}_h^{\text{int}}$ , we label the two neighboring elements by  $K_1^F$  and  $K_2^F$  and define  $\mathbf{n}^F$  as the outward unit normal vector of  $K_1^F$  toward  $K_2^F$ . This orientation follows [DPE12].

Let  $F \in \mathcal{F}_h^{\text{int}}$  and let  $u: \Omega \rightarrow \mathbb{R}$  admit well-defined traces on  $F$ . We define the average and jump by

$$\{u\}_F = \frac{1}{2} \left( (u|_{K_1^F})|_F + (u|_{K_2^F})|_F \right), \quad \llbracket u \rrbracket_F = (u|_{K_1^F})|_F - (u|_{K_2^F})|_F.$$

For an element  $K \in \mathcal{T}_h$  and a face  $F$  of  $K$ , the discrete trace inequality [DPE12, Lem. 1.46] is

$$\|u_h\|_{L^2(F)} \leq C_{\text{tr}} \|h^{-1/2} u_h\|_{L^2(K)} \quad u_h \in \mathbb{V}_h^u. \quad (\text{A.1})$$

The corresponding estimate on  $\mathbb{V}_h^v$  is analogous.

For the  $L^2$ -projection errors

$$e_{\pi,u} = u - \pi_h u, \quad e_{\pi,v} = v - \pi_h v, \quad (\text{A.2})$$

we use the following face estimate for the first field.

**Lemma A.1** ([DHK<sup>+</sup>23, Lemma 10.2]). *Let  $K \in \mathcal{T}_h$  and let  $F \in \mathcal{F}_h$  be a face of  $K$ . Then, for all  $u \in H^{\ell+1}(K)^{m_u}$  and  $0 \leq \ell \leq k$ , the projection error satisfies*

$$\|e_{\pi,u}\|_{L^2(F)} \leq C_{\pi,\partial} h^{\ell+1/2} |u|_{\ell+1,K}, \quad (\text{A.3})$$

where  $C_{\pi,\partial}$  is independent of the element  $K$  and the mesh size  $h$ .

Recall that the discrete operator domains are

$$V_h^{\mathcal{L}_w} = \mathbb{V}_h^w + \mathcal{D}(\mathcal{L}_w) \cap H^1(\mathcal{T}_h)^{m_w}, \quad w \in \{u, v\}. \quad (\text{A.4})$$

The weak formulation in Definition A.2 is motivated, as in [Stu17, §3.2], by elementwise integration by parts of  $(\mathcal{L}_{u,h}u, \phi_h)_\Omega$ , the jump-average identity

$$\llbracket u \cdot \psi \rrbracket_F = \llbracket u \rrbracket_F \cdot \{\!\!\{ \psi \}\!\!\}_F + \{\!\!\{ u \}\!\!\}_F \cdot \llbracket \psi \rrbracket_F, \quad (\text{A.5})$$

and [HK20, Lem. 4.7].

**Definition A.2** (Weak form). *We define  $\mathcal{L}_{u,h}: V_h^{\mathcal{L}^u} \rightarrow \mathbb{V}_h^v$  such that*

$$\begin{aligned} (\mathcal{L}_{u,h}u, \phi_h)_\Omega &= - \sum_{K \in \mathcal{T}_h} (u, \mathcal{L}_v \phi_h)_K + \sum_{F \in \mathcal{F}_h^{\text{int}}} (L_{\partial,u}^F \{\!\!\{ u \}\!\!\}_F, \llbracket \phi_h \rrbracket_F)_{L^2(F)} \\ &\quad + \frac{1}{2} \sum_{F \in \mathcal{F}_h^{\text{bnd}}} ((L_{\partial,u}^F + L_{\Gamma,u})u, \phi_h)_{L^2(F)} \quad \text{for all } \phi_h \in \mathbb{V}_h^v, \end{aligned} \quad (\text{A.6a})$$

and  $\mathcal{L}_{v,h}: V_h^{\mathcal{L}^v} \rightarrow \mathbb{V}_h^u$  such that

$$\begin{aligned} (\mathcal{L}_{v,h}v, \psi_h)_\Omega &= - \sum_{K \in \mathcal{T}_h} (v, \mathcal{L}_u \psi_h)_K + \sum_{F \in \mathcal{F}_h^{\text{int}}} (L_{\partial,v}^F \{\!\!\{ v \}\!\!\}_F, \llbracket \psi_h \rrbracket_F)_{L^2(F)} \\ &\quad + \frac{1}{2} \sum_{F \in \mathcal{F}_h^{\text{bnd}}} ((L_{\partial,v}^F + L_{\Gamma,v})v, \psi_h)_{L^2(F)} \quad \text{for all } \psi_h \in \mathbb{V}_h^u. \end{aligned} \quad (\text{A.6b})$$

The following strong form is equivalent. In [DHK<sup>+</sup>23, Def. 10.5], it is used as the definition.

**Lemma A.3** ([DHK<sup>+</sup>23, Def. 10.5]). *For  $u \in V_h^{\mathcal{L}^u}$  we have*

$$\begin{aligned} (\mathcal{L}_{u,h}u, \phi_h)_\Omega &= \sum_{K \in \mathcal{T}_h} (\mathcal{L}_u u, \phi_h)_K - \sum_{F \in \mathcal{F}_h^{\text{int}}} (L_{\partial,u}^F \llbracket u \rrbracket_F, \{\!\!\{ \phi_h \}\!\!\}_F)_{L^2(F)} \\ &\quad - \frac{1}{2} \sum_{F \in \mathcal{F}_h^{\text{bnd}}} ((L_{\partial,u}^F - L_{\Gamma,u})u, \phi_h)_{L^2(F)} \quad \text{for all } \phi_h \in \mathbb{V}_h^v, \end{aligned} \quad (\text{A.7a})$$

and for  $v \in V_h^{\mathcal{L}^v}$  we have

$$\begin{aligned} (\mathcal{L}_{v,h}v, \psi_h)_\Omega &= \sum_{K \in \mathcal{T}_h} (\mathcal{L}_v v, \psi_h)_K - \sum_{F \in \mathcal{F}_h^{\text{int}}} (L_{\partial,v}^F \llbracket v \rrbracket_F, \{\!\!\{ \psi_h \}\!\!\}_F)_{L^2(F)} \\ &\quad - \frac{1}{2} \sum_{F \in \mathcal{F}_h^{\text{bnd}}} ((L_{\partial,v}^F - L_{\Gamma,v})v, \psi_h)_{L^2(F)} \quad \text{for all } \psi_h \in \mathbb{V}_h^u. \end{aligned} \quad (\text{A.7b})$$

*Proof.* Integration by parts on each cell and (A.5) as well as [HK20, Lem. 4.7] yield the result.  $\square$

## CHAPTER 3

---

### Higher order local time integration for Friedrichs' systems

The content of this chapter was published as a preprint [\[HKS26\]](#).

In this paper, we construct an efficient higher-order local time integration scheme for spatially discretized linear Friedrichs' systems. In particular, our interest is in problems where only a few of the mesh elements are small while the majority of the elements is much larger. The special combination of two methods like the leapfrog method on the coarse part of the mesh and the Crank-Nicolson method on the fine part as was done in [\[HS16, HK22\]](#) is not suitable for higher-order time integration. We use implicit Gauß collocation methods, which provide higher temporal order but require solving large linear systems at each time step. Therefore, we suggest to approximate the solution of the linear systems arising in each time step by a preconditioned Krylov subspace method, e.g., the quasi-minimal residual method by Freund and Nachtigal [\[FN91\]](#). The techniques developed here for linear problems also carry over to nonlinear problems, where linear systems of the same type arise within a Newton-type iteration.

Motivated by the analysis of locally implicit methods by Hochbruck and Sturm [\[HS16\]](#), we show how to construct a preconditioner in such a way that the number of iterations required by the Krylov subspace method to achieve a certain accuracy is bounded independently of the diameter of the small mesh elements. We prove this behavior by using Faber polynomials and complex approximation theory.

The cost to apply the preconditioner consists of the solution of a small linear system, whose dimension corresponds to the degrees of freedom within the fine part of the mesh (and its next coarse neighbors). If this dimension is small compared to the size of the full mesh, the preconditioner is very efficient.

We conclude by verifying our theoretical results with numerical experiments for the fourth-order Gauß-Legendre Runge–Kutta method.

### 3.1. Introduction

Friedrichs' systems play a crucial role in the modeling of various physical phenomena. Important examples are advection equations, acoustic wave equations, Maxwell's equations, and Dirac equations. See [DHK<sup>+</sup>23] for more details. Though finite difference time-domain methods [Yee66] are still predominantly utilized to solve Maxwell's equations, numerous other methods based on finite element or finite volume space discretizations have been introduced and are gaining more and more importance.

The numerical solution of time-dependent partial differential equations by a method of lines approach involves first discretizing in space and then integrating the semi-discrete system in time. For the space discretization, discontinuous Galerkin (dG) methods (see [DPE12] and references therein) are popular due to their flexibility in treating complex geometries and discontinuous material parameters. Since dG methods lead to block-diagonal mass matrices, explicit time integration schemes can be implemented very efficiently. Unfortunately, explicit time integration methods are subject to the Courant–Friedrichs–Lewy (CFL) condition depending on the minimum diameter of all mesh elements, denoted by  $h_{\min}$ , that is, the time step  $\tau$  needs to satisfy  $\tau \lesssim h_{\min}$ . Here, we are interested in locally refined meshes, where most of the mesh elements are coarse but a very small number of mesh elements are fine. The latter require very small time steps on all mesh elements, which makes the computation inefficient. An alternative is to use implicit time integrators which can eliminate the CFL condition completely but require solving a large linear system of equations in all degrees of freedom at each time step. Unfortunately, this is expensive and might not even be feasible for large 3D problems.

To tackle this problem, locally implicit (LI) methods [Ver11, DLM13, HS16, CH16, DLM17, HS19] and local time-stepping (LTS) methods [Pip06, DG09, GM10, GMM15] or more general local time integrators (LTI) from our previous work [HS26] were introduced and studied. While there is a rigorous analysis of LTI methods of order two for linear problems, it is not clear how to prove the stability for higher-order LTI methods constructed via composition methods [HLW06, Section II.4].

In this paper, we introduce a new strategy to develop a higher-order time integration method to solve the general class of linear Friedrichs' systems possessing a two-field structure on a locally refined spatial grid in a computationally efficient way. For the time integration, we propose to use higher-order implicit Gauß collocation methods [HW96, Sec. IV.5]. It is well known that these methods are A-stable and symplectic, and have maximum order  $2s$  for nonstiff problems. Moreover, for all  $s$ , rigorous error bounds are available for a class of problems which includes Friedrichs' systems, cf. [HP15, Section 3.1]. This makes them the ideal candidates for our applications. Unfortunately, the analysis only holds for Runge-Kutta methods which are not only A-stable but also algebraically stable and coercive, properties which are not satisfied for general implicit methods such as the computationally attractive DIRK methods.

The drawback of implicit schemes is that one has to solve large linear systems of equations in each time step. For this, iterative solvers, especially (preconditioned) Krylov subspace methods, are popular choices. See [Saa03] and references therein. For instance, block preconditioners for structure-preserving discretizations of Maxwell's equations were proposed in [PSC18, AHZ17]. Other popular preconditioners include those based on multigrid [Hip99, Bañ10, AA02] and Schwarz domain decomposition [BDG<sup>+</sup>19, SCGT07]. However, these preconditioning techniques have not yet been adapted to locally refined meshes.

We show that the linear systems of equations which arise in the implementation of Gauß Runge-Kutta methods for linear Friedrichs' systems can be transformed using techniques from

[HW96, Sec. IV.5] in such a way that we obtain an equivalent linear system with a complex symmetric coefficient matrix. Thus, we use the quasi-minimal residual (QMR) method to solve it iteratively [Fre92, FN91, FN94]. Our main contribution is to construct a suitable preconditioner for the QMR method and to prove that the number of iterations required to reach a certain accuracy is independent of the fine mesh. Roughly speaking, this preconditioner only acts on the fine mesh and is thus very efficient if the number of fine mesh elements is small compared to the total degrees of freedom. Hence, the overall algorithm is unconditionally stable, obeys rigorous error bounds which do not deteriorate under mesh refinement, and has linear complexity with a constant that only depends on the coarse mesh.

The paper is organized as follows. In Section 3.2, we present our model problem and notation and recall properties of the spatial discretization of Friedrichs' systems using discontinuous Galerkin methods with central fluxes. Section 3.3 is dedicated to higher-order implicit Runge-Kutta methods. In Section 3.4, we recall known results on Krylov subspace methods, in particular error bounds. Our main contribution is to prove that the field of values of the preconditioned matrix is independent of the fine mesh and to combine this with error bounds for Krylov subspace methods. Finally, in Section 3.5, we verify our theoretical findings with numerical experiments and compare the efficiency of the higher-order locally implicit method with other local time integration schemes [HS16, HS26] and a low-storage Runge-Kutta method [NDB12].

## 3.2. Problem setting

Let  $\Omega \subset \mathbb{R}^d$ ,  $d = 1, 2, 3$ , be an open polygonal domain. The linear Friedrichs' system in a two field formulation reads

$$\partial_t u = \mathcal{L}_v v + g_u, \quad \text{on } \Omega \times \mathbb{R}_+, \quad (3.1a)$$

$$\partial_t v = \mathcal{L}_u u + g_v, \quad \text{on } \Omega \times \mathbb{R}_+, \quad (3.1b)$$

$$u(0) = u^0, \quad v(0) = v^0, \quad \text{on } \Omega, \quad (3.1c)$$

where  $\partial_t$  denotes the partial derivative with respect to time and  $g_u: \mathbb{R}_+ \rightarrow L^2(\Omega)^{m_u}$ ,  $g_v: \mathbb{R}_+ \rightarrow L^2(\Omega)^{m_v}$  are forcing terms. The Friedrichs' operators  $\mathcal{L}_u$  and  $\mathcal{L}_v$  are defined by

$$\mathcal{M}_v \mathcal{L}_u u = \sum_{i=1}^d L_i \partial_i u, \quad \mathcal{M}_u \mathcal{L}_v v = \sum_{i=1}^d L_i^T \partial_i v, \quad (3.2)$$

with coefficients  $L_i \in \mathbb{R}^{m_v \times m_u}$  for  $i = 0, \dots, d$ . Here  $\partial_i$  denotes the partial derivative with respect to the  $i$ th spatial variable. The material tensors  $\mathcal{M}_u \in L^\infty(\Omega)^{m_u \times m_u}$ ,  $\mathcal{M}_v \in L^\infty(\Omega)^{m_v \times m_v}$  are assumed to be symmetric and uniformly positive definite, and by  $m_u, m_v \in \mathbb{N}$  we count the components of  $u$  and  $v$ . To make the system (3.1) well-posed we have to prescribe boundary conditions. These are embodied into the domains  $\mathcal{D}(\mathcal{L}_u)$  and  $\mathcal{D}(\mathcal{L}_v)$ , see [DPE12, DHK<sup>+</sup>23] for details. We assume that the boundary conditions do not introduce damping, cf. [DHK<sup>+</sup>23, Assump. 11.8], which leads to the adjointness property

$$(\mathcal{L}_u u, v)_\Omega = -(u, \mathcal{L}_v v)_\Omega \quad \text{for all } u \in \mathcal{D}(\mathcal{L}_u), v \in \mathcal{D}(\mathcal{L}_v), \quad (3.3)$$

in the weighted  $L^2$  inner products defined via

$$(u, \tilde{u})_\Omega = (\mathcal{M}_u u, \tilde{u})_{L^2(\Omega)}, \quad u, \tilde{u} \in L^2(\Omega)^{m_u}. \quad (3.4)$$

The corresponding norm is denoted by  $\|\cdot\|_\Omega$ . The inner product weighted by  $\mathcal{M}_v$  is defined analogously, and it is always clear from the inserted argument which inner product or norm is meant.

For a full discretization of (3.1), we first discretize in space using a dG method with central fluxes on a suitable mesh  $\mathcal{T}_h$  [DPE12],[HK20, Section 4]. On this mesh we define the spaces

$$\mathbb{V}_h^u = \{u \in L^2(\Omega) \mid u|_K \in \mathbb{P}_d^k(K) \text{ for all } K \in \mathcal{T}_h\}^{m_u}, \quad (3.5a)$$

$$\mathbb{V}_h^v = \{v \in L^2(\Omega) \mid v|_K \in \mathbb{P}_d^k(K) \text{ for all } K \in \mathcal{T}_h\}^{m_v}, \quad (3.5b)$$

consisting of functions in  $L^2(\Omega)$  which are polynomials of total degree at most  $k$  on each mesh element. If it is clear by context which space we mean, we simply write  $\mathbb{V}_h$ . The dG method then yields

$$\partial_t u_h = \mathcal{L}_{v,h} v_h + g_{u,h}, \quad \text{on } \mathbb{R}_+, \quad (3.6a)$$

$$\partial_t v_h = \mathcal{L}_{u,h} u_h + g_{v,h}, \quad \text{on } \mathbb{R}_+, \quad (3.6b)$$

$$u_h(0) = u_h^0, \quad v_h(0) = v_h^0, \quad (3.6c)$$

where  $\mathcal{L}_{u,h}$  and  $\mathcal{L}_{v,h}$  are spatially discretized Friedrichs' operators, and  $u_h^0, v_h^0, g_{u,h}$ , and  $g_{v,h}$  are  $L^2$  projections of  $u^0, v^0, g_u$ , and  $g_v$ , respectively, onto  $\mathbb{V}_h^u, \mathbb{V}_h^v$  with respect to the weighted inner products defined in (3.4). We refer to [DPE12, HK20] for details on the dG discretization.

For  $w \in \{u, v\}$ , let  $\{\phi_1^w, \dots, \phi_{N_w}^w\}$  be a basis of  $\mathbb{V}_h^w$ . Then an element  $x_h: (0, T) \rightarrow \mathbb{V}_h^w$  can be represented as

$$x_h(t) = \sum_{j=1}^N x_j(t) \phi_j^w,$$

with coefficient vector  $\mathbf{x}(t) = (x_j(t))_{j=1}^{N_w}$ . This results in mass and stiffness matrices given by

$$(\mathbf{M}_u)_{l,j} = (\phi_j^u, \phi_l^u)_\Omega, \quad (\tilde{\mathbf{L}}_u)_{l,j} = (\mathcal{L}_{u,h} \phi_j^u, \phi_l^v)_\Omega, \quad (3.7a)$$

$$(\mathbf{M}_v)_{l,j} = (\phi_j^v, \phi_l^v)_\Omega, \quad (\tilde{\mathbf{L}}_v)_{l,j} = (\mathcal{L}_{v,h} \phi_j^v, \phi_l^u)_\Omega. \quad (3.7b)$$

Then, (3.6) is equivalent to the following system of ordinary differential equations,

$$\begin{aligned} \partial_t \mathbf{u} &= \mathbf{L}_v \mathbf{v} + \mathbf{g}_u, & \mathbf{L}_v &= \mathbf{M}_u^{-1} \tilde{\mathbf{L}}_v, \\ \partial_t \mathbf{v} &= \mathbf{L}_u \mathbf{u} + \mathbf{g}_v, & \mathbf{L}_u &= \mathbf{M}_v^{-1} \tilde{\mathbf{L}}_u, \\ \mathbf{u}(0) &= \mathbf{u}^0, & \mathbf{v}(0) &= \mathbf{v}^0. \end{aligned} \quad (3.8)$$

Here,  $\mathbf{u}^0$  and  $\mathbf{v}^0$  are the coefficient vectors of  $u_h^0$  and  $v_h^0$ , respectively.

With an abuse of notation, given  $x_h, y_h \in \mathbb{V}_h^w$ ,  $w \in \{u, v\}$ , with coefficient vectors  $\mathbf{x}, \mathbf{y} \in \mathbb{C}^{N_w}$ , we define

$$(\mathbf{x}, \mathbf{y})_\Omega := \mathbf{y}^* \mathbf{M}_w \mathbf{x} = (x_h, y_h)_\Omega, \quad (3.9)$$

and do so analogously for the induced norms in  $\mathbb{C}^{N_w}$ . Here,  $*$  denotes the conjugate transpose. For the matrix norms, we also take these weights into account since then these norms are equivalent to the operator norms of the discrete operators  $\mathcal{L}_{u,h}$  and  $\mathcal{L}_{v,h}$ , i.e., we have

$$\|\mathbf{L}_u\|_{\mathbb{V}_h^v \leftarrow \mathbb{V}_h^u} = \sup_{\mathbf{x} \in \mathbb{C}^{N_u} \setminus \{0\}} \frac{\|\mathbf{L}_u \mathbf{x}\|_\Omega}{\|\mathbf{x}\|_\Omega} = \sup_{x_h \in \mathbb{V}_h^u \setminus \{0\}} \frac{\|\mathcal{L}_{u,h} x_h\|_\Omega}{\|x_h\|_\Omega} = \|\mathcal{L}_{u,h}\|_{\mathbb{V}_h^v \leftarrow \mathbb{V}_h^u}, \quad (3.10a)$$

$$\|\mathbf{L}_v\|_{\mathbb{V}_h^u \leftarrow \mathbb{V}_h^v} = \sup_{\mathbf{x} \in \mathbb{C}^{N_v} \setminus \{0\}} \frac{\|\mathbf{L}_v \mathbf{x}\|_\Omega}{\|\mathbf{x}\|_\Omega} = \sup_{x_h \in \mathbb{V}_h^v \setminus \{0\}} \frac{\|\mathcal{L}_{v,h} x_h\|_\Omega}{\|x_h\|_\Omega} = \|\mathcal{L}_{v,h}\|_{\mathbb{V}_h^u \leftarrow \mathbb{V}_h^v}. \quad (3.10b)$$

In this paper, we are interested in locally refined meshes. We refer to our earlier papers [HS16, HS19, HS26] for detailed explanations on these meshes, but to keep this paper self-contained, we introduce the necessary notation here. A locally refined mesh is a mesh in which most of the mesh elements are coarse and very few mesh elements are fine. Let  $\mathcal{T}_{h,c}$  and  $\mathcal{T}_{h,f}$  denote the collection of all coarse and fine mesh elements, respectively. We denote by  $h_f$  and  $h_c$  the size of smallest mesh elements in  $\mathcal{T}_{h,f}$  and in  $\mathcal{T}_{h,c}$ , respectively. These two sets are related to each other via

$$h_f \ll h_c \quad \text{and} \quad \text{card}(\mathcal{T}_{h,f}) \ll \text{card}(\mathcal{T}_{h,c}).$$

Based on this decomposition of the mesh, the matrices defined in (3.8) can be split into

$$\mathbf{L}_u = \mathbf{L}_u^f + \mathbf{L}_u^c, \quad \mathbf{L}_v = \mathbf{L}_v^f + \mathbf{L}_v^c, \quad (3.11)$$

cf. [HS16, HS26] for more details. The indices  $f$  and  $c$  indicate that the elements on which  $\mathbf{L}_u^f, \mathbf{L}_v^f$  act belong to the fine submesh  $\mathcal{T}_{h,f}$  and the ones on which  $\mathbf{L}_u^c, \mathbf{L}_v^c$  act, belong to the coarse submesh  $\mathcal{T}_{h,c}$ . In fact, it was shown in [HS16] that not only the fine elements have to be considered but also their direct coarse neighbors.

Let us state some properties of these matrices which are inherited from their corresponding discrete operators, cf. [HS16]. First,  $\mathbf{L}_u$  and  $\mathbf{L}_v$  are adjoint to each other, i.e., for all  $\mathbf{u} \in \mathbb{C}^{N_u}, \mathbf{v} \in \mathbb{C}^{N_v}$ , it holds

$$(\mathbf{L}_u \mathbf{u}, \mathbf{v})_\Omega = -(\mathbf{u}, \mathbf{L}_v \mathbf{v})_\Omega. \quad (3.12)$$

It is easy to verify that these split matrices preserve the adjointness property of their respective full ones, i.e.,

$$(\mathbf{L}_u^c \mathbf{u}, \mathbf{v})_\Omega = -(\mathbf{u}, \mathbf{L}_v^c \mathbf{v})_\Omega, \quad (\mathbf{L}_u^f \mathbf{u}, \mathbf{v})_\Omega = -(\mathbf{u}, \mathbf{L}_v^f \mathbf{v})_\Omega. \quad (3.13)$$

In addition, they satisfy

$$\mathbf{L}_v^c \mathbf{L}_u^c = \mathbf{L}_v^c \mathbf{L}_u, \quad \mathbf{L}_v^f \mathbf{L}_u^f = \mathbf{L}_v^f \mathbf{L}_u. \quad (3.14)$$

Furthermore, combining the above properties, it holds

$$\|\mathbf{L}_u \mathbf{u}\|_\Omega^2 = \|\mathbf{L}_u^c \mathbf{u}\|_\Omega^2 + \|\mathbf{L}_u^f \mathbf{u}\|_\Omega^2. \quad (3.15)$$

One of the important results from [HS16] is that the explicit split matrices  $\mathbf{L}_u^c$  and  $\mathbf{L}_v^c$  can be bounded independently of the fine mesh, that is, using the weighted norms in (3.10), we have

$$\|\mathbf{L}_u^c\|_{\mathbb{V}_h^v \leftarrow \mathbb{V}_h^u} \leq ch_c^{-1}, \quad \|\mathbf{L}_v^c\|_{\mathbb{V}_h^u \leftarrow \mathbb{V}_h^v} \leq ch_c^{-1}, \quad (3.16)$$

with a constant  $c$  that is independent of  $h_f$  and  $h_c$ .

In [HS16, HS19], these split matrices were constructed to develop local time integration methods. Here, we use these split matrices in a different way, namely to construct preconditioners which improve the performance of Krylov subspace methods.

### 3.3. Higher-order implicit Runge–Kutta methods

In this section, we consider the time integration of (3.8) by an  $s$ -stage implicit Runge–Kutta (RK) methods given by its matrix  $\mathcal{Q} = (a_{ij})_{i,j=1}^s$ , weights  $b_i$  and nodes  $c_i$ ,  $i = 1, \dots, s$ , cf., [HLW06, Section II.1].

The two-field linear Friedrichs' system (3.1) fits into the general framework considered in [HP15, Eq. (2.8)] since the Friedrichs' operator  $\mathcal{L}$  is skew-adjoint on its domain  $\mathcal{D}(\mathcal{L})$ . Therefore, the time discretization error bound [HP15, Thm. 3.5] holds under certain regularity assumptions on the exact solution of (3.1). Regarding the full discretization error using a central fluxes discontinuous Galerkin discretization in space, we refer to [HP15, Thm 5.4 and Rem 5.7]. Even though these results are derived for linear Maxwell's equations, it is straightforward to extend them to two-field linear Friedrichs' systems discretized in space.

To simplify the presentation, we write (3.8) in the compact form

$$\begin{aligned}\partial_t \mathbf{x} &= \mathbf{L}\mathbf{x} + \mathbf{g}, & \text{on } (0, T), \\ \mathbf{x}^0 &= \mathbf{x}(0),\end{aligned}\tag{3.17}$$

where

$$\mathbf{x} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}, \quad \mathbf{g} = \begin{pmatrix} \mathbf{g}_u \\ \mathbf{g}_v \end{pmatrix} \in \mathbb{R}^N, \quad \text{and} \quad \mathbf{L} = \begin{pmatrix} \mathbf{0} & \mathbf{L}_v \\ \mathbf{L}_u & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{N \times N},$$

with  $N := N_u + N_v$ . Assume that we already computed an approximation  $\mathbf{x}^n \approx \mathbf{x}(t^n)$  at time  $t^n = n\tau$ , where  $\tau > 0$  denotes the step size. Then, the implicit Runge–Kutta method applied to (3.17) leads to the following coupled linear system of equations for the intermediate stages  $\mathbf{X}_i \approx \mathbf{x}(t^n + c_i\tau)$

$$\mathbf{X}_i = \mathbf{x}^n + \tau \sum_{j=1}^s a_{ij} (\mathbf{L}\mathbf{X}_j + \mathbf{G}_j), \quad i = 1, \dots, s,\tag{3.18a}$$

where  $\mathbf{G}_j = \mathbf{g}(t^n + c_j\tau)$ . The new approximation  $\mathbf{x}^{n+1} \approx \mathbf{x}(t^{n+1})$  is then given explicitly by

$$\mathbf{x}^{n+1} = \mathbf{x}^n + \tau \sum_{i=1}^s b_i (\mathbf{L}\mathbf{X}_i + \mathbf{G}_i).\tag{3.18b}$$

### 3.3.1. Gauß collocation methods

We use Gauß collocation methods to construct higher-order implicit RK methods. It is well known that these methods are algebraically stable [HW96, Theorem IV.12.9] and the RK matrix  $\mathcal{Q}$  is invertible [HW96, Section IV.14]. In addition, the error analysis for linear wave-type problems [HP15, Section 3.1] makes use of the existence of a diagonal positive definite matrix  $\hat{\mathbf{D}}$  and a positive scalar  $\eta > 0$  such that

$$\mathbf{v}^\top \hat{\mathbf{D}} \mathcal{Q}^{-1} \mathbf{v} \geq \eta \mathbf{v}^\top \hat{\mathbf{D}} \mathbf{v}, \quad \text{for all } \mathbf{v} \in \mathbb{R}^s.\tag{3.19}$$

Here,  $^\top$  denotes the transpose. For Gauß collocation methods, the coercitivity condition (3.19) is satisfied for  $\hat{\mathbf{D}} = \hat{\mathbf{B}}(\hat{\mathbf{C}}^{-1} - \mathbf{I}_s)$ , where  $\hat{\mathbf{B}} := \text{diag}(b_1, \dots, b_s)$ ,  $\hat{\mathbf{C}} := \text{diag}(c_1, \dots, c_s)$ , and  $\mathbf{I}_s$  is the identity matrix of size  $s$ , cf. [HW96, Theorem IV.14.5].

To obtain an efficient implementation for solving (3.18a), we use Kronecker products [HLW06, Section VIII.6] to rewrite it as

$$\mathbf{X} = \mathbf{1}_s \otimes \mathbf{x}^n + \tau ((\mathcal{Q} \otimes \mathbf{L})\mathbf{X} + (\mathcal{Q} \otimes \mathbf{I}_N)\mathbf{G}),\tag{3.20}$$

where  $\mathbf{X} = (\mathbf{X}_i)_{i=1}^s$ ,  $\mathbf{G} = (\mathbf{G}_i)_{i=1}^s \in \mathbb{C}^{Ns}$ ,  $\mathbf{I}_N$  is the identity matrix of size  $N$ , and the term  $\mathbf{1}_s$  denotes the vector in  $\mathbb{R}^s$  consisting of all ones. Diagonalization of  $\mathcal{Q}$  yields a nonsingular

matrix  $\mathbf{T} \in \mathbb{C}^{s \times s}$  containing the eigenvectors and a diagonal matrix  $\mathbf{\Lambda}_Q \in \mathbb{C}^{s \times s}$  containing eigenvalues  $\lambda_i$ , such that

$$\mathbf{T}^{-1} \mathbf{Q} \mathbf{T} = \mathbf{\Lambda}_Q, \quad \mathbf{\Lambda}_Q = \text{diag}(\lambda_1, \dots, \lambda_s). \quad (3.21)$$

Substituting  $\mathbf{Q} = \mathbf{T} \mathbf{\Lambda}_Q \mathbf{T}^{-1}$  into (3.20) and performing some Kronecker product operations leads to  $s$  decoupled linear systems of the form

$$(\mathbf{I}_s \otimes \mathbf{I}_N - \tau(\mathbf{\Lambda}_Q \otimes \mathbf{L})) \mathbf{Z} = \mathbf{Z}^0 + \tau(\mathbf{\Lambda}_Q \otimes \mathbf{I}_N) \tilde{\mathbf{G}}, \quad (3.22)$$

where

$$\mathbf{Z} = (\mathbf{T}^{-1} \otimes \mathbf{I}_N) \mathbf{X}, \quad \mathbf{Z}^0 = (\mathbf{T}^{-1} \otimes \mathbf{I}_N)(\mathbb{1}_s \otimes \mathbf{x}^n), \quad \tilde{\mathbf{G}} = (\mathbf{T}^{-1} \otimes \mathbf{I}_N) \mathbf{G}. \quad (3.23)$$

Note that  $\mathbf{Q}$  might have complex conjugate pairs of eigenvalues. For such eigenvalues (say  $\lambda_j = \overline{\lambda_i}$ ), the corresponding linear systems are

$$(\mathbf{I}_N - \tau \lambda_i \mathbf{L}) \mathbf{Z}_i = \mathbf{Z}_i^0 + \tau \lambda_i \tilde{\mathbf{G}}_i, \quad (3.24a)$$

$$(\mathbf{I}_N - \tau \overline{\lambda_i} \mathbf{L}) \mathbf{Z}_j = \mathbf{Z}_j^0 + \tau \overline{\lambda_i} \tilde{\mathbf{G}}_j. \quad (3.24b)$$

In the homogeneous case, i.e.,  $\mathbf{g} \equiv 0$  which leads to  $\tilde{\mathbf{G}} \equiv 0$ , the first term on the right-hand sides of (3.24a) and (3.24b) are conjugate to each other and so are the solutions.

**Lemma 3.1.** *If  $\mathbf{g} \equiv 0$  and  $\lambda_j = \overline{\lambda_i}$ , then the solutions of (3.24) satisfy  $\mathbf{Z}_j = \overline{\mathbf{Z}_i}$ .*

*Proof.* The RK matrix  $\mathbf{Q}$  is real and thus complex eigenvalues and eigenvectors appear in complex conjugate pairs. Hence, there exists a symmetric permutation matrix  $\hat{\mathbf{P}} \in \mathbb{R}^{s \times s}$  s.t.,

$$\overline{\mathbf{T}} = \mathbf{T} \hat{\mathbf{P}}, \quad \overline{\mathbf{\Lambda}_Q} = \hat{\mathbf{P}} \mathbf{\Lambda}_Q \hat{\mathbf{P}}, \quad (3.25)$$

which implies  $\mathbf{T}^{-1} = \hat{\mathbf{P}} \overline{\mathbf{T}}^{-1}$ . We choose an arbitrary index  $i \in \{0, \dots, s\}$  corresponding to a complex eigenvalue  $\lambda_i \notin \mathbb{R}$  and define the index  $j$  such that  $\mathbf{e}_j = \hat{\mathbf{P}} \mathbf{e}_i$ . By (3.23) and  $\mathbf{x}^n \in \mathbb{R}^N$ , we have

$$\overline{\mathbf{Z}}_i^0 = (\mathbf{e}_i^\top \otimes \mathbf{I}_N) \overline{\mathbf{Z}}^0 = (\mathbf{e}_i^\top \overline{\mathbf{T}}^{-1} \mathbb{1}_s \otimes \mathbf{x}^n) = ((\hat{\mathbf{P}} \mathbf{e}_i)^\top \mathbf{T}^{-1} \mathbb{1}_s) \otimes \mathbf{x}^n = \mathbf{Z}_j^0. \quad (3.26)$$

Conjugating (3.24a) proves that  $\overline{\mathbf{Z}}_i$  solves (3.24b).  $\square$

In addition to this,  $\mathbf{Z}_i$  and  $\tilde{\mathbf{G}}_i$  in (3.24a) can be further decomposed into

$$\mathbf{Z}_i = \begin{pmatrix} \mathbf{Z}_{u,i} \\ \mathbf{Z}_{v,i} \end{pmatrix}, \quad \tilde{\mathbf{G}}_i = \begin{pmatrix} \tilde{\mathbf{G}}_{u,i} \\ \tilde{\mathbf{G}}_{v,i} \end{pmatrix},$$

where  $\mathbf{Z}_{u,i}, \mathbf{Z}_{v,i}$  denote unknowns corresponding to the transformed intermediate stages of  $\mathbf{u}$  and  $\mathbf{v}$ , respectively. Taking the Schur complement, the linear systems in (3.24a) can be reduced to

$$(\mathbf{I}_{N_u} - \alpha_i \mathbf{L}_v \mathbf{L}_u) \mathbf{Z}_{u,i} = \mathbf{Z}_{u,i}^0 + \tau \lambda_i (\mathbf{L}_v \mathbf{Z}_{v,i}^0 + \tilde{\mathbf{G}}_{u,i}) + \alpha_i \mathbf{L}_v \tilde{\mathbf{G}}_{v,i}, \quad \alpha_i := \tau^2 \lambda_i^2 \in \mathbb{C}, \quad (3.27)$$

to compute the  $u$ -component of  $\mathbf{Z}_i$ . After solving this linear system of dimension  $N_u$ , the  $v$ -component of  $\mathbf{Z}_i$  can be calculated explicitly via

$$\mathbf{Z}_{v,i} = \mathbf{Z}_{v,i}^0 + \tau \lambda_i (\mathbf{L}_u \mathbf{Z}_{u,i} + \tilde{\mathbf{G}}_{v,i}). \quad (3.28)$$

An efficient implementation of an  $s$ -stage implicit Runge-Kutta method using Gauß collocation points thus requires solving a linear system of the form

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad \text{where} \quad \mathbf{A} := \mathbf{I}_{N_u} - \alpha \mathbf{L}_v \mathbf{L}_u, \quad (3.29)$$

with a possibly complex parameter  $\alpha$  in each time step. Therefore, the size of the linear system to be solved is reduced from  $s(N_u + N_v)$  in (3.20) to  $N_u$  in (3.29). The adjointness property (3.12) implies that  $\mathbf{L}_v \mathbf{L}_u$  is symmetric with respect to the discrete inner product  $(\cdot, \cdot)_\Omega$  defined in (3.9). Hence,  $\mathbf{A}$  is complex symmetric, that is,

$$(\mathbf{A}\mathbf{x}, \mathbf{x})_\Omega = (\mathbf{x}, \overline{\mathbf{A}}\mathbf{x})_\Omega, \quad \mathbf{x} \in \mathbb{C}^{N_u},$$

and also normal. If  $\alpha \in \mathbb{R}$ , then  $\mathbf{A} \in \mathbb{R}^{N \times N}$  is symmetric. However, for  $\alpha \notin \mathbb{R}$ , it follows immediately that  $\mathbf{A} \neq \mathbf{A}^*$  with respect to  $(\cdot, \cdot)_\Omega$ . Moreover, for

$$\alpha \in \mathbb{C} \setminus \{z \in \mathbb{R} : z < 0\}, \quad (3.30)$$

the matrix  $\mathbf{A}$  is invertible. For Gauß collocation methods, the coercivity condition (3.19) guarantees that the eigenvalues of  $\mathcal{Q}$  are not purely imaginary, and hence (3.30) is satisfied.

### 3.4. Preconditioned Krylov subspace methods

In this section, we aim at designing a tailored preconditioner for solving the sparse linear system (3.29) by a preconditioned Krylov subspace method. We will prove that the number of Krylov iterations to achieve a certain tolerance is independent of the fine mesh. The overall method can be considered as a locally implicit scheme because it only requires the solution of a small linear system as it is required for the second-order method in [HS16].

We remark that in Section 3.4.1, we consider the Euclidean inner product and norm in  $\mathbb{C}^N$ , but using weighted inner products would be possible as well.

#### 3.4.1. Krylov subspace methods for complex symmetric matrices

For a nonsingular, complex symmetric matrix  $\mathbf{K} = \mathbf{K}^\top \in \mathbb{C}^{N \times N}$  and a given vector  $\mathbf{f} \in \mathbb{C}^N$ , we consider the linear system

$$\mathbf{K}\mathbf{x} = \mathbf{f}. \quad (3.31)$$

Given an initial guess  $\mathbf{x}_0 \in \mathbb{C}^N$  and its initial residual vector  $\mathbf{r}_0 = \mathbf{f} - \mathbf{K}\mathbf{x}_0$ , a Krylov subspace method yields an approximation of the form

$$\mathbf{x}_m = \mathbf{x}_0 + \mathbf{W}_m \mathbf{y}_m, \quad m = 1, 2, \dots, \quad (3.32)$$

where  $\mathbf{W}_m \in \mathbb{C}^{N \times m}$  is a basis of the  $m$ th Krylov subspace

$$\mathcal{K}_m(\mathbf{K}, \mathbf{r}_0) := \text{span}(\mathbf{r}_0, \mathbf{K}\mathbf{r}_0, \dots, \mathbf{K}^{m-1}\mathbf{r}_0),$$

and  $\mathbf{y}_m \in \mathbb{C}^m$  is a suitable coefficient vector. The choices of  $\mathbf{W}_m$  and  $\mathbf{y}_m$  characterize the Krylov subspace method, cf. [FGN92, Gut97, Saa03] for more details.

To exploit the complex symmetric structure of  $\mathbf{K}$ , we suggest to use the quasi-minimal residual (QMR) algorithm for complex symmetric matrices [Fre92, Section 3], which is based on the complex symmetric Lanczos process. Here,  $\mathbf{W}_m$  satisfies

$$\mathbf{K}\mathbf{W}_m = \mathbf{W}_{m+1} \tilde{\mathbf{H}}_m, \quad \mathbf{D}_{m+1} \tilde{\mathbf{H}}_m = \mathbf{W}_{m+1}^\top \mathbf{K} \mathbf{W}_m, \quad (3.33)$$

with a diagonal matrix  $\mathbf{D}_{m+1} = \mathbf{W}_{m+1}^\top \mathbf{W}_{m+1} \in \mathbb{C}^{(m+1) \times (m+1)}$ . The complex symmetry of  $\mathbf{K}$  implies that  $\tilde{\mathbf{H}}_m \in \mathbb{C}^{(m+1) \times m}$  is tridiagonal and the upper  $m \times m$  submatrix of  $\mathbf{D}_{m+1} \tilde{\mathbf{H}}_m$  is again complex symmetric.  $\tilde{\mathbf{H}}_m$  has full column rank  $m$  until  $\mathcal{K}_m(\mathbf{K}, \mathbf{r}_0)$  becomes a  $\mathbf{K}$ -invariant subspace.

With  $\beta = \|\mathbf{r}_0\|$ , the QMR approximation is defined as

$$\mathbf{x}_m = \mathbf{x}_0 + \mathbf{W}_m \mathbf{y}_m, \quad \mathbf{y}_m = \beta \tilde{\mathbf{H}}_m^+ \mathbf{e}_1, \quad \tilde{\mathbf{H}}_m^+ = (\tilde{\mathbf{H}}_m^* \tilde{\mathbf{H}}_m)^{-1} \tilde{\mathbf{H}}_m^*, \quad (3.34)$$

where  $\mathbf{e}_1$  denotes the first canonical unit vector. The corresponding residual can be written as

$$\mathbf{r}_m = \mathbf{f} - \mathbf{K} \mathbf{x}_m = \mathbf{W}_{m+1} (\beta \mathbf{e}_1 - \tilde{\mathbf{H}}_m \mathbf{y}_m).$$

Since  $\tilde{\mathbf{H}}_m$  has full column rank,  $\mathbf{y}_m$  is equivalently characterized as the unique solution of the least-squares problem

$$\mathbf{y}_m = \operatorname{argmin}_{\mathbf{y} \in \mathbb{C}^m} \|\beta \mathbf{e}_1 - \tilde{\mathbf{H}}_m \mathbf{y}\|.$$

The advantage of this algorithm compared to methods based on the Arnoldi process (e.g., GMRES) is that it uses three-term recurrences for the computation of the basis as well as for the approximation. It can be combined with look-ahead strategies [FGN93] to prevent breakdowns of the Lanczos process, which might appear because it constructs a basis which is orthogonal w.r.t. the indefinite bilinear form  $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$ , instead of the Euclidean inner product  $(\mathbf{x}, \mathbf{y}) = \mathbf{x}^* \mathbf{y}$ , see [Fre92, Section 4]. For the sake of presentation, we assume that breakdowns do not appear until a sufficiently accurate solution is computed, but we note that with minor modifications, our analysis also holds for the (complex symmetric) look-ahead Lanczos method [FGN93]. This assumption ensures that

$$\|\mathbf{D}_{m+1}^{-1}\| \leq \delta, \quad (3.35)$$

for a given (small) tolerance  $\delta > 0$ , because otherwise, one would switch to the look-ahead version of the Lanczos process.

In the following,  $\mathbb{P}_m$  denotes the set of all polynomials over  $\mathbb{C}$  of degree at most  $m$ .

**Theorem 3.2.** *Let  $\mathbf{K}$  be a nonsingular, complex symmetric matrix and  $\mathbf{x}_m$  be the QMR approximation (3.34) after  $m$  steps. Then the error of the QMR method satisfies*

$$\|\mathbf{K}^{-1} \mathbf{f} - \mathbf{x}_m\| \leq \|\mathbf{K}^{-1} \mathbf{P}_m\| \min_{\substack{p_m \in \mathbb{P}_m \\ p_m(0)=1}} \|p_m(\mathbf{K}) \mathbf{r}_0\| \quad (3.36)$$

with a projection matrix  $\mathbf{P}_m$  given by

$$\mathbf{P}_m = \mathbf{I}_{N_u} - \mathbf{W}_{m+1} \tilde{\mathbf{H}}_m \tilde{\mathbf{H}}_m^+ \mathbf{D}_{m+1}^{-1} \mathbf{W}_{m+1}^\top.$$

Moreover, if  $\|\mathbf{W}_{m+1} \mathbf{e}_j\| = 1$ ,  $j = 1, \dots, m+1$ , and (3.35) holds, we have

$$\|\mathbf{P}_m\| \leq 1 + (m+1)\delta. \quad (3.37)$$

*Proof.* Analogously to the proof of [HL98, Theorem 2], it can be seen from (3.33) that  $\mathbf{P}_m \mathbf{K} \mathbf{W}_m = 0$ . Using (3.34), this implies

$$\mathbf{K}^{-1} \mathbf{f} - \mathbf{x}_m = \mathbf{K}^{-1} \mathbf{P}_m \mathbf{r}_0 = \mathbf{K}^{-1} \mathbf{P}_m p_m(\mathbf{K}) \mathbf{r}_0$$

for all  $p_m \in \mathbb{P}_m$  with  $p_m(0) = 1$ .

The bound on  $\|\mathbf{P}_m\|$  follows from (3.35) and  $\|\mathbf{W}_m\| \leq m^{1/2}$ .  $\square$

Since  $\|p_m(\mathbf{K})\mathbf{r}_0\| \leq \|p_m(\mathbf{K})\| \|\mathbf{r}_0\|$ , it remains to bound

$$\min_{\substack{p_m \in \mathcal{P}_m \\ p_m(0)=1}} \|p_m(\mathbf{K})\|.$$

We use a bound from complex approximation theory, based on a superset  $\mathcal{S}$  of the field of values of  $\mathbf{K}$  [Eie89], defined as

$$\mathcal{F}(\mathbf{K}) := \{\rho_{\mathbf{K}}(\mathbf{u}), \mathbf{u} \in \mathbb{C}^{N_u}, \mathbf{u} \neq \mathbf{0}\}, \quad \rho_{\mathbf{K}}(\mathbf{u}) := \frac{(\mathbf{u}, \mathbf{K}\mathbf{u})}{(\mathbf{u}, \mathbf{u})}.$$

**Theorem 3.3.** *Let  $\mathcal{S} \subset \mathbb{C}$  be a convex and bounded superset of  $\mathcal{F}(\mathbf{K})$  with  $0 \notin \mathcal{S}$  and let  $\phi$  be the conformal map which maps the exterior of  $\mathcal{S}$  onto the exterior of the unit disc with  $\phi(\infty) = \infty$ . Then*

$$\min_{\substack{p_m \in \mathcal{P}_m \\ p_m(0)=1}} \|p_m(\mathbf{K})\| \leq (1 + 2^{1/2}) \min \left\{ \frac{3}{|\phi(0)|^m}, \frac{2}{|\phi(0)|^m - 1} \right\}. \quad (3.38)$$

*Proof.* It was shown in [CP17] that

$$\|p_m(\mathbf{K})\| \leq (1 + 2^{1/2}) \max_{z \in \mathcal{S}} |p_m(z)|.$$

The statement then follows from [HL97, Eq. (2.14)] and [Eie89, Theorem 2].  $\square$

The conformal map  $\phi$  can be determined numerically by using the Schwarz-Christoffel toolbox [Dri05].

### 3.4.2. Preconditioning for locally refined grids

Our aim and the content of this section is the construction of a preconditioner such that the field of values of the preconditioned matrix with respect to the discrete weighted inner product  $(\cdot, \cdot)_{\Omega}$  defined in (3.9) is independent of the fine mesh elements. Then, by Theorem 3.3, the same holds for the error of the preconditioned Krylov method in this weighted inner product.

Motivated by locally implicit methods for Maxwell's equations in [HS16, Ver11], we suggest to precondition  $\mathbf{A}$  from (3.29) with its dominant part,

$$\mathbf{A} \approx \mathbf{B} := \mathbf{I}_{N_u} - \gamma \mathbf{L}_v^f \mathbf{L}_u^f, \quad (3.39)$$

where  $\gamma > 0$  is a suitably chosen parameter. Note that this basically boils down to replacing the Friedrichs' matrices  $\mathbf{L}_u, \mathbf{L}_v$  in (3.29) defined on the full mesh by the split matrices acting on the implicitly treated mesh elements, cf. Section 3.2 and by replacing the possibly complex factor  $\alpha$  by  $\gamma > 0$ . By (3.13) and  $\gamma > 0$ , the preconditioner  $\mathbf{B}$  is symmetric and positive definite with respect to  $(\cdot, \cdot)_{\Omega}$ , and thus it has a symmetric and positive definite square root  $\mathbf{B}^{1/2}$ . This allows us to define an equivalent preconditioned linear system

$$\tilde{\mathbf{A}} \tilde{\mathbf{x}} = \tilde{\mathbf{b}}, \quad (3.40a)$$

where

$$\tilde{\mathbf{A}} := \mathbf{B}^{-1/2} \mathbf{A} \mathbf{B}^{-1/2}, \quad \tilde{\mathbf{x}} := \mathbf{B}^{1/2} \mathbf{x} \quad \text{and} \quad \tilde{\mathbf{b}} := \mathbf{B}^{-1/2} \mathbf{b}. \quad (3.40b)$$

Since  $\mathbf{A}$  is complex symmetric and  $\mathbf{B}$  is real symmetric, the preconditioned matrix  $\tilde{\mathbf{A}}$  is again complex symmetric (with respect to  $(\cdot, \cdot)_\Omega$ ).

We now apply the QMR method for complex symmetric matrices to the preconditioned linear system (3.40) and refer to this method as the preconditioned QMR (pQMR) method, cf. [FN94, Alg. 8.1.]. Since the computation of  $\mathbf{B}^{1/2}$  is usually very expensive, it is essential that it is not required for the implementation. In fact, one only has to solve linear systems with  $\mathbf{B}$  but neither compute  $\mathbf{B}^{1/2}$  nor  $\mathbf{B}^{-1/2}$ . Solving linear systems with  $\mathbf{B}$  does not lead to too much overhead costs because  $\mathbf{L}_v^f \mathbf{L}_u^f$  only acts on the fine elements and their direct neighbors and thus is of small dimension compared to  $\mathbf{A}$ . Moreover, systems with  $\mathbf{B}$  can also be solved iteratively, e.g., using standard preconditioning techniques for symmetric positive definite matrices.

It remains to show that the error of pQMR can indeed be bounded independently of the fine mesh. Note that Theorems 3.2 and 3.3 also hold for  $\|\cdot\| = \|\cdot\|_\Omega$  if the Lanczos process and the field of values are defined w.r.t.  $(\cdot, \cdot) = (\cdot, \cdot)_\Omega$ . Using these theorems, it is sufficient to show that the field of values  $\mathcal{F}(\tilde{\mathbf{A}})$  can be bounded independent of the fine mesh.

We split the factor  $\alpha$  in (3.29) into its real and imaginary part,

$$\alpha := \alpha_R + i\alpha_I, \quad \alpha_R, \alpha_I \in \mathbb{R}, \quad (3.41)$$

and define

$$\Gamma_\zeta^e = 1 + \zeta \|\mathbf{L}_u^c\|_{\mathbb{V}_h^v \leftarrow \mathbb{V}_h^u}^2, \quad \Gamma_\zeta^i = 1 + \zeta \|\mathbf{L}_u^f\|_{\mathbb{V}_h^v \leftarrow \mathbb{V}_h^u}^2, \quad \text{for } \zeta \in \mathbb{C}. \quad (3.42)$$

Defining quadrilaterals

$$Q = \text{conv}\{1, \Gamma_\alpha^e, \frac{\alpha}{\gamma}, \frac{\alpha}{\gamma} \Gamma_\gamma^e\}, \quad (3.43a)$$

$$R = \text{conv}\{1, \Gamma_\gamma^e, 1 + (\frac{\alpha}{\gamma} - 1)(\Gamma_\gamma^e - \frac{1}{\Gamma_\gamma^i}), \Gamma_\gamma^e + (\frac{\alpha}{\gamma} - 1)(\Gamma_\gamma^e - \frac{1}{\Gamma_\gamma^i})\}, \quad (3.43b)$$

allows us to construct a superset  $\mathcal{S}$  of  $\mathcal{F}(\tilde{\mathbf{A}})$  which is independent of the fine mesh.

**Theorem 3.4.** *Let  $\alpha \neq 0$  satisfy (3.30) and let  $\tilde{\mathbf{A}}$  be defined in (3.40b), where the preconditioner  $\mathbf{B}$  is given in (3.39) for some parameter  $\gamma > 0$ . Then we have  $\mathcal{F}(\tilde{\mathbf{A}}) \subset \mathcal{S}$ , where*

$$\mathcal{S} = \begin{cases} Q \cap R, & \alpha_I \neq 0, \\ [\frac{\alpha}{\gamma}, \Gamma_\alpha^e], & \alpha_I = 0, \quad 0 < \alpha_R = \alpha \leq \gamma, \\ [1, \frac{\alpha}{\gamma} \Gamma_\gamma^e], & \alpha_I = 0, \quad 0 < \gamma \leq \alpha_R = \alpha, \end{cases}$$

is independent of the fine mesh and  $0 \notin \mathcal{S}$ .

*Proof.* Let  $\mathbf{u} \in \mathbb{C}^{N_u}$ ,  $\mathbf{u} \neq \mathbf{0}$ , and  $\tilde{\mathbf{u}} := \mathbf{B}^{1/2} \mathbf{u}$ . Then, by the symmetry of  $\mathbf{B}$  (and thus of  $\mathbf{B}^{1/2}$ ), the adjointness and split properties Equations (3.12), (3.13), and (3.15), we have

$$(\tilde{\mathbf{u}}, \tilde{\mathbf{A}} \tilde{\mathbf{u}})_\Omega = (\mathbf{u}, \mathbf{A} \mathbf{u})_\Omega = \|\mathbf{u}\|_\Omega^2 + (\alpha_R + i\alpha_I)(\|\mathbf{L}_u^c \mathbf{u}\|_\Omega^2 + \|\mathbf{L}_u^f \mathbf{u}\|_\Omega^2), \quad (3.44a)$$

$$(\tilde{\mathbf{u}}, \tilde{\mathbf{u}})_\Omega = (\mathbf{u}, \mathbf{B} \mathbf{u})_\Omega = \|\mathbf{u}\|_\Omega^2 + \gamma \|\mathbf{L}_u^f \mathbf{u}\|_\Omega^2. \quad (3.44b)$$

We now distinguish the two cases of  $\alpha$  being real or complex.

(a) For  $\alpha_I \neq 0$ , it is easy to see that

$$1 \leq \text{Re}(\rho_{\tilde{\mathbf{A}}}(\tilde{\mathbf{u}})) + \frac{\gamma - \alpha_R}{\alpha_I} \text{Im} \rho_{\tilde{\mathbf{A}}}(\tilde{\mathbf{u}}) = 1 + \frac{\gamma \|\mathbf{L}_u^c \mathbf{u}\|_\Omega^2}{\|\mathbf{u}\|_\Omega^2 + \gamma \|\mathbf{L}_u^f \mathbf{u}\|_\Omega^2} \leq \Gamma_\gamma^e. \quad (3.45a)$$

The first inequality is obvious and the second follows from the definition of the weighted matrix norm in (3.10) and  $\gamma > 0$ . In addition, we have

$$0 \leq \operatorname{Re}(\rho_{\tilde{\mathbf{A}}}(\tilde{\mathbf{u}})) - \frac{\alpha_R}{\alpha_I} \operatorname{Im} \rho_{\tilde{\mathbf{A}}}(\tilde{\mathbf{u}}) = \frac{\|\mathbf{u}\|_{\Omega}^2}{\|\mathbf{u}\|_{\Omega}^2 + \gamma \|\mathbf{L}_u^f \mathbf{u}\|_{\Omega}^2} \leq 1. \quad (3.45b)$$

A simple calculation shows that the inequalities (3.45) are satisfied if and only if  $\rho_{\tilde{\mathbf{A}}} \in Q$  with  $Q$  defined in (3.43a).

Next, we consider only the imaginary part. Using (3.44) and (3.42), we obtain

$$0 \leq \frac{\gamma}{\alpha_I} \operatorname{Im} \rho_{\tilde{\mathbf{A}}}(\tilde{\mathbf{u}}) = 1 + \frac{\gamma \|\mathbf{L}_u^c \mathbf{u}\|_{\Omega}^2}{\|\mathbf{u}\|_{\Omega}^2 + \gamma \|\mathbf{L}_u^f \mathbf{u}\|_{\Omega}^2} - \frac{\|\mathbf{u}\|_{\Omega}^2}{\|\mathbf{u}\|_{\Omega}^2 + \gamma \|\mathbf{L}_u^f \mathbf{u}\|_{\Omega}^2} \leq \Gamma_{\gamma}^e - \frac{1}{\Gamma_{\gamma}^i}. \quad (3.46)$$

The bounds (3.45a) and (3.46) are satisfied if and only if  $\rho_{\tilde{\mathbf{A}}} \in R$  with  $R$  defined in (3.43b). Hence, we proved  $\mathcal{F}(\tilde{\mathbf{A}}) \subset Q \cap R$ .

(b) For  $\alpha_I = 0$ , the matrix  $\tilde{\mathbf{A}} \in \mathbb{R}^{N_u \times N_u}$  is symmetric and thus  $\rho_{\tilde{\mathbf{A}}}(\tilde{\mathbf{u}}) \in \mathbb{R}$  for all  $\tilde{\mathbf{u}} \in \mathbb{C}^{N_u}$ . Since  $\alpha = \alpha_R$ , we have

$$\rho_{\tilde{\mathbf{A}}}(\tilde{\mathbf{u}}) = \frac{\alpha}{\gamma} + \frac{(1 - \frac{\alpha}{\gamma}) \|\mathbf{u}\|_{\Omega}^2 + \alpha \|\mathbf{L}_u^c \mathbf{u}\|_{\Omega}^2}{\|\mathbf{u}\|_{\Omega}^2 + \gamma \|\mathbf{L}_u^f \mathbf{u}\|_{\Omega}^2}. \quad (3.47)$$

If  $\alpha \geq \gamma$ , (3.47) can be bounded by

$$1 = \frac{\alpha}{\gamma} + \frac{(1 - \frac{\alpha}{\gamma}) \|\mathbf{u}\|_{\Omega}^2}{\|\mathbf{u}\|_{\Omega}^2} \leq \rho_{\tilde{\mathbf{A}}}(\tilde{\mathbf{u}}) \leq \frac{\alpha}{\gamma} + \frac{\alpha \|\mathbf{L}_u^c \mathbf{u}\|_{\Omega}^2}{\|\mathbf{u}\|_{\Omega}^2} \leq \frac{\alpha}{\gamma} \Gamma_{\gamma}^e,$$

Similarly, for  $0 < \alpha \leq \gamma$ , it is straightforward to see

$$\frac{\alpha}{\gamma} \leq \rho_{\tilde{\mathbf{A}}}(\tilde{\mathbf{u}}) \leq \frac{\alpha}{\gamma} + \frac{(1 - \frac{\alpha}{\gamma}) \|\mathbf{u}\|_{\Omega}^2 + \alpha \|\mathbf{L}_u^c \mathbf{u}\|_{\Omega}^2}{\|\mathbf{u}\|_{\Omega}^2} \leq \Gamma_{\alpha}^e.$$

Furthermore, since

$$0 \leq \Gamma_{\gamma}^e - \frac{1}{\Gamma_{\gamma}^i} \leq \Gamma_{\gamma}^e, \quad \gamma > 0,$$

all quantities defining the superset  $\mathcal{S}$  are bounded independently of the implicitly treated mesh elements and thus  $\mathcal{S}$  is independent of  $h_f$ . Finally, in all cases we have  $0 \notin \mathcal{S}$ .  $\square$

Further, we point out that  $\gamma > 0$  can be chosen freely and thus used to improve the convergence factor. For example, a natural choice would be

$$\gamma = |\alpha_R| \text{ if } \alpha_R \neq 0 \quad \text{or} \quad \gamma = |\alpha| \text{ else.} \quad (3.48)$$

In any case, one should choose  $\gamma \sim \tau^2$  so that the dominating part of  $\mathbf{A}$  is well approximated by the preconditioner  $\mathbf{B}$ .

As a special case of Theorem 3.4, we obtain an inclusion set for the field of values of  $\mathbf{A}$  itself. Hence, we can state an error bound for the complex symmetric QMR method without preconditioning.

**Corollary 3.5.** *For the matrix  $\mathbf{A}$  defined in (3.29), Theorem 3.4 holds by substituting  $\mathbf{L}_u^c = \mathbf{L}_u$  and  $\mathbf{L}_u^f = 0$  in (3.43).*

Recall that, by an inverse estimate [DPE12, Lemma 1.44], there is a constant  $c$  independent of the mesh width such that  $\|\mathbf{L}_u\|_{\mathbb{V}_h^v \leftarrow \mathbb{V}_h^u} \leq ch_{\min}^{-1}$ . Hence, without preconditioning, the superset will scale with  $h_{\min}^{-1}$ . Applying Theorems 3.2 and 3.3 to the preconditioned system (3.40a) provides the following error bound:

**Theorem 3.6.** *Let  $\mathbf{x}_m$  be the QMR approximation to the solution of the preconditioned linear system (3.40). If (3.35) is satisfied, then there is a constant  $\phi_0 > 1$  independent of the fine mesh such that the error of the  $m$ th pQMR iterate satisfies*

$$\|\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{b}} - \mathbf{x}_m\|_{\Omega} \leq c_{\tilde{\mathbf{A}}} (1 + \sqrt{2})(1 + (m+1)\delta) \min \left\{ \frac{3}{\phi_0^m}, \frac{2}{\phi_0^m - 1} \right\}, \quad (3.49)$$

where

$$\begin{cases} c_{\tilde{\mathbf{A}}} = 1, & \text{if } 0 < \gamma \leq \alpha_R, \alpha_I = 0 \text{ or } \alpha_I \neq 0, \\ c_{\tilde{\mathbf{A}}} = \frac{\gamma}{\alpha_R}, & \text{if } 0 < \alpha_R \leq \gamma, \alpha_I = 0. \end{cases} \quad (3.50)$$

*Proof.* Since  $\tilde{\mathbf{A}}$  is complex symmetric, we apply Theorem 3.2 for  $\mathbf{K} = \tilde{\mathbf{A}}$  with  $\|\cdot\| = \|\cdot\|_{\Omega}$ . By Theorem 3.4,  $\mathcal{F}(\tilde{\mathbf{A}}) \subset \mathcal{S}$  is independent of the fine mesh and the same holds for the conformal map  $\phi$  used in Theorem 3.3, in particular for  $|\phi(0)| =: \phi_0$ . Thus, the bound (3.49) follows from (3.36), (3.37), and (3.38) if we can show

$$\|\tilde{\mathbf{A}}^{-1}\mathbf{w}\|_{\Omega} \leq c_{\tilde{\mathbf{A}}} \|\mathbf{w}\|_{\Omega} \quad \text{for all } \mathbf{w} \in \mathbb{C}^{N_u}. \quad (3.51)$$

We choose an arbitrary  $\mathbf{w} \in \mathbb{C}^{N_u}$ ,  $\mathbf{w} \neq \mathbf{0}$ , and define  $\mathbf{u} = \tilde{\mathbf{A}}^{-1}\mathbf{w}$ . Then, Theorem 3.4 with  $c_{\tilde{\mathbf{A}}}$  defined in (3.50) and the Cauchy-Schwarz inequality yield

$$\frac{1}{c_{\tilde{\mathbf{A}}}} \leq \operatorname{Re}(\rho_{\tilde{\mathbf{A}}}(\mathbf{u})) = \operatorname{Re} \frac{(\mathbf{w}, \tilde{\mathbf{A}}^{-1}\mathbf{w})_{\Omega}}{\|\tilde{\mathbf{A}}^{-1}\mathbf{w}\|_{\Omega}^2} \leq \frac{\|\mathbf{w}\|_{\Omega} \|\tilde{\mathbf{A}}^{-1}\mathbf{w}\|_{\Omega}}{\|\tilde{\mathbf{A}}^{-1}\mathbf{w}\|_{\Omega}^2} = \frac{\|\mathbf{w}\|_{\Omega}}{\|\tilde{\mathbf{A}}^{-1}\mathbf{w}\|_{\Omega}}. \quad (3.52)$$

This proves (3.51).  $\square$

As an immediate consequence of Theorem 3.6, we see that the error of the pQMR method is bounded independently of the fine mesh since  $\mathcal{S}$  only depends on the coarse mesh. In particular, the number of iterations is uniformly bounded with respect to further refinement of the fine part of the mesh.

## 3.5. Numerical experiments

For all experiments, we consider the transverse-electric (TE) polarization of linear Maxwell's equations in a homogeneous medium  $\Omega = (0, 1)^2$  with  $\mu = \epsilon = 1$ , i.e.,

$$\partial_t E_x = \partial_y H_z - J_x, \quad \text{on } \Omega \times (0, T), \quad (3.53a)$$

$$\partial_t E_y = -\partial_x H_z - J_y, \quad \text{on } \Omega \times (0, T), \quad (3.53b)$$

$$\partial_t H_z = \partial_y E_x - \partial_x E_y, \quad \text{on } \Omega \times (0, T), \quad (3.53c)$$

$$E(0) = E^0, \quad H(0) = H^0, \quad \text{on } \Omega, \quad (3.53d)$$

$$E \times \mathbf{n} = 0, \quad \text{on } \partial\Omega \times (0, T). \quad (3.53e)$$

The codes to reproduce our results are available at

<https://github.com/MalikScheifinger/HigherOrderLTI>

The software is based on the FEM library deal.II [ABB<sup>+</sup>25] version 9.7 and the Maxwell toolbox TiMaxdG [CDK<sup>+</sup>].

### 3.5.1. Illustration of Theorems 3.4 and 3.6

We consider a dG discretization with polynomials of degree two and fix

$$\alpha = \left( \frac{1}{24} + i \frac{\sqrt{3}}{24} \right) \tau^2, \quad \gamma = \alpha_R = \frac{1}{24} \tau^2, \quad (3.54)$$

for the system matrix  $\mathbf{A}$  in (3.29) and the preconditioner  $\mathbf{B}$  in (3.39), respectively. This choice follows from  $\alpha_i = \tau^2 \lambda_i^2$  in (3.27), with  $\lambda_i$  one of the two complex conjugate eigenvalues of the RK matrix  $\mathcal{O}$  of the two-stage implicit Gauß–Legendre method. The parameter  $\gamma = \alpha_R$  is the choice suggested in (3.48). The spatial domain is discretized by a uniform coarse mesh which is then locally refined at the center, yielding a coarse mesh width  $h_c = 0.117851$  and fine mesh widths  $h_f^{(2)} = 0.0294628$ ,  $h_f^{(3)} = 0.0147314$ , and  $h_f^{(4)} = 0.0073657$ , where the superscript denotes the local refinement level. Local refinement level three is illustrated in Figure 3.1.

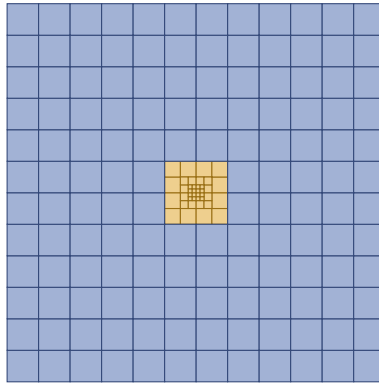


Figure 3.1.: Mesh with local refinement level three at the center of the domain.

Figure 3.2 shows the spectrum of  $\mathbf{A}$  from (3.29), the field of values of the preconditioned matrix  $\tilde{\mathbf{A}}$  from (3.40b), and the quadrilateral bounds  $Q$  and  $R$  from (3.43a) and (3.43b), for time step size  $\tau = 0.05$  and local refinement level three. By the inverse estimate used in the proof of Corollary 3.5, the eigenvalues of  $\mathbf{A}$  scale like  $h_{\min}^{-1}$ . The spectrum of the unpreconditioned system is therefore governed by the fine elements and spreads out as the mesh is refined.

Next, we apply the QMR method to the system (3.29) with and without the preconditioner (3.39). To obtain an algebraic test problem with known solution, we choose a random vector as the exact solution  $\mathbf{x}$  of the linear system (3.29) and define the right-hand side  $\mathbf{b}$  by applying the matrix  $\mathbf{A}$  to it. This allows us to compare the relative  $L^2$ -error of the QMR iterates directly. Figure 3.3 shows the results. For a fixed tolerance, the number of preconditioned QMR iterations is almost independent of the local refinement level, whereas the number of QMR iterations grows with it, as predicted by Theorem 3.6 together with Corollary 3.5.

The bound of Theorem 3.4 is fine-mesh-independent but not tight when  $\tau$  is small relative to the fine mesh width, where the linear system is only mildly stiff. In this regime the fine-mesh dependence of the superset enters solely through the term  $\Gamma_\gamma^e - 1/\Gamma_\gamma^i$  of  $R$  in (3.43b): the factor  $1/\Gamma_\gamma^i$  with  $\Gamma_\gamma^i$  from (3.42) is not negligible at the coarsest level of refinement and vanishes only as the fine mesh is refined, so  $R$  grows toward its level-independent limit. Figure 3.4 shows this effect for  $\tau = 0.01$ : preconditioned QMR still improves the convergence, and the iteration count increases with the local refinement level while remaining bounded by the fine-mesh-independent superset  $\mathcal{S}$  of Theorem 3.4. Figure 3.5 illustrates the same dependence geometrically, where

$Q$  is identical for all levels and only  $R$  grows. Such small time steps are not used in practice for the coarse spatial discretizations considered here: once the time discretization error drops below the spatial discretization error, decreasing  $\tau$  further does not reduce the total error.

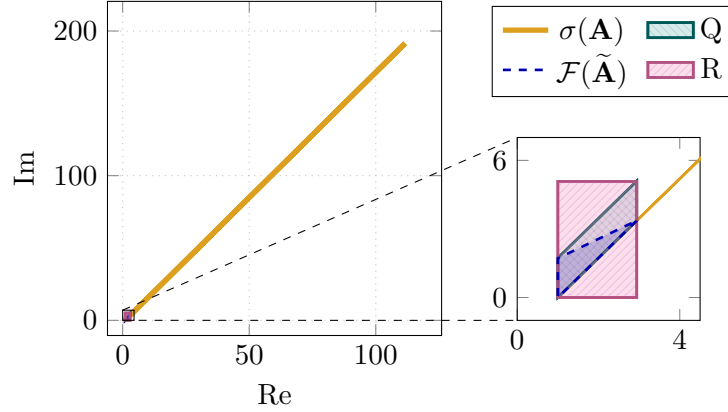


Figure 3.2.: Spectrum  $\sigma(\mathbf{A})$  (orange), field of values  $\mathcal{F}(\tilde{\mathbf{A}})$  of the preconditioned matrix (blue, dashed), and quadrilateral bounds  $Q$  (teal) and  $R$  (magenta) with  $\alpha$  chosen as in (3.54) and  $\tau = 0.05$ , local refinement level three. The inset shows a zoom near the origin.

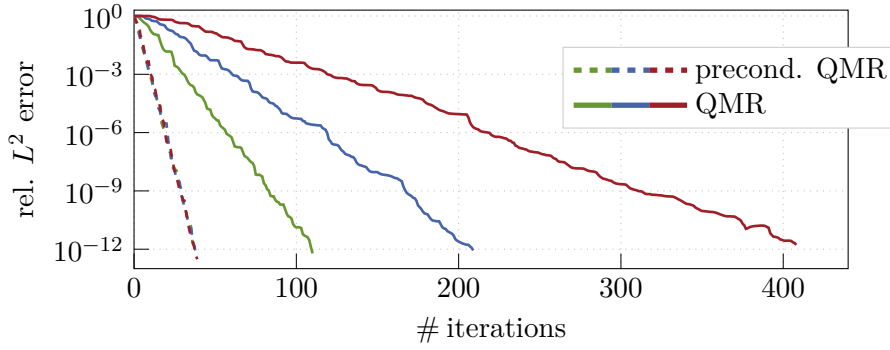


Figure 3.3.: Relative  $L^2$  error versus iteration count for QMR (solid) and preconditioned QMR (dashed) applied to (3.29) with  $\alpha$  chosen as in (3.54) and  $\tau = 0.05$ . Colors denote local refinement levels: 2 (green), 3 (blue), 4 (red).

### 3.5.2. Runtime comparison

In this section, we compare the methods with respect to runtime and time integration accuracy. We use the same TE model problem but now choose a manufactured solution which is a polynomial of degree seven in each variable. The spatial discretization uses dG polynomials of degree four. Thus, the exact solution is not represented exactly by the spatial discretization, and the measured error contains the full discretization error. To further reduce the effect of superconvergence, we randomly perturb the mesh. The locally refined mesh has  $h_f = 0.00184142$ ,  $h_c = 0.0147314$ , and  $N = 693\,900$ . The  $L^2$ -error is evaluated at the final time  $T = 1$ .

Figure 3.6 compares Gauß RK methods with two (purple) and three (orange) stages, with (dashed) and without (solid) preconditioner (3.39), the locally implicit method [HS16] (blue), a

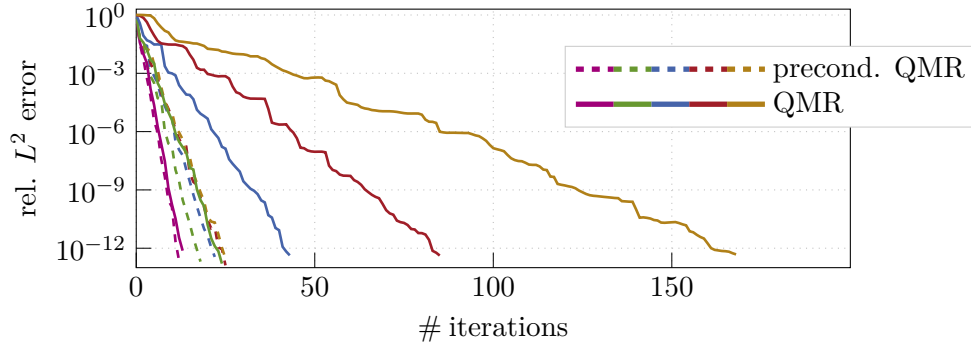


Figure 3.4.: Relative  $L^2$  error versus iteration count for QMR (solid) and preconditioned QMR (dashed) applied to (3.29) with  $\alpha$  chosen as in (3.54) and  $\tau = 0.01$ . Colors denote local refinement levels: 1 (purple), 2 (green), 3 (blue), 4 (red), 5 (orange).

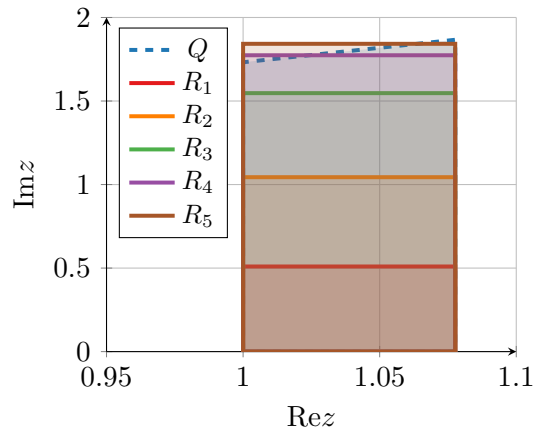


Figure 3.5.: Quadrilateral bounds  $Q$  (blue, dashed, identical for all levels) and  $R_k$  (solid) for local refinement levels  $k = 1, \dots, 5$  (red, orange, green, purple, brown) for preconditioned system (3.40b) with (3.54) and  $\tau = 0.01$ .

local time-stepping method [HS26] (green), and the popular low-storage Runge–Kutta method LSRK(12,4) [NDB12] (red). The local time-stepping method uses a Chebyshev-type iteration on the local system instead of solving a linear system like the locally implicit method, thus resulting in a fully explicit method. The degree of the Chebyshev polynomial must be chosen large enough to obtain a CFL condition independent of the stiff part of the system. For the two- and three-stage Gauß RK methods, the preconditioner significantly reduces the runtime at comparable errors. The savings come from the reduced number of QMR iterations for the preconditioned system (3.40a), while the application of the preconditioner (3.39) only requires solves on the fine part and its neighboring coarse elements. The locally implicit and local time-stepping methods are efficient for admissible time stepsizes but are still restricted by a CFL condition on the coarse mesh and become unstable for larger time stepsizes. Moreover, LSRK(12,4) is only stable for time stepsizes so small that the spatial error dominates. Though it is faster than the unpreconditioned Gauß RK methods, it is slower than the preconditioned variants at nearly the same full discretization accuracy. The Gauß RK methods are unconditionally stable and therefore remain applicable for all time stepsizes.

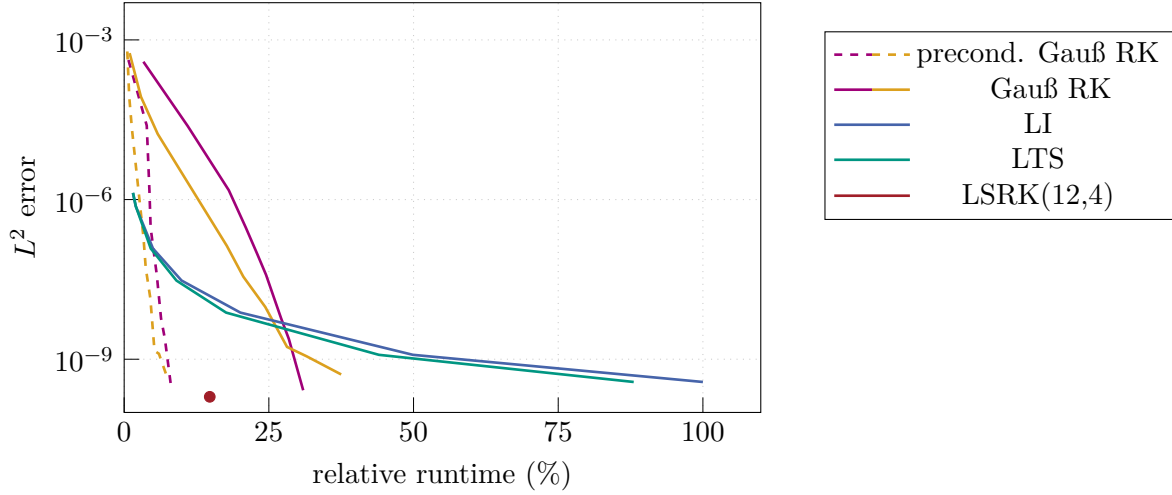


Figure 3.6.:  $L^2$  error at final time  $T = 1$  versus relative runtime for Gauß RK with two stages (purple) and three stages (orange), with preconditioner (3.39) (dashed) and without (solid), the locally implicit method (blue), the local time-stepping method (green), and LSRK(12,4) (red marker).

### 3.6. Conclusion

In this paper, we proposed and analyzed computationally efficient implicit higher-order time integration methods for solving linear Friedrichs' systems on locally refined spatial grids which consist of a small number of fine and a large number of coarse mesh elements. This is achieved by constructing a preconditioned Krylov subspace method to solve the linear systems arising in each time step of the implicit scheme. Our main result shows that the number of Krylov steps to achieve the desired accuracy can be bounded independently of the fine mesh.

Although we focused on linear Friedrichs' systems, our ideas carry over to nonlinear Friedrichs' problems like nonlinear Maxwell's equations, where linear systems of the same type appear in each iteration of a (simplified) Newton method. Moreover, instead of Gauß collocation methods, other implicit time integration schemes might be employed, and the preconditioner can also be combined with rational Krylov subspace methods.

### Acknowledgments

The authors thank Pratik Kumbhar for helpful discussions and for providing preliminary numerical examples in Matlab.



**Part II.**

**Time-integration of Gaussian variational  
approximation for the magnetic  
Schrödinger equation**



## CHAPTER 4

---

### Time-integration of Gaussian variational approximation for the magnetic Schrödinger equation

The content of this chapter is published in [\[SBH<sup>+</sup>25\]](#).

In the present paper we consider the semiclassical magnetic Schrödinger equation, which describes the dynamics of charged particles under the influence of an electro-magnetic field. The solution of the time-dependent Schrödinger equation is approximated by a single Gaussian wave packet via the time-dependent Dirac–Frenkel variational principle. For the approximation we use ordinary differential equations of motion for the parameters of the variational solution and extend the second-order Boris algorithm for classical mechanics to the quantum mechanical case. In addition, we propose a modified version of the classical fourth order Runge–Kutta method. Numerical experiments explore parameter convergence and geometric properties. Moreover, we benchmark against the analytical solution of the Penning trap.

#### 4.1. Introduction

In the present paper we study the numerical time-integration for charged quantum particles that are subjected to external magnetic and electric fields. The dynamics is governed by the semiclassical magnetic Schrödinger equation

$$i\varepsilon\partial_t\psi(t) = H(t)\psi(t), \quad \psi(0) = \psi_0, \quad t \in \mathbb{R}, \quad (4.1a)$$

on  $\mathbb{R}^d$  with magnetic Hamiltonian

$$H(t) = \frac{1}{2}(-i\varepsilon\nabla_x - A(t, \cdot))^2 + \phi(t, \cdot), \quad (4.1b)$$

and initial value  $\psi_0 \in L^2(\mathbb{R}^d)$  with semiclassical parameter  $0 < \varepsilon \ll 1$ . Here,  $A$  is a divergence-free magnetic vector potential, and  $\phi$  is the electric potential. From a numerical point of view, solving this time-dependent partial differential equation raises three major problems. First, it is a high-dimensional problem, since the space dimension is typically given by  $d = 3N$ , where

$N$  is the number of quantum particles in the system. Further, the computational domain  $\mathbb{R}^d$  is naturally unbounded, and thus most numerical methods require truncation before discretization. For the method of lines (first discretize space, then time), high dimension combined with an unbounded domain leads to inadequately if not intractably large systems that have to be integrated in time. Another challenge is given by the high oscillations induced by the small semiclassical parameter  $\varepsilon$ . For standard time integration schemes severe step-size restrictions have to be imposed and leave these methods impracticable.

We consider the case that the initial condition  $\psi_0$  is strongly localized and given by a Gaussian wave packet, and investigate the numerical time-integration of an approximate solution  $u(t) \approx \psi(t)$ , which is a Gaussian wave packet

$$u(t, x) = \exp \left( \frac{i}{\varepsilon} \left( \frac{1}{2} (x - q_t)^\top \mathcal{C}_t (x - q_t) + (x - q_t)^\top p_t + \zeta_t \right) \right).$$

The parameters to be computed are the packet's position and momentum center  $q_t, p_t$ , the complex width matrix  $\mathcal{C}_t$  of the envelope, and the complex phase and weight parameter  $\zeta_t$ . These parameters evolve according to ordinary differential equations, which are systematically derived by the Dirac–Frenkel time-dependent variational principle. By the approximation ansatz, high oscillations in time and space are captured and thus eliminated for the numerical time integration. For  $A = 0$  it is well established that variational Gaussian wave packets offer reasonable mesh-free approximations at low computational cost, see for example [LL20, Van23]. More recently, they have also been proposed for magnetic quantum dynamics [BDH<sup>+</sup>24].

##### 4.1.1. Contributions of the paper

As our main contribution we derive two fast algorithms to solve the equations of motion for the parameters of a variational Gaussian wave packet approximation such that norm and energy conservation of the quantum solution are echoed by the time integrator. On a standard laptop with a non-optimized Jupyter Notebook<sup>1</sup>, these algorithms enable us to compute approximations within minutes. First, we suggest an extension of the Boris algorithm [Bor70, BL18], which is a standard integrator for the classical equations of motion for a charged particle system in plasma physics, to the quantum mechanical setting. We furthermore modify the well-known Runge–Kutta 4 method, such that it conserves the  $L^2$ -norm of a Gaussian wave packet at every time step. Numerical experiments in two and three space dimensions underline the efficiency and expected accuracy of the proposed algorithms.

##### 4.1.2. Outline of the paper

The paper is structured as follows. In Section 4.2, we discuss the quantum dynamics of an electron and a proton in a hyperbolic Penning trap as a guiding example for a magnetic Schrödinger equation in the semiclassical regime. In Section 4.3, we compare variational and traditional Gaussian wave packet approximation and review the known error estimates. In Section 4.4, we transform the system of ordinary differential equations, that determines the variational parameter evolution, in a form that features averaged magnetic momenta and contains the magnetic vector field  $B = \nabla \times A$  on the right hand side. In particular, we derive an equation for the imaginary part of the phase parameter  $\zeta$  guaranteeing the preservation of the  $L^2$ -norm. In Section 4.5, we briefly introduce the Boris algorithm used and derive our

<sup>1</sup>Codes available at <https://gitlab.kit.edu/malik.scheifinger/magnschroedti>.

two algorithms for solving the parameters ODEs for the approximating Gaussian wave packet. Finally, in Section 4.6, we present numerical experiments for a two-dimensional magnetic system with trigonometric potentials and for the three-dimensional Penning trap. Appendix B gives formulas for the magnetic energy and other averages of Gaussian wave packets.

### 4.1.3. Notation

For a scalar function  $a: \mathbb{R}^d \rightarrow \mathbb{R}$  we denote the Hessian matrix by  $\nabla^2 a(x)$  and for a vector field  $A: \mathbb{R}^d \rightarrow \mathbb{R}^d$  we denote the Jacobi matrix by  $J_A(x) = (\partial_\ell A_k(x))_{k,\ell=1}^d$ . For a function  $W: \mathbb{R}^d \rightarrow \mathbb{R}^L$ ,  $L \geq 1$ , and more generally a linear operator  $\mathbf{A}$  acting on  $L^2(\mathbb{R}^d)$  we define the averages

$$\langle W \rangle_u := \langle u | W u \rangle, \quad \langle \mathbf{A} \rangle_u := \langle u | \mathbf{A} u \rangle,$$

if the inner products exist. We follow the convention that inner products are anti-linear in the first component. We also use the dot product of vectors  $v, w \in \mathbb{C}^L$  as  $v \cdot w := v^\top w = v_1 w_1 + \dots + v_L w_L$  and the square  $v^2 := v \cdot v$ . For complex matrices  $\mathcal{C} \in \mathbb{C}^{d \times d}$  we denote the component-wise real and imaginary parts by  $\mathcal{C}_R, \mathcal{C}_I \in \mathbb{R}^{d \times d}$ , respectively.

## 4.2. Penning trap

To illustrate semiclassical scaling for magnetic quantum dynamics we consider a charged microscopic particle, e.g., an electron or a proton, in a macroscopic hyperbolic Penning trap. Such a trap consists of an arrangement of magnetic coils and electrodes which feature a static magnetic field along the  $x_3$ -direction and a quadrupole-like static electric field that is rotationally symmetric around the  $x_3$ -axis, see, e.g., [BG84]. The corresponding vector and scalar potentials, respectively, read as

$$A(x) = \frac{1}{2} B_0 \begin{pmatrix} -x_2 \\ x_1 \\ 0 \end{pmatrix} \quad \text{and} \quad \phi(x) = \frac{\phi_0}{2\delta^2} \left( x_3^2 - \frac{1}{2} (x_1^2 + x_2^2) \right), \quad (4.2)$$

see Table 4.1 for typical trap parameters  $B_0, \phi_0, \delta$  for protons and electrons. A classical point particle (mass  $m$ , charge  $q_e$ ) moving in such an electromagnetic field configuration executes an oscillatory motion along the  $x_3$ -axis (angular frequency  $\omega_3$ ), while simultaneously executing an epicyclic motion in the  $x_1 x_2$ -plane where the low-frequency magnetron orbit (magnetron frequency  $\omega_-$ ) is overlayed with high-frequency cyclotron orbits (angular frequency  $\omega_+$ ), cf. Figure 4.1. In terms of the particle and trap parameters, these frequencies are given by

$$\omega_\pm = \frac{1}{2} (\omega_c \pm \Omega), \quad \omega_3 = \sqrt{\frac{|q_e| \phi_0}{m \delta^2}}, \quad \text{where} \quad \omega_c = \frac{|q_e| B_0}{m}, \quad \Omega = \sqrt{\omega_c^2 - 2\omega_3^2}. \quad (4.3)$$

The associated classical trajectories  $x_c(t)$  may be obtained via Newtonian [BG84] or Hamiltonian [Kre14] approaches.

The quantum dynamics of a trapped particle is governed by the time-dependent Schrödinger equation which, in SI units, reads

$$i\hbar \partial_t \psi(t, x) = \left( \frac{1}{2m} (-i\hbar \nabla_x - q_e A(x))^2 + q_e \phi(x) \right) \psi(t, x). \quad (4.4)$$

#### 4. Time-integration of Gaussian variational approximation for the magnetic Schrödinger equation

| Quantity                                                        | Electron   | Proton     |
|-----------------------------------------------------------------|------------|------------|
| $\delta$ (Trap Size)                                            | 0.00335 m  | 0.00112 m  |
| $B_0$ (Magnetic Field)                                          | 5.872 T    | 5.050 T    |
| $\phi_0$ (Electrode Potential)                                  | 10.22 V    | 53.10 V    |
| $\nu_+ = \frac{\omega_+}{2\pi}$ (Corrected Cyclotron Frequency) | 164.38 GHz | 76.299 MHz |
| $\nu_3 = \frac{\omega_3}{2\pi}$ (Axial Frequency)               | 63.698 MHz | 10.134 MHz |
| $\nu_- = \frac{\omega_-}{2\pi}$ (Magnetron Frequency)           | 12.341 kHz | 672.93 kHz |

Table 4.1.: Typical frequencies for the undulatory motion of electrons and protons in a Penning trap, cf. [BG84, Tables I and II].

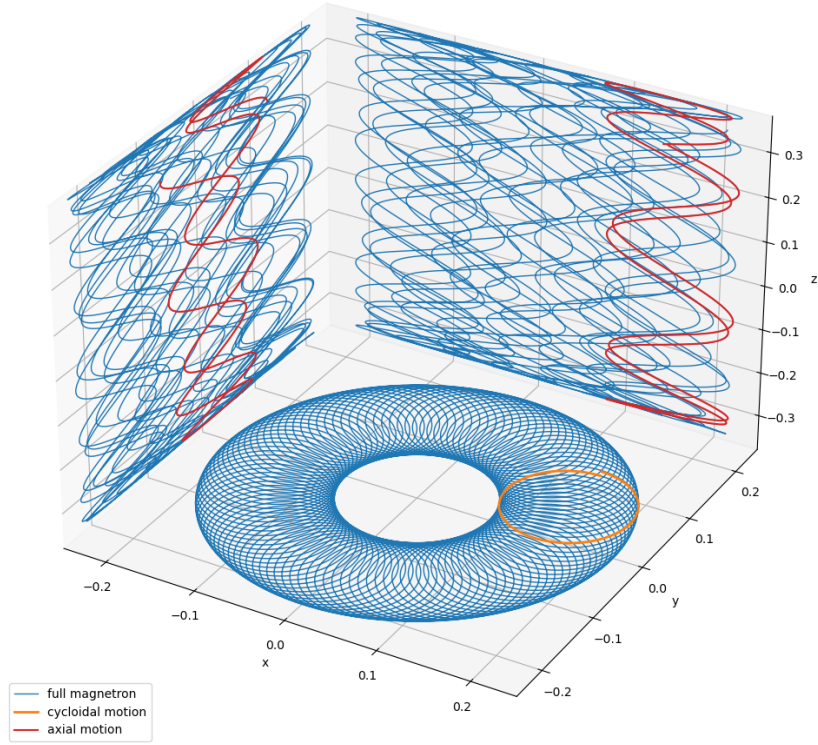


Figure 4.1.: Exact trajectory of a proton in a hyperbolic Penning trap with data from Table 4.1 and initial condition specified in Section 4.6.2 on the dimensionless time interval  $[0, 2\pi]$ .

A first comparison of the macroscopic spatial extent of the trap, described by the trap parameter  $\delta \approx 1\text{mm}$ , with the wavelength of the trapped quantum particle (e.g., a proton) corresponding to the corrected cyclotron frequency  $\nu_+ \approx 76\text{MHz}$  indicates already that mesh-based approaches to (4.4) are rather impracticable. Upon transiting to dimensionless coordinates  $x \rightarrow x/\delta$  and

introducing the dimensionless time  $t \rightarrow \omega_- t$ , equation (4.4) becomes

$$i\varepsilon \partial_t \psi(t, x) = \left( \frac{1}{2} (i\varepsilon \nabla + A_m(x))^2 + \text{sign}(q_e) \frac{\omega_+}{\omega_-} \left( x_3^2 - \frac{1}{2} (x_1^2 + x_2^2) \right) \right) \psi(t, x). \quad (4.5)$$

This is of the form (4.1), since with the effective magnetron magnetic field  $B_m = m\omega_-/q_e$  we obtain the scaled dimensionless quantities

$$\varepsilon = \hbar/(q_e B_m \delta^2) \approx 1.19 \cdot 10^{-8}, \quad \omega_+/\omega_- \approx 113.25, \quad (4.6)$$

$$A_m(x) = \frac{1}{2} \frac{B_0}{B_m} \begin{pmatrix} -x_2 \\ x_1 \\ 0 \end{pmatrix}, \quad \frac{B_0}{B_m} \approx 114.25. \quad (4.7)$$

Since the semiclassical parameter  $\varepsilon \approx 1.19 \cdot 10^{-8}$  is very small, the dynamics of the wave function are highly oscillatory in space and time, motivating the Gaussian wave packet ansatz (4.8) for eliminating high oscillations. In Figure 4.1, we illustrate the corresponding proton dynamics for the typical trap parameters of Table 4.1. The parameters of the initial Gaussian wave packet  $\psi(0, x)$  are specified in (4.76). The plot depicts the exact trajectory of the center in blue. One cycle of the cycloidal motion is highlighted in orange and one of the axial motion in red.

### 4.3. Variational approximation

In the semiclassical regime, the solution of the Schrödinger equation (4.1a) is highly oscillatory and well localized. We thus seek approximations within the manifold of complex Gaussian wave packets

$$\begin{aligned} \mathcal{M} = \left\{ u \in L^2(\mathbb{R}^d) \mid u(x) = \exp \left( \frac{i}{\varepsilon} \left( \frac{1}{2} (x - q)^\top \mathcal{C} (x - q) + (x - q)^\top p + \zeta \right) \right), \right. \\ \left. q, p \in \mathbb{R}^d, \mathcal{C} = \mathcal{C}^\top \in \mathbb{C}^{d \times d}, \text{Im } \mathcal{C} \text{ positive definite}, \zeta \in \mathbb{C} \right\}. \end{aligned} \quad (4.8)$$

We construct the optimal approximation  $u(t) \approx \psi(t)$ ,  $u(t) \in \mathcal{M}$ , in the sense that the time-derivative  $\partial_t u(t)$  minimizes the residual,

$$\|i\varepsilon \partial_t u(t) - H(t)u(t)\|_{L^2} = \min_{\partial_t u(t)}! \quad (4.9)$$

We consider initial data with  $\psi_0 = u_0 \in \mathcal{M}$  and  $\|u_0\|_{L^2} = 1$ , and mention in passing, that the variational approximation (4.9) is norm preserving in any case and energy preserving for time-independent Hamiltonians,

$$\|u(t)\|_{L^2} = \|u_0\|_{L^2}, \quad \langle H \rangle_{u(t)} = \langle H \rangle_{u_0} \quad \text{for all } t, \quad (4.10)$$

see [Lub08, §II.1.5].

#### 4.3.1. Variational equations of motion

In earlier work [BDH<sup>+</sup>24], we derived ordinary differential equations for the parameters  $q_t, p_t, \mathcal{C}_t, \zeta_t$  of the approximating Gaussian wave packet, which we reproduce here. We denote the classical Hamiltonian function for charged particles in a magnetic field by

$$h(t, x, \xi) = \frac{1}{2} (\xi - A(t, x))^2 + \phi(t, x), \quad (t, x, \xi) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d.$$

#### 4. Time-integration of Gaussian variational approximation for the magnetic Schrödinger equation

Then, the residual minimization in (4.9) implies that

$$\dot{q} = \langle \partial_\xi h \rangle_u, \quad \dot{p} = -\langle \partial_x h \rangle_u, \quad \dot{C} = -\mathcal{B}(C), \quad (4.11a)$$

$$\dot{\zeta} = -\langle h \rangle_u + \frac{\varepsilon}{4} \text{tr} \left( \mathcal{B}(C) \mathcal{C}_I^{-1} \right) + p^\top \langle \partial_\xi h \rangle_u, \quad (4.11b)$$

where the complex matrix  $\mathcal{B}(C) \in \mathbb{C}^{d \times d}$ , that depends on a Gaussian average, is given by

$$\mathcal{B}(C) = \begin{pmatrix} \text{Id} & C \end{pmatrix} \langle \nabla^2 h \rangle_u \begin{pmatrix} \text{Id} \\ C \end{pmatrix}. \quad (4.11c)$$

The averages  $\langle \partial^\alpha h \rangle_u = \langle \text{op}_{\text{Weyl}}(\partial^\alpha h) \rangle_u$  use the standard Weyl quantization of the derivatives of the Hamiltonian function. In Section 4.4, we reformulate these averages such that the equations of motion become amenable to a Boris-type time discretization.

### 4.3.2. Asymptotic accuracy

The variational Gaussian wave packet  $u(t)$  determined by (4.9) is the exact Schrödinger solution  $\psi(t)$ , if the magnetic potential  $A(t, \cdot)$  is linear and the electric potential  $\phi(t, \cdot)$  quadratic with respect to position. In particular, the dynamics in a Penning trap are exactly described. More generally, if  $A$  and  $\phi$  are sub-linear and sub-quadratic in the following sense,

$$\partial_x^\alpha A(t, \cdot), \quad \partial_x^\beta \phi(t, \cdot) \quad \text{bounded for all } |\alpha| \geq 1, \quad |\beta| \geq 2, \quad (4.12)$$

then the following bounds for the norm and the observable error can be proven, see [BDH<sup>+</sup>24, Theorems 3.8 & 3.10]. Given a finite time horizon  $[0, T]$ , the norm error and the observable error for expectation values with respect to a linear operator  $\mathbf{A}$  satisfy

$$\|\psi(t) - u(t)\|_{L^2} \leq c t \sqrt{\varepsilon}, \quad \left| \langle \mathbf{A} \rangle_{\psi(t)} - \langle \mathbf{A} \rangle_{u(t)} \right| \leq C t \varepsilon^2 \quad (4.13)$$

for all  $t \in [0, T]$ , where the constants  $c, C > 0$  are independent from  $\varepsilon$  and  $t$ , but depend on a lower bound for the eigenvalues of  $\mathcal{C}_I$  on  $[0, T]$ . It is worth pointing out the high observable accuracy.

### 4.3.3. Classical versus variational approximation

Traditional Gaussian wave packet approximation evolves the centers according to the purely classical equations of motion

$$\dot{q} = \partial_p h, \quad \dot{p} = -\partial_q h,$$

see [RC21, §4]. If the magnetic and the electric potential are linear respectively quadratic in position (as they are for a Penning trap), then the classical and the variational approximation coincide and yield the exact quantum solution. For general Hamiltonians, however, the approximations differ. The variational equations of motion (4.11) contain averages with respect to the approximating Gaussian wave packet, which are computationally more costly than the pure point evaluations of the traditional approach. However, a traditional Gaussian approximation is neither energy preserving nor as accurate as a variational one, since the classical observable error is only first order in  $\varepsilon$ , while the variational one is second order.

### 4.3.4. Hagedorn parametrization

It is convenient to write the complex symmetric width matrix  $\mathcal{C}$  in Hagedorn's parametrisation as

$$\mathcal{C} = \mathcal{C}_R + i\mathcal{C}_I = PQ^{-1} \quad \text{and} \quad \mathcal{C}_I = (QQ^*)^{-1}, \quad (4.14)$$

with two real symmetric matrices  $\mathcal{C}_R, \mathcal{C}_I$  and two invertible complex matrices  $Q, P$  that satisfy the symplecticity condition  $Q^\top P - P^\top Q = 0$ ,  $Q^*P - P^*Q = 2i\text{Id}$ . Such a decomposition of complex symmetric matrices with positive definite imaginary part is unique up to unitary factors, see [Lub08, Chapter V]. The Riccati-type equation (4.11a) for the evolution of the width matrix  $\mathcal{C}$  takes in Hagedorn's parametrisation the form

$$\dot{Q} = \langle \partial_{pq}h \rangle_u Q + \langle \partial_{pp}h \rangle_u P, \quad \dot{P} = -\langle \partial_{qq}h \rangle_u Q - \langle \partial_{qp}h \rangle_u P.$$

Unfortunately, we can express  $\mathcal{C}_R = \frac{1}{2}(PQ^{-1} + (Q^*)^{-1}P^*)$  only in terms of both  $Q$  and  $P$ , which will have implications for the integrators to be developed.

## 4.4. Transformation of equations of motion

We transform the variational equations of motion in such a way that they structurally mimic the classical equations for a charged particle.

### 4.4.1. Averaged magnetic momenta

In a first step, we rewrite the variational equations of motion such that an averaged version of the usual magnetic momenta becomes visible.

**Lemma 4.1.** *The variational equations of motion (4.11) for the parameters of a Gaussian wave packet are equivalent to*

$$\dot{q} = p - \langle A \rangle_u, \quad \dot{p} = \langle J_A^\top (\xi - A) - \nabla \phi \rangle_u, \quad (4.15a)$$

$$\dot{\mathcal{C}} = -\langle \partial_x^2 h \rangle_u + \langle J_A^\top \rangle_u \mathcal{C} + \mathcal{C} \langle J_A \rangle_u - \mathcal{C}^2, \quad (4.15b)$$

$$\dot{\xi} = -\langle h \rangle_u + \frac{\varepsilon}{4} \text{tr}(\mathcal{B}(\mathcal{C})\mathcal{C}_I^{-1}) + p^\top (p - \langle A \rangle_u), \quad (4.15c)$$

where  $\langle a \rangle_u = \langle \text{op}_{\text{Weyl}}(a) \rangle_u$  for any smooth  $a : \mathbb{R}^{2d} \rightarrow \mathbb{R}$ ,  $(x, \xi) \mapsto a(x, \xi)$ . In Hagedorn's parametrisation, the matrix evolution (4.15b) satisfies

$$\dot{Q} = P - \langle J_A \rangle_u Q, \quad (4.16a)$$

$$\dot{P} = \langle J_A^\top \rangle_u P - \langle J_A^\top J_A - \sum_{k=1}^d \nabla^2 A_k (\xi_k - A_k) + \nabla^2 \phi \rangle_u Q. \quad (4.16b)$$

*Proof.* Since the partial derivatives of the classical Hamiltonian function  $h(x, \xi) = \frac{1}{2}(\xi - A(x))^2 + \phi(x)$  satisfy

$$\partial_\xi h(x, \xi) = \xi - A(x), \quad \partial_x h(x, \xi) = -J_A(x)^\top (\xi - A(x)) + \nabla \phi(x), \quad (4.17)$$

we have (4.15a). We next turn to the second order derivatives of  $h$ . Denoting the partial derivatives of the magnetic and the electric potential shortly by  $\partial_1, \dots, \partial_d$ , we have

$$\partial_{x_m} \partial_{x_\ell} h = \sum_{k=1}^d (\partial_m A_k \partial_\ell A_k - (\xi_k - A_k) \partial_m \partial_\ell A_k) + \partial_m \partial_\ell \phi, \quad (4.18)$$

#### 4. Time-integration of Gaussian variational approximation for the magnetic Schrödinger equation

$$\partial_{x_m} \partial_{\xi_\ell} h = -\partial_m A_\ell, \quad \partial_{\xi_m} \partial_{x_\ell} h = -\partial_\ell A_m, \quad (4.19)$$

which gives a Hessian matrix

$$\nabla^2 h = \begin{pmatrix} \partial_x^2 h & \partial_{x\xi} h \\ \partial_{\xi x} h & \partial_\xi^2 h \end{pmatrix} = \begin{pmatrix} \partial_x^2 h & -J_A^\top \\ -J_A & \text{Id} \end{pmatrix}$$

with  $\partial_x^2 h = J_A^\top J_A - \sum_{k=1}^d \nabla^2 A_k (\xi_k - A_k) + \nabla^2 \phi$ . By (4.11c) we then infer that

$$\mathcal{B}(\mathcal{C}) = \langle \partial_x^2 h \rangle_u + \langle \partial_{x\xi} h \rangle_u \mathcal{C} + \mathcal{C} \langle \partial_{\xi x} h \rangle_u + \mathcal{C} \langle \partial_\xi^2 h \rangle_u \mathcal{C} \quad (4.20)$$

$$= \langle \partial_x^2 h \rangle_u - \langle J_A^\top \rangle_u \mathcal{C} - \mathcal{C} \langle J_A \rangle_u + \mathcal{C}^2. \quad (4.21)$$

Hence, the variational equations of motion (4.11) are equivalent to (4.15).  $\square$

This formulation of the equations of motion prominently features averaged magnetic momenta

$$v := p - \langle A \rangle_u, \quad \Upsilon := P - \langle J_A \rangle_u Q.$$

We aim at rewriting them in terms of the real vector  $v(t) \in \mathbb{R}^d$  and the complex matrix  $\Upsilon(t) \in \mathbb{C}^{d \times d}$ .

##### 4.4.2. Equations for the center and the width

Since the averaged magnetic momenta contain averages of the magnetic field and of its Jacobian, we need a compact way of assessing the time derivative of averages in the spirit of a magnetic Ehrenfest-type theorem.

**Proposition 4.2.** *For any smooth function  $w: \mathbb{R}^d \rightarrow \mathbb{R}$ , its average with respect to the variational Gaussian wave packet  $u = u(t)$  satisfies*

$$\frac{d}{dt} \langle w \rangle_u = \langle \nabla w \rangle_u^\top v + \frac{\varepsilon}{2} \text{tr} \left( \langle \nabla^2 w \rangle_u (\Upsilon Q^* - i \text{Id}) \right), \quad (4.22)$$

where  $\Upsilon Q^* - i \text{Id}$  is a real matrix.

*Proof.* We differentiate the position density with respect to time,

$$\partial_t |u(t, x)|^2 = \partial_t \exp \left( -\frac{1}{\varepsilon} (x - q)^\top \mathcal{C}_I (x - q) - \frac{2}{\varepsilon} \zeta_I \right) \quad (4.23)$$

$$= |u(t, x)|^2 \left( -\frac{1}{\varepsilon} (x - q)^\top \dot{\mathcal{C}}_I (x - q) + \frac{2}{\varepsilon} (x - q)^\top \mathcal{C}_I \dot{q} - \frac{2}{\varepsilon} \dot{\zeta}_I \right) \quad (4.24)$$

$$= |u(t, x)|^2 \left( \frac{2}{\varepsilon} (x - q)^\top (\mathcal{C}_R - \langle J_A^\top \rangle_u) \mathcal{C}_I (x - q) + \frac{2}{\varepsilon} (x - q)^\top \mathcal{C}_I v - \text{tr}(\mathcal{C}_R) \right), \quad (4.25)$$

where we have used that, by Lemma 4.1, the symmetry of  $\mathcal{C}_R$  and  $\mathcal{C}_I$ , and  $\text{tr}(\langle J_A \rangle_u) = 0$  since  $\nabla \cdot A = 0$ ,

$$\dot{q} = v, \quad \dot{\mathcal{C}}_I = \mathcal{C}_I (\langle J_A \rangle_u - \mathcal{C}_R) + (\langle J_A^\top \rangle_u - \mathcal{C}_R) \mathcal{C}_I, \quad \dot{\zeta}_I = \frac{\varepsilon}{2} \text{tr}(\mathcal{C}_R).$$

Therefore, we obtain

$$\begin{aligned}
 \frac{d}{dt}\langle w \rangle_u &= \frac{2}{\varepsilon} \left\langle w(x-q)^\top (\mathcal{C}_R - \langle J_A^\top \rangle_u) \mathcal{C}_I(x-q) \right\rangle_u \\
 &\quad + \frac{2}{\varepsilon} \langle w \mathcal{C}_I(x-q) \rangle_u^\top v - \langle w \rangle_u \text{tr}(\mathcal{C}_R) \\
 &= \langle w \rangle_u \text{tr}(\mathcal{C}_R - \langle J_A^\top \rangle_u) + \langle \nabla w \rangle_u^\top v - \langle w \rangle_u \text{tr}(\mathcal{C}_R) \\
 &\quad + \frac{\varepsilon}{2} \text{tr}(\langle \nabla^2 w \rangle_u \mathcal{C}_I^{-1}(\mathcal{C}_R - \langle J_A^\top \rangle_u)) \\
 &= \langle \nabla w \rangle_u^\top v + \frac{\varepsilon}{2} \text{tr}(\langle \nabla^2 w \rangle_u \mathcal{C}_I^{-1}(\mathcal{C}_R - \langle J_A^\top \rangle_u)),
 \end{aligned} \tag{4.26}$$

where the second equation relies on Lemma B.1 with matrix  $M = (\mathcal{C}_R - \langle J_A^\top \rangle_u) \mathcal{C}_I$  and the last equation on  $\nabla \cdot A = 0$ . It remains to express the occurring matrices in terms of  $Q$  and  $\Upsilon$ . Using the properties of Hagedorn's parametrization from Sec. 4.3.4 yields

$$\begin{aligned}
 \mathcal{C}_I^{-1} \mathcal{C}_R &= \frac{1}{2} Q Q^* (P Q^{-1} + (Q^*)^{-1} P^*) \\
 &= \frac{1}{2} \left( Q(P^* Q + 2i \text{Id}) Q^{-1} + Q P^* \right) \\
 &= Q(\Upsilon^* + Q^* \langle J_A^\top \rangle_u) + i \text{Id}.
 \end{aligned} \tag{4.27}$$

Hence we have  $\mathcal{C}_I^{-1}(\mathcal{C}_R - \langle J_A^\top \rangle_u) = Q \Upsilon^* + i \text{Id}$ , and use the trace identity  $\text{tr}(MN) = \text{tr}(MN^*)$  for the real symmetric matrix  $M = \langle \nabla^2 w \rangle_u$  and the real matrix  $N = Q \Upsilon^* + i \text{Id}$ .  $\square$

Equipped with an equation for the time evolution of averages, we can determine the variational equations of motion solely in terms of the averaged magnetic momenta.

**Theorem 4.3.** *We consider the averaged magnetic momenta (4.4.1) and denote the magnetic field  $B = \nabla \times A$ . The variational equations of motion for the Gaussian wave packet's center and width, equations (4.15a) and (4.16), can be transformed as*

$$\dot{q} = v, \quad \dot{v} = v \times \langle B \rangle_u + E, \tag{4.28a}$$

$$\dot{Q} = \Upsilon, \quad \dot{\Upsilon} = \Upsilon \times \langle B \rangle_u + S Q, \tag{4.28b}$$

where the matrix

$$\Upsilon \times \langle B \rangle_u := (v_1 \times \langle B \rangle_u, \dots, v_d \times \langle B \rangle_u)$$

is given by the cross product of the column vectors  $v_1, \dots, v_d$  of  $\Upsilon$  with  $\langle B \rangle_u$ . The real vector field  $E$  satisfies

$$E = -\langle \nabla \phi \rangle_u - \langle \partial_t A \rangle_u + \langle J_A^\top \rangle_u \langle A \rangle_u - \langle J_A^\top A \rangle_u \tag{4.29}$$

$$+ \frac{\varepsilon}{2} \left( \text{tr} \left( \langle \partial_k J_A^\top - \nabla^2 A_k \rangle_u (\Upsilon Q^* - i \text{Id}) + \langle \partial_k J_A^\top \rangle_u \langle J_A \rangle_u Q Q^* \right) \right)_{k=1}^d, \tag{4.30}$$

and the real matrix potential  $S$  can be written as

$$S = -\langle \nabla^2 \phi \rangle_u - \langle J_{\partial_t A} \rangle_u + \langle J_A^\top \rangle_u \langle J_A \rangle_u - \langle J_A^\top J_A \rangle_u \tag{4.31}$$

$$+ \sum_{m=1}^d \left( \langle \nabla^2 A_m \rangle_u \langle A_m \rangle_u - \langle (\nabla^2 A_m) A_m \rangle_u + (\langle \partial_k \partial_\ell A_m - \partial_m \partial_\ell A_k \rangle_u v_m)_{k,\ell=1}^d \right) \tag{4.32}$$

$$+ \frac{\varepsilon}{2} \left( \text{tr} \left( \langle \partial_k \partial_\ell J_A^\top - \nabla^2 \partial_\ell A_k \rangle_u (\Upsilon Q^* - i \text{Id}) + \langle \partial_k \partial_\ell J_A^\top \rangle_u \langle J_A \rangle_u Q Q^* \right) \right)_{k,\ell=1}^d. \tag{4.33}$$

#### 4. Time-integration of Gaussian variational approximation for the magnetic Schrödinger equation

*Proof.* By (4.15a), the time derivative of the magnetic momenta satisfies

$$\dot{v} = \dot{p} - \frac{d}{dt} \langle A \rangle_u = \langle J_A^\top (\xi - A) - \nabla \phi \rangle_u - \frac{d}{dt} \langle A \rangle_u. \quad (4.34)$$

Hence, we have to work on  $\langle J_A^\top \xi \rangle_u$ . By symbolic Weyl calculus, see for example [RC21, §2.4], we expand the operator of the product as

$$\text{op}_{\text{Weyl}}(J_A^\top \xi) = J_A^\top \text{op}_{\text{Weyl}}(\xi) + \frac{i\varepsilon}{2} \text{op}_{\text{Weyl}}(\{J_A^\top, \xi\}) = J_A^\top (-i\varepsilon \nabla), \quad (4.35)$$

where the last equation uses that  $\nabla \cdot A = 0$  implies that  $\{J_A^\top, \xi\} = -\sum_{k=1}^d \partial_{x_k} (J_A^\top) \partial_{\xi_k} \xi = -\sum_{k=1}^d \partial_{x_k} \nabla A_k = 0$ . Therefore,

$$\langle J_A^\top \xi \rangle_u = \langle u, J_A^\top (-i\varepsilon \nabla) u \rangle = \langle J_A^\top \mathcal{C}(x - q) \rangle_u + \langle J_A^\top \rangle_u p \quad (4.36)$$

due to  $-i\varepsilon \nabla u(x) = (\mathcal{C}(x - q) + p)u(x)$ . Then, we apply Lemma B.1 with  $W = J_A^\top \mathcal{C}$ ,

$$\langle J_A^\top \mathcal{C}(x - q) \rangle_u = \frac{\varepsilon}{2} \sum_{\ell=1}^d \left( \left( \langle \partial_\ell J_A^\top \rangle_u \mathcal{C} \mathcal{C}_I^{-1} \right)_{k\ell} \right)_{k=1}^d \quad (4.37)$$

$$= \frac{\varepsilon}{2} \sum_{m,\ell=1}^d \left( \langle \partial_\ell \partial_k A_m \rangle_u (\Upsilon Q^* + \langle J_A \rangle_u Q Q^* - i\text{Id})_{m\ell} \right)_{k=1}^d \quad (4.38)$$

$$= \frac{\varepsilon}{2} \left( \text{tr}(\langle \partial_k J_A \rangle_u^\top (\Upsilon Q^* - i\text{Id} + \langle J_A \rangle_u Q Q^*)) \right)_{k=1}^d, \quad (4.39)$$

since  $\mathcal{C} \mathcal{C}_I^{-1} = (PQ^{-1})(QQ^*) = (\Upsilon + \langle J_A \rangle_u Q)Q^*$  and  $\text{div } A = 0$ . We thus obtain

$$\langle J_A^\top (\xi - A) \rangle_u = \langle J_A^\top \rangle_u v + \langle J_A^\top \rangle_u \langle A \rangle_u - \langle J_A^\top A \rangle_u \quad (4.40)$$

$$+ \frac{\varepsilon}{2} \left( \text{tr}(\langle \partial_k J_A \rangle_u^\top (\Upsilon Q^* - i\text{Id} + \langle J_A \rangle_u Q Q^*)) \right)_{k=1}^d. \quad (4.41)$$

Now, we use Proposition 4.2 for each component of the magnetic potential,

$$\frac{d}{dt} \langle A \rangle_u = \langle \partial_t A \rangle_u + \langle J_A \rangle_u v + \frac{\varepsilon}{2} \left( \text{tr}(\langle \nabla^2 A_k \rangle_u (\Upsilon Q^* - i\text{Id})) \right)_{k=1}^d.$$

When collecting all the terms that originated in (4.34), we observe the occurrence of  $\langle J_A^\top \rangle_u v$  and  $-\langle J_A \rangle_u v$ , which combine according to

$$\langle (J_A^\top - J_A) \rangle_u v = v \times \langle B \rangle_u.$$

We thus arrive at the claimed equation (4.28a). As for the matrix  $\Upsilon = P - \langle J_A \rangle_u Q$ , we use the equations of motion (4.16), which contain the average of  $\partial_x^2 h = J_A^\top J_A - \sum_{m=1}^d \nabla^2 A_m (\xi_m - A_m) + \nabla^2 \phi$ . We have

$$\dot{P} = \langle J_A^\top \rangle_u (\Upsilon + \langle J_A \rangle_u Q) - \langle \partial_x^2 h \rangle_u Q.$$

For computing  $\langle \partial_x^2 h \rangle_u$ , we observe that

$$\sum_{m=1}^d \text{op}_{\text{Weyl}}(\nabla^2 A_m \xi_m) = \sum_{m=1}^d \left( \nabla^2 A_m \text{op}_{\text{Weyl}}(\xi_m) + \frac{i\varepsilon}{2} \text{op}_{\text{Weyl}}\{\nabla^2 A_m, \xi_m\} \right) \quad (4.42)$$

$$= \sum_{m=1}^d \nabla^2 A_m (-i\varepsilon \partial_m), \quad (4.43)$$

where the last equation uses that  $\sum_{m=1}^d \{\nabla^2 A_m, \xi_m\} = 0$  due to  $\nabla \cdot A = 0$ . Arguing as for (4.39), we obtain

$$\sum_{m=1}^d \langle \nabla^2 A_m (\xi_m - A_m) \rangle_u \quad (4.44)$$

$$= \sum_{m=1}^d \left( \langle \nabla^2 A_m (\mathcal{C}(x - q))_m \rangle_u - \langle (\nabla^2 A_m) A_m \rangle_u + \langle \nabla^2 A_m \rangle_u p_m \right) \quad (4.45)$$

$$= \sum_{m=1}^d \left( \langle \nabla^2 A_m \rangle_u (v_m + \langle A_m \rangle_u) - \langle (\nabla^2 A_m) A_m \rangle_u \right) \quad (4.46)$$

$$+ \frac{\varepsilon}{2} \sum_{m,n=1}^d \langle \nabla^2 \partial_n A_m \rangle_u (\Upsilon Q^* - i\text{Id} + \langle J_A \rangle_u Q Q^*)_{mn}, \quad (4.47)$$

and therefore

$$\langle \partial_x^2 h \rangle_u = \langle J_A^\top J_A + \nabla^2 \phi \rangle_u - \sum_{m=1}^d \left( \langle \nabla^2 A_m \rangle_u (v_m + \langle A_m \rangle_u) - \langle (\nabla^2 A_m) A_m \rangle_u \right) \quad (4.48)$$

$$- \frac{\varepsilon}{2} \left( \text{tr} \left( \langle \partial_k \partial_\ell J_A^\top \rangle_u (\Upsilon Q^* - i\text{Id} + \langle J_A \rangle_u Q Q^*) \right) \right)_{k,\ell=1}^d. \quad (4.49)$$

We next use Proposition 4.2 applied to each component of the Jacobian matrix  $J_A = (\partial_\ell A_k)_{k,\ell=1}^d$  and obtain

$$\frac{d}{dt} \langle J_A \rangle_u = \langle J_{\partial_t A} \rangle_u \quad (4.50)$$

$$+ \sum_{m=1}^d (\langle \partial_m \partial_\ell A_k \rangle_u v_m)_{k,\ell=1}^d + \frac{\varepsilon}{2} \left( \text{tr} \left( \langle \nabla^2 \partial_\ell A_k \rangle_u (\Upsilon Q^* - i\text{Id}) \right) \right)_{k,\ell=1}^d. \quad (4.51)$$

Now we combine and arrive at

$$\dot{\Upsilon} = \dot{P} - \left( \frac{d}{dt} \langle J_A \rangle_u \right) Q - \langle J_A \rangle_u \dot{Q} = \langle J_A^\top - J_A \rangle_u \Upsilon + S Q \quad (4.52)$$

with the matrix potential  $S$  of the claimed form.  $\square$

*Remark 4.4* (Linear potential  $A$ ). For a linear magnetic potential  $A$ , all higher order derivatives vanish and the average  $\langle A \rangle_u = A(q)$  is a point evaluation. Thus, the equations of motion of Theorem 4.3 simplify to

$$\dot{q} = v, \quad \dot{v} = v \times B(q) - (\partial_t A(q) + \langle \nabla \phi \rangle_u), \quad (4.53a)$$

$$\dot{Q} = \Upsilon, \quad \dot{\Upsilon} = \Upsilon \times B(q) - (\partial_t J_A(q) + \langle \nabla^2 \phi \rangle_u) Q. \quad (4.53b)$$

The Penning trap, see Section 4.2 and Section 4.6.2, with its quadratic electric potential, has even simpler equations of motion, since also the averages of the electric potential collapse to  $\langle \nabla \phi \rangle_u = \nabla \phi(q)$  and  $\langle \nabla^2 \phi \rangle_u = \nabla^2 \phi(q)$ .  $\diamond$

### 4.4.3. Equations for the phase

The imaginary part of the complex parameter  $\zeta = \zeta_R + i\zeta_I$  carries the normalization of the Gaussian wave packet. We derive an explicit representation, which only depends on the determinant of the complex matrix  $Q$ . This representation will support our time discretization of  $\zeta_I$  in Section 4.5.2.

**Lemma 4.5.** *For the real part  $\zeta_R = \operatorname{Re}(\zeta)$  we have*

$$\dot{\zeta}_R = \frac{1}{2}v^2 + \langle A \rangle_u^\top v - \langle \phi \rangle_u + \frac{\varepsilon}{4} \operatorname{tr} (\langle \partial_x^2 h \rangle_u Q Q^* - 2(Q Q^*)^{-1}),$$

where the average  $\langle \partial_x^2 h \rangle_u$  is given in (4.49) with respect to  $(q, v, Q, \Upsilon)$ . The imaginary part  $\zeta_I = \operatorname{Im}(\zeta)$  satisfies the normalization formula

$$\zeta_I(t) = \zeta_I(0) + \frac{\varepsilon}{2} (\ln |\det Q(t)| - \ln |\det Q(0)|).$$

*Proof.* We start with the normalization formula. Since  $\operatorname{Im} \mathcal{B}(\mathcal{C}) = -\langle J_A^\top \rangle_u \mathcal{C}_I - \mathcal{C}_I \langle J_A \rangle_u + \mathcal{C}_R \mathcal{C}_I + \mathcal{C}_I \mathcal{C}_R$  and  $\operatorname{div} A = 0$ , we have

$$\operatorname{Im} \left( \operatorname{tr} (\mathcal{B}(\mathcal{C}) \mathcal{C}_I^{-1}) \right) = 2 \operatorname{tr} (\mathcal{C}_R) = 2 \operatorname{tr} (\operatorname{Re} (P Q^{-1})) \quad (4.55)$$

$$= 2 \operatorname{tr} (\operatorname{Re} ((\dot{Q} + \langle J_A \rangle_u Q) Q^{-1})) \quad (4.56)$$

$$= 2 \operatorname{tr} (\operatorname{Re} (\dot{Q} Q^{-1})). \quad (4.57)$$

We thus obtain from (4.15c) with Jacobi's formula

$$\begin{aligned} \dot{\zeta}_I &= \frac{\varepsilon}{2} \operatorname{tr} (\operatorname{Re} (\dot{Q} Q^{-1})) \\ &= \frac{\varepsilon}{4} \frac{1}{|\det Q|^2} \left( 2 \operatorname{Re} \left( \overline{\det Q} \det Q \operatorname{tr} (\dot{Q} Q^{-1}) \right) \right) \\ &= \frac{\varepsilon}{4} \frac{1}{|\det Q|^2} \left( 2 \operatorname{Re} (\overline{\det Q} \partial_t (\det Q)) \right) \\ &= \frac{\varepsilon}{4} \partial_t \left( \ln |\det Q|^2 \right). \end{aligned} \quad (4.58)$$

Integrating (4.58) from 0 to  $t$  leads to (4.54). For the real part, we have  $\operatorname{Re} \mathcal{B}(\mathcal{C}) = \langle \partial_x^2 h \rangle_u - \langle J_A^\top \rangle_u \mathcal{C}_R - \mathcal{C}_R \langle J_A \rangle_u + \mathcal{C}_R^2 - \mathcal{C}_I^2$  and thus

$$\operatorname{tr} (\operatorname{Re} \mathcal{B}(\mathcal{C}) \mathcal{C}_I^{-1}) = \operatorname{tr} \left( \left( \langle \partial_x^2 h \rangle_u - 2 \langle J_A^\top \rangle_u \mathcal{C}_R + (\mathcal{C}_R^2 - \mathcal{C}_I^2) \right) \mathcal{C}_I^{-1} \right).$$

Combining this with  $-\langle h \rangle_u$ , we use Lemma B.2 and obtain

$$-\langle h \rangle_u + \frac{\varepsilon}{4} \operatorname{tr} (\operatorname{Re} \mathcal{B}(\mathcal{C}) \mathcal{C}_I^{-1}) \quad (4.59)$$

$$= -\frac{1}{2} \langle (p - A)^2 \rangle_u - \langle \phi \rangle_u + \frac{\varepsilon}{4} \operatorname{tr} (\langle \partial_x^2 h \rangle_u \mathcal{C}_I^{-1} - 2 \mathcal{C}_I). \quad (4.60)$$

Next we observe that

$$-\frac{1}{2} \langle (p - A)^2 \rangle_u + p^\top \langle p - A \rangle_u = \frac{1}{2} p^\top \langle p - A \rangle_u + \frac{1}{2} \langle A \rangle_u^\top \langle p - A \rangle_u \quad (4.61)$$

$$= \frac{1}{2} v^2 + \langle A \rangle_u^\top v. \quad (4.62)$$

Using that  $\mathcal{C}_I = (Q Q^*)^{-1}$ , the real part of the evolution equation (4.15c) can thus be written in the claimed form.  $\square$

## 4.5. Time integration for the equations of motion

In this section we first briefly review the classical Boris algorithm. Afterwards, we present the new Boris-type algorithm and a modification of the classical Runge–Kutta method to solve (4.28) and (4.54).

### 4.5.1. Boris algorithm for classical equations of motion

The Boris algorithm was originally proposed in [Bor70] for solving the classical equations of motion

$$\dot{q} = v, \quad \dot{v} = v \times B + E. \quad (4.63)$$

for charged particles in an electro-magnetic field. We consider a time-grid  $t^n$ ,  $n \geq 0$ , with step size  $\tau > 0$ . Given approximations  $q^n \approx q(t^n)$  and  $v^{n-\frac{1}{2}} \approx v(t^{n-\frac{1}{2}})$ , the algorithm can be written as

$$v^- = v^{n-\frac{1}{2}} + \frac{\tau}{2} E^n, \quad E^n = E(t^n, q^n), \quad (4.64a)$$

$$v^+ - v^- = \frac{\tau}{2} (v^+ + v^-) \times B^n, \quad B^n = B(t^n, q^n), \quad (4.64b)$$

$$v^{n+\frac{1}{2}} = v^+ + \frac{\tau}{2} E^n, \quad (4.64c)$$

$$q^{n+1} = q^n + \tau v^{n+\frac{1}{2}}. \quad (4.64d)$$

Note that the algorithm provides approximations on a staggered grid, where the velocities are only given at half time-steps. Approximations at  $t^n$  can be obtained by averaging

$$v^n = \frac{1}{2} (v^{n+\frac{1}{2}} + v^{n-\frac{1}{2}}). \quad (4.64e)$$

Moreover, the scheme is explicit, since one can replace (4.64b) by

$$v^+ = v^- + \left( v^- + v^- \times \frac{\tau}{2} B^n \right) \times \frac{\tau B^n}{1 + \left| \frac{\tau}{2} B^n \right|^2}, \quad (4.64f)$$

see, e.g., [BL18]. The Boris algorithm is a second-order method which is not symplectic but conserves the phase-space volume as shown in [QZX<sup>+</sup>13]. A recent analysis was presented in [HL20].

### 4.5.2. Boris-type algorithm

We aim at solving the Euler-Lagrange system (4.28) together with the two phase equations (4.54) and (4.54). The former are closely related to the classical equations of motion of a charged particle (4.63) except that  $B$  is replaced by an averaged field  $\langle B \rangle_u$  and that the fields  $E$  and  $SQ$  do not only depend on  $q$  and  $Q$  but also in a nontrivial way on  $v$  and  $\Upsilon$ . While the means  $\langle B \rangle_u$  can be approximated by a suitable Gauß-Hermite quadrature formula, evaluating the fields  $E$  and  $S$  is more involved since they require approximations to  $v$  and  $\Upsilon$  at time  $t^n$ . Unfortunately, these quantities are only defined on the staggered grid  $t^{n \pm \frac{1}{2}}$  and even worse, the update of  $\Upsilon$  is coupled to the evaluation of  $SQ$ , rendering the scheme implicit.

#### 4. Time-integration of Gaussian variational approximation for the magnetic Schrödinger equation

To be more precise, it is the matrix  $\Upsilon$  which is necessary to compute the fields  $E$  and  $S$  in Theorem 4.3. Averaging  $\Upsilon$  as in (4.64e) would lead to a nonlinear system in  $\Upsilon^{n+\frac{1}{2}}$ . Therefore we propose the second-order extrapolation

$$\Upsilon^n = \frac{3}{2}\Upsilon^{n-\frac{1}{2}} - \frac{1}{2}\Upsilon^{n-\frac{3}{2}}, \quad (4.65)$$

and compute also  $E^n$  and  $S^n$  from it. In our numerical experiments we saw that using fixed-point iterations to improve the accuracy of  $\Upsilon^n$  did not change the errors.

##### 4.5.3. Discretization of the phase

Motivated by (4.54), we define the approximation to  $\zeta_I(t^{n+1})$  as

$$\zeta_I^{n+1} = \zeta_I^n + \frac{\varepsilon}{2} \left( \ln |\det Q^{n+1}| - \ln |\det Q^n| \right). \quad (4.66)$$

This update formula ensures norm preservation of the wave packet.

**Lemma 4.6.** *Let  $u^n$  and  $u^{n+1}$  denote two Gaussian wave packets with parameters  $(q^n, v^n, Q^n, \Upsilon^n, \zeta^n)$  and  $(q^{n+1}, v^{n+1}, Q^{n+1}, \Upsilon^{n+1}, \zeta^{n+1})$ , respectively. Then (4.66) is equivalent to*

$$\|u^n\|_{L^2} = \|u^{n+1}\|_{L^2}. \quad (4.67)$$

*Proof.* Because of

$$\|u^n\|_{L^2}^2 = \exp(-\frac{2}{\varepsilon}\zeta_I^n)(\varepsilon\pi)^{\frac{d}{2}} |\det Q^n|, \quad (4.68)$$

the norm preservation is equivalent to

$$\exp(\frac{2}{\varepsilon}(\zeta_I^n - \zeta_I^{n+1})) = \frac{|\det Q^n|}{|\det Q^{n+1}|}. \quad (4.69)$$

This proves the statement.  $\square$

Furthermore, for the integration of (4.54), we propose to use the midpoint rule with time step size  $2\tau$ . This results in a two-step method. Since  $v^{n\pm\frac{1}{2}}$  and  $\Upsilon^{n\pm\frac{1}{2}}$  are already available, we apply averaging (4.64e) instead of extrapolation to obtain approximations at  $t^n$ . This results in

$$\zeta_R^{n+1} = \zeta_R^{n-1} + 2\tau \left( \frac{1}{2}(v^n)^2 + (\langle A \rangle_u^n)^\top v^n - \langle \phi \rangle_u^n \right) \quad (4.70)$$

$$+ \frac{\varepsilon\tau}{2} \operatorname{tr} \left( \langle \partial_x^2 h \rangle_u^n Q^n (Q^n)^* - 2(Q^n Q^{n*})^{-1} \right) \quad (4.71)$$

##### 4.5.4. Complete Boris-type algorithm

Overall, combining the time discretization of center, width and phase of the Gaussian variational wave packet, we get the following algorithm for solving the equations of motion (4.28) and (4.54).

**Algorithm 4.7** (One time step with the Boris algorithm)

*Input:* Last steps  $(q^n, v^{n-\frac{1}{2}}, Q^n, \Upsilon^{n-\frac{1}{2}}, \zeta^n, \zeta^{n-1})$ .

*Output:* Approximations  $(q^{n+1}, v^{n+\frac{1}{2}}, Q^{n+1}, \Upsilon^{n+\frac{1}{2}}, \zeta^{n+1})$ .

- compute  $q^{n+1}, v^{n+\frac{1}{2}}$  with the Boris algorithm (4.64) applied to (4.28a)
- compute  $Q^{n+1}, \Upsilon^{n+\frac{1}{2}}$  with the Boris algorithm (4.64) applied column-wise to (4.28b),
- compute  $\zeta_I^{n+1}$  and  $\zeta_R^{n+1}$  from (4.66) and (4.70), respectively.

In general, one has to apply Gauss-Hermite quadrature rule to approximate the averages, see [LL20, Section 8] for details. Note that for strong magnetic fields, filtered variants of the Boris algorithms might be more efficient, cf. [HLW20].

#### 4.5.5. Modified classical Runge–Kutta method

As an alternative to the Boris-type algorithm, we propose a modification of the classical Runge–Kutta method (RK4) of order 4. The modification consists of updating the component  $\zeta_I$  in each intermediate step by using (4.66) to get the approximations to  $\zeta_I(t_{n+1/2})$  and  $\zeta_I(t_{n+1})$ . All other components are updated by the standard RK4 procedure. In contrast to the original RK4 scheme, the modification automatically conserves the  $L^2$ -norm of a Gaussian wave packet.

### 4.6. Numerical experiments

In the following section, we present some numerical examples.

#### 4.6.1. Sublinear magnetic potential in two dimensions

If we consider the potentials

$$A(x, t) = \begin{pmatrix} \sin(x_1 + x_2 + \alpha t) \\ -\sin(x_1 + x_2 + \alpha t) \end{pmatrix}, \quad \phi(x, t) = \sin(x_1 + x_2) \quad (4.72)$$

with  $\alpha \in \{0, 1\}$ , we can calculate the occurring averages  $\langle A \rangle_u, \langle \phi \rangle_u$  analytically as

$$\int_{\mathbb{R}^d} \sin(x_1 + x_2 + \alpha t) |u(x)|^2 dx \quad (4.73)$$

$$= (\pi\varepsilon)^{d/2} \det(L^{-1}) \exp(-\frac{2}{\varepsilon} \zeta_I - \frac{\varepsilon}{4} \mathbb{1}^\top Q Q^* \mathbb{1}) \sin(q_1 + q_2 + \alpha t) \quad (4.74)$$

where  $\mathbb{1} = (1 \ 1)^\top$  and  $(Q Q^*)^{-1} = L L^\top$  is the Cholesky decomposition. Since we compare the new time-integrators with the standard RK4 method, which is not norm-conserving, we do not assume normalization of  $u$ . We use for the curl of a 2d vector potential  $A$  the convention  $\nabla \times A = \partial_1 A_2 - \partial_2 A_1$ . Initial values are chosen as

$$q^0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad p^0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad Q^0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad P^0 = \begin{pmatrix} i & 0 \\ 0 & i \end{pmatrix}, \quad \zeta_R^0 = 0. \quad (4.75)$$

The imaginary part of the phase  $\zeta_I^0$  is chosen such that the corresponding initial Gaussian wave packet  $u^0$  is normalized. Note that the normalization is  $\varepsilon$  dependent and thus the initial value  $\zeta_I^0$  changes for different values of  $\varepsilon$ .

In Figure 4.2 we depict the component errors of a Gaussian wave packet where we solve (4.28) with the Boris-type algorithm 4.7 (left) and the modified fourth order Runge–Kutta (mRK4) method (right). As end time we choose  $T = 8$  and set  $\alpha = 1, \varepsilon = 10^{-3}$ . We compare the

#### 4. Time-integration of Gaussian variational approximation for the magnetic Schrödinger equation

numerical solution to a reference solution calculated by the standard RK4 method applied to (4.15) and (4.16) with time step-size  $\tau = 10^{-4}$ . As we see, the error decreases by order two for the Boris-type method and by order four for the mRK4 method with decreasing time step-size.

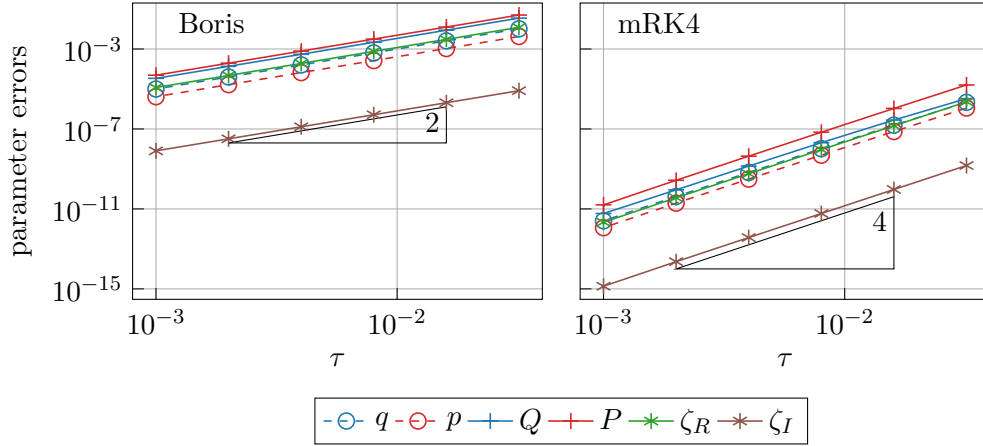


Figure 4.2.: Simulation of the motion of a particle in a magnetic field in dimension two with potentials (4.72) and initial values (4.75). Errors of the numerical solution to (4.28) approximated by the Boris-type algorithm (left) and the mRK4 method (right) measured in the Frobenius norm scaled with the inverse number of entries.

In Figure 4.3 we illustrate the  $L^2$ -error of a Gaussian wave packet with parameters calculated by the Boris-type method (left) and a Gaussian wave packet with parameters calculated by the mRK4 method (right). Again we compute the error against a reference solution calculated with the standard RK4 method applied to (4.15) and (4.16) with time stepsize  $\tau = 10^{-4}$ . The  $L^2$ -norm between the two Gaussian wave packets is computed with a Gauss–Hermite quadrature rule. We see a reduction of order two for the Boris-type method and of order four for the mRK4 method in the  $L^2$ -norm. Moreover, we see that the error constant scales with as  $\varepsilon^{-1}$  in both cases, which is supported by the theoretical result [LL20, Thm. 7.7]. Further note that the  $L^2$ -error between two normed Gaussian wave packet is bounded by 2, which explains the upper plateau in the left plot. The lower plateau in the right plot corresponds to the numerical computation of the underlying integral at almost machine precision.

In Figure 4.4 we compare the error between the energy (B.9) and the initial energy (top) of a Gaussian wave packet between the three methods, Boris-type, mRK4, and standard RK4. We have fixed  $\alpha = 0$  in (4.72) such that we have time-independent potentials and therefore theoretical energy conservation. As end time we chose  $T = 200$  and  $\varepsilon = 10^{-3}$ . On the left, we plot the maximal energy error at each time stamp to the initial energy for different time step-sizes  $\tau$ . We see a reduction of order two of the error for the Boris-type method and of order four for the mRK4 and standard RK4 methods. Note, however, that the error constant of the standard RK4 method is worse than that of the mRK4 method and the Boris-type method. In the right plot, we plot the energy error against the initial error at each time stamp for a fixed time step-size  $\tau = 10^{-1}$ . As we see, the error of the Boris-type method oscillates at the same level, while we see for the standard RK4 and mRK4 methods a drift. Note, however, that the drift of the mRK4 method is way smaller than that of the standard RK4 method. Moreover, at the bottom of Figure 4.4 we illustrate the error of the  $L^2$  norm of a Gaussian

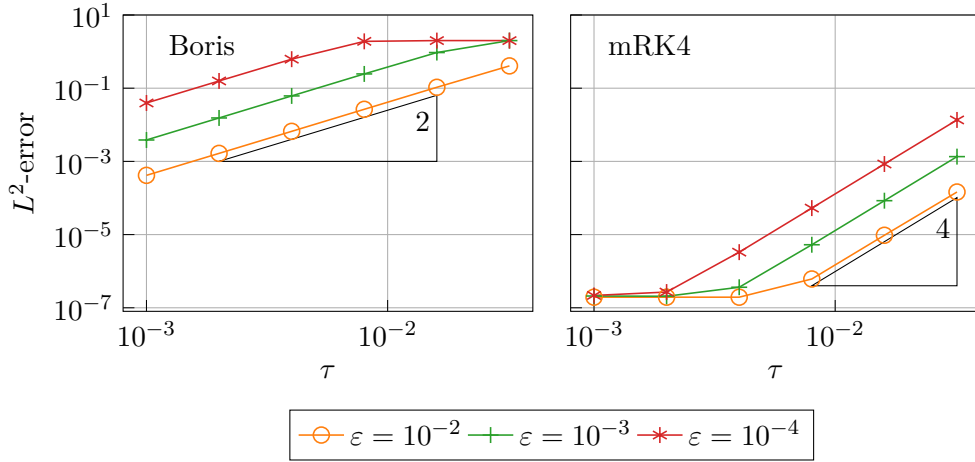


Figure 4.3.:  $L^2$ -error of a Gaussian wave packet approximated by the Boris-type method (4.7) (left) and the mRK4 method (right) against a reference Gaussian wave packet with coefficients approximated by the classical RK4 method with time stepsize  $\tau = 10^{-4}$ . The potentials are given by (4.72) and the initial values by (4.75). Different values for  $\epsilon$  are considered.

wave packet to norm conservation using time stepsize  $\tau = 10^{-1}$ . The error for the Boris-type method and the mRK4 method is close to the machine precision and hence shows conservation of the  $L^2$ -norm, whereby using the standard RK4 method results in a deviation. The  $L^2$ -norm of a Gaussian wave packet given its parameters can be calculated analytically by (4.68).

#### 4.6.2. Penning trap

As a second example, we apply the time-integrators to the three-dimensional quantum dynamics of a proton in a hyperbolic Penning trap, see Section 4.2. We consider the Schrödinger equation (4.5) with the trap parameters of Table 4.1 and an initial Gaussian wave packet with parameters (in dimensionless units)

$$\begin{aligned} q^0 &= \begin{pmatrix} 0.133 & 0.133 & 0.258 \end{pmatrix}^\top, & Q^0 &= \text{diag}(q^0), \\ p^0 &= \begin{pmatrix} 0.133 & 7.492 & 3.879 \end{pmatrix}^\top, & P^0 &= i \cdot \text{diag}(7.492, 7.492, 3.879), \\ \zeta^0 &= 1.009 - 1.84 \cdot 10^{-7}i. \end{aligned} \tag{4.76}$$

The initial condition is chosen such that the dynamics are coherent in the sense, that the width of the packet does not change over time. The phase parameter has a non-vanishing real part to be aligned with the analytic expressions for the center motion that were recently given in [BBB24, eqs. (12) and (17)] (with  $\omega_\perp = \Omega/2$ ). As previously mentioned, see for example Section 4.3.2, the variational approximation is exact in this case, and the evolution does not require averages but only point evaluations, see Remark 4.4. In this set-up, we see the exact errors of our numerical schemes to integrate the equations of motion (4.28) and (4.54).

In Figure 4.1, we showed the exact trajectory of the position center. Virtually the same trajectory is obtained by our numerical simulations for  $\tau \lesssim 10^{-3}$ , as confirmed by the small errors illustrated in Figure 4.5. There, we present the maximal error over all time steps of the

#### 4. Time-integration of Gaussian variational approximation for the magnetic Schrödinger equation

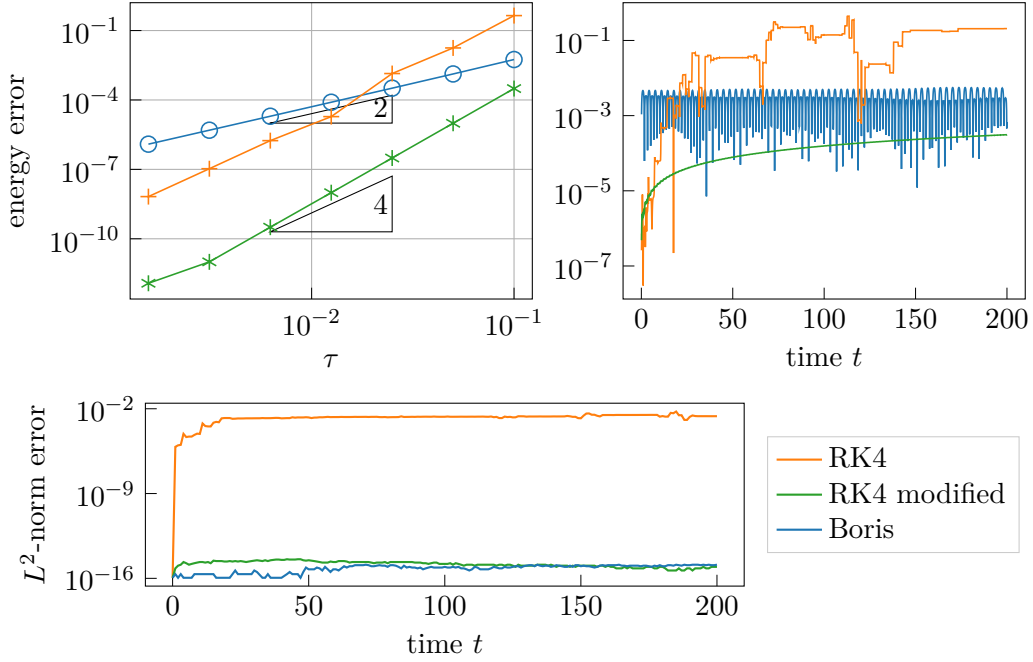


Figure 4.4.: Energy error against the initial energy (top) and  $L^2$ -norm error (bottom) using the potentials and initial values in Section 4.6.1. The endtime is chosen as  $T = 200$  and  $\varepsilon = 10^{-3}$ . On the left, the maximal energy error is illustrated over all time stamps for different time stepsizes  $\tau$ . On the right, the energy error over the time interval  $[0, 200]$  is plotted for a fixed time stepsize  $\tau = 10^{-1}$ .

parameters against the exact solution for different step-sizes  $\tau$  in the Frobenius norm scaled by the inverse number of the component. As expected, the Boris-type method converges with order two while the RK4 and mRK4 methods converge with order four. We are convinced that the plateau at larger time step-sizes occurs since the parameter evolution is mildly oscillatory because the quotients  $\omega_+/\omega_- \approx 113.25$  and  $B/B_m \approx 114.25$  of the potentials in (4.5) are not small, and since we did not observe such plateaus when setting these quotients close to one. Moreover, the error for the imaginary phase  $\zeta_I$  using the RK4 method applied to (4.15) is close to the machine precision since the exact solution is constant. In contrast, in the mRK4 method, the error of  $\zeta_I$  directly relates to the error of  $Q$  by (4.66) and thus shows order four. Finally, in Figure 4.6 we compare the energy error of the three methods, which shows a drift for the RK4 methods but not for the Boris-type algorithm.

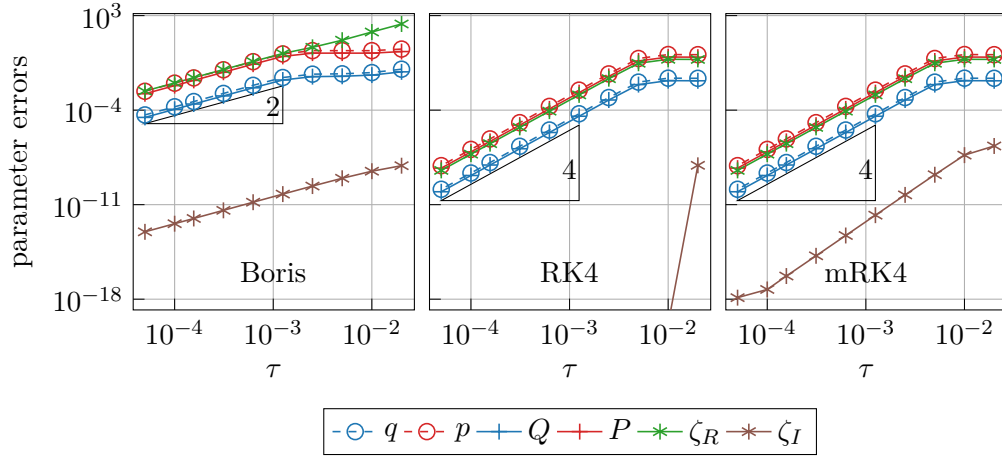


Figure 4.5.: Maximum errors of the parameters approximated by the Boris-type method (4.7) (left), the RK4 method (middle) and the mRK4 method (right) against the exact solution using the potentials (4.2) with data given in Table 4.1 and initial values (4.76). Measured in the Frobenius norm scaled with the inverse number of components.

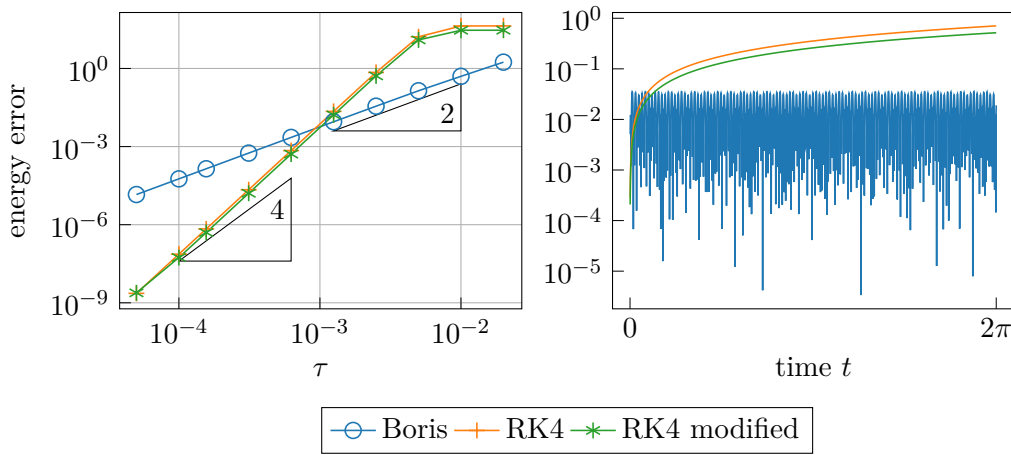


Figure 4.6.: Energy error using the potentials (4.2) and initial values (4.76). The endtime is chosen as  $T = 2\pi$ . On the left, the maximal energy error is plotted for different time stepsizes  $\tau$ . On the right, we plotted the energy error over time for  $\tau = 2.5 \cdot 10^{-3}$ .



## APPENDIX B

---

### Gaussian calculus and energy formula

**Lemma B.1.** *Consider smooth functions  $W: \mathbb{R}^d \rightarrow \mathbb{C}^{d \times d}$  and  $w: \mathbb{R}^d \rightarrow \mathbb{C}$ , and an arbitrary matrix  $M \in \mathbb{C}^{d \times d}$ . Let  $u \in L^2(\mathbb{R}^d)$  be a Gaussian wave packet with position  $q$  and width  $\mathcal{C} = \mathcal{C}_R + i\mathcal{C}_I$ . Then,*

$$\langle W(x - q) \rangle_u = \frac{\varepsilon}{2} \sum_{\ell=1}^d \left( \langle (\partial_\ell W)_u \mathcal{C}_I^{-1} \rangle_{k\ell} \right)_{k=1}^d, \quad (\text{B.1})$$

$$\langle (x - q)^\top w M (x - q) \rangle_u = \frac{\varepsilon}{2} \langle w \rangle_u \text{tr}(M \mathcal{C}_I^{-1}) + \frac{\varepsilon^2}{4} \text{tr}(\langle \nabla^2 w \rangle_u \mathcal{C}_I^{-1} M \mathcal{C}_I^{-1}). \quad (\text{B.2})$$

In particular,  $\langle w(x - q) \rangle_u = \frac{\varepsilon}{2} \mathcal{C}_I^{-1} \langle \nabla w \rangle_u$ .

*Proof.* The position density

$$|u(t, x)|^2 = \exp\left(-\frac{1}{\varepsilon}(x - q)^\top \mathcal{C}_I(x - q) - \frac{2}{\varepsilon} \zeta_I\right)$$

satisfies  $\nabla |u(t, x)|^2 = -\frac{2}{\varepsilon} \mathcal{C}_I(x - q) |u(t, x)|^2$ . Therefore, partial integration yields that

$$\langle W(x - q) \rangle_u = -\frac{\varepsilon}{2} \int_{\mathbb{R}^d} W(x) \mathcal{C}_I^{-1} \nabla |u(t, x)|^2 dx \quad (\text{B.3})$$

$$= \frac{\varepsilon}{2} \sum_{\ell=1}^d \left( \langle \partial_\ell (W \mathcal{C}_I^{-1})_{k,\ell} \rangle_u \right)_{k=1}^d. \quad (\text{B.4})$$

Similarly, two partial integrations imply that

$$\langle w(x - q)^\top M (x - q) \rangle_u = \sum_{k,\ell=1}^d \left\langle w(\mathcal{C}_I(x - q))_k (\mathcal{C}_I^{-1} M \mathcal{C}_I^{-1})_{k\ell} (\mathcal{C}_I(x - q))_\ell \right\rangle_u \quad (\text{B.5})$$

$$= \frac{\varepsilon}{2} \sum_{k,\ell=1}^d \left( \langle w \rangle_u (\mathcal{C}_I)_{k\ell} + \langle \partial_\ell w(\mathcal{C}_I(x - q))_k \rangle_u \right) (\mathcal{C}_I^{-1} M \mathcal{C}_I^{-1})_{k\ell} \quad (\text{B.6})$$

$$= \sum_{k,\ell=1}^d \left( \frac{\varepsilon}{2} \langle w \rangle_u (\mathcal{C}_I)_{k\ell} + \frac{\varepsilon^2}{4} \langle \partial_k \partial_\ell w \rangle_u \right) (\mathcal{C}_I^{-1} M \mathcal{C}_I^{-1})_{k\ell} \quad (\text{B.7})$$

$$= \frac{\varepsilon}{2} \langle w \rangle_u \operatorname{tr} (M \mathcal{C}_I^{-1}) + \frac{\varepsilon^2}{4} \operatorname{tr} (\langle \nabla^2 w \rangle_u \mathcal{C}_I^{-1} M \mathcal{C}_I^{-1}). \quad (\text{B.8})$$

□

**Lemma B.2.** *Let  $u$  be a normalized Gaussian wave packet with center  $(q, p)$  and width  $\mathcal{C} = \mathcal{C}_R + i\mathcal{C}_I$ . Then it holds for the energy*

$$\langle H \rangle_u = \frac{1}{2} p^2 - p^\top \langle A \rangle_u + \frac{1}{2} \langle A^2 \rangle_u + \langle \phi \rangle_u + \frac{\varepsilon}{4} R_H \quad (\text{B.9})$$

with remainder

$$R_H = \operatorname{tr} \left( (\mathcal{C}_R^2 + \mathcal{C}_I^2 - 2 \langle J_A^\top \rangle_u \mathcal{C}_R) \mathcal{C}_I^{-1} \right).$$

If the magnetic potential  $A$  is linear in  $x$ , then  $\langle A \rangle_u = A(q)$  and

$$\langle A^2 \rangle_u = A(q)^2 + \frac{\varepsilon}{2} \operatorname{tr} \left( J_A(q)^\top J_A(q) \mathcal{C}_I^{-1} \right) \quad (\text{B.10})$$

*Proof.* In the following we ignore the dependence on  $t$  to simplify the notation. Since  $\langle H \rangle_u = \langle h \rangle_u = \frac{1}{2} \langle (\xi - A)^2 \rangle_u + \langle \phi \rangle_u$ , we only need to work on the mean of  $\operatorname{op}_{\text{Weyl}}((\xi - A)^2) = (\operatorname{op}_{\text{Weyl}}(\xi - A))^2$ . We have

$$\begin{aligned} \langle (\xi - A)^2 \rangle_u &= \langle (\xi - A)u, (\xi - A)u \rangle \\ &= \langle (\mathcal{C}(x - q) + (p - A))u, (\mathcal{C}(x - q) + (p - A))u \rangle \\ &= \langle (x - q)^\top \mathcal{C}^* \mathcal{C} (x - q) \rangle_u + \langle (p - A)^\top (\mathcal{C} + \mathcal{C}^*) (x - q) \rangle_u + \langle (p - A)^2 \rangle_u \\ &= \frac{\varepsilon}{2} \operatorname{tr} (\mathcal{C}^* \mathcal{C} \mathcal{C}_I^{-1}) - \frac{\varepsilon}{2} \operatorname{tr} (\langle J_A^\top \rangle_u (\mathcal{C} + \mathcal{C}^*) \mathcal{C}_I^{-1}) + \langle (p - A)^2 \rangle_u \end{aligned}$$

due to Lemma B.1. For the traces, we have

$$\begin{aligned} \operatorname{tr} (\mathcal{C}^* \mathcal{C} \mathcal{C}_I^{-1}) &= \operatorname{tr} ((\mathcal{C}_R - i\mathcal{C}_I)(\mathcal{C}_R + i\mathcal{C}_I) \mathcal{C}_I^{-1}) = \operatorname{tr} ((\mathcal{C}_R^2 + \mathcal{C}_I^2) \mathcal{C}_I^{-1}) \\ \operatorname{tr} (\langle J_A^\top \rangle_u (\mathcal{C} + \mathcal{C}^*) \mathcal{C}_I^{-1}) &= 2 \operatorname{tr} (\langle J_A^\top \rangle_u \mathcal{C}_R \mathcal{C}_I^{-1}). \end{aligned}$$

Combining the terms, we obtain (B.9). In the linear case, we expand  $A(x) = A(q) + J_A(q)^\top (x - q)$  and use Lemma B.1 to prove (B.10). □

## Bibliography

- [AA02] D.A. ARULIAH and U.M. ASCHER. Multigrid preconditioning for Krylov methods for time-harmonic Maxwell's equations in three dimensions. *SIAM J. Sci. Comput.*, 2002, vol. 24(2), pp. 702–718. ISSN 1064-8275. <http://dx.doi.org/10.1137/S1064827501387358>.
- [ABB<sup>+</sup>23] D. ARNDT, W. BANGERTH, M. BERGBAUER, M. FEDER, M. FEHLING, J. HEINZ, T. HEISTER, L. HELTAI, M. KRONBICHLER, M. MAIER, P. MUNCH, J.P. PELTERET, B. TURCK SIN, D. WELLS, and S. ZAMPINI. The deal.II Library, Version 9.5. *Journal of Numerical Mathematics*, 2023, vol. 31(3), pp. 231–246. <http://dx.doi.org/10.1515/jnma-2023-0089>.
- [ABB<sup>+</sup>25] D. ARNDT, W. BANGERTH, M. BERGBAUER, B. BLAIS, M. FEHLING, R. GASSMÖLLER, T. HEISTER, L. HELTAI, M. KRONBICHLER, M. MAIER, P. MUNCH, S. SCHEUERMAN, B. TURCK SIN, S. UZUNBAJAKAU, D. WELLS, and M. WICHROWSKI. The deal.II library, Version 9.7. *Journal of Numerical Mathematics*, 2025, vol. 33(4), pp. 403–415. <http://dx.doi.org/10.1515/jnma-2025-0115>.
- [AHZ17] J.H. ADLER, X. HU, and L.T. ZIKATANOV. Robust solvers for Maxwell's equations with dissipative boundary conditions. *SIAM J. Sci. Comput.*, 2017, vol. 39(5), pp. S3–S23. ISSN 1064-8275. <http://dx.doi.org/10.1137/16M1073339>.
- [AM17] M. ALMQUIST and M. MEHLIN. Multilevel local time-stepping methods of Runge-Kutta-type for wave equations. *SIAM J. Sci. Comput.*, 2017, vol. 39(5), pp. A2020–A2048. ISSN 1064-8275. <http://dx.doi.org/10.1137/16M1084407>.
- [Bañ10] L. BAÑAS. An efficient multigrid preconditioner for Maxwell's equations in micromagnetism. *Math. Comput. Simulation*, 2010, vol. 80(8), pp. 1657–1663. ISSN 0378-4754. <http://dx.doi.org/10.1016/j.matcom.2009.02.009>.
- [BBBB24] I. BIALYNICKI-BIRULA and Z. BIALYNICKA-BIRULA. Classical-quantum correspondence for particles in the Penning trap. *Physica Scripta*, 2024, vol. 99(4), p. 045103. <http://dx.doi.org/10.1088/1402-4896/ad2c43>.
- [BDG<sup>+</sup>19] M. BONAZZOLI, V. DOLEAN, I.G. GRAHAM, E.A. SPENCE, and P.H. TOURNIER. Domain decomposition preconditioning for the high-frequency time-harmonic Maxwell equations with absorption. *Math. Comp.*, 2019, vol. 88(320), pp. 2559–2604. ISSN 0025-5718. <http://dx.doi.org/10.1090/mcom/3447>.
- [BDH<sup>+</sup>24] S. BURKHARD, B. DÖRICH, M. HOCHBRUCK, and C. LASSER. Variational Gaussian approximation for the magnetic Schrödinger equation. *J. Phys. A: Math. Theor.*, 2024, vol. 57(29), p. 295202. ISSN 1751-8121. <http://dx.doi.org/10.1088/1751-8121/ad591e>.
- [BG84] L.S. BROWN and G. GABRIELSE. Geonium theory: Physics of a single electron or ion in a Penning trap. *Rev. Mod. Phys.*, 1984, vol. 58(1), pp. 233–311. <http://dx.doi.org/10.1103/RevModPhys.58.233>.
- [BL18] C. BIRDSALL and A. LANGDON. *Plasma Physics via Computer Simulation*. Series in Plasma Physics. CRC Press, Boca Raton, FL, 2018, 1st ed. ISBN 978-1-4822-6306-0. <http://dx.doi.org/10.1201/9781315275048>.

- [Bor70] J.P. BORIS. Relativistic plasma simulation—optimization of a hybrid code. In: J.P. BORIS and R.A. SHANNY (Eds.), *Proceedings of the Fourth Conference on Numerical Simulation of Plasmas*. Naval Research Laboratory, Washington, DC, 1970 pp. 3–67. <https://apps.dtic.mil/sti/citations/ADA023511>.
- [Car21] C. CARLE. *On leapfrog-Chebyshev schemes for second-order differential equations*. Ph.D. thesis, Karlsruhe Institute of Technology (KIT), 2021. <https://doi.org/10.5445/IR/1000147725>.
- [CDK<sup>+</sup>] C. CARLE, J. DÖRNER, J. KÖHLER, J. LEIBOLD, and B. MAIER. TiMaxdG. <https://gitlab.kit.edu/kit/ianm/ag-numerik/projects/dg-maxwell/timaxdg>.
- [CH22] C. CARLE and M. HOCHBRUCK. Error analysis of multirate leapfrog-type methods for second-order semilinear ODEs. *SIAM J. Numer. Anal.*, 2022, vol. 60(5), pp. 2897–2924. ISSN 0036-1429. <http://dx.doi.org/10.1137/21M1427255>.
- [CH24] C. CARLE and M. HOCHBRUCK. Error analysis of second-order local time integration methods for discontinuous Galerkin discretizations of linear wave equations. *Math. Comp.*, 2024, vol. 93(350), pp. 2611–2641. ISSN 0025-5718,1088-6842. <http://dx.doi.org/10.1090/mcom/3952>.
- [CHS20] C. CARLE, M. HOCHBRUCK, and A. STURM. On leapfrog-Chebyshev schemes. *SIAM J. Numer. Anal.*, 2020, vol. 58(4), pp. 2404–2433. ISSN 0036-1429. <http://dx.doi.org/10.1137/18M1209453>.
- [CI16] J. CHABASSIER and S. IMPERIALE. Fourth-order energy-preserving locally implicit time discretization for linear wave equations. *Internat. J. Numer. Methods Engrg.*, 2016, vol. 106(8), pp. 593–622. ISSN 0029-5981. <http://dx.doi.org/10.1002/nme.5130>.
- [CP17] M. CROUZEIX and C. PALENCIA. The numerical range is a  $(1 + \sqrt{2})$ -spectral set. *SIAM J. Matrix Anal. Appl.*, 2017, vol. 38(2), pp. 649–655. ISSN 0895-4798. <http://dx.doi.org/10.1137/17M1116672>.
- [CW17] J. CHAN and T. WARBURTON. On the penalty stabilization mechanism for upwind discontinuous Galerkin formulations of first order hyperbolic systems. *Comput. Math. Appl.*, 2017, vol. 74(12), pp. 3099–3110. ISSN 0898-1221. <http://dx.doi.org/10.1016/j.camwa.2017.08.003>.
- [DFF<sup>+</sup>10] V. DOLEAN, H. FAHS, L. FEZOU, and S. LANTERI. Locally implicit discontinuous Galerkin method for time domain electromagnetics. *J. Comput. Phys.*, 2010, vol. 229(2), pp. 512–526. ISSN 0021-9991. <http://dx.doi.org/10.1016/j.jcp.2009.09.038>.
- [DG09] J. DIAZ and M.J. GROTE. Energy conserving explicit local time stepping for second-order wave equations. *SIAM J. Sci. Comput.*, 2009, vol. 31(3), pp. 1985–2014. ISSN 1064-8275,1095-7197. <http://dx.doi.org/10.1137/070709414>.
- [DG15] J. DIAZ and M.J. GROTE. Multi-level explicit local time-stepping methods for second-order wave equations. *Comput. Methods Appl. Mech. Engrg.*, 2015, vol. 291, pp. 240–265. ISSN 0045-7825. <http://dx.doi.org/10.1016/j.cma.2015.03.027>.
- [DHK<sup>+</sup>23] W. DÖRFLER, M. HOCHBRUCK, J. KÖHLER, A. RIEDER, R. SCHNAUBELT, and C. WIENERS. *Wave Phenomena: Mathematical Analysis and Numerical Approximation, Oberwolfach Seminars*, vol. 49. Springer International Publishing, Cham, 2023. ISBN 978-3-031-05792-2 978-3-031-05793-9. <https://link.springer.com/10.1007/978-3-031-05793-9>.
- [Dir30] P.A.M. DIRAC. Note on Exchange Phenomena in the Thomas Atom. *Mathematical Proceedings of the Cambridge Philosophical Society*, 1930, vol. 26(3), p. 376–385. <http://dx.doi.org/10.1017/S0305004100016108>.

- [DLM13] S. DESCOMBES, S. LANTERI, and L. MOYA. Locally implicit time integration strategies in a discontinuous Galerkin method for Maxwell’s equations. *J. Sci. Comput.*, 2013, vol. 56(1), pp. 190–218. ISSN 0885-7474. <http://dx.doi.org/10.1007/s10915-012-9669-5>.
- [DLM14] S. DESCOMBES, S. LANTERI, and L. MOYA. High-order locally implicit time integration strategies in a discontinuous Galerkin method for Maxwell’s equations. In: *Spectral and high order methods for partial differential equations—ICOSAHOM 2012, Lect. Notes Comput. Sci. Eng.*, vol. 95. Springer, Cham, 2014, pp. 205–215. [http://dx.doi.org/10.1007/978-3-319-01601-6\\_16](http://dx.doi.org/10.1007/978-3-319-01601-6_16).
- [DLM17] S. DESCOMBES, S. LANTERI, and L. MOYA. Temporal convergence analysis of a locally implicit discontinuous Galerkin time domain method for electromagnetic wave propagation in dispersive media. *J. Comput. Appl. Math.*, 2017, vol. 316, pp. 122–132. ISSN 0377-0427. <http://dx.doi.org/10.1016/j.cam.2016.09.038>.
- [DPE12] D.A. DI PIETRO and A. ERN. *Mathematical aspects of discontinuous Galerkin methods, Mathématiques & Applications (Berlin) [Mathematics & Applications]*, vol. 69. Springer, Heidelberg, 2012, pp. xviii+384. ISBN 978-3-642-22979-4. <http://dx.doi.org/10.1007/978-3-642-22980-0>.
- [Dri05] T.A. DRISCOLL. Algorithm 843: improvements to the Schwarz-Christoffel toolbox for MATLAB. *ACM Trans. Math. Software*, 2005, vol. 31(2), pp. 239–251. ISSN 0098-3500. <http://dx.doi.org/10.1145/1067967.1067971>.
- [EG06a] A. ERN and J.L. GUERMOND. Discontinuous Galerkin methods for Friedrichs’ systems. I. General theory. *SIAM J. Numer. Anal.*, 2006, vol. 44(2), pp. 753–778. ISSN 0036-1429. <http://dx.doi.org/10.1137/050624133>.
- [EG06b] A. ERN and J.L. GUERMOND. Discontinuous Galerkin methods for Friedrichs’ systems. II. Second-order elliptic PDEs. *SIAM J. Numer. Anal.*, 2006, vol. 44(6), pp. 2363–2388. ISSN 0036-1429. <http://dx.doi.org/10.1137/05063831X>.
- [EG21] A. ERN and J.L. GUERMOND. *Finite elements III—first-order and time-dependent PDEs, Texts in Applied Mathematics*, vol. 74. Springer, Cham, [2021] ©2021, pp. viii+417. ISBN 978-3-030-57347-8; 978-3-030-57348-5. <http://dx.doi.org/10.1007/978-3-030-57348-5>.
- [Eie89] M. EIERMANN. On semiiterative methods generated by Faber polynomials. *Numer. Math.*, 1989, vol. 56(2-3), pp. 139–156. ISSN 0029-599X. <http://dx.doi.org/10.1007/BF01409782>.
- [FGN92] R.W. FREUND, G.H. GOLUB, and N.M. NACHTIGAL. Iterative solution of linear systems. In: *Acta numerica, 1992*, Acta Numer. Cambridge Univ. Press, Cambridge, 1992, pp. 57–100. <http://dx.doi.org/10.1017/s0962492900002245>.
- [FGN93] R.W. FREUND, M.H. GUTKNECHT, and N.M. NACHTIGAL. An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices. *SIAM J. Sci. Comput.*, 1993, vol. 14(1), pp. 137–158. ISSN 1064-8275. <http://dx.doi.org/10.1137/0914009>.
- [FML<sup>+</sup>26] M. FUNK, S. MERK, C. LASSER, M. HOCHBRUCK, and N. TALEBI. Meshfree versus grid-based Schrödinger solvers for modeling the interactions between free-electron wave packets and light. CRC 1173 Preprint 2026/24, Karlsruhe Institute of Technology, 2026. <http://dx.doi.org/10.5445/IR/1000194959>.
- [FN91] R.W. FREUND and N.M. NACHTIGAL. QMR: a quasi-minimal residual method for non-Hermitian linear systems. *Numer. Math.*, 1991, vol. 60(3), pp. 315–339. ISSN 0029-599X. <http://dx.doi.org/10.1007/BF01385726>.
- [FN94] R.W. FREUND and N.M. NACHTIGAL. An implementation of the QMR method based on coupled two-term recurrences. *SIAM J. Sci. Comput.*, 1994, vol. 15(2), pp. 313–337. ISSN 1064-8275. Iterative methods in numerical linear algebra (Copper Mountain Resort, CO, 1992), <http://dx.doi.org/10.1137/0915022>.

- [Fre34] J. FRENKEL. *Wave mechanics: Advanced general theory*. Clarendon Press, Oxford, 1934.
- [Fre92] R.W. FREUND. Conjugate gradient-type methods for linear systems with complex symmetric coefficient matrices. *SIAM J. Sci. Statist. Comput.*, 1992, vol. 13(1), pp. 425–448. ISSN 0196-5204. <http://dx.doi.org/10.1137/0913023>.
- [Fri58] K.O. FRIEDRICHS. Symmetric Positive Linear Differential Equations. *Communications on Pure and Applied Mathematics*, 1958, vol. 11(3), pp. 333–418. ISSN 1097-0312. <https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160110306>.
- [Gau97] W. GAUTSCHI. *Numerical analysis*. Birkhäuser Boston, Inc., Boston, MA, 1997, pp. xiv+506. An introduction. ISBN 0-8176-3895-4.
- [GM10] M.J. GROTE and T. MITKOVA. Explicit local time-stepping methods for Maxwell’s equations. *J. Comput. Appl. Math.*, 2010, vol. 234(12), pp. 3283–3302. ISSN 0377-0427,1879-1778. <http://dx.doi.org/10.1016/j.cam.2010.04.028>.
- [GM13a] M.J. GROTE and T. MITKOVA. Explicit local time-stepping methods for time-dependent wave propagation. In: *Direct and inverse problems in wave propagation and applications, Radon Ser. Comput. Appl. Math.*, vol. 14. De Gruyter, Berlin, 2013, pp. 187–218. <http://dx.doi.org/10.1515/9783110282283.187>.
- [GM13b] M.J. GROTE and T. MITKOVA. High-order explicit local time-stepping methods for damped wave equations. *J. Comput. Appl. Math.*, 2013, vol. 239, pp. 270–289. ISSN 0377-0427. <http://dx.doi.org/10.1016/j.cam.2012.09.046>.
- [GMM15] M.J. GROTE, M. MEHLIN, and T. MITKOVA. Runge-Kutta-based explicit local time-stepping methods for wave propagation. *SIAM J. Sci. Comput.*, 2015, vol. 37(2), pp. A747–A775. ISSN 1064-8275,1095-7197. <http://dx.doi.org/10.1137/140958293>.
- [GMS21] M.J. GROTE, S. MICHEL, and S.A. SAUTER. Stabilized leapfrog based local time-stepping method for the wave equation. *Math. Comp.*, 2021, vol. 90(332), pp. 2603–2643. ISSN 0025-5718,1088-6842. <http://dx.doi.org/10.1090/mcom/3650>.
- [GR21] V. GRADINARU and O. RIETMANN. Hagedorn wavepackets and Schrödinger equation with time-dependent, homogeneous magnetic field. *J. Comput. Phys.*, 2021, vol. 445, pp. Paper No. 110581, 18. ISSN 0021-9991. <http://dx.doi.org/10.1016/j.jcp.2021.110581>.
- [Gut97] M.H. GUTKNECHT. Lanczos-type solvers for nonsymmetric linear systems of equations. In: *Acta numerica, 1997, Acta Numer.*, vol. 6. Cambridge Univ. Press, Cambridge, 1997, pp. 271–397. <http://dx.doi.org/10.1017/S0962492900002737>.
- [Hac16] W. HACKBUSCH. *Iterative solution of large sparse systems of equations, Applied Mathematical Sciences*, vol. 95. Springer, [Cham], 2016, second ed., pp. xxiii+509. ISBN 978-3-319-28481-1; 978-3-319-28483-5. <http://dx.doi.org/10.1007/978-3-319-28483-5>.
- [Hag98] G.A. HAGEDORN. Raising and lowering operators for semiclassical wave packets. *Ann. Physics*, 1998, vol. 269(1), pp. 77–104. ISSN 0003-4916. <http://dx.doi.org/10.1006/aphy.1998.5843>.
- [Hel76] E.J. HELLER. Time dependent variational approach to semiclassical dynamics. *J. Chem. Phys.*, 1976, vol. 64(1), pp. 63–73. ISSN 0021-9606. <http://dx.doi.org/10.1063/1.431911>.
- [Hel81] E.J. HELLER. Frozen Gaussians: a very simple semiclassical approximation. *J. Chem. Phys.*, 1981, vol. 75(6), pp. 2923–2931. ISSN 0021-9606. <http://dx.doi.org/10.1063/1.442382>.
- [Hig08] N.J. HIGHAM. *Functions of matrices*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008, pp. xx+425. Theory and computation. ISBN 978-0-89871-646-7. <http://dx.doi.org/10.1137/1.9780898717778>.

- [Hip99] R. HIPTMAIR. Multigrid method for Maxwell's equations. *SIAM J. Numer. Anal.*, 1999, vol. 36(1), pp. 204–225. ISSN 0036-1429. <http://dx.doi.org/10.1137/S0036142997326203>.
- [HK84] M.F. HERMAN and E. KLUK. A semiclassical justification for the use of non-spreading wavepackets in dynamics calculations. *Chem. Phys.*, 1984, vol. 91(1), pp. 27–34. ISSN 0301-0104. [http://dx.doi.org/10.1016/0301-0104\(84\)80039-7](http://dx.doi.org/10.1016/0301-0104(84)80039-7).
- [HK19] M. HOCHBRUCK and J. KÖHLER. On the efficiency of the Peaceman-Rachford ADI-dG method for wave-type problems. In: *Numerical mathematics and advanced applications—ENUMATH 2017, Lect. Notes Comput. Sci. Eng.*, vol. 126. Springer, Cham, 2019, pp. 135–144. [http://dx.doi.org/10.1007/978-3-319-96415-7\\_10](http://dx.doi.org/10.1007/978-3-319-96415-7_10).
- [HK20] M. HOCHBRUCK and J. KÖHLER. Error analysis of discontinuous Galerkin discretizations of a class of linear wave-type problems. In: W. DÖRFLER, M. HOCHBRUCK, D. HUNDERTMARK, W. REICHEL, A. RIEDER, R. SCHNAUBELT, and B. SCHÖRKHUBER (Eds.), *Mathematics of Wave Phenomena*, Trends in Mathematics. Birkhäuser Basel, 2020 pp. 197–218. [http://dx.doi.org/10.1007/978-3-030-47174-3\\_12](http://dx.doi.org/10.1007/978-3-030-47174-3_12).
- [HK22] M. HOCHBRUCK and J. KÖHLER. Error analysis of a fully discrete discontinuous Galerkin alternating direction implicit discretization of a class of linear wave-type problems. *Numer. Math.*, 2022, vol. 150(3), pp. 893–927. ISSN 0029-599X. <http://dx.doi.org/10.1007/s00211-021-01262-z>.
- [HKK22] M. HOCHBRUCK, J. KÖHLER, and P.M. KUMBHAR. Preconditioned implicit time integration schemes for Maxwell's equations on locally refined grids. CRC 1173 Preprint 2022/29, Karlsruhe Institute of Technology, 2022. <http://dx.doi.org/10.5445/IR/1000148078>.
- [HKS26] M. HOCHBRUCK, J. KÖHLER, and M. SCHEIFINGER. Efficient higher-order local time integration for Friedrichs' systems, 2026, arXiv:2607.15192 [math.NA]. <http://arxiv.org/abs/2607.15192>, <https://arxiv.org/abs/2607.15192>.
- [HL97] M. HOCHBRUCK and C. LUBICH. On Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.*, 1997, vol. 34(5), pp. 1911–1925. ISSN 0036-1429. <http://dx.doi.org/10.1137/S0036142995280572>.
- [HL98] M. HOCHBRUCK and C. LUBICH. Error analysis of Krylov methods in a nutshell. *SIAM J. Sci. Comput.*, 1998, vol. 19(2), pp. 695–701. ISSN 1064-8275. <http://dx.doi.org/10.1137/S1064827595290450>.
- [HL20] E. HAIRER and C. LUBICH. Long-term analysis of a variational integrator for charged-particle dynamics in a strong magnetic field. *Numer. Math.*, 2020, vol. 144(3), pp. 699–728. ISSN 0029-599X. <http://dx.doi.org/10.1007/s00211-019-01093-z>.
- [HLW06] E. HAIRER, C. LUBICH, and G. WANNER. *Geometric numerical integration, Springer Series in Computational Mathematics*, vol. 31. Springer-Verlag, Berlin, 2006, second ed., pp. xviii+644. Structure-preserving algorithms for ordinary differential equations. ISBN 3-540-30663-3; 978-3-540-30663-4.
- [HLW20] E. HAIRER, C. LUBICH, and B. WANG. A filtered Boris algorithm for charged-particle dynamics in a strong magnetic field. *Numer. Math.*, 2020, vol. 144(4), pp. 787–809. ISSN 0029-599X, 0945-3245. <http://dx.doi.org/10.1007/s00211-020-01105-3>.
- [HP15] M. HOCHBRUCK and T. PAŽUR. Implicit Runge-Kutta methods and discontinuous Galerkin discretizations for linear Maxwell's equations. *SIAM J. Numer. Anal.*, 2015, vol. 53(1), pp. 485–507. ISSN 0036-1429. <http://dx.doi.org/10.1137/130944114>.
- [HPS18] M. HOCHBRUCK, T. PAŽUR, and R. SCHNAUBELT. Error analysis of implicit Runge-Kutta methods for quasilinear hyperbolic evolution equations. *Numer. Math.*, 2018, vol. 138(3), pp. 557–579. ISSN 0029-599X, 0945-3245. <http://dx.doi.org/10.1007/s00211-017-0914-6>.

- [HS16] M. HOCHBRUCK and A. STURM. Error analysis of a second-order locally implicit method for linear Maxwell’s equations. *SIAM J. Numer. Anal.*, 2016, vol. 54(5), pp. 3167–3191. ISSN 0036-1429. <http://dx.doi.org/10.1137/15M1038037>.
- [HS19] M. HOCHBRUCK and A. STURM. Upwind discontinuous Galerkin space discretization and locally implicit time integration for linear Maxwell’s equations. *Math. Comp.*, 2019, vol. 88(317), pp. 1121–1153. ISSN 0025-5718,1088-6842. <http://dx.doi.org/10.1090/mcom/3365>.
- [HS26] M. HOCHBRUCK and M. SCHEIFINGER. Local Time Integration for Friedrichs’ Systems. *SIAM J. Numer. Anal.*, 2026, vol. 64(2), pp. 370–390. ISSN 0036-1429,1095-7170. <http://dx.doi.org/10.1137/25M1735627>.
- [HW96] E. HAIRER and G. WANNER. *Solving ordinary differential equations. II, Springer Series in Computational Mathematics*, vol. 14. Springer-Verlag, Berlin, 1996, second ed., pp. xvi+614. Stiff and differential-algebraic problems. ISBN 3-540-60452-9. <http://dx.doi.org/10.1007/978-3-642-05221-7>.
- [HW08] J.S. HESTHAVEN and T. WARBURTON. *Nodal discontinuous Galerkin methods, Texts in Applied Mathematics*, vol. 54. Springer, New York, 2008. Algorithms, analysis, and applications. ISBN 978-0-387-72065-4. <http://dx.doi.org/10.1007/978-0-387-72067-8>.
- [KFL20] M. KOTOVSHCHIKOVA, D.K. FIRSOV, and S.H. LUI. A third-order multirate Runge-Kutta scheme for finite volume solution of 3D time-dependent Maxwell’s equations. *Commun. Appl. Math. Comput. Sci.*, 2020, vol. 15(1), pp. 65–87. ISSN 1559-3940. <http://dx.doi.org/10.2140/camcos.2020.15.65>.
- [Köh19] J. KÖHLER. *The Peaceman–Rachford ADI-dG method for linear wave-type problems*. Dissertation, Karlsruhe Institute of Technology (KIT), 2019. <http://dx.doi.org/10.5445/IR/1000089271>.
- [Kre14] M. KRETZSCHMAR. Octupolar excitation of ion motion in a Penning trap: A theoretical study. *International Journal of Mass Spectrometry*, 2014, vol. 357, pp. 1–21. ISSN 1387-3806. <http://dx.doi.org/https://doi.org/10.1016/j.ijms.2013.09.007>.
- [LL20] C. LASSER and C. LUBICH. Computing quantum dynamics in the semiclassical regime. *Acta Numer.*, 2020, vol. 29, pp. 229–401. ISSN 0962-4929. <http://dx.doi.org/10.1017/s0962492920000033>.
- [LR74] P. LESAIN and P.A. RAVIART. On a finite element method for solving the neutron transport equation. In: C. DE BOOR (Ed.), *Mathematical aspects of finite elements in partial differential equations*. Academic Press, New York, 1974, pp. 89–123.
- [LS17] C. LASSER and D. SATTLEGGGER. Discretising the Herman-Kluk propagator. *Numer. Math.*, 2017, vol. 137(1), pp. 119–157. ISSN 0029-599X,0945-3245. <http://dx.doi.org/10.1007/s00211-017-0871-0>.
- [Lub08] C. LUBICH. *From quantum to classical molecular dynamics: reduced models and numerical analysis*. Zurich Lectures in Advanced Mathematics. European Mathematical Society (EMS), Zürich, 2008, pp. x+144. ISBN 978-3-03719-067-8. <http://dx.doi.org/10.4171/067>.
- [ML26] S. MERK and C. LASSER. Structure-Preserving Integration for Magnetic Gaussian Wave Packet Dynamics, 2026, arXiv:2603.25596 [math.NA]. <http://arxiv.org/abs/2603.25596>, <https://arxiv.org/abs/2603.25596>.
- [MPF<sup>+</sup>08] E. MONTSENY, S. PERNET, X. FERRIÈRES, and G. COHEN. Dissipative terms and local time-stepping improvements in a spatial high order discontinuous Galerkin scheme for the time-domain Maxwell’s equations. *J. Comput. Phys.*, 2008, vol. 227(14), pp. 6795–6820. ISSN 0021-9991. <http://dx.doi.org/10.1016/j.jcp.2008.03.032>.

- [NDB12] J. NIEGEMANN, R. DIEHL, and K. BUSCH. Efficient low-storage Runge-Kutta schemes with optimized stability regions. *J. Comput. Phys.*, 2012, vol. 231(2), pp. 364–372. ISSN 0021-9991,1090-2716. <http://dx.doi.org/10.1016/j.jcp.2011.09.003>.
- [Nie09] J. NIEGEMANN. *Higher-Order Methods for Solving Maxwell's Equations in the Time-Domain*. Dissertation, Universität Karlsruhe (TH), 2009. <http://dx.doi.org/10.5445/IR/1000011812>.
- [Pip06] S. PIPERNO. Symplectic local time-stepping in non-dissipative DGTD methods applied to wave propagation problems. *M2AN Math. Model. Numer. Anal.*, 2006, vol. 40(5), pp. 815–841. ISSN 0764-583X. <http://dx.doi.org/10.1051/m2an:2006035>.
- [PSC18] E.G. PHILLIPS, J.N. SHADID, and E.C. CYR. Scalable preconditioners for structure preserving discretizations of Maxwell equations in first order form. *SIAM J. Sci. Comput.*, 2018, vol. 40(3), pp. B723–B742. ISSN 1064-8275. <http://dx.doi.org/10.1137/17M1135827>.
- [QZX<sup>+</sup>13] H. QIN, S. ZHANG, J. XIAO, J. LIU, Y. SUN, and W.M. TANG. Why is Boris algorithm so good? *Physics of Plasmas*, 2013, vol. 20(8), 084503. <http://dx.doi.org/10.1063/1.4818428>.
- [RB02] T. RYLANDER and A. BONDESON. Stability of explicit–implicit hybrid time-stepping schemes for Maxwell's equations. *J. Comput. Phys.*, 2002, vol. 179(2), pp. 426–438. ISSN 0021-9991. <http://dx.doi.org/https://doi.org/10.1006/jcph.2002.7063>.
- [RC21] D. ROBERT and M. COMBESCURE. *Coherent states and applications in mathematical physics*. Theoretical and Mathematical Physics. Springer, Cham, 2021, pp. xvii+576. Second edition [of 2952171]. ISBN 978-3-030-70844-3; 978-3-030-70845-0. <http://dx.doi.org/10.1007/978-3-030-70845-0>.
- [RH73] W.H. REED and T.R. HILL. Triangular mesh methods for the neutron transport equation. Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.
- [Saa03] Y. SAAD. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2003, second ed., pp. xviii+528. ISBN 0-89871-534-2. <http://dx.doi.org/10.1137/1.9780898718003>.
- [SBH<sup>+</sup>25] M. SCHEIFINGER, K. BUSCH, M. HOCHBRUCK, and C. LASSER. Time-integration of Gaussian variational approximation for the magnetic Schrödinger equation. *J. Comput. Phys.*, 2025, vol. 541, pp. Paper No. 114349, 16. ISSN 0021-9991,1090-2716. <http://dx.doi.org/10.1016/j.jcp.2025.114349>.
- [SCGT07] A. ST-CYR, M.J. GANDER, and S.J. THOMAS. Optimized multiplicative, additive, and restricted additive Schwarz preconditioning. *SIAM J. Sci. Comput.*, 2007, vol. 29(6), pp. 2402–2425. ISSN 1064-8275. <http://dx.doi.org/10.1137/060652610>.
- [Stu17] A. STURM. *Locally Implicit Time Integration for Linear Maxwell's Equations*. Ph.D. thesis, Karlsruher Institut für Technologie (KIT), 2017. <http://dx.doi.org/10.5445/IR/1000069341>.
- [Van23] J. VANICEK. Family of Gaussian wavepacket dynamics methods from the perspective of a nonlinear Schrödinger equation. *The Journal of Chemical Physics*, 2023, vol. 159(1), p. 014114. ISSN 0021-9606. <http://dx.doi.org/10.1063/5.0146680>.
- [Ver11] J.G. VERWER. Component splitting for semi-discrete Maxwell equations. *BIT*, 2011, vol. 51(2), pp. 427–445. ISSN 0006-3835,1572-9125. <http://dx.doi.org/10.1007/s10543-010-0296-y>.
- [Yee66] K. YEE. Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. *IEEE Transactions on Antennas and Propagation*, 1966, vol. 14(3), pp. 302–307. <https://ieeexplore.ieee.org/document/1138693>.

## Bibliography

- [Zho14] Z. ZHOU. Numerical approximation of the Schrödinger equation with the electromagnetic field by the Hagedorn wave packets. *J. Comput. Phys.*, 2014, vol. 272, pp. 386–407. ISSN 0021-9991. <http://dx.doi.org/10.1016/j.jcp.2014.04.041>.
- [Zie20] D.A. ZIEGLER. *A parallel and adaptive space-time discontinuous Galerkin method for visco-elastic and visco-acoustic waves*. Dissertation, Karlsruhe Institute of Technology (KIT), 2020. <http://dx.doi.org/10.5445/IR/1000110469>.