Dual-Norm Least-Squares Finite Element Methods for Hyperbolic Problems

by

Delyan Zhelev Kalchev

Bachelor, Sofia University, Bulgaria, 2009
Master, Sofia University, Bulgaria, 2012
M.S., University of Colorado at Boulder, 2015

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Applied Mathematics

2018

This thesis entitled:
Dual-Norm Least-Squares Finite Element Methods for Hyperbolic Problems
written by Delyan Zhelev Kalchev
has been approved for the Department of Applied Mathematics

Thomas A. Manteuffel

Stephen Becker

Date

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Kalchev, Delyan Zhelev (Ph.D., Applied Mathematics)

Dual-Norm Least-Squares Finite Element Methods for Hyperbolic Problems

Thesis directed by Professor Thomas A. Manteuffel

Least-squares finite element discretizations of first-order hyperbolic partial differential equations (PDEs) are proposed and studied. Hyperbolic problems are notorious for possessing solutions with jump discontinuities, like contact discontinuities and shocks, and steep exponential layers. Furthermore, nonlinear equations can have rarefaction waves as solutions. All these contribute to the challenges in the numerical treatment of hyperbolic PDEs.

The approach here is to obtain appropriate least-squares formulations based on suitable minimization principles. Typically, such formulations can be reduced to one or more (e.g., by employing a Newton-type linearization procedure) quadratic minimization problems. Both theory and numerical results are presented.

A method for nonlinear hyperbolic balance and conservation laws is proposed. The formulation is based on a Helmholtz decomposition and closely related to the notion of a weak solution and a H^{-1}-type least-squares principle. Accordingly, the respective important conservation properties are studied in detail and the theoretically challenging convergence properties, with respect to the L^2 norm, are discussed.

In the linear case, the convergence in the L^2 norm is explicitly and naturally guaranteed by suitable formulations that are founded upon the original $\mathcal{L}\mathcal{L}^*$ method developed for elliptic PDEs. The approaches considered here are the $\mathcal{L}\mathcal{L}^*$-based and $(\mathcal{L}\mathcal{L}^*)^{-1}$ methods, where the latter utilizes a special negative-norm least-squares minimization principle. These methods can be viewed as specific approximations of the generally infeasible quadratic minimization that determines the L^2-orthogonal projection of the exact solution. The formulations are analyzed and studied in detail.

Key words: first-order hyperbolic problems; hyperbolic balance laws; Burgers equation; weak solutions; Helmholtz decomposition; conservation properties; space-time discretization; least-squares methods; dual methods; adjoint methods; negative-norm methods; finite element methods; discontinuous coefficients; exponential layers

AMS subject classifications: 65N30, 65N15

To George and my family

Acknowledgements

I extend my gratitude to my advisor, Tom Manteuffel, for his constant support, guidance, teaching, and patience. I also thank the Department of Applied Mathematics and the Grandview Computation Group for facilitating everything that was necessary to complete this work and providing me with the opportunity to learn.

I also wish to thank all my teachers and professors. I am grateful to Panayot Vassilevski for directing me towards science and research and giving me the opportunity to obtain experience at Lawrence Livermore National Laboratory – an institution that also deserves my gratitude for the good and motivating environment.

Finally, I thank my family and everyone that stood by me, had confidence in me, and offered help, support, and encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This dissertation focuses on the development and study of novel finite element discretization techniques of least-squares type for first-order hyperbolic partial differential equations (PDEs). Such equations are of high interest in practice and their numerical treatment is quite challenging. In particular, applications of hyperbolic-type PDEs arise in problems of fluid dynamics [1, 2, 3, 4, 5], particle transport [6, 7], and plasma modeling via the Vlasov equation [8].

Solutions of practical interest to the considered hyperbolic PDEs are often rather irregular, possessing steep layers or jump discontinuities. Moreover, discontinuities can be associated with nonlinear wave behaviors [9] like shocks and their interactions. Other nonlinear wave phenomena arising in solutions to hyperbolic equations of interest are rarefaction waves. The notion of a weak solution is rather important for nonlinear hyperbolic PDEs since it allows the consideration of non-smooth solutions. Further difficulties are caused by the fact that, in general, the weak solution to a nonlinear hyperbolic equation is not uniquely determined. Namely, a problem can have multiple (even infinitely many) weak solutions, thus giving rise to the additional notion of so-called entropy (or admissibility) that facilitates the determination of a unique physically admissible weak solution [1]; that is, an additional entropy (or admissibility) condition is imposed together with the differential equation to guarantee uniqueness of the weak solution to the problem.

The conservation properties of a numerical scheme are very important in the numerical treatment of these PDEs. In the context of this thesis, conservation is regarded as the property that the limit of converging approximations provided by a method is a weak solution to the given equation. This is of fundamental importance for the ability of a scheme to correctly approximate weak solutions (i.e., non-smooth solutions) to nonlinear hyperbolic PDEs. In particular, this is associ-

ated with the shock capturing capabilities of the discretization. Ideally, a method would provide approximations that correctly capture the locations of discontinuities, resolve them sharply without spurious oscillations, and, also, converge to an admissible weak solution. All this would be obtained via a computationally efficient numerical procedure. Moreover, an ideal method would provide high-order approximations in the regions where the solution is smooth.

A considerable number of methods for hyperbolic equations have been developed and studied. They include finite difference and, quite notably, finite volume methods [1, 2, 4, 3, 10, 11, 5]. In the context of those methods, Lax and Wendroff established the importance of utilizing so-called conservative schemes (i.e., those that can be expressed in a conservative form). Such schemes are based on an integral form (and integration by parts) of the equation and satisfy an exact discrete conservation property for an appropriate numerical flux [2, 3, 12]. Lax and Wendroff showed that, due to the exact discrete conservation, the above mentioned property (having a weak solution as a limit) holds.

Finite element methods have also been proposed and are actively developed and investigated. Quite notably, discontinuous Galerkin (DG; see [13] and the references therein) methods are often applied to the solution of hyperbolic PDEs as well as SUPG (streamline-upwind/Petrov-Galerkin) methods [14, 15, 16, 17]. DG methods also count on the exact discrete conservation property and the Lax-Wendroff theorem. In fact, the DG approach can be viewed as a finite element generalization of the finite volume method, that can naturally use high-oder elements and unstructured meshes. The spaces of piecewise discontinuous functions allow for formulations with upwind numerical fluxes [18]. The SUPG method is a stabilized Petrov-Galerkin modification of the standard Galerkin formulation. Some derivations of SUPG methods use variational formulations, in which an additional term is added. Numerical diffusion in the streamline direction is added artificially by a mesh-dependent perturbation of the test functions. SUPG methods show good results for advection-dominated elliptic problems [17, 19, 20, 21, 22], as well as DG methods [23]. SUPG formulations are additionally augmented by shock capturing terms [14, 24] to improve the quality of the solution near discontinuities.

An intriguing development in the the field of hyperbolic solvers is the BLAST library [25, 26, 27]. BLAST is a parallel code targeting applications in hydrodynamics. In particular, it solves the Euler equation of compressible hydrodynamics in a moving Lagrangian frame. It implements a

general finite element framework that uses curvilinear elements with continuous bases for the spacial discretization of some of the variables and discontinuous bases for the rest of the variables, providing high-order methods. A related work is presented in [28], where a high-order DG method for linear hyperbolic equations is described that utilizes appropriate flux limiters to diminish oscillations and avoid nonphysical values in the approximate solutions.

Least-squares finite element methods (see [29, 30, 31]) have also been applied to the solution of hyperbolic problems [32, 33, 34, 35, 15]; see also [36, 37, 7, 38]. However, compared to the previously mentioned approaches, least-squares methods are less developed in the context of hyperbolic PDEs, which provides a field of many challenges and novelties to be addressed and studied. The PhD dissertation by Luke Olson [32] and the related articles [33, 34] constitute an important development and are major references in this thesis. In particular, [32, 33] study a least-squares formulation that is related to the method of characteristics for linear advection hyperbolic PDEs. Moreover, they propose an intriguing discontinuous least-squares (DLS) method and investigate the performance of an algebraic multigrid (AMG; see [39]) linear systems solver for hyperbolic problems. The DLS formulation resembles the DG approach, but the numerical flux that normally appears in DG and finite volume methods is replaced by interface terms in the functional that build the connecting components between the elements of the mesh in the minimization formulation. In [32, 34], $H(\mathrm{div})$-conforming least-squares formulations are proposed for nonlinear hyperbolic conservation laws that satisfy the above mentioned conservation property. In principle, those formulations are related to a specific Helmholtz decomposition or, alternatively, can be associated with the characterization of divergence-free fields within the de Rham complex [40, 41, 42].

This thesis is focused on the development and study of new least-squares finite element discretizations for hyperbolic PDEs. It can be viewed as a continuation and extension of the work in [32, 33, 34], although not all methods that are proposed here are incremental developments from the considerations in [32, 33, 34]. Here, a rather concise overview is presented, while more detailed view on the connections and differences between methods appears in the coming chapters. In particular, the method for nonlinear hyperbolic balance laws based on the Helmholtz decomposition in Chapter 3 is an extension of the approach for conservation laws proposed in [32, 34] and it also satisfies the desired conservation property. The $\mathcal{L}\mathcal{L}^*$-type and $(\mathcal{L}\mathcal{L}^*)^{-1}$ approaches for linear problems in Chapter 4 can be seen as more related to the $\mathcal{L}\mathcal{L}^*$ [43], hybrid [44], and H^{-1} [45]

least-squares methods, while the theoretical results in [32, 33] are important for the applicability of the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method to hyperbolic problems. In fact, the $(\mathcal{L}\mathcal{L}^*)^{-1}$ approach can be derived as an improvement (in terms of approximation quality) of the $\mathcal{L}\mathcal{L}^*$-type methods, based on the known specifics of these methods but utilized and combined differently. Alternatively, it can be motivated as an improvement of an H^{-1} formulation and the method based on the Helmholtz decomposition, which is closely related to an H^{-1}-type formulation, for linear hyperbolic problems. Namely, the H^{-1} norm is replaced, in the linear case, by a dual norm that better conforms to the particular PDE and naturally provides convergence in the L^2 norm, which is very challenging to establish for the methods in Chapter 3 and [34]. This view furnishes a relation between the methods studied in this thesis. Another connection is provided by uniting all approaches under the group of "dual" methods. Here, "dual" means that the methods utilize or can be associated with adjoint operators or dual spaces (i.e., spaces of functionals) and their respective norms. Moreover, all considered formulations are typically viewed as unconstrained minimization problems. Nevertheless, it is somewhat unique and interesting that "saddle-point" problems and inf-sup conditions arise. Actually, the saddle-point formulation of the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method can be related to an equality constrained quadratic minimization and the original unconstrained form of the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method can be associated with the respective dual problem, in view of the theory of mathematical optimization. This adds to the "duality" point of view on these methods and, in fact, even the standard $\mathcal{L}\mathcal{L}^*$ approach of [43] can, in theory (and somewhat artificially), be posed as a constrained problem and the respective duality can be considered. While this is interesting, we do not elaborate much on it, since it is currently unclear what the possible benefits may be of such a view on the formulations.

Overall, the novelty in this dissertation can be summarized as the introduction and study of the method based on the Helmholtz decomposition for balance laws and the $(\mathcal{L}\mathcal{L}^*)^{-1}$ approach, as well as the further study and analysis of the $\mathcal{L}\mathcal{L}^*$-type methods. Note that, similar to [32, 33, 34], space-time discretizations are considered, without employing a particular time-stepping strategy, whereas the previously mentioned (finite difference, finite volume, DG, and SUPG) methods typically utilize a time-stepping scheme.

The outline of the rest of the thesis follows.

Chapter 2 provides a short and basic overview of some notions associated with least-squares methods and hyperbolic equations. The main purpose is to set the stage for the following chapters

that describe in detail the actual new contributions of this dissertation.

Chapter 3 is devoted to a method based on the Helmholtz decomposition for nonlinear scalar hyperbolic balance laws. Whereas [32, 34] consider conservation laws, which only have zero source terms, this chapter extends their ideas to balance laws, which allow nonzero sources. The fundamental idea is similar in the sense that a Helmholtz decomposition is used. However, the method here utilizes a different Helmholtz decomposition compared to that of [32, 34], which allows not only the accommodation of source terms but also a natural treatment of the inflow boundary conditions. The method is analyzed and its conservation properties are shown. The formulation satisfies the desired conservation property essentially by design due to the use of the particular Helmholtz decomposition. It is closely related to the well-known notion of a weak solution to a hyperbolic problem. Also, the convergence properties of the method with respect to the L^2 norm are discussed.

In Chapter 4, the $(\mathcal{L}\mathcal{L}^*)^{-1}$ and $\mathcal{L}\mathcal{L}^*$-type (single- and two-stage methods, as well as the standard $\mathcal{L}\mathcal{L}^*$ formulation) approaches for linear hyperbolic problems are investigated. The $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation is a novel method. Also, the idea of reformulating a negative-norm least-squares principle as a "saddle-point" problem does not exist in the literature. It allows the utilization of the original (unmodified) $(\mathcal{L}\mathcal{L}^*)^{-1}$ minimization principle. The standard $\mathcal{L}\mathcal{L}^*$ method is not new; it is formulated in [43] in the context of elliptic problems. The single- and two-stage methods are simple extensions of the original $\mathcal{L}\mathcal{L}^*$ approach and can be seen as a part of the hybrid method in [44]. The application of the $\mathcal{L}\mathcal{L}^*$, single-, and two-stage methods to hyperbolic problems is, however, a new development. In particular, the error analysis of the single- and two-stage methods in terms of the approximation properties of the involved finite element spaces was not previously known. The analysis is developed in a general setting and is, thus, applicable for more general linear PDEs, beyond hyperbolic problems.

Chapter 5 contains conclusions, final remarks, and future work.

Chapter 2

Overview of Least-Squares Methods and Hyperbolic Equations

Least-squares methods have been applied successfully to a variety of problems. In particular, they are well-studied in the context of partial differential equations (PDEs) of elliptic and parabolic types; see, e.g., [29, 31, 46, 47, 30, 48, 49, 50, 51]. This dissertation is focused on hyperbolic-type problems. To aid the exposition, the present chapter serves to set the stage for the chapters that follow. This is just a short overview and nothing novel is presented in the chapter. Everything that is discussed is either well known, or can be directly and easily obtained from known facts. A good general presentation of least-squares methods is provided in [29, 30, 31, 52], some references related to the PDE topic are [10, 3, 1, 2, 11, 53, 54, 55, 56, 57, 9], and a general detailed description of the finite element method is presented in [58, 59, 60, 61, 62, 41, 40]. It should be noted that the considerations in Chapters 3 and 4 do not fall entirely within the commonly utilized framework. This contributes to the fact that those chapters are somewhat self-contained. Nevertheless, it is beneficial to have a preparation in the form of the present chapter. A short overview of the least-squares ideas is provided in Section 2.1. Section 2.2 is a concise presentation of hyperbolic equations.

2.1 On least-squares methods

This is a general, abstract, and quite standard overview of least-squares principles and their application to approximating solutions to linear differential equations. The main idea is obtaining a Rayleigh-Ritz-type finite element formulation by considering the unconstrained minimization

of convex quadratic functionals associated with least-squares principles. Only conforming finite element methods are considered, i.e., "variational crimes" are not outlined.

2.1.1 General formulation

Let $\mathcal{H}$ and $\mathcal{V}$ be two real Hilbert spaces, where $\mathcal{H}$ is endowed with the inner product $\langle\cdot,\cdot\rangle$ and the respective norm $\|\cdot\|$, while the norm in $\mathcal{V}$ is denoted by $\|\cdot\|_{\mathcal{V}}$. Consider a linear operator $L\colon \mathcal{V}\to\mathcal{H}$ and some given $f\in\mathcal{H}$. The purpose is to solve the linear operator equation

$$Lv = f, \tag{2.1.1}$$

for the unknown $v\in\mathcal{V}$. In least-squares methods, the following quadratic functional is defined:

$$\mathcal{F}(v;\ f) = \|Lv - f\|^2, \quad \forall v\in\mathcal{V},$$

and the goal is to minimize this functional; that is, the following problem is considered:

$$u = \operatorname*{argmin}_{v\in\mathcal{V}} \mathcal{F}(v;\ f) = \operatorname*{argmin}_{v\in\mathcal{V}} \|Lv - f\|^2. \tag{2.1.2}$$

Typically, the minimization in (2.1.2) is associated with the weak formulation

$$\text{Find } u\in\mathcal{V}\colon a(u,v) = \ell(v), \quad \forall v\in\mathcal{V}, \tag{2.1.3}$$

where $a\colon \mathcal{V}\times\mathcal{V}\to\mathbb{R}$ is a *symmetric* bilinear form and $\ell\colon \mathcal{V}\to\mathbb{R}$ is a functional defined as

$$a(u,v) = \langle Lu, Lv\rangle, \quad \ell(v) = \langle f, Lv\rangle.$$

Clearly, the bilinear form a is *positive semidefinite*, i.e., $a(v,v)\geq 0$ for all $v\in\mathcal{V}$. Also, by dropping constant additive terms and using the just introduced notation, the minimization in (2.1.2) can be seen in the more stereotypical Rayleigh-Ritz setting by expressing it as

$$u = \operatorname*{argmin}_{v\in\mathcal{V}} [a(v,v) - 2\ell(v)]. \tag{2.1.4}$$

Questions of interest are uniqueness and existence of solutions to (2.1.2) (or, equivalently, (2.1.4)) and (2.1.3), as well as when (2.1.2) and (2.1.3) are equivalent, i.e., when they provide the same solutions. These questions turn out to be interrelated. In particular, it is not difficult to see that $\mathcal{F}$ (or the functional in (2.1.4)) is always Gâteaux differentiable. Thus, a necessary

condition for $u \in \mathcal{V}$ to minimize (2.1.2) is u to be a zero of the Gâteaux derivative (also called *first variation* in this context), which is the same[1] as to be a solution to (2.1.3); that is, any minimizer of (2.1.2) (or (2.1.4)) is a solution to (2.1.3). The converse is more delicate. However, this simple observation clearly shows that the uniqueness of the solution to (2.1.3) implies the uniqueness of the minimizer of (2.1.2). Particularly, if $a(\cdot,\cdot)$ is *positive definite*, i.e., $a(v,v) > 0$ for all $v \in \mathcal{V}\setminus\{0\}$, then it is easy to see that (2.1.3) can have at most one solution, implying the uniqueness of the minimizer of (2.1.2). This is to be expected, since the uniqueness of the minimizer of (2.1.2) follows from the strict convexity of the functional $\mathcal{F}$, which is equivalent to the positive definiteness of $a(\cdot,\cdot)$. Furthermore, the strict convexity of $\mathcal{F}$ implies that the unique minimizer (if it exists) is characterized as the unique zero of the Gâteaux derivative. Hence, (2.1.2) and (2.1.3) are equivalent. In summary, if $a(\cdot,\cdot)$ is positive definite, then (2.1.2) and (2.1.3) are equivalent and they have at most one solution.

Existence of a solution is typically established via the Riesz representation theorem (which also implies uniqueness). Here, positive definiteness of $a(\cdot,\cdot)$ shows that it is an inner product in $\mathcal{V}$, but the Riesz theorem does not apply, in general, since $\mathcal{V}$ may not be a Hilbert space with respect to the induced norm, denoted by $\|\cdot\|_a$. Thus, the positive definiteness of $a(\cdot,\cdot)$ alone and the Riesz representation theorem[2] only imply the existence of a solution in the completion (closure) of $\mathcal{V}$ with respect to $\|\cdot\|_a$. Therefore, the following additional assumption on $a(\cdot,\cdot)$, that it is $\mathcal{V}$-*coercive* (in short, *coercive*), is needed:

$$\alpha\|v\|_{\mathcal{V}}^2 \le a(v,v), \quad \forall v \in \mathcal{V}, \tag{2.1.5}$$

for some constant $\alpha > 0$. This states that the norm $\|\cdot\|_a$ is stronger than $\|\cdot\|_{\mathcal{V}}$. Hence, $\mathcal{V}$ is closed, and thus a Hilbert space, with respect to $\|\cdot\|_a$. Now, the Riesz representation theorem provides the desired existence. In summary, the coercivity (2.1.5) implies equivalence between (2.1.2) and (2.1.3), as well as the respective existence and uniqueness of a solution in $\mathcal{V}$.

In such a general setting, the equation (2.1.1) may be overdetermined in the sense that $L\colon \mathcal{V} \to \mathcal{H}$ may not be surjective, (2.1.1) may not possess general existence, and the minimal value of $\mathcal{F}$ may not be zero. Note that (2.1.5) only implies that $L\colon \mathcal{V} \to \mathcal{H}$ is injective (i.e., (2.1.1) can have at most

[1] In the literature, (2.1.3) is sometimes called an *Euler-Lagrange equation* associated with the minimization (2.1.2) (or (2.1.4)).

[2] It is a simple matter to see that the functional ℓ is bounded with respect $\|\cdot\|_a$.

one solution) and has a closed range. In general, the usual behavior of least-squares methods is obtained; that is, the minimizer of (2.1.2) is precisely the unique solution to the equation $Lv = f_p$, where $f_p \in L(\mathcal{V}) \subset \mathcal{H}$ is the $\mathcal{H}$-orthogonal projection of f onto the range of L, $L(\mathcal{V})$. However, in practice, this is usually not an issue, since L is often surjective or f is taken in the range of L and, thus, the minimal value of $\mathcal{F}$ is zero. This thesis concentrates on surjective operators. Therefore, from now on we assume that L is surjective. Sometimes it is instructive to express the least-squares problem in terms of the *normal equation* $L^*Lv = L^*f$, where L^* is the adjoint of L. Actually, (2.1.3) can be viewed as the weak form (the Galerkin closure) of the normal equation.

2.1.2 Discrete formulation

The purpose now is to lead to finite element methods. The approach is to restrict (2.1.2) and (2.1.3) to a (discrete) finite element space $\mathcal{V}^h \subset \mathcal{V}$ with a basis $\{\, \phi_i^h \,\}_{i=1}^N$:

$$u^h = \operatorname*{argmin}_{v^h \in \mathcal{V}^h} \mathcal{F}(v^h;\ f) = \operatorname*{argmin}_{v^h \in \mathcal{V}^h} \|Lv^h - f\|^2, \tag{2.1.6}$$

$$\text{Find } u^h \in \mathcal{V}^h \colon a(u^h, v^h) = \ell(v^h), \quad \forall v^h \in \mathcal{V}^h. \tag{2.1.7}$$

All considerations in the continuous case continue to hold in the discrete case. Additionally, just the positive definiteness of $a(\cdot,\cdot)$ implies the existence of a discrete least-squares solution. This is easy, but instructive, to see. Namely, consider the symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ and the vector $\boldsymbol{f} \in \mathbb{R}^N$ defined as

$$(\boldsymbol{A})_{ij} = a(\phi_j^h, \phi_i^h), \quad (\boldsymbol{f})_i = \ell(\phi_i^h).$$

Then (2.1.7) induces the algebraic system of linear equations

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f},$$

where the solution $\boldsymbol{u} \in \mathbb{R}^N$ is the vector of coefficients of the solution $u^h \in \mathcal{V}^h$ to (2.1.7) with respect to the basis functions $\{\, \phi_i^h \,\}_{i=1}^N$. If $a(\cdot,\cdot)$ is positive definite, then $\boldsymbol{A}$ is positive definite and, thus, nonsingular. In particular, this implies the existence and uniqueness of a solution to (2.1.7) and, by equivalence, (2.1.6). Furthermore, this reveals an important feature of standard least-squares formulations in that they lead to linear systems with symmetric positive definite matrices, which generally holds for convex quadratic minimization problems.

It is important to establish that appropriate approximations of the solution to the continuous problem are obtained. This is provided by the Céa's lemma and its particular version for formulations with symmetric bilinear forms, which is a basic result in the theory of finite element methods. First, observe that, by combining (2.1.3) and (2.1.7), the following orthogonality property is obtained:

$$a(u^h - u, v^h) = 0, \quad \forall v^h \in \mathcal{V}^h,$$

where $u^h \in \mathcal{V}^h$ is the solution to (2.1.7) and $u \in \mathcal{V}$ is the solution to (2.1.3). Thus, u^h is the a-orthogonal projection of u onto $\mathcal{V}^h$, i.e., it is the best approximation of u with respect to the norm $\|\cdot\|_a$; that is,

$$u^h = \underset{v^h \in \mathcal{V}^h}{\operatorname{argmin}} \|v^h - u\|_a^2 = \underset{v^h \in \mathcal{V}^h}{\operatorname{argmin}} \|L(v^h - u)\|^2, \tag{2.1.8}$$

which, since $f = Lu$, is precisely (2.1.6).

In practice, error estimates with respect to the $\mathcal{V}$ norm are desired. To this end, the following additional assumption, that $a(\cdot, \cdot)$ is $\mathcal{V}$-*continuous* (shortly, *continuous*), is needed:

$$a(v, v) \leq \beta \|v\|_{\mathcal{V}}^2, \quad \forall v \in \mathcal{V}, \tag{2.1.9}$$

where $\beta > 0$ is some constant. Notice that (2.1.9) is equivalent to the boundedness of $L \colon \mathcal{V} \to \mathcal{H}$ and implies the continuity of ℓ with respect to $\|\cdot\|_{\mathcal{V}}$, i.e., $\ell \in \mathcal{V}'$ – the dual space of $\mathcal{V}$. Combining (2.1.5), (2.1.8), and (2.1.9) provides the (quasi-)optimal estimate

$$\|u^h - u\|_{\mathcal{V}} \leq \left(\frac{\beta}{\alpha}\right)^{\frac{1}{2}} \|v^h - u\|_{\mathcal{V}}, \quad \forall v^h \in \mathcal{V}^h. \tag{2.1.10}$$

In practice, $\mathcal{V}$ is replaced by a suitable Sobolev space and plugging an appropriate interpolant in place of v^h in (2.1.10) allows the utilization of the interpolation bounds of the polynomial approximation theory to deduce the rate of convergence of u^h to u in terms the particular Sobolev norm. In summary, the continuity (2.1.9) and the approximation properties of $\mathcal{V}^h$ yield a bound on the rate at which the functional values, $\mathcal{F}(u^h;\ f)$, converge to zero, while the functional convergence together with the coercivity (2.1.5) determine the rate at which the $\mathcal{V}$ norm of the error approaches zero. The combination guarantees that the error of the method in the $\mathcal{V}$ norm decays with the optimal rate provided by the approximation properties of $\mathcal{V}^h$.

The coercivity (2.1.5) is further important, since it provides the stability of the variational formulations (2.1.3) and (2.1.7); that is, the following a priori estimates hold:

$$\|u\|_{\mathcal{V}} \leq \frac{1}{\alpha}\|\ell\|_{\mathcal{V}'}, \quad \|u^h\|_{\mathcal{V}} \leq \frac{1}{\alpha}\|\ell\|_{\mathcal{V}'}.$$

Observe that the coercivity and continuity relations (2.1.5) and (2.1.9) can be conveniently expressed as

$$\alpha\|v\|_{\mathcal{V}}^2 \leq \mathcal{F}(v;\ 0) \leq \beta\|v\|_{\mathcal{V}}^2, \quad \forall v \in \mathcal{V}, \tag{2.1.11}$$

and it is said that $\mathcal{F}$ and the respective bilinear form, $a(\cdot,\cdot)$, are $\mathcal{V}$-*elliptic* (or $\mathcal{V}$-*equivalent*). It is instructive to derive the following from (2.1.11):

$$\alpha\|u^h - u\|_{\mathcal{V}}^2 \leq \mathcal{F}(u^h - u;\ 0) = \mathcal{F}(u^h;\ f) \leq \beta\|u^h - u\|_{\mathcal{V}}^2,$$

demonstrating one of the important features of standard least-squares formulations. Namely, the functional value $\mathcal{F}(u^h;\ f)$, which is computable, is a good candidate for an *a-posteriori* error estimate [63, 64, 65]. In practice, the value of the functional is computed locally (i.e., on each element), providing a *reliable* global upper bound of the error (due to the coercivity) and a *sharp* local lower bound of the error (due to a local version of the continuity). Particularly, in the common setting of a small coercivity constant, α, and a non-large continuity constant, β, this means that small global values of the functional may not necessary indicate that the error is correspondingly small in the $\mathcal{V}$ norm, while large local values indicate the substantial presence of a local error. Thus, the least-squares functional, $\mathcal{F}$, can be used as a local error estimate for adaptive mesh refinement.

Typically, in practice, when PDEs are considered, L is a first-order differential operator, (2.1.1) is a PDE (a scalar equation or, more often, a system), and $\mathcal{H}$,$\mathcal{V}$ are Sobolev spaces. The boundary conditions can be treated in a variety of ways – they can be imposed weakly as a part of the functional, strongly in the spaces, or a combination of both. Usually, $\mathcal{H}$ is the L^2 function space, but it can also be a negative-order Sobolev space, like H^{-1}, or a product of L^2 and H^{-1} spaces [52]. The L^2 version of FOSLS (first-order system least-squares) is somewhat easier to implement and work with, which contributes to it being more common.

A substantial effort, especially for elliptic PDEs, has been invested in obtaining least-squares formulations and proving appropriate ellipticity results like (2.1.11). Usually, the continuity is not

difficult, whereas the coercivity can be quite challenging to show. Particularly, in the context of L^2 functionals with first-order systems, a core component of the approach is the derivation of a suitable first-order system of the general form (2.1.1) that induces a least-squares principle with the desired properties. Considerable work has been devoted to developing formulations that are H^1-equivalent for elliptic problems. This way, approximations with respect to the H^1 norm, which is somewhat natural for elliptic PDEs, are guaranteed and good performance of algebraic multigrid (AMG) linear system solvers is obtained. A major advantage and, at the same time, disadvantage of least-squares methods is the great flexibility that they offer, which includes a variety of treatments of boundaries and weighting of the terms in the functional. The weighting has been particularly useful for elliptic problems with singularities. It allows utilizing weighted Sobolev spaces and obtaining ellipticity with respect to weighted norms, leading to methods that properly treat singularities and prohibit their pollution effect on the entire solution. Full elliptic regularity (also called H^2 regularity) in terms of the weighted Sobolev spaces can be recovered, providing optimal convergence rates with respect to the weighted norms and even an enhanced L^2 convergence (away from the singularities) due to the Aubin-Nitsche duality argument.

Furthermore, due to the utilization of unconstrained minimization, FOSLS has the positive feature that it provides a natural treatment of the respective elliptic first-order systems without the special need for inf-sup conditions and compatible finite elements spaces, as well as a simpler setting for using nonconforming finite elements [52]. However, even though only unconstrained minimization problems are considered in this thesis, inf-sup conditions appear in the considerations in the chapters that follow.

2.1.3 Hybrid and $\mathcal{L}\mathcal{L}^*$ formulations

This subsection contains a short overview of least-squares methods that, in the terminology used in this thesis, can be called dual methods. The exposition is basic and intuitive to provide introduction. Chapter 4 presents a more rigorous view as necessary.

Consider (2.1.1), where $L\colon \mathcal{D}(L) \subset L^2(\Omega) \to L^2(\Omega)$ is a first-order differential operator, Ω is a domain in $\mathbb{R}^d$ (d is the dimension), and $\mathcal{D}(L)$ is the domain of L. Note that $\mathcal{D}(L)$ is a space of functions that satisfy certain suitable homogeneous boundary conditions (for more details, see Chapter 4). In general, as it is well known, linear problems can be reduced to equations with

homogeneous boundary conditions via superposition. Thus, seeking a solution to (2.1.1) in $\mathcal{D}(L)$ does not lead to any loss of generality. Consider also the differential operator $L^*\colon \mathcal{D}(L^*) \subset L^2(\Omega) \to L^2(\Omega)$ – the L^2-adjoint of L. Here, $\mathcal{D}(L^*)$ is the domain of L^* and it also consists of functions that satisfy appropriate homogeneous boundary conditions, which are in a sense adjoint to the boundary conditions in $\mathcal{D}(L)$. The $\mathcal{LL}^*$ method (also called FOSLL*) is proposed in [43] for elliptic problems with singularities, targeting approximations in the L^2 norm. The intuitive idea is to seek $w \in \mathcal{D}(L^*)$ that solves

$$LL^*w = f. \tag{2.1.12}$$

The final solution to (2.1.1) is obtained as $u = L^*w$. More precisely, consider a finite element space, $\mathcal{Z}^h \subset \mathcal{D}(L^*)$, and the quadratic minimization,

$$w^h = \operatorname*{argmin}_{z^h \in \mathcal{Z}^h} \|L^*z^h - u\|^2, \tag{2.1.13}$$

where $\|\cdot\|$ denotes the L^2 norm, $\langle \cdot, \cdot \rangle$ is the respective inner product, and $u \in \mathcal{D}(L)$ is the exact solution to (2.1.1). The obtained approximation is $u^h = L^*w^h \in L^*(\mathcal{Z}^h)$, which clearly is the L^2-orthogonal projection of the exact solution, u, onto $L^*(\mathcal{Z}^h)$, i.e., u^h is the best approximation of u in the L^2 norm on $L^*(\mathcal{Z}^h)$. A valuable feature of (2.1.13) is that the associated weak formulation is computationally feasible, since the exact solution, u, is not explicitly needed. Indeed, the weak form is:

$$\text{Find } w^h \in \mathcal{Z}^h\colon \langle L^*w^h, L^*z^h \rangle = \langle f, z^h \rangle, \quad \forall z^h \in \mathcal{Z}^h, \tag{2.1.14}$$

which contains only available information. Observe that (2.1.14) can be formally seen as the Galerkin closure (weak formulation) of (2.1.12), while (2.1.13) represents the least-squares closure of $u = L^*w$.

It is a curious fact that the equation (2.1.12) can be related to known approaches in numerical linear algebra and convex optimization, as explained in [43]. Using known facts from linear algebra and optimization (see [66, Section 6.2]), one can see that the $\mathcal{LL}^*$ approach can be (formally, at least) associated, via duality, with the following constrained quadratic minimization:

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|u\|^2, \\ &\text{subject to} && Lu = f. \end{aligned} \tag{2.1.15}$$

This is sketched as follows; see also [66, Section 5.1.5]. Note that the Lagrangian associated with (2.1.15) is

$$\mathscr{L}(u,w) = \frac{1}{2}\langle u,u\rangle + \langle w, f - Lu\rangle,$$

for $u \in \mathcal{D}(L)$ and $w \in \mathcal{D}(L^*)$. Then, the so called (Lagrange) dual function is given by $\mathscr{D}(w) = \inf_{u\in\mathcal{D}(L)} \mathscr{L}(u,w)$. Observe that $\mathscr{L}(u,w)$ is a convex quadratic functional in u. Hence, assuming that L^* is surjective, the minimizer can be expressed as $u = L^*w$, providing

$$\mathscr{D}(w) = -\frac{1}{2}\langle L^*w, L^*w\rangle + \langle f, w\rangle = -\frac{1}{2}\langle L^*w, L^*w\rangle + \langle u, L^*w\rangle,$$

where $f = Lu$. Generally, the Lagrange dual problem of (2.1.15) is the unconstrained problem of maximizing the dual function $\mathscr{D}(w)$ over $w \in \mathcal{D}(L^*)$, or, equivalently, the minimization of $-\mathscr{D}(w)$. Thus, the dual optimization problem (Lagrange dual) of (2.1.15) is equivalent, in terms of minimizers, to

$$\text{minimize} \quad \|L^*w - u\|^2,$$

which recovers the $\mathcal{L}\mathcal{L}^*$ minimization (2.1.13) (in a slightly different notation) and w is the dual variable (Lagrange multiplier) associated with the constraint in (2.1.15). Also, the Karush-Kuhn-Tucker optimality conditions for (2.1.15) are

$$u - L^*w = 0, \quad Lu = f,$$

recovering (2.1.12) and the relation $u = L^*w$. In summary, the $\mathcal{L}\mathcal{L}^*$ formulation can be related to the minimal norm solution to the least-squares problem $\min_{u\in\mathcal{D}(L)} \|Lu - f\|^2$ (i.e., the minimal norm solution to $Lu = f$).

In this thesis, no explicit considerations of constrained minimization problems are presented. The focus is on unconstrained least-squares principles and we do not provide a detailed investigation of relations between these principles and constrained problems, even if such simple connections exist as in the $\mathcal{L}\mathcal{L}^*$ method. Note that there are specialized constrained least-squares finite element methods in the literature; see [67, 68].

The hybrid least-squares method is introduced in [44]. The idea is to combine the $\mathcal{L}\mathcal{L}^*$ method, which provides a certain best approximation in the L^2 norm, and the FOSLS method, which yields the best approximation in the operator norm, $\|\cdot\|_a$, to obtain a formulation that aims at approximations with respect to the *graph norm*, $(\|v\|^2 + \|Lv\|^2)^{1/2}$. This is achieved by combining,

in a single functional, the FOSLL* and FOSLS functionals together with a connecting *intermediate term* as follows:

$$(u^h, w^h) = \underset{(v^h, z^h) \in \mathcal{V}^h \times \mathcal{Z}^h}{\operatorname{argmin}} \left[\|L^* z^h - u\|^2 + \|v^h - L^* z^h\|^2 + \|L v^h - f\|^2 \right], \tag{2.1.16}$$

where $\mathcal{V}^h \subset \mathcal{D}(L)$, $\mathcal{Z}^h \subset \mathcal{D}(L^*)$ are some finite element spaces and $u \in \mathcal{D}(L)$ is the exact solution to (2.1.1). Note that one of the major features of FOSLS, the natural error measure, is lost in the FOSLL* method, but it is partially recovered in the hybrid formulation. Namely, the functional in (2.1.16) can be used as an error estimate by removing the FOSLL* term, i.e., by only using the FOSLS and intermediate terms.

In Chapter 4, a couple of formulations are studied that resemble the hybrid method (2.1.16) but without the FOSLS term. Actually, the method of Chapter 3 can also be loosely viewed as a hybrid-type formulation, but it is considerably more specialized for the particular hyperbolic PDE at hand.

2.2 On hyperbolic equations

This dissertation is in the general field of numerical analysis, and novel least-squares methods for hyperbolic equations are proposed. While research in the theory of hyperbolic PDEs is in no way a goal in this thesis, the rather basic introduction in this section is useful as a preparation for the considerations in the coming chapters.

Probably the most widely known topic regarding the types of PDEs is the classification of linear second-order PDEs. We are only concerned with first-order equations of hyperbolic type here, and this is the only class we discuss. This is, clearly, related to the widely known notion of hyperbolicity in the second-order equations, which intuitively is associated with the possession of a "full" set of characteristics.

For simplicity and illustration, as is customary, only homogeneous equations (i.e., *conservation laws*) and initial value (Cauchy) problems are considered. This should not be an issues for the purpose of this section even though there is interest in equations with sources and this thesis is focused on such problems. Generally, initial value problems are formulated on the whole space and for non-negative time, involving only initial conditions without any boundary conditions. In the coming chapters, as practical, bounded computational domains are used and there is no significant

distinction between space and time, as well as between initial and boundary conditions and they are collectively called "(inflow) boundary conditions". Nevertheless, a more stereotypical path is followed in this section.

Consider a partial differential equation of the form, for some integer dimensions $d, m > 0$,

$$\boldsymbol{u}_t + \sum_{i=1}^{d} \boldsymbol{f}_i(\boldsymbol{u})_{x_i} = \boldsymbol{0}, \tag{2.2.1}$$

where[1] $\boldsymbol{u}\colon \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^m$ is the vector of unknowns (dependent variables) also called *state variables*, which represent some (physical) quantities to be conserved, t, x_i are the independent variables, where t is time, and $\boldsymbol{f}_i\colon \mathbb{R}^m \to \mathbb{R}^m$ are called *flux functions* or *flux vectors* as they describe the flow of the conserved quantities, $\boldsymbol{u}$. It becomes clear below why PDEs like (2.2.1) are called *(systems of) conservation laws*, when they are derived from basic conservation principles. However, this is also clear from expressing (2.2.1) as

$$\boldsymbol{u}_t + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{F}(\boldsymbol{u}) = \boldsymbol{0},$$

where $\boldsymbol{F}\colon \mathbb{R}^m \to (\mathbb{R}^m)^d$, $\boldsymbol{F}(\boldsymbol{u}) = [\boldsymbol{f}_1(\boldsymbol{u}), \dots, \boldsymbol{f}_d(\boldsymbol{u})]$ and $\nabla_{\boldsymbol{x}}\cdot$ is the divergence with respect to the x_i variables, i.e., the spatial divergence. Furthermore, (2.2.1) can be expressed as

$$\nabla_{t,\boldsymbol{x}} \cdot \boldsymbol{\Phi}(\boldsymbol{u}) = \boldsymbol{0},$$

which is a preferred form in this dissertation. Here, $\boldsymbol{\Phi}\colon \mathbb{R}^m \to (\mathbb{R}^m)^{d+1}$, $\boldsymbol{\Phi}(\boldsymbol{u}) = [\boldsymbol{\iota}(\boldsymbol{u}), \boldsymbol{f}_1(\boldsymbol{u}), \dots, \boldsymbol{f}_d(\boldsymbol{u})]$, where $\boldsymbol{\iota}\colon \mathbb{R}^m \to \mathbb{R}^m$ is the identity map, and $\nabla_{t,\boldsymbol{x}}\cdot$ is the divergence with respect to t and the x_i variables, i.e., the space-time divergence.

The basic conservation principle states that, for any space-time volume (domain), $D \subset \mathbb{R}_+ \times \mathbb{R}^d$, the net flux across its boundary, ∂D, vanishes, indicating that no source or sink of the conserved quantities is present in the volume; that is,

$$\int_{\partial D} \boldsymbol{n} \cdot \boldsymbol{\Phi}(\boldsymbol{u}) \,\mathrm{d}\boldsymbol{\sigma} = \boldsymbol{0}, \quad \forall D \subset \mathbb{R}_+ \times \mathbb{R}^d, \tag{2.2.2}$$

where $\boldsymbol{n}$ denotes the unit outward normal to ∂D. Alternatively, this states that the total values of the conserved quantities in any spatial region change only due to flux through the respective spatial boundaries [1]. Green's formula (integration by parts) gives

$$\int_D \nabla_{t,\boldsymbol{x}} \cdot \boldsymbol{\Phi}(\boldsymbol{u}) \,\mathrm{d}t \,\mathrm{d}\boldsymbol{x} = \int_{\partial D} \boldsymbol{n} \cdot \boldsymbol{\Phi}(\boldsymbol{u}) \,\mathrm{d}\boldsymbol{\sigma} = \boldsymbol{0}, \quad \forall D \subset \mathbb{R}_+ \times \mathbb{R}^d.$$

[1] Here, $\mathbb{R}_+ = \{\, x \in \mathbb{R};\ x \geq 0 \,\}$.

This equality implies the *differential form* (2.2.1) of the conservation law, since it holds on any domain D.

The PDE (2.2.1) is said to be *hyperbolic* if, for each value of $\boldsymbol{v} \in \mathbb{R}^m$, every real linear combination $\sum_{i=1}^{d} \omega_i \boldsymbol{f}_i'(\boldsymbol{v})$ is diagonalizable with real eigenvalues and eigenvectors, where $\boldsymbol{f}_i'$ denote the respective Jacobian matrices. It is *strictly hyperbolic* if all eigenvalues are distinct. In particular, scalar PDEs (i.e., the case $m = 1$) of the form (2.2.1) are hyperbolic. In more detail, (2.2.1) is hyperbolic in the direction of time (or t-hyperbolic), which essentially means that the characteristics are never perpendicular to the time direction and, thus, allowing the consideration of initial value problems, where initial conditions can be imposed on hyperplanes in $\mathbb{R}_+ \times \mathbb{R}^d$ that are perpendicular to the t-axis. Typically, the initial conditions are provided for $t = 0$ as

$$\boldsymbol{u}(0, \boldsymbol{x}) = \boldsymbol{u}_0(\boldsymbol{x}), \tag{2.2.3}$$

for a given initial state $\boldsymbol{u}_0 \colon \mathbb{R}^d \to \mathbb{R}^m$. Together (2.2.1) and (2.2.3) constitute a Cauchy hyperbolic problem.

In the linear case, the PDE becomes

$$\boldsymbol{u}_t + \sum_{i=1}^{d} \boldsymbol{A}_i \boldsymbol{u}_{x_i} = \boldsymbol{0},$$

where $\boldsymbol{A}_i \colon \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^{m \times m}$ are given matrix-valued functions. Two alternative ways to write this equation are

$$\boldsymbol{u}_t + \mathbb{A} \cdot \nabla_{\boldsymbol{x}} \boldsymbol{u} = \boldsymbol{0},$$

$$\mathbb{B} \cdot \nabla_{t,\boldsymbol{x}} \boldsymbol{u} = \boldsymbol{0},$$

where $\mathbb{A} \colon \mathbb{R}_+ \times \mathbb{R}^d \to (\mathbb{R}^{m \times m})^d$, $\mathbb{A}(t, \boldsymbol{x}) = [\boldsymbol{A}_1(t, \boldsymbol{x}), \ldots, \boldsymbol{A}_d(t, \boldsymbol{x})]$, $\nabla_{\boldsymbol{x}}$ is the spatial gradient, $\mathbb{B} \colon \mathbb{R}_+ \times \mathbb{R}^d \to (\mathbb{R}^{m \times m})^{d+1}$, $\mathbb{B}(t, \boldsymbol{x}) = [\boldsymbol{I}, \boldsymbol{A}_1(t, \boldsymbol{x}), \ldots, \boldsymbol{A}_d(t, \boldsymbol{x})]$, and $\nabla_{t,\boldsymbol{x}}$ is the space-time gradient. The linear system is hyperbolic if every real linear combination $\sum_{i=1}^{d} \omega_i \boldsymbol{A}_i$ is diagonalizable with real eigenvalues and eigenvectors, and strictly hyperbolic if all eigenvalues are distinct. Note that the linear hyperbolic PDE above can be seen as a special case of (2.2.1) as long as the flux functions are allowed to explicitly depend on t and $\boldsymbol{x}$, which should not be a source of any issues.

Clearly, (2.2.1) can be expressed as a first-order *quasilinear* PDE. The *method of characteristics* is a well-known approach for solving and studying first-order quasilinear PDEs. It reduces

the problem to (autonomous) systems of ordinary differential equations. The classical method of characteristics mostly addresses *classical solutions* to the PDE, i.e., solutions that are continuously differentiable and satisfy (2.2.1) and (2.2.3) in the classical pointwise sense. However, one of the major difficulties associated with equations like (2.2.1) is that they model important behaviors that can be caused by the nonlinearity of the problem and result in solutions that have discontinuities, i.e., solutions that are not classical. In the linear case, this is less of an issue since linear hyperbolic PDEs can only have *contact discontinuities*, i.e., the discontinuities propagate only along characteristics and are not associated with any collisions between the characteristics. Thus, even though such discontinuous solutions are not classical, they can still be obtained unambiguously by the method of characteristics. In contrast, in the nonlinear case, characteristics can collide, even if all the given data is smooth, resulting in ambiguities in the method of characteristics since it leads to multivalued solutions. Such colliding characteristics lead to discontinuities called *shocks*, which are a typical of a nonlinear behavior. Moreover, another common nonlinear behavior is the characteristics "spreading apart", which leads to entire regions where the method of characteristics cannot provide information on the solution. This setting is associated with *rarefaction waves* in the solution.

The solutions discussed above model important behaviors and while, due to their irregularity, they do not satisfy the differential form of the conservation law (2.2.1) in the classical sense, they satisfy the *integral form* of the conservation law in (2.2.2) and, thus, they model valuable physical behaviors. However, working with (2.2.2) is difficult. Therefore, the notion of a *weak solution* to the Cauchy problem (2.2.1) and (2.2.3) is introduced using a more convenient integral form of the conservation law. Namely, the PDE (2.2.1) is multiplied by continuously differentiable functions $\boldsymbol{\phi}\colon \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^m$ that is compactly supported in $\mathbb{R}_+ \times \mathbb{R}^d$ and then the Green's formula (integration by parts) together with the initial condition (2.2.3) provide the weak formulation

$$\int_{\mathbb{R}_+ \times \mathbb{R}^d} \boldsymbol{\Phi}(\boldsymbol{u}) \cdot \nabla_{t,\boldsymbol{x}} \boldsymbol{\phi} \,\mathrm{d}t \,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^d} (\boldsymbol{n} \cdot \boldsymbol{\Phi}(\boldsymbol{u}_0)) \cdot \boldsymbol{\phi}(0, \boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \left[= - \int_{\mathbb{R}^d} \boldsymbol{u}_0(\boldsymbol{x}) \cdot \boldsymbol{\phi}(0, \boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \right], \tag{2.2.4}$$

for all continuously differentiable and compactly supported test functions $\boldsymbol{\phi}\colon \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^m$. The notation here is intuitive, noting that some of the dot products are sums of other (lower-dimensional) dot products, effectively representing matrix multiplications. A weak solution to (2.2.1) and (2.2.3) is defined as a solution to the weak form (2.2.4). Notice that restricting the test

functions in (2.2.4) to only compactly supported in the open set $(\mathbb{R}_+ \setminus \{0\}) \times \mathbb{R}^d$ shows that any weak solution satisfies the PDE (2.2.1) in the *sense of distributions*, a terminology that we avoid in the rest of the thesis, but the weak form (2.2.4) also contains information on the initial condition (2.2.3). Moreover, weak solutions allow discontinuities and satisfy the integral form (2.2.2), thus maintaining the basic conservation principle.

In practice, there is a major interest in piecewise continuously differentiable weak solutions. Since the notion of a weak solution is naturally related to the conservation expressed in (2.2.2), piecewise continuously differentiable weak solutions satisfy the so-called *Rankine-Hugoniot jump condition*, which can be expressed here as

$$[\![\boldsymbol{n} \cdot \boldsymbol{\Phi}(\boldsymbol{u})]\!]_{\mathscr{C}} = \mathbf{0}$$

almost everywhere along any surface of discontinuity $\mathscr{C} \subset \mathbb{R}_+ \times \mathbb{R}^d$, where $[\![\cdot]\!]_{\mathscr{C}}$ denotes the jump across $\mathscr{C}$ and $\boldsymbol{n}$ is the unit normal to $\mathscr{C}$, whose orientation is insignificant. In fact, the Rankine-Hugoniot condition is satisfied along any (orientable) surface, which may or may not be associated with a discontinuity. A converse also holds. Namely, [3, Theorem 2.1 on page 16] shows that if a piecewise continuously differentiable function satisfies the Rankine-Hugoniot condition along any surface of discontinuity and satisfies the PDE (2.2.1) in the classical sense where it is continuously differentiable, then it solves (2.2.1) in the sense of distributions. In general, as in Chapter 3, the Rankine-Hugoniot condition can be associated with the generalized flux, $\boldsymbol{\Phi}(\boldsymbol{u})$, belonging in an appropriate $H(\mathrm{div})$ Sobolev space. In fact, this theoretically extends the Rankine-Hugoniot condition, since it generally holds even if the weak solution is not piecewise continuously differentiable.

The approach to obtaining the weak formulation (2.2.4) and the considerations and facts above are quite general. Therefore, they can be applied when source terms are present (i.e., when (2.2.1) becomes a *balance law*) or bounded domains are used and initial-boundary value problems are considered.

While the general equation (2.2.1) does not seem to result in a very complicated form of a PDE, the numerical treatment of hyperbolic problems of this type is quite challenging. A major source of the challenges is that the weak solutions of interest are nonsmooth, resulting in difficulties in capturing discontinuities and their speeds, smearing and spurious oscillations close to the discontinuities. The conservation properties of the methods are quite important for obtaining correct

approximations to weak solutions. As discussed above, conservation is associated with the integral forms of the laws. In fact, a naive treatment of the differential form (2.2.1) of the conservation law can lead to inadequate methods. Thus, the notion of conservation and the integral forms of the law need to be taken into account. In particular, the method in Chapter 3 is closely related to the integral formulation (2.2.4), providing the desired conservation properties. In comparison, the conservative form introduced by Lax and Wendroff in [12] is a discrete representation of the integral form expressing the basic principle that the total values of the conserved quantities in any spatial region change only due to flux through the respective spatial boundaries. The resulting exact discrete conservation property, which is common in finite volume and discontinuous Galerkin methods, can be viewed as the following discretization of the integral form (2.2.2):

$$\int_{\partial E} \boldsymbol{n} \cdot \boldsymbol{\Phi}^h(\boldsymbol{u}^h)\, \mathrm{d}\boldsymbol{\sigma} = \mathbf{0}, \tag{2.2.5}$$

for all space-time computational cells (or elements) E, where $\boldsymbol{\Phi}^h$ is an appropriate *numerical flux*. In fact, (2.2.5) holds when E is replaced by any union of computational cells (or elements) [1, 2]. This is related to an important component of the conservation, which is natural and general (i.e., also valid for balance laws). Namely, as in [3, page 362], the numerical flux needs to satisfy the Rankine-Hugoniot condition on the interfaces between cells (or elements) to obtain discrete conservation. This is intuitive since the continuity of the normal components of the flux guarantees that no quantity of the conserved variables is produced or consumed across the interfaces between cells (or elements). It is practical and common to study the behavior and properties of new methods by applying them to scalar PDEs. This path is followed in this dissertation, while the application and extension of the formulations to systems of PDEs is a subject of future work.

A further difficulty arises due to the possible multiplicity of the weak solutions, which is specific to nonlinear problems. Typically, in practice, only one solution is physically relevant [1]. Intuitively, this means that the fundamental conservation that defines the weak solution may lead to, in a sense, an underdetermined problem, since it may not be sufficient for unambiguously singling out the physically important behavior that is being modeled by the PDE. Therefore, additional physical principles are needed to select the *admissible* solution (also called *entropy* solution). This is achieved by introducing additional *entropy* (or *admissibility*) conditions that, in principle, recover the missing physical effects in the model. The name comes from gas dynamics, which is a major

field that uses hyperbolic models. The intuitive idea is that the entropy in the system cannot decrease over time. Thus, admissible discontinuities are either ones that follow the characteristics, or shocks in which the characteristics collide and information disappears, hence, increasing the entropy. In contrast, discontinuities from which characteristics emerge and, thus, information is generated are not admissible, since the entropy decreases. Another justification is related to the fact that hyperbolic conservation and balance laws are often seen as the limits of parabolic-type equations as the viscosity or diffusion vanishes. The vanishing viscosity (or diffusion) idea is rigorously, and in detail, studied in the theory of partial differential equations. In this case, the entropy solutions model the limiting behavior of the vanishing viscosity (or diffusion) solutions. That is, the hyperbolic models can be seen as simplifications of parabolic-type equations that model physical effects with a negligible contribution from diffusion or viscosity. However, the lack of viscous or diffusive terms leads to a possible loss of uniqueness and introduction of nonphysical behaviors, which is remedied by the entropy conditions that provide information on the desired physically relevant solution for the model. A sample setting where non-uniqueness of the weak solution can be observed is associated with rarefaction waves. In that case, as mentioned, the method of characteristics leaves gaps where the solution is not determined by the method. Those gaps can be filled in more than one way to obtain a weak solution. Only one approach is physically valid. Namely, filling the gap with a characteristic "fan" which represents the rarefaction. All other solutions involve inadmissible behaviors, like discontinuities from which characteristics emerge, that are unstable with respect to the presence of any small amount of viscosity or diffusion, which recovers the rarefaction wave as an admissible physical behavior. Entropy and admissibility are not particularly studied in this dissertation. They are discussed here only for completeness in pointing out the challenges associated with hyperbolic equations.

Chapter 3

A Weak Method Based on the Helmholtz Decomposition for Nonlinear Balance Laws

In this chapter, a least-squares finite element method for scalar nonlinear hyperbolic balance laws is proposed and studied. The main focus is on a formulation that utilizes an appropriate Helmholtz decomposition and is closely related to the standard notion of a weak solution. This relationship, together with a corresponding connection to negative-norm least-squares, is described in detail. As a consequence, an important numerical conservation theorem is obtained, similar to the famous Lax-Wendroff theorem. The numerical conservation properties of the method in this chapter do not fall precisely in the framework introduced by Lax and Wendroff, but they are similar in spirit as they guarantee that when certain convergence holds, the resulting approximations approach a weak solution to the hyperbolic problem. The convergence properties of the method are also discussed. Numerical results for the inviscid Burgers equation with discontinuous sources are shown. The numerical method utilizes a least-squares functional and a Gauss-Newton quadraticization technique.

3.1 Introduction

Hyperbolic conservation and balance laws arise often in practice, especially in problems of fluid mechanics [1, 2, 3, 4, 5]. The notion of a weak solution [69] is rather important for this type of partial differential equation (PDE) since it allows the consideration of solutions that posses discontinuities, which are of practical interest. In the numerical treatment of these problems, a related important

property is that the obtained approximations, if they converge, approach a weak solution [12] of the respective hyperbolic PDE. Such a property is related to the ability of the numerical method to correctly approximate weak solutions (i.e., solutions with discontinuities) to nonlinear problems [70, 1, 2, 3]. This is associated with the famous Lax-Wendroff theorem established in [12]. Based on that result, it has become standard, especially in the context of finite volume [2, 3] and discontinuous Galerkin (DG) finite element [13] methods, to consider so-called conservative schemes that posses a certain discrete conservation property. Such a conservation property in the Lax-Wendroff theorem provides a sufficient condition for approximating weak solutions to nonlinear hyperbolic PDEs. As demonstrated in [34], the discrete conservation property, while sufficient, is not necessary for obtaining convergence to a weak solution – a fact that is also utilized in this chapter. As in [34], the considerations here do not precisely abide by the framework provided by Lax and Wendroff in [12]. However, similar to [12, 34], we establish the important and desired numerical conservation property that approximations obtained by the method of this chapter approach a weak solution to the hyperbolic PDE of interest. Instead of the discrete conservation, here, similar to [34], this is due to the utilization of an appropriate least-squares minimization principle that is closely related to the notion of a weak solution. This largely motivates the consideration of the particular formulation in this chapter.

A variety of numerical schemes have been developed for the solution of hyperbolic conservation and balance laws. This includes finite difference and finite volume [1, 2, 4, 3, 10, 11, 5] as well as finite element methods. In the field of finite elements, notably, DG methods (see [13] and the references therein) are often utilized for the solution of hyperbolic PDEs as well as SUPG (streamline-upwind/Petrov-Galerkin) methods [14, 15, 16]. This chapter focuses on least-squares finite element techniques. Least-squares methods [29] have been developed for a variety of problems, including linear [33, 35, 15] and nonlinear [34] first-order hyperbolic PDEs; see also [32, 36, 37, 7, 38]. These approaches utilize appropriate least-squares minimization principles to obtain finite element discretizations of PDEs. Computationally, the problem is reduced to solving linear algebraic systems with symmetric positive definite matrices associated with quadratic minimization problems. Least-squares formulations provide natural error estimates for adaptive mesh refinement [63, 64, 65].

The main contributions of this chapter are summarized as follows. This work proposes and studies a general least-squares finite element formulation for scalar hyperbolic balance laws, which

is based on the standard notion of a weak solution. The fundamental idea is related to the considerations in [34, 32]. Whereas [34, 32] introduce methods that are particularly tailored to *conservation laws*, which only have zero source terms, this chapter extends their ideas to *balance laws*, which allow nonzero sources. The approaches here and in [34, 32] are related in the sense that a Helmholtz decomposition is used to obtain the final formulation. However, the method of this chapter utilizes a different version of the Helmholtz decomposition compared to that of [34, 32], which allows not only the accommodation of source terms but also a natural treatment of the inflow boundary conditions. The differences are discussed in more detail at the end of Subsection 3.4.2. Moreover, the proposed formulation is analyzed and, most notably, an important weak numerical conservation property, similar to [12], is established that, in a sense, extends the respective result in [34, 32]. The method here satisfies such a weak conservation property essentially by design due to the use of the particular Helmholtz decomposition and the resulting relationship of the obtained least-squares principle with the notion of a weak solution. For clarity and simplicity of the considerations, the basic connection between the definition of a weak solution and an H^{-1}-type formulation is identified. Also, the convergence properties of the method with respect to the L^2 norm are discussed. The close and natural relation to the definition of a weak solution, however, contributes to considerable difficulties in showing the desired norm convergence of the obtained approximations.

The approach here extends the ideas in [34, 32]. However, the particular method here and the ones introduced in [34, 32] are different and do not coincide even when applied to conservation laws, since different Helmholtz decompositions are used; see the end of Subsection 3.4.2. Furthermore, a related method for balance laws can be obtained here by introducing an additional vector field variable similar to the so-called "flux vector" formulation for conservation laws in [34, 32]. Such an approach is not particularly considered in this chapter since it is quite closely related to the proposed formulation with analogous properties.

The outline of the rest of the chapter is the following. Basic notions and the utilized Helmholtz decomposition are presented in Section 3.2. Section 3.3 contains a general overview of scalar hyperbolic balance laws. In Section 3.4, the least-squares formulations of interest are introduced, and they are analyzed and studied in more detail in Section 3.5, including numerical conservation and convergence properties. Section 3.6 is devoted to numerical results. Section 3.7 presents additional considerations that are particularly specialized on linear hyperbolic problems. The

conclusions and future work are in the final Section 3.8.

3.2 Basic definitions and the Helmholtz decomposition

Here, basic notation and definitions are presented and the Helmholtz decomposition that is relevant to the considerations in this chapter is stated.

Let Ω be an open, bounded, and simply connected subset of $\mathbb{R}^2$ with a Lipschitz-continuous boundary, $\Gamma = \partial\Omega$, as defined in [71]. In the context of time-dependent hyperbolic problems, $\mathbb{R}^2$ represents the space-time, i.e., it is the tx-space, where t and x are the independent variables. Accordingly, $\nabla\cdot$ denotes the space-time divergence, i.e., $\nabla \cdot \boldsymbol{v} = \partial_t v_1 + \partial_x v_2$, for any appropriate vector field $\boldsymbol{v}\colon \Omega \to \mathbb{R}^2$, $\boldsymbol{v} = [v_1, v_2]$. Similarly, ∇ and $\nabla^\perp$ are space-time differential operators defined as $\nabla v = [\partial_t v, \partial_x v]$ and $\nabla^\perp v = [\partial_x v, -\partial_t v]$, for any appropriate scalar function $v\colon \Omega \to \mathbb{R}$.

Let $\Gamma_S \subset \Gamma$ be a portion of the boundary, Γ, of Ω with a nonzero surface measure. The following denotes the space of $H^1(\Omega)$ functions with vanishing traces on Γ_S:

$$H^1_{0,\Gamma_S}(\Omega) = \Big\{ v \in H^1(\Omega);\ v = 0 \text{ on } \Gamma_S \Big\}.$$

The space $H^1_{0,\Gamma_S}(\Omega)$ can be endowed with the $H^1(\Omega)$ norm: $\|v\|_1^2 = \|v\|^2 + \|\nabla v\|^2$, for $v \in H^1(\Omega)$, where $\|\cdot\|$ denotes the norms on both $L^2(\Omega)$ and $[L^2(\Omega)]^2$. It is convenient to also consider the $H^1(\Omega)$ seminorm: $|v|_1 = \|\nabla v\|$, for all $v \in H^1(\Omega)$. Owing to Poincaré's inequality, using that Γ_S has a nonzero surface measure, $|\cdot|_1$ is a norm in $H^1_{0,\Gamma_S}(\Omega)$, equivalent to $\|\cdot\|_1$; cf., [40, Lemma 3.1 in Chapter I]. Therefore, in this chapter, $H^1_{0,\Gamma_S}(\Omega)$ is endowed with the norm $|\cdot|_1$ and, clearly, it is a Hilbert space with respect to that norm.

It is customary to define the dual of a positive-order Sobolev space as a "negative-order" Sobolev space. Following this practice, the dual space of $H^1_{0,\Gamma_S}(\Omega)$ is denoted by $H^{-1}_{\Gamma_S}(\Omega)$ and it is endowed with the respective functional norm

$$\|\ell\|_{-1,\Gamma_S} = \sup_{v \in H^1_{0,\Gamma_S}(\Omega)} \frac{|\ell(v)|}{|v|_1}, \quad \forall \ell \in H^{-1}_{\Gamma_S}(\Omega),$$

where, to simplify notation, it is understood that $v \neq 0$ in the supremum. In particular, in the special case when $\Gamma_S \equiv \Gamma$, the commonly used notation is $H^1_0(\Omega) = H^1_{0,\Gamma}(\Omega)$ and $H^{-1}(\Omega) = H^{-1}_{\Gamma}(\Omega)$.

The inner products in both $L^2(\Omega)$ and $[L^2(\Omega)]^2$, which are associated with the respective norms $\|\cdot\|$, are denoted by $(\cdot,\cdot)$. Following the notation in [40], the inner product in $L^2(\Gamma)$ is denoted by

$\langle \cdot, \cdot \rangle_\Gamma$. By extending the $L^2(\Gamma)$ inner product into a duality pairing, as is customary, $\langle \cdot, \cdot \rangle_\Gamma$ is also used (as in [40]) to denote the duality pairing between $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$, where $H^{1/2}(\Gamma)$ is the space of traces on Γ of functions in $H^1(\Omega)$ and $H^{-1/2}(\Gamma)$ is its dual; see [40, 41, 72, 73, 74].

The Sobolev space of square integrable vector fields on Ω with square integrable divergence (see [40, 41]) is defined as

$$H(\mathrm{div};\ \Omega) = \Big\{ \boldsymbol{v} \in [L^2(\Omega)]^2;\ \nabla \cdot \boldsymbol{v} \in L^2(\Omega) \Big\},$$

where (cf., [72, 62]) $\nabla \cdot \boldsymbol{v} \in L^2(\Omega)$, for $\boldsymbol{v} \in [L^2(\Omega)]^2$, is understood in the sense that there exists a (unique) function $v \in L^2(\Omega)$ such that

$$-(\boldsymbol{v}, \nabla\phi) = (v, \phi), \quad \forall \phi \in H^1_0(\Omega), \tag{3.2.1}$$

in which case $\nabla \cdot \boldsymbol{v} = v \in L^2(\Omega)$.

Using the notation above, we can write the following Green's formula [40]:

$$(\boldsymbol{v}, \nabla\phi) + (\nabla \cdot \boldsymbol{v}, \phi) = \langle \boldsymbol{v} \cdot \boldsymbol{n}, \phi \rangle_\Gamma, \quad \forall \boldsymbol{v} \in H(\mathrm{div};\ \Omega),\ \forall \phi \in H^1(\Omega), \tag{3.2.2}$$

where $\boldsymbol{n}$ is the unit outward normal to Γ.

Finally, assume that Γ is split into two non-overlapping relatively open subsurfaces Γ_1 and Γ_2 of nonzero surface measures, i.e., $\Gamma = \overline{\Gamma_1} \cup \overline{\Gamma_2}$ and $\Gamma_1 \cap \Gamma_2 = \emptyset$. Also, Γ_1 and Γ_2 are assumed to consist of finite numbers of connected components. Similar to [40, Sections 2 and 3 of Chapter I], the following Helmholtz decomposition can be obtained; cf., [40, Theorem 3.2 in Chapter I].

Theorem 3.2.1 (Helmholtz decomposition) *For every $\boldsymbol{v} \in [L^2(\Omega)]^2$, the following L^2-orthogonal decomposition holds:*

$$\boldsymbol{v} = \nabla q + \nabla^\perp \psi, \tag{3.2.3}$$

where $q \in H^1_{0,\Gamma_1}(\Omega)$ is the unique solution to

$$\textit{Find } q \in H^1_{0,\Gamma_1}(\Omega) \colon (\nabla q, \nabla\phi) = (\boldsymbol{v}, \nabla\phi), \quad \forall \phi \in H^1_{0,\Gamma_1}(\Omega), \tag{3.2.4}$$

and $\psi \in H^1_{0,\Gamma_2}(\Omega)$ is the unique solution to

$$\textit{Find } \psi \in H^1_{0,\Gamma_2}(\Omega) \colon (\nabla^\perp \psi, \nabla^\perp \nu) = (\boldsymbol{v}, \nabla^\perp \nu), \quad \forall \nu \in H^1_{0,\Gamma_2}(\Omega). \tag{3.2.5}$$

Remark 3.2.2 The weak problem (3.2.5) can be interpreted, formally, as the following elliptic PDE for ψ:

$$\begin{aligned} -\Delta\psi &= \operatorname{curl} \boldsymbol{v} && \text{in } \Omega, \\ \psi &= 0 && \text{on } \Gamma_2, \\ \frac{\partial\psi}{\partial \boldsymbol{n}} &= -(\boldsymbol{v} - \nabla q)\cdot\boldsymbol{\tau} = -\boldsymbol{v}\cdot\boldsymbol{\tau} && \text{on } \Gamma_1. \end{aligned}$$

Here, Δ denotes the Laplace operator, $\Delta\psi = \partial_{tt}\psi + \partial_{xx}\psi$, $\operatorname{curl}\boldsymbol{v} = \partial_t v_2 - \partial_x v_1$, and $\boldsymbol{\tau} = [-n_2, n_1]$, where $\boldsymbol{n} = [n_1, n_2]$, is the unit tangent to the boundary, Γ. ◇

Remark 3.2.3 In particular, when additionally $\boldsymbol{v} \in H(\operatorname{div};\,\Omega)$, then, using (3.2.2), (3.2.4) can be expressed as

$$\text{Find } q \in H^1_{0,\Gamma_1}(\Omega)\colon (\nabla q, \nabla\phi) = -(\nabla\cdot\boldsymbol{v}, \phi) + \langle \boldsymbol{v}\cdot\boldsymbol{n}, \phi\rangle_\Gamma, \quad \forall \phi \in H^1_{0,\Gamma_1}(\Omega), \tag{3.2.6}$$

which is interpreted (cf., [40, Corollary 2.6 in Chapter I], see also [74]) as the weak formulation of the following elliptic PDE for q:

$$\begin{aligned} \Delta q &= \nabla\cdot\boldsymbol{v} \text{ in } \Omega, \\ q &= 0 \qquad \text{on } \Gamma_1, \\ \frac{\partial q}{\partial \boldsymbol{n}} &= \boldsymbol{v}\cdot\boldsymbol{n} \text{ on } \Gamma_2. \end{aligned}$$

For general $\boldsymbol{v} \in [L^2(\Omega)]^2$, (3.2.4) can be interpreted the same way, but only formally. ◇

3.3 Scalar hyperbolic balance laws

This section provides an overview of the basic notions and properties associated with hyperbolic balance laws. This serves as a foundation for the sections that follow.

In this chapter, we consider scalar *hyperbolic balance laws* (see [2]) of the form

$$\nabla\cdot\boldsymbol{f}(u) = r \text{ in } \Omega, \tag{3.3.1a}$$

$$u = g \text{ on } \Gamma_I, \tag{3.3.1b}$$

where the generally nonlinear flux vector $\boldsymbol{f}\colon \mathbb{R} \to \mathbb{R}^2$, $\boldsymbol{f} \in [L^\infty_{\mathrm{loc}}(\mathbb{R})]^2$, $\boldsymbol{f} = [f_1, f_2]$, the source term $r \in L^2(\Omega)$, the inflow boundary data $g \in L^\infty(\Gamma_I)$ are given, and u is the unknown dependent

variable. Recall that $L^\infty_{\mathrm{loc}}(\mathbb{R})$ is the space of measurable functions that are in $L^\infty(J)$, for all compact subsets $J \subset \mathbb{R}$. Clearly, under the assumptions on $\boldsymbol{f}$ below, (3.3.1a) can be represented as a first-order quasilinear PDE for u. When $r \equiv 0$, (3.3.1) becomes a *hyperbolic conservation law*. Here, Γ_I denotes the inflow portion of the boundary, Γ, to be considered in more detail below. In problems of type (3.3.1) that are of practical interest, $\boldsymbol{f}$ is often continuously differentiable and $f_1 \equiv \iota$, where $\iota\colon \mathbb{R} \to \mathbb{R}$ is the identity function $\iota(v) = v$, $v \in \mathbb{R}$. Nevertheless, it is convenient here to consider scalar balance laws in the general form (3.3.1). Since the focus is on weak solutions to (3.3.1) (defined below), we only assume that the components of the flux vector, $\boldsymbol{f}$, are locally Lipschitz-continuous on $\mathbb{R}$; that is, for every compact subset $J \subset \mathbb{R}$, there exists a constant $K_{\boldsymbol{f},J} > 0$, which generally depends on $\boldsymbol{f}$ and the set J, such that

$$|f_i(v_1) - f_i(v_2)| \leq K_{\boldsymbol{f},J}|v_1 - v_2|, \quad \forall v_1, v_2 \in J, \quad i = 1, 2. \tag{3.3.2}$$

Note that, by Rademacher's theorem (see, e.g., [54, Theorem 6 in Subsection 5.8.3]), (3.3.2) implies that $\boldsymbol{f}$ is differentiable, in the classical sense, almost everywhere (a.e.) in $\mathbb{R}$ and $\boldsymbol{f}' \in [L^\infty_{\mathrm{loc}}(\mathbb{R})]^2$.

The assumption (3.3.2) is reasonable and mild since it includes a wide class of problems. In particular, any continuously differentiable $\boldsymbol{f}$ clearly satisfies (3.3.2). Moreover, certain requirements on the Lipschitz-continuity of the "numerical flux" are also encountered in the literature on finite volume methods; see, e.g., [2, Subsection 4.3.1][3, Subsection 4.1.2 of Chapter IV]. In general, there are PDEs of interest with discontinuous flux functions; see, e.g., [75] and the references therein. In this chapter, for simplicity of the analysis, we concentrate on problems that satisfy (3.3.2). Nevertheless, the considered formulations are also sensible in the general case of discontinuous $\boldsymbol{f}$. A further investigation in that direction is a subject of future work. Currently, (3.3.2) admits discontinuous $\boldsymbol{f}'$, which allows a quasilinear PDE with discontinuous coefficients.

Remark 3.3.1 In view of [54, Subsection 5.8.2b], the above assumptions on the flux vector, $\boldsymbol{f}$, are equivalent to the simple assumption $\boldsymbol{f} \in [W^{1,\infty}_{\mathrm{loc}}(\mathbb{R})]^2$. The implied a.e. differentiability of $\boldsymbol{f}$ is sufficient for the notions considered in this section. Further assumptions are made as they become necessary. ◇

According to the method of characteristics (see, e.g., [56, 55, 54, 57]), the characteristics of (3.3.1a) have directions determined by $\boldsymbol{f}'(\hat{u})$, where $\hat{u}$ is an exact (weak) solution to (3.3.1) (defined

below) of interest; i.e., in the nonlinear case, the characteristics depend on the solution. This motivates the following definition of portions of the boundary, Γ, which also depend on the solution, in the nonlinear case.

Definition 3.3.2 Let $\hat{u}$ be an exact (weak) solution to (3.3.1) (defined below) of interest. The *inflow* portion of the boundary, Γ, of Ω is defined as (see [32, 34, 35])

$$\Gamma_I = \{\, \boldsymbol{x} \in \Gamma;\ \boldsymbol{f}'(\hat{u}) \cdot \boldsymbol{n} < 0 \,\}.$$

Similarly, the *outflow* portion of the boundary and the portion that is tangential to the flow are, respectively,

$$\Gamma_O = \{\, \boldsymbol{x} \in \Gamma;\ \boldsymbol{f}'(\hat{u}) \cdot \boldsymbol{n} > 0 \,\}, \quad \Gamma_T = \{\, \boldsymbol{x} \in \Gamma;\ \boldsymbol{f}'(\hat{u}) \cdot \boldsymbol{n} = 0 \,\}.$$

The complement (essentially) in Γ of Γ_I is $\Gamma_C = \Gamma_O \cup \Gamma_T = \{\, \boldsymbol{x} \in \Gamma;\ \boldsymbol{f}'(\hat{u}) \cdot \boldsymbol{n} \geq 0 \,\}$. ◇

This motivates the consideration of boundary conditions that are posed on the inflow boundary, Γ_I, in (3.3.1b). Furthermore, a consistency requirement on the inflow data, g, is that $\boldsymbol{f}'(g) \cdot \boldsymbol{n} < 0$ on Γ_I.

Note that weak solutions (defined below) to (3.3.1) of practical interest possess a certain structure. Namely, solutions to (3.3.1), that are important in practical cases, are piecewise continuously differentiable; see [3, 54, 2, 1, 10]. For completeness, we include the definition of piecewise continuously differentiable functions on Ω, which is useful in the context here; see also [3].

Definition 3.3.3 A function u on Ω is *piecewise continuously differentiable* (or, in short, *piecewise* $\mathcal{C}^1$) if Ω can be partitioned by a finite number of continuous curves into a finite number of open simply connected subdomains, Ω_i, for $i = 1, \ldots, m$, with Lipschitz-continuous boundaries, $\partial\Omega_i$, such that $u \in \mathcal{C}^1(\Omega_i) \cap \mathcal{C}(\overline{\Omega_i})$, for $i = 1, \ldots, m$. Hence, u is allowed to have only jump discontinuities (with finite jumps) across the interfaces of the subdomains. ◇

Remark 3.3.4 Observe that any piecewise $\mathcal{C}^1$ function on Ω is uniformly continuous on the respective subdomains, Ω_i. Thus, piecewise $\mathcal{C}^1$ functions are in $L^\infty(\Omega)$ and also possess a simple notion of restrictions (traces) on Γ, since they are clearly piecewise continuous on Γ. Indeed, any such trace is well-defined in a pointwise sense on Γ excluding a finite number of points, i.e., excluding a set

of zero surface measure. Clearly, sets of zero measure are irrelevant for the weak formulations considered below. In particular, for cases of practical interest, when piecewise $\mathcal{C}^1$ solutions are sought, the boundary data in (3.3.1b), g, is piecewise continuous. These considerations provide a meaning behind (3.3.1b), i.e., the boundary condition can be understood in the sense of the above discussed traces. In terms of Sobolev spaces, for any piecewise $\mathcal{C}^1$ function u, it holds that $u \in H^{1/2-\epsilon}(\Omega)$ and $u|_{\Gamma_I} \in H^{1/2-\epsilon}(\Gamma_I)$, for any real $\epsilon > 0$. Hence, for boundary data, g, of practical interest, it also holds that $g \in H^{1/2-\epsilon}(\Gamma_I)$. $\diamond$

Remark 3.3.5 Under the assumption of local Lipschitz-continuity (3.3.2), $\boldsymbol{f}(u)$ is piecewise continuous when u is piecewise $\mathcal{C}^1$. Thus, $\boldsymbol{f}(u) \in [L^\infty(\Omega)]^2$ and $\boldsymbol{f}(u)$, similar to Remark 3.3.4, has a clear notion of a trace on Γ. Furthermore, $\boldsymbol{f}(u) \cdot \boldsymbol{n}$ makes sense a.e. on Γ and $\boldsymbol{f}(u) \cdot \boldsymbol{n} \in L^\infty(\Gamma)$. Similarly, $\boldsymbol{f}'(u)$ and $\boldsymbol{f}'(u) \cdot \boldsymbol{n}$ make sense a.e. on Γ and $\boldsymbol{f}'(u) \cdot \boldsymbol{n} \in L^\infty(\Gamma)$. This provides some substance into the above definition of portions of the boundary, Definition 3.3.2, when piecewise $\mathcal{C}^1$ solutions are considered. Namely, the portions of the boundary in Definition 3.3.2, for a piecewise $\mathcal{C}^1$ solution $\hat{u}$, are defined up to sets of zero surface measure. This is sufficient for the weak formulations considered below. $\diamond$

The notion of a classical solution to (3.3.1) is rather simple; see, e.g., [3]. Namely, $\hat{u} \in \mathcal{C}(\overline{\Omega}) \cap \mathcal{C}^1(\Omega)$ is a *classical solution* to (3.3.1) if it satisfies (3.3.1) in a pointwise sense, where the derivatives in (3.3.1a) are interpreted in the classical sense. This needs the additional assumption that $\boldsymbol{f}$ is continuously differentiable, which often holds, so that the PDE (3.3.1a) would be well-defined everywhere on Ω in the classical sense.

As already discussed, it is common in practical applications of balance laws of the type (3.3.1) to consider solutions that are piecewise $\mathcal{C}^1$. Clearly, in general, such solutions are not classical. Therefore, we are not particularly interested in classical solutions and concentrate on the important notion of a weak solution that extends the notion of a classical solution (i.e., every classical solution is also a weak one) and allows for piecewise $\mathcal{C}^1$ solutions; cf., [3, 2, 1, 54, 55, 11, 12, 69, 34, 10]. In particular, any piecewise $\mathcal{C}^1$ function is a weak solution to (3.3.1) if it satisfies the equation in the classical sense a.e. in the regions where it is continuously differentiable, whereas across jumps it needs to conform to additional jump conditions, which guarantee that it satisfies the equation

(3.3.1a) in a "weak sense" [3]. This is further discussed below after the following definition of a weak solution based on integration by parts.

Definition 3.3.6 A function $\hat{u} \in L^\infty(\Omega)$ is a *weak solution* to (3.3.1) if it satisfies

$$-\int_\Omega \boldsymbol{f}(\hat{u}) \cdot \nabla\phi \, \mathrm{d}\boldsymbol{x} = \int_\Omega r\phi \, \mathrm{d}\boldsymbol{x} - \int_{\Gamma_I} \boldsymbol{f}(g) \cdot \boldsymbol{n}\phi \, \mathrm{d}\sigma, \quad \forall \phi \in \mathcal{C}^1_{0,\Gamma_C}(\overline{\Omega}),$$

where $\mathcal{C}^1_{0,\Gamma_C}(\overline{\Omega}) = \{\, \phi \in \mathcal{C}^1(\overline{\Omega});\ \phi = 0 \text{ on } \Gamma_C \,\}$. In terms of the notation in Section 3.2, this can be equivalently expressed as

$$-(\boldsymbol{f}(\hat{u}), \nabla\phi) = (r, \phi) - \langle \boldsymbol{f}(g) \cdot \boldsymbol{n}, \phi \rangle_\Gamma, \quad \forall \phi \in H^1_{0,\Gamma_C}(\Omega), \tag{3.3.3}$$

using the density (cf., [76]) of $\mathcal{C}^1_{0,\Gamma_C}(\overline{\Omega})$ in $H^1_{0,\Gamma_C}(\Omega)$. ◇

The following lemma is obtained easily from the above definitions; see also [3, 34].

Lemma 3.3.7 *It holds that $\boldsymbol{f}(\hat{u}) \in H(\mathrm{div};\ \Omega)$ and $\nabla \cdot \boldsymbol{f}(\hat{u}) = r$, for any weak solution (in the sense of* (3.3.3)*), $\hat{u} \in L^\infty(\Omega)$, to* (3.3.1)*.*

Proof. It is easy to see, using (3.3.2), that $\boldsymbol{f}(u) \in [L^\infty(\Omega)]^2 \subset [L^2(\Omega)]^2$, for every $u \in L^\infty(\Omega)$. Furthermore, in view of (3.3.3) and (3.2.1), it holds that $\nabla \cdot \boldsymbol{f}(\hat{u}) \in L^2(\Omega)$ and $\nabla \cdot \boldsymbol{f}(\hat{u}) = r \in L^2(\Omega)$, for any weak solution, $\hat{u} \in L^\infty(\Omega)$, to (3.3.1). □

In other words, the PDE (3.3.1a) is satisfied by $\hat{u}$ in an L^2 sense. Also, in view of Lemma 3.3.7, the weak formulation (3.3.3) can be seen, in a sense, as the result of applying the Green's formula (3.2.2) to (3.3.1).

Note that Lemma 3.3.7 holds under quite general assumptions and $\hat{u}$ does not need to be piecewise $\mathcal{C}^1$. In particular, in view of [34, Lemma 2.4] (see also [62, Lemma 5.3(3)]), for any piecewise $\mathcal{C}^1$ function u, using that $\boldsymbol{f}(u)$ is piecewise Lipschitz-continuous, $\boldsymbol{f}(u) \in H(\mathrm{div};\ \Omega)$ is equivalent to the *Rankine-Hugoniot jump condition*

$$[\![\boldsymbol{f}(u) \cdot \boldsymbol{n}]\!]_{\mathscr{C}} = 0 \text{ a.e. on } \mathscr{C},$$

for every curve $\mathscr{C} \subset \Omega$; cf., [34, 3, 54, 55]. Here, $[\![\cdot]\!]_{\mathscr{C}}$ is the jump across the curve $\mathscr{C}$ and $\boldsymbol{n}$ denotes the unit normal to $\mathscr{C}$, where the particular orientation is irrelevant for the jump condition. Furthermore, it is not difficult to generalize the equivalence in [3, Theorem 2.1 on page 16] to the

case of balance laws of the form (3.3.1), utilizing the argument in [3, Theorem 2.1 on page 16] with minor modifications. However, these results are not needed for the considerations in this chapter and are only mentioned to highlight the connection between $\boldsymbol{f}(\hat{u})$ being in $H(\mathrm{div};\,\Omega)$ and the more often encountered, in the context of hyperbolic conservation and balance laws, Rankine-Hugoniot jump condition, when piecewise $\mathcal{C}^1$ weak solutions are considered; see also [34].

3.4 Least-squares formulations

This section is devoted to the least-squares principles, related to the weak formulation (3.3.3), that can be used for deriving finite element methods for balance laws of the form (3.3.1). First, an H^{-1}-based formulation is discussed. Then, the approach based on the Helmholtz decomposition in Theorem 3.2.1, which is a main focus of this chapter, is described.

3.4.1 A H^{-1}-based formulation

Here, we comment on the relation between the definition (3.3.3), of a weak solution to (3.3.1), and the H^{-1}-type spaces defined in Section 3.2. First, "weak-weak divergence" is introduced.

Definition 3.4.1 For any vector field $\boldsymbol{v} \in [L^2(\Omega)]^2$, the "weak-weak divergence" operator

$$[\nabla_{\mathfrak{w}}\cdot]\colon [L^2(\Omega)]^2 \to H^{-1}_{\Gamma_C}(\Omega)$$

is defined as $\nabla_{\mathfrak{w}} \cdot \boldsymbol{v} = \ell_{\boldsymbol{v}} \in H^{-1}_{\Gamma_C}(\Omega)$, where $\ell_{\boldsymbol{v}}(\phi) = -(\boldsymbol{v}, \nabla\phi)$, for all $\phi \in H^1_{0,\Gamma_C}(\Omega)$. ◇

Remark 3.4.2 For the case when $\boldsymbol{v} \in H(\mathrm{div};\,\Omega)$, owing to (3.2.2), the relation between the "weak-weak" and the "standard", $[\nabla\cdot]\colon H(\mathrm{div};\,\Omega) \to L^2(\Omega)$ (defined via (3.2.1)), divergence operators is

$$[\nabla_{\mathfrak{w}} \cdot \boldsymbol{v}](\phi) = -(\boldsymbol{v}, \nabla\phi) = (\nabla \cdot \boldsymbol{v}, \phi) - \langle \boldsymbol{v}\cdot\boldsymbol{n}, \phi\rangle_\Gamma, \quad \forall \phi \in H^1_{0,\Gamma_C}(\Omega).$$

Thus, if additionally $\boldsymbol{v}\cdot\boldsymbol{n} = 0$ on Γ_I, then $\nabla_{\mathfrak{w}} \cdot \boldsymbol{v}$ and $\nabla \cdot \boldsymbol{v}$ can be equated via the standard embedding of $L^2(\Omega)$ into $H^{-1}_{\Gamma_C}(\Omega)$; see [77, Remark 3 in Section 5.2]. Otherwise, $\nabla_{\mathfrak{w}} \cdot \boldsymbol{v}$ and $\nabla \cdot \boldsymbol{v}$ cannot be identified, since $\nabla_{\mathfrak{w}} \cdot \boldsymbol{v}$ treats the terms on the inflow boundary, Γ_I, differently, which is convenient when we return to the weak formulation (3.3.3) below. ◇

Next, consider the linear functional ℓ_d, defined as

$$\ell_d(\phi) = (r, \phi) - \langle \boldsymbol{f}(g) \cdot \boldsymbol{n}, \phi \rangle_\Gamma, \quad \forall \phi \in H^1_{0,\Gamma_C}(\Omega), \tag{3.4.1}$$

where r and g represent the given data in (3.3.1). It is easy to see that $\ell_d \in H^{-1}_{\Gamma_C}(\Omega)$. This functional, $\ell_d \in H^{-1}_{\Gamma_C}(\Omega)$, contains all the given data in (3.3.1) and (3.3.3), i.e., it contains both the source and inflow boundary data. Then, in view of the weak formulation (3.3.3), which defines a weak solution to (3.3.1), and the definition of $\nabla_{\mathfrak{w}}\cdot$, the problem of finding weak solutions to (3.3.1) is equivalent to the problem of finding solutions, in $L^\infty(\Omega)$, to the equation

$$\nabla_{\mathfrak{w}} \cdot \boldsymbol{f}(u) = \ell_d, \tag{3.4.2}$$

where the equality is understood in $H^{-1}_{\Gamma_C}(\Omega)$ sense (i.e., it is understood as the equality of functionals in $H^{-1}_{\Gamma_C}(\Omega)$: $[\nabla_{\mathfrak{w}} \cdot \boldsymbol{f}(u)](\phi) = \ell_d(\phi)$, for all $\phi \in H^1_{0,\Gamma_C}(\Omega)$). Equivalently, this can be expressed as $\|\nabla_{\mathfrak{w}} \cdot \boldsymbol{f}(u) - \ell_d\|_{-1,\Gamma_C} = 0$.

Thus, a natural discrete least-squares formulation is

$$u^h = \underset{v^h \in \mathcal{U}^h}{\operatorname{argmin}} \|\nabla_{\mathfrak{w}} \cdot \boldsymbol{f}(v^h) - \ell_d\|^2_{-1,\Gamma_C}, \tag{3.4.3}$$

for some finite element space $\mathcal{U}^h \subset L^\infty(\Omega)$. In the next subsection, a related (as it is also associated with the weak formulation (3.3.3)) least-squares principle is considered, based on the Helmholtz decomposition in Theorem 3.2.1, which is more convenient for computation.

Finally, the following lemma is useful for the considerations below. It, in a sense, already hints at the relationship between the Helmholtz decomposition and the H^{-1}-based formulations (3.4.2) and (3.4.3).

Lemma 3.4.3 *Let $\boldsymbol{v} \in [L^2(\Omega)]^2$ have the Helmholtz decomposition (provided by Theorem 3.2.1, with $\Gamma_1 = \Gamma_C$ and $\Gamma_2 = \Gamma_I$) $\boldsymbol{v} = \nabla q + \nabla^\perp \psi$, for the respective unique $q \in H^1_{0,\Gamma_C}(\Omega)$ and $\psi \in H^1_{0,\Gamma_I}(\Omega)$. Then it holds that*

$$[\nabla_{\mathfrak{w}} \cdot \boldsymbol{v}](\phi) = -(\nabla q, \nabla \phi), \quad \forall \phi \in H^1_{0,\Gamma_C}(\Omega),$$

$$\|\nabla_{\mathfrak{w}} \cdot \boldsymbol{v}\|_{-1,\Gamma_C} = \|\nabla q\|.$$

Proof. Using the definition of $\nabla_{\mathfrak{w}}\cdot$ and (3.2.4), for any $\phi \in H^1_{0,\Gamma_C}(\Omega)$, it holds that

$$[\nabla_{\mathfrak{w}} \cdot \boldsymbol{v}](\phi) = -(\boldsymbol{v}, \nabla \phi) = -(\nabla q, \nabla \phi),$$

$$\|\nabla_{\mathfrak{w}} \cdot \boldsymbol{v}\|_{-1,\Gamma_C} = \sup_{\phi \in H^1_{0,\Gamma_C}(\Omega)} \frac{|[\nabla_{\mathfrak{w}} \cdot \boldsymbol{v}](\phi)|}{|\phi|_1} = \sup_{\phi \in H^1_{0,\Gamma_C}(\Omega)} \frac{|(\nabla q, \nabla \phi)|}{|\phi|_1} = \|\nabla q\|. \qquad \square$$

As a consequence of Lemma 3.4.3, $\|\nabla_{\mathfrak{w}} \cdot \boldsymbol{v}\|_{-1,\Gamma_C} \le \|\boldsymbol{v}\| = (\|\nabla q\|^2 + \|\nabla^\perp \psi\|^2)^{1/2}$, for any $\boldsymbol{v} \in [L^2(\Omega)]^2$. This shows that $[\nabla_{\mathfrak{w}} \cdot]\colon [L^2(\Omega)]^2 \to H^{-1}_{\Gamma_C}(\Omega)$ is a bounded linear operator. Also, the null space of $\nabla_{\mathfrak{w}} \cdot$ is

$$\mathcal{N}(\nabla_{\mathfrak{w}} \cdot) = \Big\{ \boldsymbol{v} \in [L^2(\Omega)]^2;\ \boldsymbol{v} = \nabla^\perp \nu \text{ for some } \nu \in H^1_{0,\Gamma_I}(\Omega) \Big\}. \tag{3.4.4}$$

3.4.2 A formulation based on the Helmholtz decomposition

In this subsection, the formulation based on the Helmholtz decomposition in Theorem 3.2.1, which is a main focus of this chapter, is described. Here, Theorem 3.2.1 is applied, using the notation $\Gamma_1 = \Gamma_C$ and $\Gamma_2 = \Gamma_I$.

Let $\hat{u} \in L^\infty(\Omega)$ be a weak solution of interest to (3.3.1), as defined by (3.3.3). In view of Lemma 3.3.7, $\boldsymbol{f}(\hat{u}) \in H(\mathrm{div};\ \Omega) \subset [L^2(\Omega)]^2$. Thus, owing to Theorem 3.2.1, consider the Helmholtz decomposition

$$\boldsymbol{f}(\hat{u}) = \nabla q + \nabla^\perp \psi, \tag{3.4.5}$$

for the respective, uniquely determined (by $\boldsymbol{f}(\hat{u})$), $q \in H^1_{0,\Gamma_C}(\Omega)$ and $\psi \in H^1_{0,\Gamma_I}(\Omega)$. Owing to (3.2.4), the definition of $\nabla_{\mathfrak{w}} \cdot$, and (3.4.2) (or, using directly (3.3.3)), it holds that

$$(\nabla q, \nabla \phi) = (\boldsymbol{f}(\hat{u}), \nabla \phi) = -[\nabla_{\mathfrak{w}} \cdot \boldsymbol{f}(\hat{u})](\phi) = -\ell_d(\phi), \quad \forall \phi \in H^1_{0,\Gamma_C}(\Omega), \tag{3.4.6}$$

where $\ell_d \in H^{-1}_{\Gamma_C}(\Omega)$ is defined in (3.4.1). Alternatively, this follows from Lemma 3.3.7 and (3.2.6). Furthermore, in view of Remark 3.2.3, Lemma 3.3.7, and (3.3.1), (3.4.6) is interpreted as the weak formulation of the following elliptic PDE for q:

$$\begin{aligned} \Delta q &= r && \text{in } \Omega, \\ q &= 0 && \text{on } \Gamma_C, \\ \frac{\partial q}{\partial \boldsymbol{n}} &= \boldsymbol{f}(g) \cdot \boldsymbol{n} && \text{on } \Gamma_I. \end{aligned} \tag{3.4.7}$$

Remark 3.4.4 It is known (cf., [1, 2]) that, in general, nonlinear PDEs of the type (3.3.1) can have multiple weak solutions. This is specific to nonlinear problems, since the method of characteristics provides the uniqueness of the solution to first-order linear hyperbolic PDEs. The well-posedness of

linear hyperbolic problems is studied in detail, in a least-squares context, in [33]; see also [32, 35] and Chapter 4. Recall that, as discussed in the previous subsection, the functional $\ell_d \in H^{-1}_{\Gamma_C}(\Omega)$ contains all the given data in the problem (3.3.1). Thus, in view of the weak form (3.4.6), $q \in H^1_{0,\Gamma_C}(\Omega)$, in the decomposition (3.4.5), is uniquely determined by the given data as the solution, interestingly, to the *elliptic* PDE (3.4.7). In contrast, $\psi \in H^1_{0,\Gamma_I}(\Omega)$, in (3.4.5), is uniquely determined once the weak solution, $\hat{u}$, is fixed and, by (3.2.5), it satisfies $(\nabla^\perp \psi, \nabla^\perp \nu) = (\boldsymbol{f}(\hat{u}), \nabla^\perp \nu)$, for all $\nu \in H^1_{0,\Gamma_I}(\Omega)$, but it is not directly determined by the given data in (3.3.1). Hence, if $\tilde{u} \in L^\infty(\Omega)$ is another weak solution to (3.3.1), then $\boldsymbol{f}(\tilde{u}) = \nabla q + \nabla^\perp \tilde{\psi}$, for some $\tilde{\psi} \in H^1_{0,\Gamma_I}(\Omega)$, whereas $q \in H^1_{0,\Gamma_C}(\Omega)$ is the same as in (3.4.5), since it must satisfy (3.4.6). In view of the operator $\nabla_{\mathfrak{w}} \cdot$ and the formulation (3.4.2), this is to be expected, since, from (3.4.4), $[\boldsymbol{f}(\hat{u}) - \boldsymbol{f}(\tilde{u})] \in \mathcal{N}(\nabla_{\mathfrak{w}} \cdot)$. In theory, another possible source of non-uniqueness is if $\boldsymbol{f}(\hat{u}) = \boldsymbol{f}(\tilde{u})$ can hold, even when $\hat{u} \neq \tilde{u}$. However, this cannot happen in practical problems, since, as discussed in Section 3.3, the first component of $\boldsymbol{f}$ is the identity function on $\mathbb{R}$, $f_1 \equiv \iota$. Therefore, the only practically possible source of non-uniqueness of the weak solution to (3.3.1) is the potential non-uniqueness of the $H^1_{0,\Gamma_I}(\Omega)$ component of the decomposition in (3.4.5), since it is determined only implicitly by the given data; that is, once the component $q \in H^1_{0,\Gamma_C}(\Omega)$ in (3.4.5) is fixed by the given data, through the weak problem (3.4.6), any $\psi \in H^1_{0,\Gamma_I}(\Omega)$ can be selected, as long as the equality (3.4.5) would hold for some $\hat{u} \in L^\infty(\Omega)$, which is the only constraint on the $H^1_{0,\Gamma_I}(\Omega)$ component in (3.4.5) that the PDE (3.3.1) and the weak formulation (3.3.3) (or its equivalent (3.4.2)) provide. $\diamond$

Now, based on the Helmholtz decomposition (3.4.5) and the above discussion, (3.3.1) is reformulated as the following first-order system of PDEs, for the unknowns v, p, and μ:

$$\begin{aligned} \boldsymbol{f}(v) - \nabla p - \nabla^\perp \mu &= \boldsymbol{0} \quad \text{in } \Omega, \\ \nabla p &= \nabla q \text{ in } \Omega, \\ p &= 0 \quad \text{on } \Gamma_C, \\ \mu &= 0 \quad \text{on } \Gamma_I, \end{aligned} \tag{3.4.8}$$

where $q \in H^1_{0,\Gamma_C}(\Omega)$ is considered given as the unique solution to (3.4.6). The least-squares functional derived from (3.4.8) is

$$\mathcal{F}(v, p, \mu;\ q) = \|\boldsymbol{f}(v) - \nabla p - \nabla^\perp \mu\|^2 + \|\nabla p - \nabla q\|^2, \tag{3.4.9}$$

defined for $v \in L^\infty(\Omega)$, $p \in H^1_{0,\Gamma_C}(\Omega)$, and $\mu \in H^1_{0,\Gamma_I}(\Omega)$. This results in a least-squares principle for the minimization of $\mathcal{F}$. Simply setting $p = q$ and removing the second term in (3.4.9) is not an option in practice since $q \in H^1_{0,\Gamma_C}(\Omega)$ is only given implicitly as the solution to (3.4.6) and it is a function from the infinite-dimensional Sobolev space $H^1_{0,\Gamma_C}(\Omega)$, that cannot be exactly represented in a practical computation. One approach to addressing such a minimization is by reformulating it as a two-stage process, where the first stage obtains an approximation to $q \in H^1_{0,\Gamma_C}(\Omega)$ by solving (3.4.7) and then, using this approximation, the minimization of $\mathcal{F}$ is addressed in the second stage. Alternatively, owing to (3.4.6), we consider the functional, for $v \in L^\infty(\Omega)$, $p \in H^1_{0,\Gamma_C}(\Omega)$, and $\mu \in H^1_{0,\Gamma_I}(\Omega)$,

$$\begin{aligned} \hat{\mathcal{F}}(v,p,\mu;\ r,g) &= \|\boldsymbol{f}(v) - \nabla p - \nabla^\perp \mu\|^2 + \|\nabla p\|^2 + 2\ell_d(p) \\ &= \|\boldsymbol{f}(v) - \nabla p - \nabla^\perp \mu\|^2 + \|\nabla p\|^2 + 2\left[(r,p) - \langle \boldsymbol{f}(g)\cdot\boldsymbol{n}, p\rangle_\Gamma\right], \end{aligned} \tag{3.4.10}$$

which utilizes only available data and is more convenient for a direct implementation. The minimization of $\mathcal{F}$ is equivalent (in terms of minimizers) to the minimization of $\hat{\mathcal{F}}$. Notice that the minimal value of $\mathcal{F}$ is 0, whereas the minimal value of $\hat{\mathcal{F}}$ is $-\|\nabla q\|^2$. Unlike $\mathcal{F}$, $\hat{\mathcal{F}}$ can be evaluated for any given $(v,p,\mu) \in L^\infty(\Omega) \times H^1_{0,\Gamma_C}(\Omega) \times H^1_{0,\Gamma_I}(\Omega)$ using only available information, without requiring explicit knowledge of q.

Consider finite element spaces $\mathcal{U}^h \subset L^\infty(\Omega)$, $\mathcal{V}^h_{\Gamma_C} \subset H^1_{0,\Gamma_C}(\Omega)$, and $\mathcal{V}^h_{\Gamma_I} \subset H^1_{0,\Gamma_I}(\Omega)$. In general, these spaces do not need to be on the same mesh or of the same order. Nevertheless, for simplicity of notation, we use h to denote the mesh parameters on all spaces and it is not difficult to extend the results and formulations in this chapter to the general case of different mesh parameters. The discrete least-squares formulation of interest in this chapter is

$$\begin{aligned} \text{minimize} \quad & \mathcal{F}(v^h,p^h,\mu^h;\ q) \ \text{ or } \ \hat{\mathcal{F}}(v^h,p^h,\mu^h;\ r,g), \\ \text{for} \quad & v^h \in \mathcal{U}^h,\ p^h \in \mathcal{V}^h_{\Gamma_C},\ \mu^h \in \mathcal{V}^h_{\Gamma_I}. \end{aligned} \tag{3.4.11}$$

If $(u^h,q^h,\psi^h) \in \mathcal{U}^h \times \mathcal{V}^h_{\Gamma_C} \times \mathcal{V}^h_{\Gamma_I}$ is a minimizer of (3.4.11), then $u^h \in \mathcal{U}^h$ is the obtained approximation of a weak solution to (3.3.1).

The discrete least-squares problem (3.4.11) can be approached by methods for solving differentiable unconstrained optimization problems. A common and simple choice, which is tailored to non-quadratic least-squares problems, is the *Gauss-Newton method* (see [78, Chapter 10], [79, Section 10.3]), where the system (more precisely, the first equation in the system) (3.4.8) is linearized

by the Newton method and the resulting linear first-order system is reformulated to a quadratic least-squares method. Namely, let $v_0 \in L^\infty(\Omega)$ be a current iterate. The aim is to obtain a next iterate $v \in L^\infty(\Omega)$. To this purpose, (3.4.8) is linearized around v_0. In general, for a (Fréchet) differentiable nonlinear operator $F(v)$, the Newton linearization of the equation $F(v) = 0$ around v_0 is $F(v_0) + F'(v_0)\delta v = 0$, where $F'(v_0)\delta v$ is, generally, the Gâteaux (i.e., directional) derivative (cf., [72]) of F at v_0 in the direction δv. This is an equation for the *update (step)* δv, where the new iterate is obtained as $v = v_0 + \delta v$, which can be reformulated as a least-squares method. For a more detailed description see [78, Chapter 10], [79, Section 10.3]. In particular, the Newton linearization of (3.4.8) is

$$\begin{aligned} \boldsymbol{f}'(v_0)\delta v - \nabla\delta p - \nabla^\perp\delta\mu &= \nabla p_0 + \nabla^\perp\mu_0 - \boldsymbol{f}(v_0) && \text{in } \Omega, \\ \nabla\delta p &= \nabla q - \nabla p_0 && \text{in } \Omega, \\ \delta p &= 0 && \text{on } \Gamma_C, \\ \delta\mu &= 0 && \text{on } \Gamma_I, \end{aligned}$$

for the unknowns δv, δp, and $\delta\mu$, where $q \in H^1_{0,\Gamma_C}(\Omega)$ is only given implicitly as the solution to (3.4.6). The corresponding quadratic (a quadraticization of $\mathcal{F}$) least-squares functional, for $\delta v \in L^\infty(\Omega)$, $\delta p \in H^1_{0,\Gamma_C}(\Omega)$, and $\delta\mu \in H^1_{0,\Gamma_I}(\Omega)$, is

$$\begin{aligned} \mathcal{F}_l(\delta v, \delta p, \delta\mu;\ v_0, p_0, \mu_0;\ q) = \|\boldsymbol{f}'(v_0)\delta v - \nabla\delta p - \nabla^\perp\delta\mu - \nabla p_0 - \nabla^\perp\mu_0 + \boldsymbol{f}(v_0)\|^2 \\ + \|\nabla\delta p - \nabla q + \nabla p_0\|^2, \end{aligned} \tag{3.4.12}$$

or, alternatively, we consider the quadratic (a quadraticization of $\hat{\mathcal{F}}$) functional

$$\begin{aligned} \hat{\mathcal{F}}_l(\delta v, \delta p, \delta\mu;\ v_0, p_0, \mu_0;\ r, g) = \|\boldsymbol{f}'(v_0)\delta v - \nabla\delta p - \nabla^\perp\delta\mu - \nabla p_0 - \nabla^\perp\mu_0 + \boldsymbol{f}(v_0)\|^2 \\ + |p_0 + \delta p|_1^2 + 2\left[(r, p_0 + \delta p) - \langle \boldsymbol{f}(g)\cdot\boldsymbol{n}, p_0 + \delta p\rangle_\Gamma\right]. \end{aligned} \tag{3.4.13}$$

Thus, for current iterates $u_0^h \in \mathcal{U}^h$, $q_0^h \in \mathcal{V}^h_{\Gamma_C}$, and $\psi_0^h \in \mathcal{V}^h_{\Gamma_I}$, the quadraticized discrete least-squares problem (a quadraticization of (3.4.11)) is

$$\begin{aligned} &\text{minimize} \quad \mathcal{F}_l(\delta u^h, \delta q^h, \delta\psi^h;\ u_0^h, q_0^h, \psi_0^h;\ q) \ \text{ or } \ \hat{\mathcal{F}}_l(\delta u^h, \delta q^h, \delta\psi^h;\ u_0^h, q_0^h, \psi_0^h;\ r, g), \\ &\text{for} \quad \delta u^h \in \mathcal{U}^h,\ \delta q^h \in \mathcal{V}^h_{\Gamma_C},\ \delta\psi_h \in \mathcal{V}^h_{\Gamma_I}. \end{aligned} \tag{3.4.14}$$

Since it utilizes only given data, the computationally feasible weak formulation[1] associated with

[1] It can be derived directly, using the functional $\hat{\mathcal{F}}_l$, or using the functional $\mathcal{F}_l$ in a combination with (3.4.6).

(3.4.14) is: Find $(\delta u^h, \delta q^h, \delta\psi^h) \in \mathcal{U}^h \times \mathcal{V}^h_{\Gamma_C} \times \mathcal{V}^h_{\Gamma_I}$ such that

$$\begin{cases} (\boldsymbol{f}'(u_0^h)\delta u^h - \ \nabla\delta q^h - \nabla^\perp\delta\psi^h,\ \boldsymbol{f}'(u_0^h)v^h) = (\nabla q_0^h + \nabla^\perp\psi_0^h - \boldsymbol{f}(u_0^h),\ \boldsymbol{f}'(u_0^h)v^h), \\ -(\boldsymbol{f}'(u_0^h)\delta u^h - 2\nabla\delta q^h - \nabla^\perp\delta\psi^h,\ \nabla p^h) \qquad = (\boldsymbol{f}(u_0^h) - 2\nabla q_0^h - \nabla^\perp\psi_0^h,\ \nabla p^h) - \ell_d(p^h), \\ -(\boldsymbol{f}'(u_0^h)\delta u^h - \ \nabla\delta q^h - \nabla^\perp\delta\psi^h,\ \nabla^\perp\mu^h) \qquad = (\boldsymbol{f}(u_0^h) - \nabla q_0^h - \nabla^\perp\psi_0^h,\ \nabla^\perp\mu^h), \end{cases}$$

for all $(v^h, p^h, \mu^h) \in \mathcal{U}^h \times \mathcal{V}^h_{\Gamma_C} \times \mathcal{V}^h_{\Gamma_I}$. The final approximation is obtained by iteratively repeating the above procedure. In practice, as in Section 3.6, a *damped* Gauss-Newton approach is preferred by combining the Gauss-Newton method with a *line search*, where the ability to evaluate the functional $\hat{\mathcal{F}}$ is utilized; see [78, 79, 66].

Similar to [34], by viewing, for the moment, $\boldsymbol{f}$ as the nonlinear map $\boldsymbol{f}\colon L^\infty(\Omega) \to [L^2(\Omega)]^2$ and using (3.3.2), one can show that, for any $v_0 \in L^\infty(\Omega)$, $\boldsymbol{f}'(v_0)\colon L^\infty(\Omega) \to [L^2(\Omega)]^2$ is a bounded linear operator, i.e., $\|\boldsymbol{f}'(v_0)v\| \le C_{\boldsymbol{f},v_0}\|v\|_{L^\infty(\Omega)}$, for all $v \in L^\infty(\Omega)$, where the constant $C_{\boldsymbol{f},v_0} > 0$ can generally depend on $\boldsymbol{f}$ and $\|v_0\|_{L^\infty(\Omega)}$. This demonstrates that the mapping $\boldsymbol{f}\colon L^\infty(\Omega) \to [L^2(\Omega)]^2$ is (Fréchet) differentiable on $L^\infty(\Omega)$, which is of basic importance for the applicability of the Gauss-Newton method to the solution of the discrete problem (3.4.11). The convergence of the Gauss-Newton process can be guaranteed under additional assumptions as described in [78, Section 10.2], [79, Section 10.3]; see also [72, Section 7.7].

The first-order system (3.4.8), based on the Helmholtz decomposition (3.4.5), possesses the convenient property that the nonlinearity (i.e., $\boldsymbol{f}(v)$) is only in a zeroth-order term (i.e., a term that does not involve differential operators). Actually, this is more than convenience, since, as observed in [34, 32] for the case of conservation laws, the L^2-type least-squares principle obtained directly from the first-order equation (3.3.1) poses additional difficulties; see [34, 32] for more details. This justifies the utilization of a formulation like (3.4.11), based on the Helmholtz decomposition, which is related to the weak formulation (3.3.3) and the H^{-1}-type least-squares principle (3.4.3). Moreover, this relationship between the formulations provides the desirable numerical conservation properties of (3.4.11); see Section 3.5.

The method here is general and can be applied to balance laws of the type (3.3.1). Particularly, it can be used for conservation laws, $r \equiv 0$. However, although the approach here is related to the ideas in [34, 32], developed for conservation laws, it differs from the methods in [34, 32], even when applied to conservation laws. There, the methods are specially tailored to conservation laws

utilizing a different Helmholtz decomposition. Namely, for any $\boldsymbol{v} \in [L^2(\Omega)]^2$, $\boldsymbol{v} = \nabla\tilde{q} + \nabla^\perp\tilde{\psi}$, for uniquely determined $\tilde{q} \in H^1_0(\Omega)$ and $\tilde{\psi} \in H^1(\Omega)/\mathbb{R}$. Whence, $\nabla \cdot \boldsymbol{v} = 0$ if and only if $\boldsymbol{v} = \nabla^\perp\tilde{\psi}$ for some $\tilde{\psi} \in H^1(\Omega)$; see [40, Theorem 3.1 in Chapter I]. In particular, this is convenient for conservation laws, since, in that case, $\boldsymbol{f}(\tilde{u})$ is divergence free, for any respective weak solution $\tilde{u}$; see Lemma 3.3.7. Another consequence of using the Helmholtz decomposition of Theorem 3.2.1 is that $q \in H^1_{0,\Gamma_C}(\Omega)$ in (3.4.5) carries all given data (both the source term and the inflow condition), whereas an equality like $\boldsymbol{f}(\tilde{u}) = \nabla^\perp\tilde{\psi}$, specific to conservation laws, only provides $\nabla \cdot \boldsymbol{f}(\tilde{u}) = 0$ and does not contain any information on the inflow boundary condition. This motivates the boundary terms $\|\boldsymbol{n} \cdot (\nabla^\perp\tilde{\psi} - \boldsymbol{f}(g))\|^2_{\Gamma_I}$ and $\|v - g\|^2_{\Gamma_I}$ in the functionals in [34, 32]. Here, $\|\cdot\|_{\Gamma_I}$ denotes the norm in $L^2(\Gamma_I)$. Notice that $\boldsymbol{n} \cdot \nabla^\perp\mu = 0$ on Γ_I, for any $\mu \in H^1_{0,\Gamma_I}(\Omega)$, so a term like $\|\boldsymbol{n}\cdot(\nabla^\perp\mu - \boldsymbol{f}(g))\|^2_{\Gamma_I}$ is not useful here. However, although a term like $\|v - g\|^2_{\Gamma_I}$ is meaningless for a general function $v \in L^\infty(\Omega)$, it makes sense for piecewise continuous functions and, particularly, for finite element functions in place of v; see Remarks 3.3.4 and 3.3.5. Thus, in practical computations, the formulations (3.4.11), (3.4.14) can be augmented with such a boundary term or a scaled version of it, i.e., it can be added to the functionals in (3.4.9), (3.4.10), (3.4.12), and (3.4.13). This is utilized in the numerical experiments in Section 3.6. Nonetheless, it is not needed for the study of the analytical properties of the formulation, in Section 3.5, since, as already explained, the boundary data is incorporated in the formulation due to the particular Helmholtz decomposition in Theorem 3.2.1 and the resulting relation to (3.3.3).

3.5 Analysis

This section is devoted to the further study and analysis of the formulation in Subsection 3.4.2. In particular, we address the important numerical conservation properties of the method. Also, the norm convergence of the approximations is discussed. It is convenient to concentrate on the functional $\mathcal{F}$ in (3.4.9). All considerations, with minor modifications, carry over to the functional $\hat{\mathcal{F}}$ in (3.4.10) since, as discussed in Subsection 3.4.2, they provide equivalent formulations. The notation in Subsection 3.4.2 is reused here. In particular, $\hat{u}$ denotes a weak solution to (3.3.1) and the decomposition (3.4.5) is utilized.

3.5.1 Weak solutions and the conservation property

Here, we investigate in more detail the relationship between (3.4.11) and the weak formulation (3.3.3). As a consequence, the numerical conservation property of the method is obtained.

The following approximation bounds are used as assumptions in the results below:

$$\exists\, \hat{u}^h \in \mathcal{U}^h \ : \|\hat{u}^h - \hat{u}\| \qquad \leq Ch^s \|\hat{u}\|_s, \qquad 0 < s \leq \beta_{\hat{u}}, \tag{3.5.1}$$

$$\exists\, \hat{q}^h \in \mathcal{V}^h_{\Gamma_C} : \|\nabla(\hat{q}^h - q)\| \ \leq Ch^s \|q\|_{s+1}, \quad 0 < s \leq \beta_q, \tag{3.5.2}$$

$$\exists\, \hat{\psi}^h \in \mathcal{V}^h_{\Gamma_I} : \|\nabla(\hat{\psi}^h - \psi)\| \leq Ch^s \|\psi\|_{s+1}, \quad 0 < s \leq \beta_\psi, \tag{3.5.3}$$

where $C > 0$ and $\beta_{\hat{u}}, \beta_q, \beta_\psi > 0$ are some constants that do not depend on h, the functions $\hat{u}$, q, and ψ are defined in (3.4.5), and $\|\cdot\|_s$, for $s \in \mathbb{R}_+$, denotes the norm in the Sobolev space $H^s(\Omega)$. The assumptions (3.5.1)–(3.5.3) are associated with well known interpolation bounds of polynomial approximation theory; see [58, 59, 60, 61, 40]. As known, the approximation orders, $\beta_{\hat{u}}$, β_q, and β_ψ, depend on the smoothness (in the Sobolev sense) of the functions $\hat{u}$, q, and ψ, respectively, and the polynomial orders of the respective finite element spaces, $\mathcal{U}^h$, $\mathcal{V}^h_{\Gamma_C}$, and $\mathcal{V}^h_{\Gamma_I}$.

As already discussed, the formulations (3.4.11) and (3.4.3) are related as they both are based on the notion of a weak solution defined by (3.3.3). In fact, this relationship can be seen more directly. For $v^h \in \mathcal{U}^h$, consider the corresponding Helmholtz decomposition $\boldsymbol{f}(v^h) = \nabla q_{v^h} + \nabla^\perp \psi_{v^h}$, where $q_{v^h} \in H^1_{0,\Gamma_C}(\Omega)$ and $\psi_{v^h} \in H^1_{0,\Gamma_I}(\Omega)$. Then, using Lemma 3.4.3, (3.4.5), and (3.4.2), it follows that

$$\begin{aligned}
\min_{\substack{p\in H^1_{0,\Gamma_C}(\Omega)\\ \mu\in H^1_{0,\Gamma_I}(\Omega)}} \mathcal{F}(v^h, p, \mu;\ q) &= \min_{p\in H^1_{0,\Gamma_C}(\Omega)} \Big[\|\nabla q_{v^h} - \nabla p\|^2 + \|\nabla p - \nabla q\|^2\Big] \\
&= \frac{1}{2}\|\nabla q_{v^h} - \nabla q\|^2 = \frac{1}{2}\|\nabla_{\mathfrak{w}} \cdot [\boldsymbol{f}(v^h) - \boldsymbol{f}(\hat{u})]\|^2_{-1,\Gamma_C} \\
&= \frac{1}{2}\|\nabla_{\mathfrak{w}} \cdot \boldsymbol{f}(v^h) - \ell_d\|^2_{-1,\Gamma_C}.
\end{aligned} \tag{3.5.4}$$

Thus, the minimization (3.4.3) is equivalent to the minimization

$$\begin{aligned}
\text{minimize} \quad & \mathcal{F}(v^h, p, \mu;\ q) \ \text{ or } \ \hat{\mathcal{F}}(v^h, p, \mu;\ r, g), \\
\text{for} \quad & v^h \in \mathcal{U}^h,\ p \in H^1_{0,\Gamma_C}(\Omega),\ \mu \in H^1_{0,\Gamma_I}(\Omega).
\end{aligned}$$

However, in a practical, fully discrete formulation like (3.4.11), (discrete) finite element spaces $\mathcal{V}^h_{\Gamma_C} \subset H^1_{0,\Gamma_C}(\Omega)$ and $\mathcal{V}^h_{\Gamma_I} \subset H^1_{0,\Gamma_I}(\Omega)$ are utilized. Therefore, a more detailed study of the relationship between (3.4.3) and the fully discrete (3.4.11) is in the sequel. It is important since it,

essentially, represents the relationship between (3.4.11) and the weak formulation (3.3.3), which, in turn, is important for the numerical conservation properties, considered below, of the formulation (3.4.11). To this end, for simplicity of exposition, the following functional is introduced, for $v \in L^\infty(\Omega)$:

$$\mathcal{G}^h(v;\, q) = \min_{\substack{p^h \in \mathcal{V}^h_{\Gamma_C} \\ \mu^h \in \mathcal{V}^h_{\Gamma_I}}} \mathcal{F}(v, p^h, \mu^h;\, q).$$

Considering, for $v \in L^\infty(\Omega)$, the corresponding Helmholtz decomposition $\boldsymbol{f}(v) = \nabla q_v + \nabla^\perp \psi_v$, where $q_v \in H^1_{0,\Gamma_C}(\Omega)$ and $\psi_v \in H^1_{0,\Gamma_I}(\Omega)$, it is easy to see that

$$\mathcal{G}^h(v;\, q) = \min_{\substack{p^h \in \mathcal{V}^h_{\Gamma_C} \\ \mu^h \in \mathcal{V}^h_{\Gamma_I}}} \left[\|\nabla p^h - \nabla q_v\|^2 + \|\nabla^\perp \mu^h - \nabla^\perp \psi_v\|^2 + \|\nabla p^h - \nabla q\|^2 \right]. \tag{3.5.5}$$

Notice that the minimization in (3.5.5), which defines the functional $\mathcal{G}^h$, is a discrete least-squares problem, where v and q are considered given. It is trivial to check that the respective formulation is $[H^1_{0,\Gamma_C}(\Omega) \times H^1_{0,\Gamma_I}(\Omega)]$-equivalent, implying the existence and uniqueness of a minimizer; cf., [72, Theorem 6.1-1]. Thus, the functional $\mathcal{G}^h$ is well-defined. Also, problem (3.4.11) can be equivalently expressed as

$$u^h = \operatorname*{argmin}_{v^h \in \mathcal{U}^h} \mathcal{G}^h(v^h;\, q). \tag{3.5.6}$$

Now, the relationship between (3.4.3) and (3.4.11) (or (3.5.6)) as well as other properties of the formulation (3.4.11) (or (3.5.6)) are shown in the following results; see [32, 34] for a related discussion on conservation laws.

Theorem 3.5.1 *For any $v^h \in \mathcal{U}^h$, the following estimate holds:*

$$\frac{1}{2}\|\nabla_{\mathfrak{w}} \cdot \boldsymbol{f}(v^h) - \ell_d\|^2_{-1,\Gamma_C} \leq \mathcal{G}^h(v^h;\, q).$$

Proof. By (3.5.4) and the definition of $\mathcal{G}^h$,

$$\frac{1}{2}\|\nabla_{\mathfrak{w}} \cdot \boldsymbol{f}(v^h) - \ell_d\|^2_{-1,\Gamma_C} = \min_{\substack{p \in H^1_{0,\Gamma_C}(\Omega) \\ \mu \in H^1_{0,\Gamma_I}(\Omega)}} \mathcal{F}(v^h, p, \mu;\, q) \leq \mathcal{G}^h(v^h;\, q).$$

∎

Theorem 3.5.2 *Assume the approximation bounds* (3.5.2) *and* (3.5.3). *For $v^h \in \mathcal{U}^h$, consider the corresponding Helmholtz decomposition $\boldsymbol{f}(v^h) = \nabla q_{v^h} + \nabla^\perp \psi_{v^h}$, where $q_{v^h} \in H^1_{0,\Gamma_C}(\Omega)$ and*

$\psi_{v^h} \in H^1_{0,\Gamma_I}(\Omega)$. *Then, using the notation introduced in* (3.4.5) *and* (3.4.1)*, the following estimates hold, for some constant* $C > 0$*:*

$$\begin{aligned}\mathcal{G}^h(v^h;\, q) &\le 2\|\nabla_{\mathfrak{w}} \cdot \boldsymbol{f}(v^h) - \ell_d\|^2_{-1,\Gamma_C} \\ &\quad + Ch^{2\beta_q}\|q\|^2_{\beta_q+1} + \min_{\mu^h \in \mathcal{V}^h_{\Gamma_I}} \|\nabla(\mu^h - \psi_{v^h})\|^2,\end{aligned}$$

$$\mathcal{G}^h(v^h;\, q) \le 2\|\boldsymbol{f}(v^h) - \boldsymbol{f}(\hat{u})\|^2 + Ch^{2\beta_q}\|q\|^2_{\beta_q+1} + Ch^{2\beta_\psi}\|\psi\|^2_{\beta_\psi+1}. \tag{3.5.7}$$

Proof. Recall that $\hat{u}$ is a weak solution to (3.3.1), i.e., it solves the equation (3.4.2). By (3.5.5), Lemma 3.4.3, (3.4.5), (3.4.2), (3.5.2), and the obvious $\|\nabla^\perp \psi\| = \|\nabla \psi\|$, it follows that

$$\begin{aligned}\mathcal{G}^h(v^h;\, q) &= \min_{\substack{p^h \in \mathcal{V}^h_{\Gamma_C} \\ \mu^h \in \mathcal{V}^h_{\Gamma_I}}} [\|\nabla p^h - \nabla q_{v^h}\|^2 + \|\nabla^\perp \mu^h - \nabla^\perp \psi_{v^h}\|^2 + \|\nabla p^h - \nabla q\|^2] \\ &= \min_{\substack{p^h \in \mathcal{V}^h_{\Gamma_C} \\ \mu^h \in \mathcal{V}^h_{\Gamma_I}}} [\|\nabla(p^h - q) + \nabla(q - q_{v^h})\|^2 + \|\nabla^\perp(\mu^h - \psi_{v^h})\|^2 + \|\nabla(p^h - q)\|^2] \\ &\le 2\|\nabla(q_{v^h} - q)\|^2 + \min_{\mu^h \in \mathcal{V}^h_{\Gamma_I}} \|\nabla^\perp(\mu^h - \psi_{v^h})\|^2 + 3 \min_{p^h \in \mathcal{V}^h_{\Gamma_C}} \|\nabla(p^h - q)\|^2 \\ &\le 2\|\nabla_{\mathfrak{w}} \cdot [\boldsymbol{f}(v^h) - \boldsymbol{f}(\hat{u})]\|^2_{-1,\Gamma_C} + \min_{\mu^h \in \mathcal{V}^h_{\Gamma_I}} \|\nabla(\mu^h - \psi_{v^h})\|^2 + Ch^{2\beta_q}\|q\|^2_{\beta_q+1} \\ &= 2\|\nabla_{\mathfrak{w}} \cdot \boldsymbol{f}(v^h) - \ell_d\|^2_{-1,\Gamma_C} + \min_{\mu^h \in \mathcal{V}^h_{\Gamma_I}} \|\nabla(\mu^h - \psi_{v^h})\|^2 + Ch^{2\beta_q}\|q\|^2_{\beta_q+1}.\end{aligned}$$

Finally, owing to the definition of $\mathcal{G}^h$, (3.5.2), (3.4.5), and (3.5.3), it holds that

$$\begin{aligned}\mathcal{G}^h(v^h;\, q) &= \min_{\substack{p^h \in \mathcal{V}^h_{\Gamma_C} \\ \mu^h \in \mathcal{V}^h_{\Gamma_I}}} \left[\|\boldsymbol{f}(v^h) - \nabla p^h - \nabla^\perp \mu^h\|^2 + \|\nabla(p^h - q)\|^2\right] \\ &\le \|\boldsymbol{f}(v^h) - \nabla \hat{q}^h - \nabla^\perp \hat{\psi}^h\|^2 + \|\nabla(\hat{q}^h - q)\|^2 \\ &\le \|\boldsymbol{f}(v^h) - \boldsymbol{f}(\hat{u}) + \nabla(q - \hat{q}^h) + \nabla^\perp(\psi - \hat{\psi}^h)\|^2 + Ch^{2\beta_q}\|q\|^2_{\beta_q+1} \\ &\le 2\|\boldsymbol{f}(v^h) - \boldsymbol{f}(\hat{u})\|^2 + Ch^{2\beta_q}\|q\|^2_{\beta_q+1} + Ch^{2\beta_\psi}\|\psi\|^2_{\beta_\psi+1},\end{aligned}$$

where $\hat{q}^h \in \mathcal{V}^h_{\Gamma_C}$ and $\hat{\psi}^h \in \mathcal{V}^h_{\Gamma_I}$ satisfy the bounds (3.5.2) and (3.5.3), respectively. ∎

Remark 3.5.3 The term $\min_{\mu^h \in \mathcal{V}^h_{\Gamma_I}} \|\nabla(\mu^h - \psi_{v^h})\|^2$ above can be further treated by considering an approximation bound similar to (3.5.3), but for the function ψ_{v^h}. This provides the estimate, with the respective approximation order $\beta_{\psi_h} > 0$,

$$\mathcal{G}^h(v^h;\, q) \le 2\|\nabla_{\mathfrak{w}} \cdot \boldsymbol{f}(v^h) - \ell_d\|^2_{-1,\Gamma_C} + Ch^{2\beta_q}\|q\|^2_{\beta_q+1} + Ch^{2\beta_{\psi_h}}\|\psi_{v^h}\|^2_{\beta_{\psi_h}+1}.$$

We should note that the bounds (3.5.1)–(3.5.3) are for some a priori fixed weak solution, $\hat{u}$, and they are invariant with respect to the arguments of the functionals above. In contrast, a similar bound for ψ_{v^h} depends on the argument, v^h, of the functional. $\diamond$

Corollary 3.5.4 *Assume the approximation bounds* (3.5.2) *and* (3.5.3)*. Furthermore, consider a subset $\mathcal{Q}^h \subset \mathcal{U}^h$ that is bounded in the $L^\infty(\Omega)$ norm, i.e., there is a constant $B > 0$ such that $\|\hat{u}\|_{L^\infty(\Omega)} \leq B$ and $\|v^h\|_{L^\infty(\Omega)} \leq B$, for all $v^h \in \mathcal{Q}^h$. Then, for some constants $C > 0$ and $C_{\boldsymbol{f},B} > 0$, where $C_{\boldsymbol{f},B}$ generally depends on $\boldsymbol{f}$ and B, it holds that*

$$\mathcal{G}^h(v^h;\, q) \leq C_{\boldsymbol{f},B}\|v^h - \hat{u}\|^2 + Ch^{2\beta_q}\|q\|^2_{\beta_q+1} + Ch^{2\beta_\psi}\|\psi\|^2_{\beta_\psi+1}.$$

Proof. Consider the compact interval $J = [-B, B]$ in (3.3.2). By (3.5.7) and (3.3.2),

$$\begin{aligned}\mathcal{G}^h(v^h;\, q) &\leq 2\|\boldsymbol{f}(v^h) - \boldsymbol{f}(\hat{u})\|^2 + Ch^{2\beta_q}\|q\|^2_{\beta_q+1} + Ch^{2\beta_\psi}\|\psi\|^2_{\beta_\psi+1} \\ &\leq 4K^2_{\boldsymbol{f},J}\|v^h - \hat{u}\|^2 + Ch^{2\beta_q}\|q\|^2_{\beta_q+1} + Ch^{2\beta_\psi}\|\psi\|^2_{\beta_\psi+1}.\end{aligned}$$

$\square$

Corollary 3.5.5 *Assume the approximation bounds* (3.5.1)–(3.5.3) *and that* (3.5.6) *has a minimizer $u^h \in \mathcal{U}^h$. Furthermore, assume that $\hat{u}^h \in \mathcal{U}^h$, which satisfies the bound in* (3.5.1)*, can be selected such that it forms a bounded sequence in $L^\infty(\Omega)$ as $h \to 0$, i.e., there is a constant $B > 0$ such that $\|\hat{u}\|_{L^\infty(\Omega)} \leq B$ and $\|\hat{u}^h\|_{L^\infty(\Omega)} \leq B$ as $h \to 0$. Then, for some constants $C > 0$ and $C_{\boldsymbol{f},B} > 0$, where $C_{\boldsymbol{f},B}$ generally depends on $\boldsymbol{f}$ and B, it holds that*

$$\mathcal{G}^h(u^h;\, q) \leq C_{\boldsymbol{f},B}h^{2\beta_{\hat{u}}}\|\hat{u}\|^2_{\beta_{\hat{u}}} + Ch^{2\beta_q}\|q\|^2_{\beta_q+1} + Ch^{2\beta_\psi}\|\psi\|^2_{\beta_\psi+1}.$$

In particular, this implies that $\mathcal{G}^h(u^h;\, q) \to 0$ as $h \to 0$.

Proof. Similar to Corollary 3.5.4, using (3.5.7), (3.3.2), and (3.5.1), it holds that

$$\begin{aligned}\mathcal{G}^h(u^h;\, q) \leq \mathcal{G}^h(\hat{u}^h;\, q) &\leq C_{\boldsymbol{f},B}\|\hat{u}^h - \hat{u}\|^2 + Ch^{2\beta_q}\|q\|^2_{\beta_q+1} + Ch^{2\beta_\psi}\|\psi\|^2_{\beta_\psi+1} \\ &\leq C_{\boldsymbol{f},B}h^{2\beta_{\hat{u}}}\|\hat{u}\|^2_{\beta_{\hat{u}}} + Ch^{2\beta_q}\|q\|^2_{\beta_q+1} + Ch^{2\beta_\psi}\|\psi\|^2_{\beta_\psi+1}.\end{aligned}$$

$\square$

Remark 3.5.6 It may be more instructive to express the estimate in Corollary 3.5.5, for the corresponding minimizer $(u^h, q^h, \psi^h) \in \mathcal{U}^h \times \mathcal{V}^h_{\Gamma_C} \times \mathcal{V}^h_{\Gamma_I}$ of (3.4.11), as

$$\mathcal{F}(u^h, q^h, \psi^h;\, q) \leq C_{\boldsymbol{f},B}h^{2\beta_{\hat{u}}}\|\hat{u}\|^2_{\beta_{\hat{u}}} + Ch^{2\beta_q}\|q\|^2_{\beta_q+1} + Ch^{2\beta_\psi}\|\psi\|^2_{\beta_\psi+1},$$

which implies that $\mathcal{F}(u^h, q^h, \psi^h;\ q) \to 0$ as $h \to 0$; that is, the minimal value of $\mathcal{F}$ in (3.4.11) tends to 0 (its lower bound) as $h \to 0$. $\diamond$

The above results suggest that asymptotically (i.e., as $h \to 0$) the formulations in (3.4.3) and (3.4.11) approach each other. This can be interpreted that, in a sense, (3.4.11) behaves like the weak formulation (3.3.3) in the limit, since (3.4.3) is closely related to (3.3.3). A rather important consequence of this is the "numerical conservation" property of the least-squares formulation (3.4.11), which is the topic of the next theorem. The classical notion of "numerical conservation" (or "conservative schemes") is associated with the result in [12] in the context of hyperbolic conservation laws; cf., [3, page 168 and Subsections 4.1.2, 4.2.2 of Chapter IV][2, Sections 12.9 and 12.10], see also [11, 1, 69]. This has been an important guiding principle in the design of numerical schemes for conservation laws, since it guarantees that, when certain convergence occurs, the limit is a weak solution. As observed in [34, 32], also in the context of conservation laws, the discrete conservation property in the Lax-Wendroff theorem [12], while sufficient, is not necessary for obtaining weak solutions. As in [34, 32], the considerations here do not fall precisely into the framework introduced in [12], but they are similar in spirit. Namely, assuming appropriate convergence of the discrete solutions, the limit is guaranteed to be a weak solution to (3.3.1). Similar to [12] (and also to [34, 32]), the theorem below does not guarantee that the convergence holds; instead, it assumes that it holds, and does not provide information on which weak solution is obtained, if more than one exist; cf., Remark 3.4.4. Nevertheless, this result is very important and it largely motivates the consideration of the formulation (3.4.11). In fact, it is not surprising that (3.4.11) possesses such a property, since it is closely related to (3.3.3) by design. In particular, this relationship to the notion of a weak solution is associated with the ability of the method to provide correct approximations to piecewise $\mathcal{C}^1$ (i.e., discontinuous) weak solutions to (3.3.1).

Theorem 3.5.7 (numerical conservation property) *Let* (3.5.6) *possess a minimizer* $u^h \in \mathcal{U}^h$ *(or, equivalently, let* (3.4.11) *possess a minimizer* $(u^h, q^h, \psi^h) \in \mathcal{U}^h \times \mathcal{V}^h_{\Gamma_C} \times \mathcal{V}^h_{\Gamma_I}$*) and let the assumptions in Corollary 3.5.5 hold. Assume, in addition,* $L^2(\Omega)$ *convergence,*

$$\lim_{h\to 0} \|u^h - \tilde{u}\| = 0, \tag{3.5.8}$$

for some function $\tilde{u} \in L^\infty(\Omega)$*, and that* u^h *forms a bounded sequence in* $L^\infty(\Omega)$ *as* $h \to 0$*, i.e.,*

there is a constant $B > 0$ *such that* $\|\tilde{u}\|_{L^\infty(\Omega)} \leq B$ *and* $\|u^h\|_{L^\infty(\Omega)} \leq B$ *as* $h \to 0$. *Then* $\tilde{u}$ *is a weak solution of* (3.3.1) *in the sense of* (3.3.3).

Proof. As in the proof of Corollary 3.5.4, by (3.3.2), (3.5.8) implies that

$$\lim_{h\to 0} \|\boldsymbol{f}(u^h) - \boldsymbol{f}(\tilde{u})\| = 0.$$

Thus,

$$(\boldsymbol{f}(u^h), \nabla\phi) \xrightarrow{h\to 0} (\boldsymbol{f}(\tilde{u}), \nabla\phi), \quad \forall \phi \in H^1_{0,\Gamma_C}(\Omega).$$

Owing to Corollary 3.5.5 and Theorem 3.5.1, it holds that

$$\lim_{h\to 0} \|\nabla_{\mathfrak{w}} \cdot \boldsymbol{f}(u^h) - \ell_d\|_{-1,\Gamma_C} = 0.$$

This implies, using the definitions of $\nabla_{\mathfrak{w}}\cdot$ and ℓ_d, in (3.4.1), that

$$-(\boldsymbol{f}(u^h), \nabla\phi) \xrightarrow{h\to 0} (r, \phi) - \langle \boldsymbol{f}(g)\cdot\boldsymbol{n}, \phi\rangle_\Gamma, \quad \forall \phi \in H^1_{0,\Gamma_C}(\Omega).$$

Combining the above results provides (cf., (3.3.3))

$$-(\boldsymbol{f}(\tilde{u}), \nabla\phi) = (r, \phi) - \langle \boldsymbol{f}(g)\cdot\boldsymbol{n}, \phi\rangle_\Gamma, \quad \forall \phi \in H^1_{0,\Gamma_C}(\Omega). \qquad \square$$

The assumptions that $\hat{u}^h$ and u^h in Corollary 3.5.5 and Theorem 3.5.7 are bounded in $L^\infty(\Omega)$ sense are reasonable, especially when approximating piecewise $\mathcal{C}^1$ weak solutions to (3.3.1). Furthermore, similar assumptions can be seen in the classical result in [12]; see also [11, Theorem 10.17][3, Proposition 4.1 on page 378]. Additionally, it is easy to see that the convergence assumption (3.5.8) can be replaced by a convergence in the $L^1(\Omega)$ norm or in a pointwise a.e. sense, i.e., the result is similar to the one in [12] (also in [11, 3]).

Note that in Theorem 3.5.7, unlike the formulations specific to conservation laws in [34, 32], no special treatment and assumptions associated with the boundary conditions are necessary. This is due to the Helmholtz decomposition in Theorem 3.2.1, as discussed in the end of Subsection 3.4.2, which incorporates the boundary conditions into the formulation in a natural way, i.e., in a way related to (3.3.3).

3.5.2 Convergence discussion

The conservation property in Theorem 3.5.7 is natural for the discrete least-squares formulation (3.4.11) since it is based on (3.3.3) and (3.4.3). A norm convergence like (3.5.8) is more challenging to show and it may potentially be too strong for formulations closely related to the notion of a weak solution (3.3.3). Here, we comment on the convergence properties of (3.4.11) (or (3.5.6)). To simplify the considerations, notice that the assumptions on the minimizer, u^h, in Theorem 3.5.7 are utilized to obtain the convergence in $[L^2(\Omega)]^2$ of $\boldsymbol{f}(u^h)$ to $\boldsymbol{f}(\tilde{u})$, i.e., of the nonlinear term. Conversely, it is reasonable to assume that

$$c\|v_1 - v_2\| \leq \|\boldsymbol{f}(v_1) - \boldsymbol{f}(v_2)\|, \quad \forall v_1, v_2 \in L^\infty(\Omega), \tag{3.5.9}$$

for some $c > 0$, since, as discussed, in practice $f_1 \equiv \iota$, the identity function on $\mathbb{R}$. Thus, under (3.5.9), L^2-convergence of $\boldsymbol{f}(u^h)$ implies convergence of u^h.

Therefore, the question reduces to the convergence properties of (3.4.11) with respect to $\boldsymbol{f}(u^h)$. It is not difficult to observe that a uniform coercivity, which provides the respective control of the $[L^2(\Omega)]^2$ norm, is not an innate property of the functional $\mathcal{F}$ in (3.4.9). Indeed, using the decomposition $\boldsymbol{f}(v) = \nabla q_v + \nabla^\perp \psi_v$,

$$\|\nabla q_v\|^2 + \|\nabla^\perp(\psi_v - \mu)\|^2 + \|\nabla p\|^2 = \|\boldsymbol{f}(v) - \nabla^\perp \mu\|^2 + \|\nabla p\|^2 \leq 3\,\mathcal{F}(v, p, \mu;\, 0),$$

for all $(v, p, \mu) \in L^\infty(\Omega) \times H^1_{0,\Gamma_C}(\Omega) \times H^1_{0,\Gamma_I}(\Omega)$. This, as well as (3.5.4), suggests that the functional $\mathcal{F}$ provides only partial control of the $[L^2(\Omega)]^2$ norm. Namely, it explicitly controls only the $H^1_{0,\Gamma_C}(\Omega)$ component of the Helmholtz decomposition, not the $H^1_{0,\Gamma_I}(\Omega)$ component. This is due to the close relation to the formulations (3.3.3) and (3.4.2); cf., (3.4.4) and Remark 3.4.4. Furthermore, similar to [32], one can construct, even in the linear case, an oscillatory counterexample (see Example 3.7.1 below) to demonstrate the lack of an appropriate uniform coercivity that controls the $[L^2(\Omega)]^2$ norm, or simply observe that such coercivity, together with (3.5.9), would imply the uniqueness of the weak solution to (3.3.1), which, as discussed in Remark 3.4.4, does not hold in general.

Proving a norm convergence like (3.5.8) is often challenging for conservative methods that are based on a weak formulation like (3.3.3); see [3, the comments preceding Proposition 4.1 on page 378 and Remark 4.5 on page 388][32, 34, 10]. On the other hand, it is desirable, in the context of finite

element methods, to obtain L^2- or L^1-convergence of the respective finite element approximations. Currently, Theorem 3.5.1 and Corollary 3.5.5 only imply the convergence of $\boldsymbol{f}(u^h)$ with respect to the seminorm $\|\nabla_{\mathfrak{w}} \cdot \boldsymbol{v}\|_{-1,\Gamma_C}$, for $\boldsymbol{v} \in [L^2(\Omega)]^2$.

The above discussion suggests that L^2-convergence may not require a uniform L^2-coercivity, but it may possibly be due to the utilization of the functional $\mathcal{F}$ in the discrete setting in formulation (3.4.11), since an L^2-coercivity, while sufficient, may not be necessary for convergence. A complete analysis of the convergence is potentially rather deep and might involve multiple factors that contribute to such a behavior. Our purpose here is to discuss the possibility of L^2-convergence despite the lack of a uniform L^2-coercivity and consider some factors that can contribute to that convergence. Numerical results that demonstrate the desired L^2-convergence are shown in Section 3.6.

First, the simplest setting that can provide L^2-convergence is discussed. Let (3.5.6) have a minimizer $\tilde{u}^h \in \mathcal{U}^h$ and $\boldsymbol{f}(\tilde{u}^h) = \nabla q_{\tilde{u}^h} + \nabla^\perp \psi_{\tilde{u}^h}$. Consider $q \in H^1_{0,\Gamma_C}(\Omega)$ as defined in (3.4.5)–(3.4.7) and assume, for some $C > 0$ and $\beta > 0$, that

$$\|\nabla_{\mathfrak{w}} \cdot \boldsymbol{f}(\tilde{u}^h) - \ell_d\|_{-1,\Gamma_C} = \|\nabla(q_{\tilde{u}^h} - q)\| \leq Ch^\beta. \tag{3.5.10}$$

In view of Theorem 3.5.1 and Corollary 3.5.5, under the respective assumptions, it holds that $\beta \geq \min\{\, \beta_{\hat{u}}, \beta_q, \beta_\psi \,\}$. Further, assume the (semi-discrete) inf-sup condition, for some $c, \alpha > 0$,

$$\inf_{\boldsymbol{s}^h \in \boldsymbol{\mathcal{S}}^h} \sup_{p \in H^1_{0,\Gamma_C}(\Omega)} \frac{|(\boldsymbol{s}^h, \nabla p)|}{\|\boldsymbol{s}^h\| \|\nabla p\|} \geq ch^\alpha, \tag{3.5.11}$$

where $\boldsymbol{\mathcal{S}}^h = \operatorname{span}\{\, \boldsymbol{f}(v^h);\ v^h \in \mathcal{U}^h \,\} \subset [L^2(\Omega)]^2$.

Theorem 3.5.8 *Let* (3.5.10) *and* (3.5.11) *hold with* $\beta > \alpha$ *and suppose that* $\mathcal{U}^h$ *forms an increasing sequence of nested spaces as* $h \to 0$. *Then* $\boldsymbol{f}(\tilde{u}^h)$ *converges in* $[L^2(\Omega)]^2$ *as* $h \to 0$.

Proof. Assumption (3.5.10) implies that $\|\nabla(q_{\tilde{u}^h} - q_{\tilde{u}^{h/2}})\| \leq Ch^\beta$. Clearly, from the definition of $\nabla_{\mathfrak{w}}\cdot$, (3.5.11) is equivalent to

$$\|\nabla_{\mathfrak{w}} \cdot \boldsymbol{s}^h\|_{-1,\Gamma_C} \geq ch^\alpha \|\boldsymbol{s}^h\|, \quad \forall \boldsymbol{s}^h \in \boldsymbol{\mathcal{S}}^h. \tag{3.5.12}$$

The nestedness of the $\mathcal{U}^h$ spaces provides $[\boldsymbol{f}(\tilde{u}^h) - \boldsymbol{f}(\tilde{u}^{h/2})] \in \boldsymbol{\mathcal{S}}^{h/2}$ and, by (3.5.12),

$$\|\nabla_{\mathfrak{w}} \cdot (\boldsymbol{f}(\tilde{u}^h) - \boldsymbol{f}(\tilde{u}^{h/2}))\|_{-1,\Gamma_C} \geq ch^\alpha \|\boldsymbol{f}(\tilde{u}^h) - \boldsymbol{f}(\tilde{u}^{h/2})\|.$$

Combining the above estimates with Lemma 3.4.3 implies that $\|\boldsymbol{f}(\tilde{u}^h) - \boldsymbol{f}(\tilde{u}^{h/2})\| \leq Ch^{\beta-\alpha}$, which, together with $\beta > \alpha$, can be used to show that $\boldsymbol{f}(\tilde{u}^h)$ forms a Cauchy sequence in $[L^2(\Omega)]^2$. ∎

Remark 3.5.9 Here, the inf-sup condition (3.5.11) is utilized differently compared to the more usual setting of standard mixed finite element methods [41, 58] or of Chapter 4. Typically, inf-sup conditions express a relation that is easily satisfied in the continuous (i.e., infinite-dimensional) case and finite element spaces are appropriately selected to maintain the property in the discrete setting. In our considerations, generally, a continuous version of (3.5.11) does not hold since it is essentially the uniform coercivity that was already discussed; that is, (3.5.11) is a relation that can hold only discretely and fails in the continuous setting. Namely, (3.5.11) is an assumption on the "discrete coercivity" (3.5.12). Moreover, the condition (3.5.11) restrains the space $\mathcal{U}^h$ (based on the flux $\boldsymbol{f}$), thus potentially limiting the freedom of choice of $\mathcal{U}^h$. Actually, this is partially related to the heuristic considerations in [32, Subsection 6.4.2] and provides certain justification for using $\mathcal{C}^0$ (i.e., Lagrangian) finite element spaces as $\mathcal{U}^h$. In that case, owing to (3.3.2), $\boldsymbol{\mathcal{S}}^h \subset [H^1(\Omega)]^2 \subset H(\mathrm{div}; \Omega)$, which is a convenient property, since, in view of Lemma 3.3.7 and (3.3.3), it guarantees that every $v^h \in \mathcal{U}^h$ is a weak solution to some balance law of the type (3.3.1). ◇

Assumption (3.5.11) involves only the discrete space $\mathcal{U}^h$ and not $\mathcal{V}^h_{\Gamma_C}$, $\mathcal{V}^h_{\Gamma_I}$ in (3.4.11). In fact, (3.5.11) is more closely related to the H^{-1}-based formulation (3.4.3). Indeed, similar to Theorem 3.5.8, the respective L^2-convergence for the minimizers of (3.4.3) can be shown, using the corresponding assumptions (3.5.10) and (3.5.11). Now, we consider assumptions that are more suitable for the discrete formulation (3.4.11). In particular, especially due to the space $\mathcal{V}^h_{\Gamma_I}$, (3.4.11) can potentially provide better "control" of the $[L^2(\Omega)]^2$ norm compared to (3.4.3). First, define the following distance and a corresponding subset of $\boldsymbol{\mathcal{S}}^h$:

$$R_{\tilde{u}^h} = \min_{\mu^h \in \mathcal{V}^h_{\Gamma_I}} \|\nabla(\mu^h - \psi_{\tilde{u}^h})\|,$$

$$\boldsymbol{\mathcal{R}}^h = \left\{ \boldsymbol{s}^h \in \boldsymbol{\mathcal{S}}^h;\ \min_{\mu^h \in \mathcal{V}^h_{\Gamma_I}} \|\nabla(\mu^h - \psi_{\boldsymbol{s}^h})\| \leq 2R_{\tilde{u}^{2h}},\ \text{where } \boldsymbol{s}^h = \nabla q_{\boldsymbol{s}^h} + \nabla^\perp \psi_{\boldsymbol{s}^h} \right\},$$

where $\tilde{u}^h$ and $\tilde{u}^{2h}$ denote minimizers of (3.5.6) for respective mesh parameters h and $2h$. Assume the following "restricted" version of the inf-sup condition, for some $c, \gamma > 0$:

$$\inf_{\boldsymbol{r}^h \in \boldsymbol{\mathcal{R}}^h} \sup_{p \in H^1_{0,\Gamma_C}(\Omega)} \frac{|(\boldsymbol{r}^h, \nabla p)|}{\|\boldsymbol{r}^h\| \|\nabla p\|} \geq ch^\gamma. \tag{3.5.13}$$

Notice that if (3.5.11) holds, then (3.5.13) also holds and, generally, $\gamma \leq \alpha$. The following convergence result is obtained, which can potentially be stronger than Theorem 3.5.8.

Theorem 3.5.10 *Let* (3.5.10) *and* (3.5.13) *hold with* $\beta > \gamma$ *and suppose that* $\mathcal{U}^h$, $\mathcal{V}^h_{\Gamma_I}$ *form respective increasing sequences of nested spaces as* $h \to 0$. *Assume also that* $R_{\tilde{u}^h} \leq R_{\tilde{u}^{2h}}$, *for any value of the mesh parameter,* h. *Then* $\boldsymbol{f}(\tilde{u}^h)$ *converges in* $[L^2(\Omega)]^2$ *as* $h \to 0$.

Proof. By (3.5.10), it holds that $\|\nabla(q_{\tilde{u}^{2h}} - q_{\tilde{u}^h})\| \leq Ch^\beta$. Also, (3.5.13) is equivalent to

$$\|\nabla_{\mathfrak{w}} \cdot \boldsymbol{r}^h\|_{-1,\Gamma_C} \geq ch^\gamma \|\boldsymbol{r}^h\|, \quad \forall \boldsymbol{r}^h \in \boldsymbol{\mathcal{R}}^h. \tag{3.5.14}$$

The nestedness of the $\mathcal{U}^h$ spaces provides $[\boldsymbol{f}(\tilde{u}^{2h}) - \boldsymbol{f}(\tilde{u}^h)] \in \boldsymbol{\mathcal{S}}^h$. Consider

$$\nu^{2h} = \underset{\mu^{2h} \in \mathcal{V}^{2h}_{\Gamma_I}}{\operatorname{argmin}} \|\nabla(\mu^{2h} - \psi_{\tilde{u}^{2h}})\|, \quad \nu^h = \underset{\mu^h \in \mathcal{V}^h_{\Gamma_I}}{\operatorname{argmin}} \|\nabla(\mu^h - \psi_{\tilde{u}^h})\|.$$

The nestedness of the $\mathcal{V}^h_{\Gamma_I}$ spaces implies $\nu^{2h} \in \mathcal{V}^h_{\Gamma_I}$. Then

$$\begin{aligned} \min_{\mu^h \in \mathcal{V}^h_{\Gamma_I}} \|\nabla(\mu^h - (\psi_{\tilde{u}^{2h}} - \psi_{\tilde{u}^h}))\| &\leq \|\nabla(\nu^{2h} - \nu^h - \psi_{\tilde{u}^{2h}} + \psi_{\tilde{u}^h})\| \\ &\leq \|\nabla(\nu^{2h} - \psi_{\tilde{u}^{2h}})\| + \|\nabla(\nu^h - \psi_{\tilde{u}^h})\| \\ &= R_{\tilde{u}^{2h}} + R_{\tilde{u}^h} \leq 2R_{\tilde{u}^{2h}}. \end{aligned}$$

Thus, $[\boldsymbol{f}(\tilde{u}^{2h}) - \boldsymbol{f}(\tilde{u}^h)] \in \boldsymbol{\mathcal{R}}^h$ and, by (3.5.14),

$$\|\nabla_{\mathfrak{w}} \cdot (\boldsymbol{f}(\tilde{u}^{2h}) - \boldsymbol{f}(\tilde{u}^h))\|_{-1,\Gamma_C} \geq ch^\gamma \|\boldsymbol{f}(\tilde{u}^{2h}) - \boldsymbol{f}(\tilde{u}^h)\|.$$

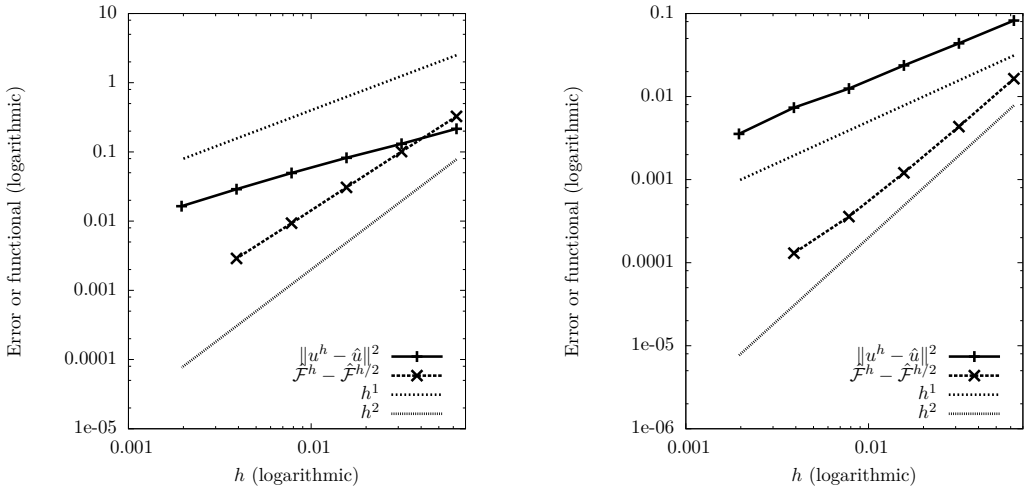
Combining the above estimates with Lemma 3.4.3 implies that $\|\boldsymbol{f}(\tilde{u}^{2h}) - \boldsymbol{f}(\tilde{u}^h)\| \leq Ch^{\beta-\gamma}$, which, together with $\beta > \gamma$, can be used to show that $\boldsymbol{f}(\tilde{u}^h)$ forms a Cauchy sequence in $[L^2(\Omega)]^2$. ∎

Remark 3.5.11 The assumption $R_{\tilde{u}^h} \leq R_{\tilde{u}^{2h}}$ in Theorem 3.5.10 is reasonable since $R_{\tilde{u}^h} \leq Ch^\delta$, for some $C, \delta > 0$. In view of (3.5.5) and Corollary 3.5.5, under the respective assumptions, it holds that $\delta \geq \min\{\, \beta_{\hat{u}}, \beta_q, \beta_\psi \,\}$. More precisely, δ depends on the smoothness of $\psi_{\tilde{u}^h}$ in relation to an approximation bound like (3.5.3). ◇

Theorems 3.5.8 and 3.5.10, together with (3.5.9), show that $\tilde{u}^h$, a minimizer of (3.5.6), converges in the $L^2(\Omega)$ norm to some function $\tilde{u} \in L^2(\Omega)$. These theorems imply that the rate of convergence

is, respectively, $\mathcal{O}(h^{\beta-\alpha})$ or $\mathcal{O}(h^{\beta-\gamma})$. Under the additional assumption in Theorem 3.5.7 that $\tilde{u}^h$ forms a bounded sequence in $L^\infty(\Omega)$, it can be shown that $\tilde{u} \in L^\infty(\Omega)$ and, by Theorem 3.5.7, $\tilde{u}$ is a weak solution to (3.3.1). This analysis does not determine which weak solution is obtained.

The complete study of the convergence properties is still an open and challenging question. The main purpose here is to justify that the convergence (3.5.8) is plausible due to some "weak control" of the $L^2(\Omega)$ norm in the discrete setting, even when a uniform coercivity does not hold, and present basic tools that can aid the analysis of such $L^2(\Omega)$ norm convergence. The possible multiplicity of the weak solutions may contribute to additional difficulties in any further investigation. The above results are stated in the context of formulation (3.4.11) (or (3.5.6)). However, as mentioned, Theorem 3.5.8 more naturally corresponds to (3.4.3), whereas Theorem 3.5.10 is further specialized to reflect the specifics of (3.4.11). In particular, the order α in (3.5.11) (or the equivalent discrete coercivity (3.5.12)) reflects a certain weak control of the $L^2(\Omega)$ norm. Notice that (3.5.11) takes into account the worst case in terms of control and the proper handling of that case is required in Theorem 3.5.8 to obtain the convergence of (3.4.3) and, accordingly, of (3.4.11). In contrast, Theorem 3.5.10 demonstrates that handling the globally worst case may not be necessary, since, in view of (3.5.5), formulation (3.4.11) (or (3.5.6)) enforces certain "proximity" of $\psi_{\tilde{u}^h}$ to the discrete space $\mathcal{V}^h_{\Gamma_I}$. Thus, (3.4.11) may provide better control of the $L^2(\Omega)$ norm compared to (3.4.3). Intuitively, this can be interpreted that $\mathcal{V}^h_{\Gamma_I}$ may act as a "filter" that diminishes certain modes and behaviors that may hinder the convergence; e.g., it may "filter out" or dampen oscillatory modes, like the ones in [32, Subsection 6.4.1] and Example 3.7.1, that are utilized to demonstrate the lack of uniform coercivity and represent components that can jeopardize the L^2-convergence. Here, we suggest that the proximity of $\psi_{\tilde{u}^h}$ to $\mathcal{V}^h_{\Gamma_I}$ may "enhance" the control of the $H^1_{0,\Gamma_I}(\Omega)$ component of the Helmholtz decomposition (3.4.5), whereas the $H^1_{0,\Gamma_C}(\Omega)$ component is naturally controlled by $\mathcal{F}$, thus providing (or "enhancing") the L^2-convergence. In summary, the convergence (3.5.8) may stem from a complex relationship between the spaces involved in the formulation (3.4.11) that depends on the flux vector, $\boldsymbol{f}$.



(a) $\mathcal{U}^h, \mathcal{V}^h_{\Gamma_C}, \mathcal{V}^h_{\Gamma_I}$ – linear (b) $\mathcal{U}^h$ – linear, $\mathcal{V}^h_{\Gamma_C}, \mathcal{V}^h_{\Gamma_I}$ – quadratic

Figure 3.1: Convergence results for Example 1.

3.6 Numerical examples

This section is devoted to numerical results for formulation (3.4.11), utilizing a damped Gauss-Newton procedure, applied to the inviscid Burgers equation, which is of the form (3.3.1) for $\boldsymbol{f}(v) = [v, v^2/2]$, with a discontinuous source term, r. The examples are inspired by [80], which also provides the exact solutions for computing errors. As mentioned in the end of Subsection 3.4.2, the functional $\hat{\mathcal{F}}$ in (3.4.10) is replaced, for practical purposes, by the following "augmented" version (for simplicity, the notation is reused):

$$\hat{\mathcal{F}}(v^h, p^h, \mu^h;\ r, g) = \|\boldsymbol{f}(v^h) - \nabla p^h - \nabla^\perp \mu^h\|^2 + \|\nabla p^h\|^2 + 2\ell_d(p^h) + \|h^{1/2}(v^h - g)\|^2_{\Gamma_I},$$

where g is given in (3.3.1b) and $\|\cdot\|_{\Gamma_I}$ denotes the norm in $L^2(\Gamma_I)$. In all cases, continuous finite element spaces on structured triangular meshes are used. Here, u^h denotes the obtained approximation, $\hat{u}$ is the exact solution, $\hat{\mathcal{F}}^h$ denotes the obtained minimal value of the functional, $\hat{\mathcal{F}}$, on a mesh with a parameter h. The meshes consist of right-crossed squares, ◪, where the coarsest mesh has 16 squares in t and 32 squares in x, while the finer meshes are obtained by consecutive uniform refinements.

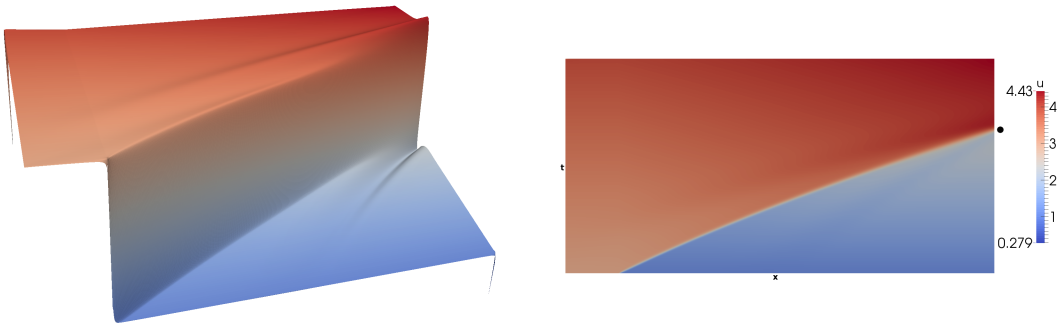


Figure 3.2: The approximation, u^h, obtained from Example 1 on the finest mesh, when all spaces, $\mathcal{U}^h$, $\mathcal{V}^h_{\Gamma_C}$, and $\mathcal{V}^h_{\Gamma_I}$, are linear. The black dot, $\bullet$, shows where the shock exists the domain in the exact solution, $\hat{u}$.

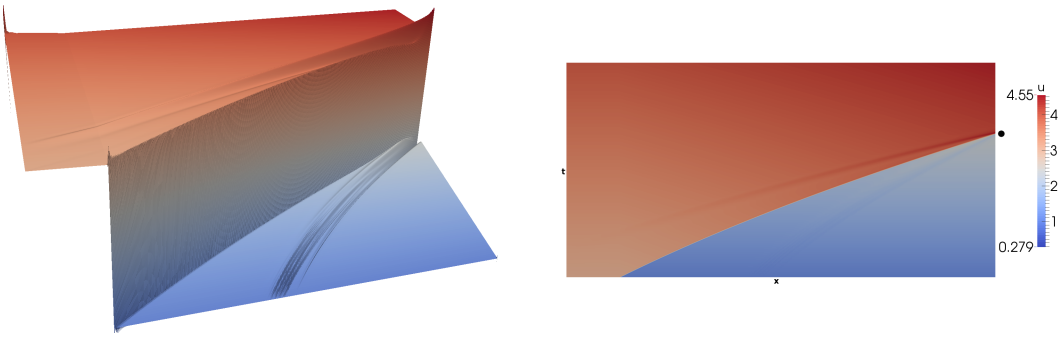


Figure 3.3: The approximation, u^h, obtained from Example 1 on the finest mesh, when $\mathcal{U}^h$ is linear and $\mathcal{V}^h_{\Gamma_C}, \mathcal{V}^h_{\Gamma_I}$ are quadratic. The black dot, $\bullet$, shows where the shock exists the domain in the exact solution, $\hat{u}$.

Example 1 (a single shock) Consider (3.3.1) with

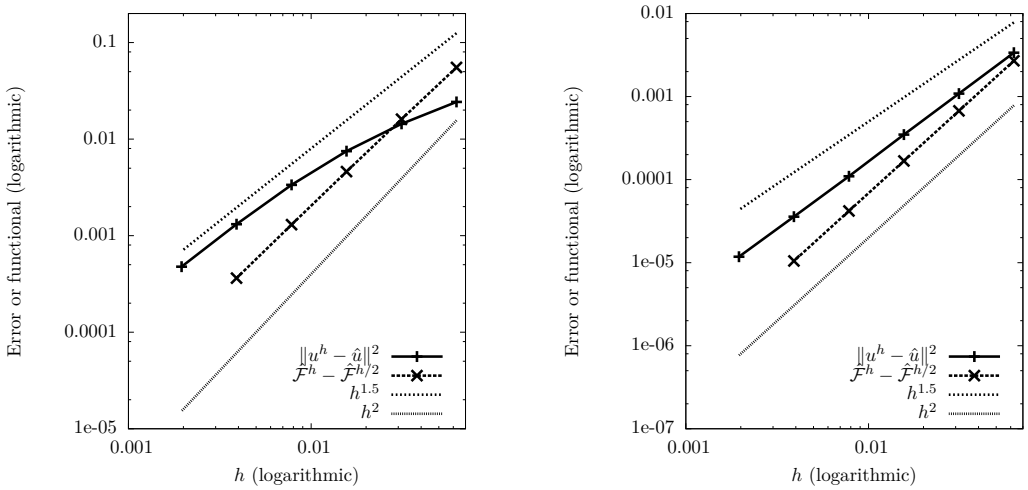
$$\Omega = \{0 < t < 1,\ -0.25 < x < 1.75\},\quad r = \begin{cases} 1, & x \leq 0 \\ 2, & x > 0 \end{cases},\quad g = \begin{cases} 3, & t = 0,\ x \leq 0 \\ 1, & t = 0,\ x > 0 \\ t + 3, & x = -0.25 \end{cases}.$$

Convergence of the functional values and the approximations obtained by the method are demonstrated in Figure 3.1 for a couple of choices of finite element orders. Notice that, in both cases, similar to the methods for conservation laws in [34, 32], the squared $L^2(\Omega)$ norm of the error approaches $\mathcal{O}(h)$, which is the theoretically optimal rate [58, 60, 59, 61]. The functional values converge with a higher rate on the tested meshes, similar to [34, 32]. These results align with the discussion in Subsection 3.5.2 that, in general, the functional can only provide a "weak control" of

the $L^2(\Omega)$ norm and a respective uniform coercivity does not generally hold.

Figures 3.2 and 3.3 show the resulting approximations in the two cases. Note that the method correctly captures the shock speed and its curved trajectory, which can be expected considering the convergence in Figure 3.1. It is worth discussing the spikes in the corners of the domain. Theoretically, such behavior can be linked to the fact that the functional can only provide a sufficient control of certain "weaker" norms (e.g., the convergence in the $L^2(\Omega)$ norm is not significantly affected by such spikes) and does not generally provide a substantial control of "stronger" Sobolev norms involving derivatives (or their "fractions") of the solution. This is associated with the fact that formulation (3.4.11) is closely related to the notion of a weak solution. A more particular inspection of the corner spikes suggests that they can be linked to the specific Helmholtz decomposition and the associated elliptic PDEs in Remarks 3.2.2 and 3.2.3, while they are clearly not the result of any surprising behavior of the exact solution to the hyperbolic PDE. Namely, in view of Remarks 3.2.2 and 3.2.3, the two corners with the spikes are precisely where the Neumann and Dirichlet boundary conditions meet in the respective elliptic problems that define the components of the Helmholtz decomposition, resulting in a decreased quality of the approximations close to the corners of these components, which are important parts of formulation (3.4.11). This is supported by the fact that increasing the order of the spaces for the components of the Helmholtz decomposition, $\mathcal{V}^h_{\Gamma_C}$ and $\mathcal{V}^h_{\Gamma_I}$, in Figure 3.3, compared to Figure 3.2, substantially decreases the spikes, since better approximations of these components are obtained. Observe that the corner spikes do not "pollute" the rest of the solution.

Furthermore, the oscillations at the initial and exit points of the shock in Figures 3.2 and 3.3 can be associated with both the singularity (discontinuity) in the solution of the hyperbolic PDE, as well as with the Helmholtz decomposition and the respective elliptic PDEs, since, in view of Remarks 3.2.2 and 3.2.3, at these points the Neumann boundary conditions in the respective elliptic problems exhibit jump discontinuities. Very similar oscillations are already observed in the methods for conservation laws in [34, 32], whereas the corner spikes are specific to the method of this chapter due to the utilization of the particular Helmholtz decomposition, which is different from the one in [34, 32], as discussed in the end of Subsection 3.4.2. Our experience shows that the oscillations around the shock become narrower to accommodate the $L^2(\Omega)$ convergence and remain bounded in amplitude as h is decreased. The backward propagation of such oscillations results from formulation



(a) $\mathcal{U}^h, \mathcal{V}^h_{\Gamma_C}, \mathcal{V}^h_{\Gamma_I}$ – linear

(b) $\mathcal{U}^h$ – linear, $\mathcal{V}^h_{\Gamma_C}, \mathcal{V}^h_{\Gamma_I}$ – quadratic

Figure 3.4: Convergence results for Example 2.

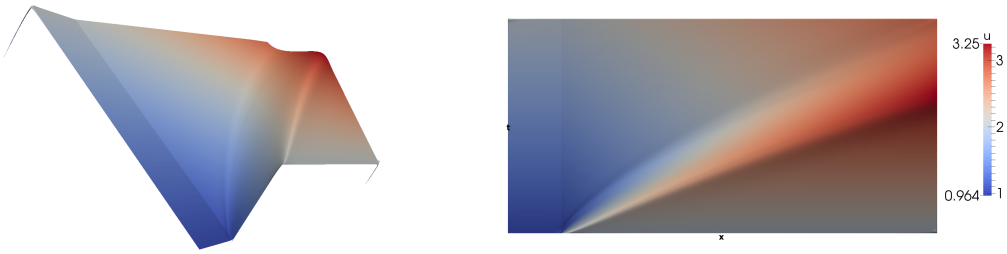


Figure 3.5: The approximation, u^h, obtained from Example 2 on the finest mesh, when all spaces, $\mathcal{U}^h$, $\mathcal{V}^h_{\Gamma_C}$, and $\mathcal{V}^h_{\Gamma_I}$, are linear.

(3.4.11) being a global (space-time) minimization that currently does not employ any upwinding techniques.

Observe that the shock in Figure 3.2 is noticeably more smeared than the one in Figure 3.3 and, accordingly, the backward propagating oscillations from the shock exit point are better dissipated in Figure 3.2. In our experience, the reduced numerical dissipation in Figure 3.3 is mostly due to the utilization of higher-order elements for $\mathcal{V}^h_{\Gamma_I}$ and not so much due to the space $\mathcal{V}^h_{\Gamma_C}$, whereas the reduction of the corner spikes benefits substantially from both $\mathcal{V}^h_{\Gamma_I}$ and $\mathcal{V}^h_{\Gamma_C}$ being of higher order. This, to some extent, aligns with the discussion about the "regularizing effect" of the space $\mathcal{V}^h_{\Gamma_I}$ in the end of Subsection 3.5.2.

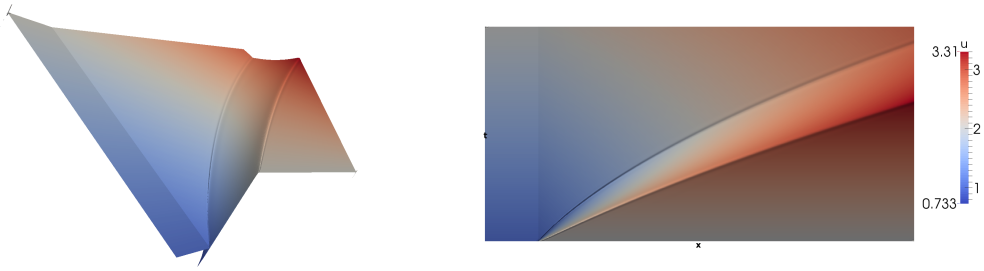


Figure 3.6: The approximation, u^h, obtained from Example 2 on the finest mesh, when $\mathcal{U}^h$ is linear and $\mathcal{V}^h_{\Gamma_C}, \mathcal{V}^h_{\Gamma_I}$ are quadratic.

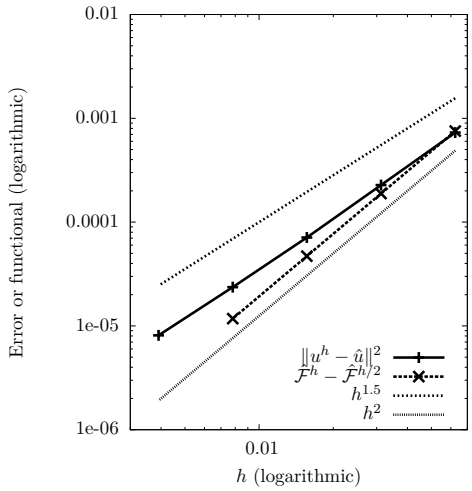
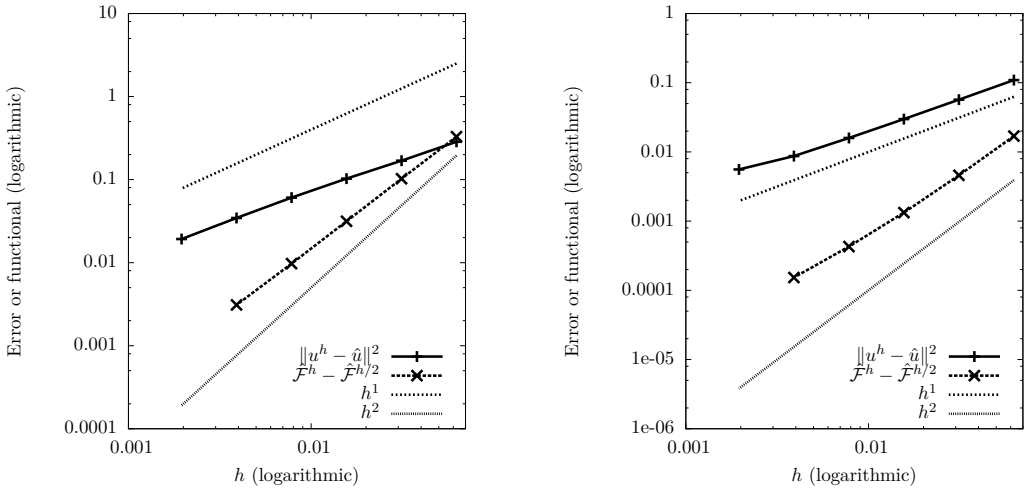


Figure 3.7: Convergence results for Example 2 with quadratic $\mathcal{U}^h$ and cubic $\mathcal{V}^h_{\Gamma_C}, \mathcal{V}^h_{\Gamma_I}$.

Example 2 (a rarefaction wave) Consider (3.3.1) with

$$\Omega = \{0 < t < 1,\ -0.25 < x < 1.75\},\quad r = \begin{cases} 1, & x \le 0 \\ 2, & x > 0 \end{cases},\quad g = \begin{cases} 1, & t = 0,\ x \le 0 \\ 2, & t = 0,\ x > 0 \\ t+1, & x = -0.25 \end{cases}.$$

Results are shown in Figures 3.4–3.6. The main challenge is that such a setting is associated with an infinite multiplicity of the weak solutions [1], where the rarefaction wave (associated with the respective "characteristic fan") is the unique admissible (or entropy) solution, which is of physical significance. It is a positive indication that the method recovers the physically admissible solution. However, it is currently unclear if this is an innate property of the formulation for all cases or if special entropy fixes may be necessary in general. This is a topic of further investigation. The



(a) $\mathcal{U}^h, \mathcal{V}^h_{\Gamma_C}, \mathcal{V}^h_{\Gamma_I}$ – linear

(b) $\mathcal{U}^h$ – linear, $\mathcal{V}^h_{\Gamma_C}, \mathcal{V}^h_{\Gamma_I}$ – quadratic

Figure 3.8: Convergence results for Example 3.

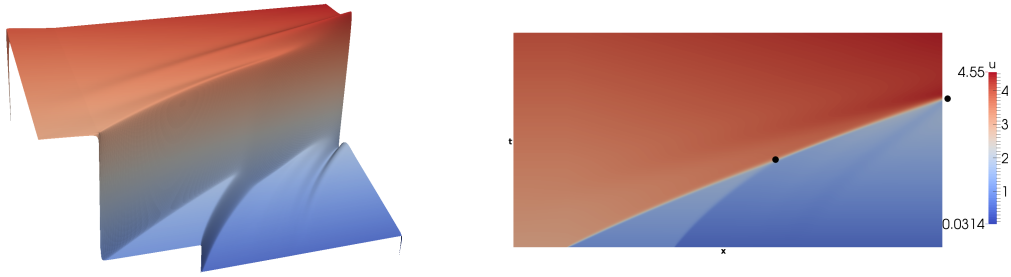


Figure 3.9: The approximation, u^h, obtained from Example 3 on the finest mesh, when all spaces, $\mathcal{U}^h$, $\mathcal{V}^h_{\Gamma_C}$, and $\mathcal{V}^h_{\Gamma_I}$, are linear. The black dots, •, show where the shocks collide and the resulting shock exists the domain in the exact solution, $\hat{u}$.

convergence rate is possibly suboptimal. In theory the decay rate of the squared $L^2(\Omega)$ norm of the error cannot be faster than $\mathcal{O}(h^{2-\epsilon})$, for any small $\epsilon > 0$. As shown in Figure 3.7, increasing the order of the spaces does not improve the rate of convergence.

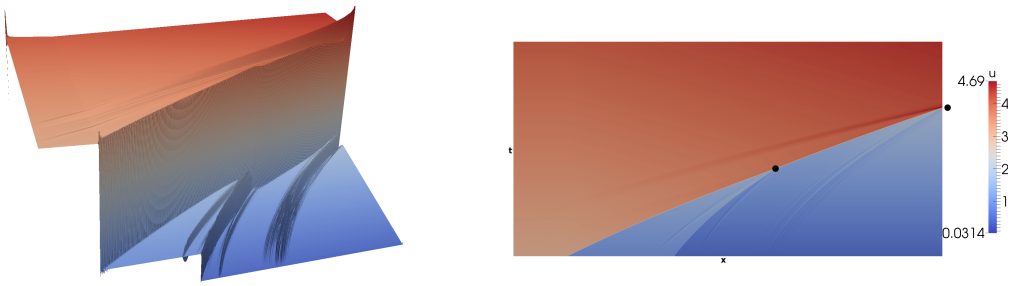


Figure 3.10: The approximation, u^h, obtained from Example 3 on the finest mesh, when $\mathcal{U}^h$ is linear and $\mathcal{V}^h_{\Gamma_C}, \mathcal{V}^h_{\Gamma_I}$ are quadratic. The black dots, $\bullet$, show where the shocks collide and the resulting shock exists the domain in the exact solution, $\hat{u}$.

Example 3 (colliding shocks) Consider (3.3.1) with

$$\Omega = \{0 < t < 1,\ -0.25 < x < 1.75\},\ r = \begin{cases} 1, x \leq 0 \\ \\ 2, x > 0 \end{cases},\ g = \begin{cases} 3, & t = 0,\ x \leq 0 \\ \\ 1, & t = 0,\ 0 < x \leq 0.5 \\ \\ 0.5, & t = 0,\ x \geq 0.5 \\ \\ t + 3, & x = -0.25 \end{cases}.$$

The respective results are shown in Figures 3.8–3.10. Note that the method accurately captures the shocks. However, the collision point, while correctly obtained, is substantially smeared in Figure 3.9, whereas this is not an issue in Figure 3.10.

Finally, the Gauss-Newton procedure utilizes a simple constant function as an initial guess on the coarsest mesh and, for every uniform refinement, the solution on the previous mesh is used as an initial guess. The number of Gauss-Newton iterations, for all cases and refinement levels in this chapter, are shown in Table 3.1. Note that the performance is expected to substantially improve by implementing adaptive mesh refinement in a nested iteration framework, which is a subject of future work.

3.7 About the linear case

This section concentrates on certain particulars associated with linear hyperbolic problems. First, some basic considerations and counterexamples are provided. Then, discrete coercivity results are shown. Finally, the potential impossibility of improving the discrete coercivity estimates is

Example 1	$\mathcal{V}^h_{\Gamma_C}, \mathcal{V}^h_{\Gamma_I}$ – linear	6, 4, 4, 4, 4, 5
	$\mathcal{V}^h_{\Gamma_C}, \mathcal{V}^h_{\Gamma_I}$ – quadratic	8, 4, 4, 4, 5, 10
Example 2	$\mathcal{V}^h_{\Gamma_C}, \mathcal{V}^h_{\Gamma_I}$ – linear	5, 3, 3, 3, 3, 3
	$\mathcal{V}^h_{\Gamma_C}, \mathcal{V}^h_{\Gamma_I}$ – quadratic	5, 3, 3, 3, 2, 2
Example 3	$\mathcal{V}^h_{\Gamma_C}, \mathcal{V}^h_{\Gamma_I}$ – linear	7, 4, 4, 5, 5, 6
	$\mathcal{V}^h_{\Gamma_C}, \mathcal{V}^h_{\Gamma_I}$ – quadratic	10, 5, 6, 8, 10, 5

Table 3.1: Number of Gauss-Newton iterations for all cases and refinement levels. The third column contains the number of iterations as the mesh is refined, from left to right. The space $\mathcal{U}^h$ is linear in all cases.

discussed formally. This serves the additional purpose to furnish a relation and transition to the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method in Chapter 4.

3.7.1 Basics

Here, a short introduction and basic considerations are presented. Reusing the notation in (3.3.1), consider the scalar linear hyperbolic problem

$$\begin{aligned} \nabla \cdot \boldsymbol{b}u &= r \text{ in } \Omega, \\ u &= g \text{ on } \Gamma_I, \end{aligned} \tag{3.7.1}$$

where $\boldsymbol{b} \in [L^\infty(\Omega)]^2$, $\nabla \cdot \boldsymbol{b} \in L^\infty(\Omega)$ is given. Observe that (3.7.1) can be seen, in a sense, as a special case of (3.3.1), where $\boldsymbol{f}(u, \boldsymbol{x}) = \boldsymbol{b}(\boldsymbol{x})u(\boldsymbol{x})$; that is, the flux vector, $\boldsymbol{f}$, depends also on $\boldsymbol{x}$ and not just on u. This should not result in any confusion. Notice in this case that $\frac{\partial \boldsymbol{f}}{\partial u}(u, \boldsymbol{x}) = \boldsymbol{b}(\boldsymbol{x})$, which is the direction of the characteristics of (3.7.1). Thus, clearly, the boundary portions in Definition 3.3.2 take the usual form for linear problems:

$$\begin{aligned} \Gamma_I &= \{\, \boldsymbol{x} \in \Gamma;\ \boldsymbol{b}(\boldsymbol{x}) \cdot \boldsymbol{n}(\boldsymbol{x}) < 0 \,\}, \\ \Gamma_O &= \{\, \boldsymbol{x} \in \Gamma;\ \boldsymbol{b}(\boldsymbol{x}) \cdot \boldsymbol{n}(\boldsymbol{x}) > 0 \,\}, \\ \Gamma_T &= \{\, \boldsymbol{x} \in \Gamma;\ \boldsymbol{b}(\boldsymbol{x}) \cdot \boldsymbol{n}(\boldsymbol{x}) = 0 \,\}, \\ \Gamma_C &= \Gamma_O \cup \Gamma_T = \{\, \boldsymbol{x} \in \Gamma;\ \boldsymbol{b}(\boldsymbol{x}) \cdot \boldsymbol{n}(\boldsymbol{x}) \geq 0 \,\}. \end{aligned}$$

It is well known that the characteristics of (3.7.1) are precisely the streamlines (trajectories) of the following autonomous system of ODE:

$$\dot{\boldsymbol{x}} = \boldsymbol{b}(\boldsymbol{x}),$$

where $\boldsymbol{x}$ is viewed as a function $\boldsymbol{x}(s)$ (i.e., depending on a parameter $s \in \mathbb{R}_+$), the points on Γ_I act as initial conditions, and $\Omega \subset \mathbb{R}^2$ is the phase space[1] (state space) of the autonomous system. Therefore, the further assumptions on $\boldsymbol{b}$, as in [33], are based on the properties of this ODE system. Namely, assume that $|\boldsymbol{b}|$ is bounded away from zero a.e. on Ω, no two streamlines intersect, Ω is entirely covered by streamlines, the streamlines in Ω are of finite length, and there exists a transformation (of the phase space) with a bounded Jacobian that transforms the flow field, $\boldsymbol{b}$, to one that is aligned with a coordinate axis. Moreover, these assumptions imply an estimate like (3.5.9) for the linear case.

Now, a simple counterexample is presented, demonstrating that, in general, the functional $\mathcal{F}$ in (3.4.9) cannot be expected to provide "control" of the $L^2(\Omega)$ norm.

Example 3.7.1 Let $\Omega = (0,\pi)^2$, $\boldsymbol{b} = [1,0]$, $r \equiv 0$, and $g \equiv 0$. Thus, the exact solution to (3.7.1) is $\hat{u} = 0$. Consider

$$v_n(x_1,x_2) = \sin\frac{x_1}{2}\sin(nx_2), \quad n \geq 1.$$

Then $v_n \in H^1_{0,\Gamma_I}(\Omega)$ and $\|v_n\| = \pi/2$. Observe that

$$\nabla\cdot\boldsymbol{b}v_n = (v_n)'_{x_1} = \frac{1}{2}\cos\frac{x_1}{2}\sin(nx_2) \in H^1_{0,\Gamma_C}(\Omega),$$

and $\|(v_n)'_{x_1}\| = \pi/4$. Considering the Helmholtz decomposition $\boldsymbol{b}v_n = \nabla q_{v_n} + \nabla^\perp\psi_{v_n}$, it is easy to verify that

$$q_{v_n}(x_1,x_2) = -\frac{4}{4n^2+1}(v_n)'_{x_1} \in H^1_{0,\Gamma_C}(\Omega),$$

since, in view of Remark 3.2.3, $\partial q_{v_n}/\partial\boldsymbol{n} = 0 = (\boldsymbol{b}v_n)\cdot\boldsymbol{n}$ on Γ_I and $\Delta q_{v_n} = (v_n)'_{x_1} = \nabla\cdot\boldsymbol{b}v_n$. Using (3.5.4) for this case gives

$$\begin{aligned}\min_{\substack{p\in H^1_{0,\Gamma_C}(\Omega)\\ \mu\in H^1_{0,\Gamma_I}(\Omega)}} \mathcal{F}(v_n,p,\mu;\,0) &= \frac{1}{2}\|\nabla q_{v_n}\|^2 = \frac{\pi^2}{4(4n^2+1)}\\ &= \frac{1}{4n^2+1}\|v_n\|^2 \xrightarrow{n\to\infty} 0. \qquad \diamond\end{aligned}$$

Consider some norm $\|\cdot\|_{\mathfrak{F}}$ that is defined on $L^2(\Omega)$ and assume that X is a subset of $L^2(\Omega)$ such that the $L^2(\Omega)$ norm and $\|\cdot\|_{\mathfrak{F}}$ are equivalent on X, i.e., there are constants $a, b > 0$ such that

$$a\|v\| \leq \|v\|_{\mathfrak{F}} \leq b\|v\|, \quad \forall v \in X.$$

[1]The set where the state vector $\boldsymbol{x}$ takes values.

Let $\overline{X}$ be the closure of X with respect to the $L^2(\Omega)$ norm. It is not difficult to show that the norm equivalence continues to hold on $\overline{X}$ and that $\overline{X}$ is also the closure of X with respect to $\|\cdot\|_{\mathfrak{F}}$. In particular, if X is dense in $L^2(\Omega)$, then the norms are equivalent on the whole $L^2(\Omega)$.

In the case of interest, when $\|\cdot\|_{\mathfrak{F}}$ is the functional norm

$$\|v\|_{\mathfrak{F}}^2 = \min_{\substack{p\in H^1_{0,\Gamma_C}(\Omega)\\ \mu\in H^1_{0,\Gamma_I}(\Omega)}} \mathcal{F}(v,p,\mu;\,0), \quad v\in L^2(\Omega),$$

this, together with Example 3.7.1, implies that it cannot be expected that the functional, $\mathcal{F}$, controls the $L^2(\Omega)$ norm on dense subsets of $L^2(\Omega)$. This unfortunately excludes any of the common Sobolev and finite element spaces. Meaningful candidates are only proper closed subsets of $L^2(\Omega)$, that is, a uniform discrete coercivity with respect to the $L^2(\Omega)$ norm cannot hold on any family of finite element spaces, since this would imply the respective uniform coercivity that is rejected by Example 3.7.1. Therefore, the reasonable coercivity is the discrete one in (3.5.12) and the respective inf-sup condition (3.5.11). This is studied in some detail in Subsection 3.7.2 for the linear case.

3.7.2 Discrete coercivity and inf-sup conditions

This subsection presents proofs, in a few cases, of inf-sup conditions of the form (3.5.11) for the linear case (3.7.1). More precisely, we show discrete coercivity estimates of the form (3.5.12), which are equivalent to inf-sup conditions like (3.5.11). All results here yield $\alpha = 1$ and Subsection 3.7.3 provides a formal discussion on why a smaller α may not be obtainable on standard finite element spaces.

In the linear case, it holds that $\boldsymbol{\mathcal{S}}^h = \{\, \boldsymbol{b}v^h;\ v^h \in \mathcal{U}^h \,\}$ for a given finite element space $\mathcal{U}^h \subset L^\infty(\Omega)$. The following discrete coercivity estimate is shown for a few cases:

$$\|\nabla_{\mathfrak{w}}\cdot \boldsymbol{b}v^h\|_{-1,\Gamma_C} \geq ch\|\boldsymbol{b}v^h\|, \quad \forall v^h\in\mathcal{U}^h. \tag{3.7.2}$$

for a constant $c>0$. The general assumptions are that Ω is a polygonal (or polyhedral) domain, $0<h\leq 1$, and quasi-uniform (see [58], e.g.) meshes are used. Note that, for brevity, the standard abuse of notation is utilized by regarding every function $v\in L^2(\Omega)$ as a functional in any desired negative Sobolev space. This is achieved by the standard implicit identification of v with the functional $(v,\cdot)$.

The first result is for finite element spaces that we are used to seeing in the context of $\mathcal{LL}^*$ methods.

Proposition 3.7.2 *Consider a finite element space $\mathcal{Z}^h \subset H^1_{0,\Gamma_C}(\Omega)$ and let $\mathcal{U}^h = \{\, \boldsymbol{b}\cdot\nabla z^h;\ z^h \in \mathcal{Z}^h \,\}$. Then (3.7.2) holds with $c>0$ depending on Ω, $\boldsymbol{b}$, and the quasi-uniformity of the mesh.*

Proof. By [58, Theorem 4.5.11], the following inverse estimate holds, with a constant that depends on the quasi-uniformity of the mesh:

$$|z^h|_1 \le Ch^{-1}\|z^h\|, \quad \forall z^h \in \mathcal{Z}^h.$$

Owing to the Poincaré-type inequality in [33, Lemma 2.4], the following estimate holds, with a constant depending on Ω and $\boldsymbol{b}$:

$$\|z^h\| \le C\|\boldsymbol{b}\cdot\nabla z^h\|, \quad \forall z^h \in \mathcal{Z}^h.$$

Combining the two estimates above provides

$$|z^h|_1 \le Ch^{-1}\|\boldsymbol{b}\cdot\nabla z^h\|, \quad \forall z^h \in \mathcal{Z}^h.$$

Using this, for any $v^h \in \mathcal{U}^h$, gives

$$\begin{aligned}
\|\boldsymbol{b}v^h\| \le C\|v^h\| &= C\sup_{\phi^h\in\mathcal{U}^h}\frac{|(v^h,\phi^h)|}{\|\phi^h\|}\\
&= C\sup_{z^h\in\mathcal{Z}^h}\frac{|(v^h,\boldsymbol{b}\cdot\nabla z^h)|}{\|\boldsymbol{b}\cdot\nabla z^h\|} \le Ch^{-1}\sup_{z^h\in\mathcal{Z}^h}\frac{|(v^h,\boldsymbol{b}\cdot\nabla z^h)|}{|z^h|_1}\\
&\le Ch^{-1}\sup_{z\in H^1_{0,\Gamma_C}(\Omega)}\frac{|(\boldsymbol{b}v^h,\nabla z)|}{|z|_1} = Ch^{-1}\|\nabla_{\mathfrak{w}}\cdot\boldsymbol{b}v^h\|_{-1,\Gamma_C}.
\end{aligned}$$

∎

Next, certain H^1-conforming finite element spaces are considered in the following result.

Proposition 3.7.3 *If $\mathcal{U}^h \subset H^1_{0,\Gamma_C}(\Omega)$, then (3.7.2) holds with $c>0$ depending on Ω, $\boldsymbol{b}$, and the quasi-uniformity of the mesh.*

Proof. Owing to the coercivity estimate[1] [32, Lemma 6.8] and the inverse estimate in [58, Theorem 4.5.11], it holds, for $v^h \in \mathcal{U}^h$, that

$$\|\nabla_{\mathfrak{w}}\cdot\boldsymbol{b}v^h\|_{-1,\Gamma_C} \ge c\|v^h\|_{-1,\Gamma_C} = c\sup_{\phi\in H^1_{0,\Gamma_C}(\Omega)}\frac{|(v^h,\phi)|}{|\phi|_1}$$

[1]Note that [32, Lemma 6.8] is obtained under the assumption that that Γ_I and Γ_O do not touch each other. Similarly, the result can be obtained under the assumption that Γ_I and Γ_O each consist of a single connected component of the boundary. Our purpose is mainly to present ideas in this section. So, to keep thing simple, we do not elaborate too much on this. Nevertheless, more details appear in Lemma 3.7.4.

$$\geq c \sup_{\phi^h \in \mathcal{U}^h} \frac{|(v^h, \phi^h)|}{|\phi^h|_1} \geq ch \sup_{\phi^h \in \mathcal{U}^h} \frac{|(v^h, \phi^h)|}{\|\phi^h\|} = ch\|v^h\| \geq ch\|\boldsymbol{b}v^h\|. \qquad \square$$

Stronger discrete coercivity estimates (i.e., ones that include a larger class of Lagrangian finite element spaces) are obtained by strengthening, in a sense, the coercivity result in [32, Lemma 6.8]. This is outlined in the following lemma. The argument is based on the proof of [32, Lemma 6.8]; see also [33].

Lemma 3.7.4 *Assume that Γ_I and Γ_O each consist of a single connected component of the boundary. Then there exists a constant $c > 0$, depending on Ω and $\boldsymbol{b}$, such that*

$$c\|v\|_{-1,\Gamma_T} \leq \|\nabla_{\mathfrak{w}} \cdot \boldsymbol{b}v\|_{-1,\Gamma_C}, \quad \forall v \in L^2(\Omega).$$

In particular, the estimate holds when Γ_T is of zero surface measure.

Proof. First, by a standard density argument, it is sufficient to restrict the considerations to the case of $v \in \mathcal{C}_c^\infty(\Omega)$, the space of infinitely smooth compactly supported functions on Ω. Hence, consider $v \in \mathcal{C}_c^\infty(\Omega)$. In this case, in view of Remark 3.4.2, $\nabla \cdot \boldsymbol{b}v$ and $\nabla_{\mathfrak{w}} \cdot \boldsymbol{b}v$ can be identified via standard embeddings. Thus, it is sufficient to show that

$$c\|v\|_{-1,\Gamma_T} \leq \|\nabla \cdot \boldsymbol{b}v\|_{-1,\Gamma_C}.$$

By assumption, $\boldsymbol{b}$ can be aligned with one of the coordinate axes by a transformation with a bounded Jacobian. Thus, assume that such a transformation is applied and Ω is in a (τ, s)-coordinate system, and $\boldsymbol{b} = [1, 0]$. The proof is carried for this case and the general inequality follows by considering the contribution of the Jacobian.

Consider Γ_O as the graph of $\tau_O(s)$ for $s \in J$, where $J \subset \mathbb{R}$ is some appropriate set. Since Γ_O is a Lipschitz-continuous boundary, $\tau_O \in W^{1,\infty}(J)$. In particular, τ_O' is defined in classical sense almost everywhere in J and $\|\tau_O'\|_{L^\infty(J)} < \infty$. Similarly, Γ_I is the graph of $\tau_I(s)$ for $s \in J$.

Now, it holds that

$$v(\tau, s) = \int_{\tau_I(s)}^{\tau} \frac{\partial}{\partial \rho} v(\rho, s) \, \mathrm{d}\rho.$$

Denote $X = \mathcal{C}^\infty(\overline{\Omega}) \cap H^1_{0,\Gamma_T}(\Omega)$ and let $p \in X$, which is motivated by the density of X in $H^1_{0,\Gamma_T}(\Omega)$. The last equality and changing the order of integration yield

$$\int_{\tau_I(s)}^{\tau_O(s)} v(\tau, s) p(\tau, s) \, \mathrm{d}\tau = \int_{\tau_I(s)}^{\tau_O(s)} p(\tau, s) \int_{\tau_I(s)}^{\tau} \frac{\partial}{\partial \rho} v(\rho, s) \, \mathrm{d}\rho \, \mathrm{d}\tau$$

$$= \int_{\tau_I(s)}^{\tau_O(s)} \frac{\partial}{\partial \rho} v(\rho, s) \int_{\rho}^{\tau_O(s)} p(\tau, s)\, \mathrm{d}\tau\, \mathrm{d}\rho$$
$$= \int_{\tau_I(s)}^{\tau_O(s)} \frac{\partial}{\partial \tau} v(\tau, s) \int_{\tau}^{\tau_O(s)} p(\rho, s)\, \mathrm{d}\rho\, \mathrm{d}\tau$$

Denote

$$q_p(\tau, s) = \int_{\tau}^{\tau_O(s)} p(\rho, s)\, \mathrm{d}\rho. \tag{3.7.3}$$

Thus,

$$\int_J \int_{\tau_I(s)}^{\tau_O(s)} v(\tau, s) p(\tau, s)\, \mathrm{d}\tau\, \mathrm{d}s = \int_J \int_{\tau_I(s)}^{\tau_O(s)} \frac{\partial}{\partial \tau} v(\tau, s) q_p(\tau, s)\, \mathrm{d}\tau\, \mathrm{d}s.$$

Here, this can be expressed as

$$(v, p) = (\nabla \cdot \boldsymbol{b} v, q_p).$$

Notice that

$$\frac{\partial q_p}{\partial \tau} = -p,$$

whence

$$\left\| \frac{\partial q_p}{\partial \tau} \right\|^2 = \|p\|^2 \leq \|p\|_1^2.$$

Also,

$$\frac{\partial q_p}{\partial s} = \int_{\tau}^{\tau_O(s)} \frac{\partial}{\partial s} p(\rho, s)\, \mathrm{d}\rho + p(\tau_O(s), s) \tau_O'(s).$$

Now, the simple inequality $(a+b)^2 \leq 2(a^2+b^2)$ and Jensen's inequality yield

$$\begin{aligned}
\left\| \frac{\partial q_p}{\partial s} \right\|_{0,\Omega}^2 &\leq C \left[\int_J \int_{\tau_I(s)}^{\tau_O(s)} \left(\int_{\tau}^{\tau_O(s)} \frac{\partial}{\partial s} p(\rho, s)\, \mathrm{d}\rho \right)^2 \mathrm{d}\tau\, \mathrm{d}s \right. \\
&\qquad \left. + \int_J \int_{\tau_I(s)}^{\tau_O(s)} \left(p(\tau_O(s), s) \tau_O'(s) \right)^2 \mathrm{d}\tau\, \mathrm{d}s \right] \\
&\leq C \left[\int_J \int_{\tau_I(s)}^{\tau_O(s)} \int_{\tau}^{\tau_O(s)} \left(\frac{\partial}{\partial s} p(\rho, s) \right)^2 \mathrm{d}\rho\, \mathrm{d}\tau\, \mathrm{d}s \right. \\
&\qquad \left. + \|\tau_O'(s)\|_{L^\infty(J)}^2 \int_J p^2(\tau_O(s), s)\, \mathrm{d}s \right] \\
&\leq C \left[\int_J \int_{\tau_I(s)}^{\tau_O(s)} \int_{\tau_I(s)}^{\tau_O(s)} \left(\frac{\partial}{\partial s} p(\rho, s) \right)^2 \mathrm{d}\rho\, \mathrm{d}\tau\, \mathrm{d}s \right. \\
&\qquad \left. + \int_J p^2(\tau_O(s), s) \sqrt{1 + (\tau_O'(s))^2}\, \mathrm{d}s \right] \\
&\leq C \left[\int_J \int_{\tau_I(s)}^{\tau_O(s)} \left(\frac{\partial}{\partial s} p(\tau, s) \right)^2 \mathrm{d}\tau\, \mathrm{d}s + \|p\|_{\Gamma_O}^2 \right]
\end{aligned}$$

$$\leq C \left[\left\| \frac{\partial p}{\partial s} \right\|^2 + \|p\|^2_{1/2,\Gamma} \right] \leq C\|p\|^2_1.$$

Thus,

$$|q_p|_1 \leq C\|p\|_1. \tag{3.7.4}$$

Also, (3.7.3) clearly shows that $q_p = 0$ on Γ_O. Hence, $q_p \in H^1_{0,\Gamma_O}(\Omega)$. In general, it is possible that $q_p \notin H^1_{0,\Gamma_C}(\Omega)$. This is where the assumption that Γ_I and Γ_O each consist of a single connected component of the boundary becomes useful. Actually, it is sufficient to have every connected component of Γ_T touching Γ_O. Since $p = 0$ on Γ_T, then (3.7.3) implies that $q_p = 0$ on Γ_T and hence $q_p \in H^1_{0,\Gamma_C}(\Omega)$.

Combining the above results, the following duality argument is clear:

$$\sup_{p\in X} \frac{|(v,p)|}{\|p\|_1} \leq C \sup_{p\in X} \frac{|(\nabla\cdot \boldsymbol{b}v, q_p)|}{|q_p|_1} \leq C \sup_{q\in H^1_{0,\Gamma_C}(\Omega)} \frac{|(\nabla\cdot \boldsymbol{b}v, q)|}{|q|_1} = C\|\nabla\cdot \boldsymbol{b}v\|_{-1,\Gamma_C}.$$

This implies the desired estimate

$$c\|v\|_{-1,\Gamma_T} \leq \|\nabla\cdot \boldsymbol{b}v\|_{-1,\Gamma_C}.$$

The case when Γ_T is of zero surface measure is analogous. ∎

Now, a broader discrete coercivity result can be obtained for H^1 finite element spaces.

Proposition 3.7.5 *If $\mathcal{U}^h \subset H^1_{0,\Gamma_T}(\Omega)$, then (3.7.2) holds with $c > 0$ depending on Ω, $\boldsymbol{b}$, and the quasi-uniformity of the mesh. In particular, when Γ_T is of zero surface measure, then $\mathcal{U}^h$ can be any H^1 finite element space, i.e., any $\mathcal{U}^h \subset H^1(\Omega)$.*

Proof. The argument is similar to the proof of Proposition 3.7.3, using Lemma 3.7.4 and [58, Theorem 4.5.11]. ∎

More general discrete coercivity results are currently not established. Also, it is not completely clear if better powers of h can be obtained. Subsection 3.7.3 argues that it may be impossible in many cases or at least very hard. Note that all versions of $\mathcal{U}^h$ above form dense sets in $L^2(\Omega)$ as $h \to 0$.

3.7.3 Limitations on the discrete coercivity

It is a quite deep question whether it is possible to improve on the power of h (i.e., make it as close to zero as possible) in (3.7.2) for the results in Subsection 3.7.2. Here, we brush over this topic.

One main idea in Subsection 3.7.2 is to utilize a coercivity estimate like the one in Lemma 3.7.4 to obtain a relationship between $\|\nabla_{\mathfrak{w}} \cdot \boldsymbol{b}v\|_{-1,\Gamma_C}$ and some dual norm, and then invoke an inverse estimate with respect to the corresponding primal norm to conclude the final discrete coercivity estimate. It is not difficult, using operator interpolation theory (see [58, Chapter 14],[73]), to extend the inverse estimate in [58, Theorem 4.5.11] for the $H^\theta(\Omega)$ norm, obtaining a $h^{-\theta}$ rate, where $0 < \theta < 1$. Also, the estimate in Lemma 3.7.4 can be further strengthened for a norm that is stronger than the $H^{-1}_{\Gamma_T}(\Omega)$ norm. However, no matter how tempting, this turns out to be insufficient for improving the final discrete estimate.

The key to strengthening the dual estimate in Lemma 3.7.4 is to improve on the primal estimate in (3.7.4) by utilizing a weaker norm of p in the bound on $|q_p|_1$. While this is not difficult in general, the issue is that one cannot essentially get rid of or weaken the term $\|\partial p/\partial s\|$ in the norm of p. This is suggested by the following observation based on Example 3.7.1. Consider $\Omega = (0,\pi)^2$, $\boldsymbol{b} = [1,0]$, and

$$\begin{aligned} p_n(\tau,s) &= \cos\tau\sin(ns) \in X, \\ q_{p_n}(\tau,s) &= -\sin\tau\sin(ns) \in H^1_{0,\Gamma_C}(\Omega), \\ \nabla q_{p_n} &= \left[-\cos\tau\sin(ns), -n\sin\tau\cos(ns)\right], \\ \frac{\partial p_n}{\partial s} &= n\cos\tau\cos(ns), \end{aligned}$$

where $n \geq 1$. Then

$$\begin{aligned} |q_{p_n}|_1^2 &= \|\cos\tau\sin(ns)\|^2 + n^2\|\sin\tau\cos(ns)\|^2 = \frac{\pi^2}{4} + n^2\frac{\pi^2}{4}, \\ \left\|\frac{\partial p_n}{\partial s}\right\|^2 &= n^2\|\cos\tau\cos(ns)\|^2 = n^2\frac{\pi^2}{4}. \end{aligned}$$

Therefore, any norm of p_n that is "weaker" than $\mathcal{O}(n)$ (i.e., weaker than $\|\partial p/\partial s\|$) cannot provide an estimate of the type (3.7.4). Intuitively, this may be interpreted that a "full control" of the partial derivative with respect to s (i.e., in the direction perpendicular to $\boldsymbol{b}$, which is the cross-stream direction) is needed. The particular issue is that an inverse estimate in which $\|\partial v^h/\partial s\|$ is

bounded by $\|v^h\|$, derived by the usual finite element local analysis (i.e., element-by-element using a reference element), already results in h^{-1}, since having just one "whole" partial derivative involved prohibits any improvements on the power of h; see [58, Section 4.5]. Note that this is in no way a rigorous proof of the impossibility of obtaining better discrete estimates, but it outlines a major issue and provides some justification. An intuitive interpretation is that $\|\nabla_{\mathfrak{w}} \cdot \boldsymbol{b}v\|_{-1,\Gamma_C}$ is based on an H^{-1} norm, which in turn is related to the "action" of an inverse Laplace operator, $(-\Delta)^{-1}$, which is a Riesz isomorphism in this context. The operator $(-\Delta)^{-1}$ has a "smoothing effect" in all directions, including the cross-stream direction, which particularly results in the presence of h in the discrete coercivity estimates. This is already exploited in the last example as well as in Example 3.7.1. In contrast, the operator $(-\nabla \cdot \boldsymbol{b}\boldsymbol{b}^T\nabla)^{-1}$, sometimes called "total anisotropy", has a "smoothing effect" only in the direction of the stream (i.e., in the direction of $\boldsymbol{b}$).

In short, the above formal discussion suggests that $\|\nabla_{\mathfrak{w}} \cdot \boldsymbol{b}v\|_{-1,\Gamma_C}$ "behaves" like an H^{-1} norm in the cross-stream direction, while it is stronger in the streamline direction. This is the fundamental reason for failing the uniform coercivity in Example 3.7.1 and failing to improve the discrete coercivity estimates, as discussed above. This further justifies the need for considering conditions like (3.5.13). Additionally, it motivates the development of the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method in Chapter 4 for linear hyperbolic problems. The $(\mathcal{L}\mathcal{L}^*)^{-1}$ approach utilizes a more specialized isomorphism in the place of $(-\Delta)^{-1}$ (in fact, the isomorphism in the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method is associated with the "total anisotropy" operator) and provides a natural relation to the $L^2(\Omega)$ norm.

3.8 Conclusions and future work

We proposed and studied a least-squares finite element formulation for hyperbolic balance laws that is based on the Helmholtz decomposition and is closely related to the notion of a weak solution. The ability of this approach to correctly approximate weak solutions and its convergence properties were discussed, and numerical results were provided. The method demonstrates good convergence, shock capturing capabilities, and correctly obtains rarefaction solutions to a nonlinear PDE.

There are many directions of future development. Particularly, adaptive mesh refinement in a nested iteration setting constitutes important follow-up work as it would contribute to the practical applicability of the method; while rarefaction waves are accurately obtained by the method, it is

still unclear if it naturally produces approximations to admissible (entropy) solutions or it may need to explicitly impose appropriate entropy conditions; extending the method to systems by utilizing a suitable Helmholtz decomposition is an important topic of future investigation; and generalizing the formulation for problems where the source term, r, in (3.3.1a) is allowed to depend linearly or nonlinearly on the unknown variable, u, would allow the consideration of more general hyperbolic equations.

Chapter 4

Mixed $(\mathcal{L}\mathcal{L}^*)^{-1}$ and $\mathcal{L}\mathcal{L}^*$ Methods for Linear Problems

In this chapter, a few dual least-squares finite element methods and their application to scalar linear hyperbolic problems are studied. The purpose is to obtain L^2-norm approximations on finite element spaces of the exact solutions to hyperbolic partial differential equations of interest. This is approached by approximating the generally infeasible quadratic minimization that defines the L^2-orthogonal projection of the exact solution by feasible least-squares principles using the ideas of the original $\mathcal{L}\mathcal{L}^*$ method [43] proposed in the context of elliptic equations. All methods in this chapter are founded upon and extend the $\mathcal{L}\mathcal{L}^*$ approach, which is rather general and applicable beyond the setting of elliptic problems. Error bounds are shown that point to the factors affecting convergence and provide conditions that guarantee optimal rates. Furthermore, preconditioning of the resulting linear systems is discussed. Numerical results are provided to illustrate the behavior of the methods on common finite element spaces. Note that the methods, analysis, and all considerations are general and applicable to partial differential equations (PDEs), both scalar and systems, of a larger class than the hyperbolic equations discussed in this chapter. Nevertheless, following the subject of this thesis, the concentration is on linear hyperbolic PDEs and, for simplicity, scalar equations.

4.1 Introduction

Consider a scalar linear hyperbolic partial differential equation of the form

$$\begin{aligned} \nabla \cdot \boldsymbol{b}\psi + \sigma\psi &= r \quad \text{in } \Omega, \\ \psi &= g \quad \text{on } \Gamma_I, \end{aligned} \tag{4.1.1}$$

where the simply connected domain, $\Omega \subset \mathbb{R}^d$ (d is the dimension of the Euclidean space), flow field, $\boldsymbol{b} \in [L^\infty(\Omega)]^d$, absorption coefficient, $\sigma \in L^\infty(\Omega)$, source term, $r \in L^2(\Omega)$, and inflow boundary data[1], $g \in L^2(\Gamma_I)$, are given and ψ is the unknown dependent variable. Here, Γ_I denotes the inflow portion of the boundary, $\partial\Omega$, i.e.,

$$\Gamma_I = \{\, \boldsymbol{x} \in \partial\Omega;\ \boldsymbol{n}(\boldsymbol{x}) \cdot \boldsymbol{b}(\boldsymbol{x}) < 0 \,\},$$

where $\boldsymbol{n}$ is the unit outward normal to $\partial\Omega$.

Equations like (4.1.1) arise often in applications and can also serve as model problems towards solving more elaborate hyperbolic PDEs [6, 1, 2, 3, 11, 7].

The solution to (4.1.1) can be quite irregular – exhibiting jump discontinuities or, depending on the contrast in σ, extremely steep exponential layers, leading to large variations of the solution in neighboring subregions of Ω. We are interested in obtaining approximations of the solution without utilizing any additional information on its features and using only information provided by the differential operator in (4.1.1). In particular, we consider general unstructured meshes that are not aligned with the flow, $\boldsymbol{b}$, i.e., the mesh does not follow the characteristics of (4.1.1). Also, the mesh does not need to resolve steep exponential layers, i.e., on the scale of the mesh such layers can appear as jump discontinuities. Moreover, we aim at solving (4.1.1) as a global space-time problem (if one of the independent variables represents time) without applying any time-stepping scheme, i.e., Ω is a domain in the space-time.

Least-squares finite element methods have been extensively studied for problems of elliptic and parabolic types; see, e.g., [29, 31, 46, 47, 30, 48, 49, 50, 51]. They have also been applied to hyperbolic problems, including of the type (4.1.1); cf., [33, 35, 15, 37, 32, 34, 36, 38]. These methods exhibit substantial numerical diffusion, unless proper scaling is implemented, which may include utilizing information about the characteristics of the problem and the respective features of the solution [81]. Diffusion results in stable methods (less oscillatory approximations) and least-squares have been used to augment Galerkin formulations to stabilize them; see, e.g, [38]. However, excessive diffusion can lead to unsatisfactory quality of the approximation. In our experience, this especially holds when large jumps in σ cause very steep exponential layers in the solution that are not resolved by the mesh.

[1] In general, the function g is in a space on Γ_I that can be larger than $L^2(\Gamma_I)$; see the trace results in [33]. For our considerations, the space $L^2(\Gamma_I)$ is sufficiently rich for inflow boundary conditions.

In this chapter, we address these issues (the solution irregularity, unstructured meshes not resolving steep exponential layers, and the excessive numerical diffusion) by seeking approximations in the $L^2(\Omega)$ norm. Note that the least-squares methods [33, 35] possess coercivity in a norm stronger than the $L^2(\Omega)$ norm, so they control the L^2-norm error, but the error can remain relatively large until the mesh size is sufficiently small to begin resolving the features of the solution. This contributes to the amount of numerical diffusion in the least-squares methods. In contrast, we approach the L^2-norm approximation more directly. The $(\mathcal{L}\mathcal{L}^*)^{-1}$ and $\mathcal{L}\mathcal{L}^*$ methods considered in this chapter are based on least-squares principles, which, in a sense, approximate the minimization that defines the best L^2-norm approximation.

Generally, given $f \in L^2(\Omega)$ and a linear first-order differential operator, L, our goal is to solve an equation of the form

$$Lu = f, \tag{4.1.2}$$

for the unknown $u \in \mathcal{D}(L)$, where $\mathcal{D}(L)$ denotes the domain of L. The general definition of $\mathcal{D}(L)$ is provided in Section 4.2 and in Section 4.7 the particular definition for (4.1.1) is shown. Equation (4.1.1) can be reduced to (4.1.2) using superposition, since, in this case, the functions in $\mathcal{D}(L)$ vanish on Γ_I. In practice, solving (4.1.2) is addressed by numerically approximating the exact solution, $\hat{u} \in \mathcal{D}(L)$, of equation (4.1.2). The focus of this chapter is on obtaining finite element approximations of $\hat{u}$ with respect to the $L^2(\Omega)$ norm, denoted $\|\cdot\|$. Given a finite element space $\mathcal{U}^h$, the best, in $\mathcal{U}^h$, L^2-norm approximation of $\hat{u}$ is defined by the minimization

$$u^h = \underset{v^h \in \mathcal{U}^h}{\operatorname{argmin}} \|v^h - \hat{u}\|^2, \tag{4.1.3}$$

where the minimizer, u^h, is the L^2-orthogonal projection of $\hat{u}$ onto $\mathcal{U}^h$. The minimization problem (4.1.3) can be reformulated as a standard $\mathcal{L}\mathcal{L}^*$ method [43, 82], but only for a special choice of the finite element space. However, for general $\mathcal{U}^h$, the L^2-orthogonal projection of $\hat{u}$ onto $\mathcal{U}^h$ cannot be directly computed, unless the exact solution, $\hat{u}$, is readily known. The idea here is to replace (4.1.3) with a similar, but computationally feasible, minimization problems using an additional (auxiliary) finite element space and applying the ideas of the standard $\mathcal{L}\mathcal{L}^*$ and negative-norm methods; see, e.g., [45] for an H^{-1} approach to elliptic problems. In comparison, the standard $\mathcal{L}\mathcal{L}^*$ method obtains the best $L^2(\Omega)$ approximation under the compromise of using a particular

and nonstandard finite element space, whereas the $(\mathcal{L}\mathcal{L}^*)^{-1}$ and $\mathcal{L}\mathcal{L}^*$-type methods studied in this chapter allow utilizing standard finite element spaces but generally do not provide precisely the L^2-orthogonal projection of the exact solution.

Several methods are studied and compared in this chapter. In particular, the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method is in the class of negative-norm least-squares methods. However, unlike a more standard H^{-1} approach, the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method is better tailored to the particular problem (4.1.2). Namely, the isomorphism $(-\Delta)^{-1}$ in the H^{-1} method is replaced[1] by the isomorphism $(L_w L^*)^{-1}$. Here, L_w is a special "weak version" of the operator L that is rigorously defined below. In general, the norm $\|L(\cdot)\|_{-1}$ (here, $\|\cdot\|_{-1}$ denotes the H^{-1} norm) does not control the $L^2(\Omega)$ norm, when L is a hyperbolic operator; in fact, it is not even discretely (i.e., on any collection of finite element spaces) L^2-coercive [32]; see also Chapter 3. This is associated with the difficulty in analyzing the L^2-convergence of the H^{-1}-based methods in [34, 32] (and its related $H(\mathrm{div})$-conforming method) and in Chapter 3. In contrast, we observe that replacing $\|\cdot\|_{-1}$ with the dual norm corresponding to $(L_w L^*)^{-1}$ precisely recovers the $L^2(\Omega)$ norm. In practice, this desirable property of $(L_w L^*)^{-1}$ is lost when the operator is approximated by a discrete version. We demonstrate that under certain conditions a discrete L^2-coercivity of the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method remains valid, which is sufficient for obtaining optimal convergence rates. All methods studied in this chapter converge in the $L^2(\Omega)$ norm. Since operators play such an important role in our considerations, we provide a systematic analysis of the properties of the operators of interest here.

Negative-norm least-squares methods can be viewed as particular Petrov-Galerkin finite element methods, since Petrov-Galerkin methods constitute a very wide class; see [83, 84] and the references therein. This chapter follows a slightly different path, in a sense, more in the spirit of least-squares methods. Namely, we extend the standard $\mathcal{L}\mathcal{L}^*$ method of [43] either by further projections onto $\mathcal{U}^h$ constituting the $\mathcal{L}\mathcal{L}^*$-type methods, or by employing a related negative-norm minimization resulting in the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method. All methods of this chapter are fundamentally based on the original $\mathcal{L}\mathcal{L}^*$ minimization principle in [43]. The relation to Petrov-Galerkin methods is interesting in its own right. The potential of further extending the $\mathcal{L}\mathcal{L}^*$ approach using the (discontinuous) Petrov-Galerkin framework is a subject of future work.

[1] In view of the weak formulations of these isomorphisms, this can be stated as: the gradient, ∇, is replaced by L^* – the L^2-adjoint of L.

The main contributions of this chapter are summarized as follows. The $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation is introduced and analyzed, which is a new approach. Also, the idea of formulating a negative-norm least-squares method as a "saddle-point problem", to our knowledge, does not exist in the literature. A more typical approach is the one in [45], where the conjugate gradient method is directly applied to minimize the functional of interest. For practical purposes, they use a preconditioner (i.e., an approximate inverse of an operator) that effectively modifies the least-squares principle. In contrast, the approach here allows utilization of the original (unmodified) minimization principle. Note that the norm for the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method is different from the one in [45]. Moreover, additional difficulties arise when using the conjugate gradient method for a modified least-squares principle in the context of hyperbolic PDEs; see Section 4.6. The standard $\mathcal{L}\mathcal{L}^*$ method is not new; it is formulated in [43] in the context of elliptic problems. The single- and two-stage methods are simple extensions of the original $\mathcal{L}\mathcal{L}^*$ approach. Although not in such a pure form, they can be seen as a part of the hybrid method in [44]. The application of the $\mathcal{L}\mathcal{L}^*$, single-, and two-stage methods to hyperbolic problems is, however, a new development. Most notably, the error analysis in Section 4.5 of the single- and two-stage methods in terms of the approximation properties of the involved finite element spaces was not previously known.

The outline of the rest of the chapter is the following. Basic notions and assumptions are presented in Section 4.2. Section 4.3 contains a systematic overview of the properties of the operators of interest. In Section 4.4, the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method is formulated and analyzed. Section 4.5 is devoted to the $\mathcal{L}\mathcal{L}^*$-type methods and their comparison to the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method. In Section 4.6, we comment on the implementation of the methods and the preconditioning of the respective linear systems. The specifics of applying the methods to (4.1.1) are discussed in Section 4.7. Particular numerical results are collected in Section 4.8. Section 4.9 discusses certain regularizations of the $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation. Conclusions and possible future work are in Section 4.10. Section 4.A is an appendix that, for convenience, provides an overview of the generalization of the considerations in the chapter. Note that this mainly contributes to a detailed exposition, while the methods are naturally implemented and used in such a general framework.

4.2 Notation, definitions, and assumptions

In this section, useful notation and definitions are presented, along with a pair of basic assumptions.

Consider a domain, $\Omega \subset \mathbb{R}^d$, and a linear first-order differential operator, L (i.e., closed unbounded). The norm and inner product on $L^2(\Omega)$ are denoted by $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$, respectively. The domain of L is defined as

$$\mathcal{D}(L) = \{\, u \in L^2(\Omega);\ Lu \in L^2(\Omega) \text{ and } Bu = 0 \,\},$$

where $Bu = 0$ represents appropriate homogeneous boundary conditions. Note that L is densely defined in the sense that $\mathcal{D}(L)$ is dense in $L^2(\Omega)$. This is easy to see since, clearly, the infinitely smooth compactly supported functions on Ω, $\mathcal{C}_c^\infty(\Omega)$, are contained in $\mathcal{D}(L)$ and it is well known that $\mathcal{C}_c^\infty(\Omega)$ is dense in $L^2(\Omega)$. Thus, L^*, the L^2-adjoint of L, is a well-defined closed linear operator [85]. In general, the adjoint operator, L^*, and its domain, $\mathcal{D}(L^*)$, are defined as follows: if, for $w \in L^2(\Omega)$, there exists $q \in L^2(\Omega)$ such that

$$\langle Lu, w \rangle = \langle u, q \rangle, \quad \forall u \in \mathcal{D}(L),$$

then we say that $w \in \mathcal{D}(L^*)$ and $L^* w = q$. For our considerations, it is convenient to express $\mathcal{D}(L^*)$ as

$$\mathcal{D}(L^*) = \{\, w \in L^2(\Omega);\ L^* w \in L^2(\Omega) \text{ and } B^* w = 0 \,\},$$

where $B^* w = 0$ is the adjoint homogeneous boundary condition. Furthermore, it is known from functional analysis (see [85]) that, in general, L being densely defined and closed implies that L^* is also densely defined and $(L^*)^* = L$.

Assume that L^* satisfies a Poincaré-type inequality and that it is surjective; that is, for some constant $c_P^* > 0$,

(ASM 1) $$c_P^* \|w\| \leq \|L^* w\|, \quad \forall w \in \mathcal{D}(L^*),$$

(ASM 2) $$L^*(\mathcal{D}(L^*)) = L^2(\Omega).$$

The motivation behind these assumptions is that they are important for the theory in Section 4.3 and they are satisfied by the problem of interest (4.1.1). This is discussed in Section 4.7.

Notice that assumption (ASM 1) implies that $\mathcal{D}(L^*)$ is a Hilbert space with respect to the norm $\|\cdot\|_{\mathcal{D}(L^*)} = \|L^*(\cdot)\|$ and this norm is equivalent to the respective graph norm on $\mathcal{D}(L^*)$; that is, there exists a constant $c_G^* > 0$ such that

$$c_G^*(\|w\|^2 + \|L^*w\|^2) \leq \|L^*w\|^2 \leq \|w\|^2 + \|L^*w\|^2, \quad \forall w \in \mathcal{D}(L^*).$$

Denote the dual space of $\mathcal{D}(L^*)$ by $\mathcal{D}'(L^*)$. The associated functional norm is

$$\|\ell\|_{\mathcal{D}'(L^*)} = \sup_{w \in \mathcal{D}(L^*)} \frac{|\ell(w)|}{\|L^*w\|}, \quad \forall \ell \in \mathcal{D}'(L^*).$$

To simplify notation, it is understood that $w \neq 0$ in the supremum and this convention is used throughout the chapter. This leads to the following definitions.

Definition 4.2.1 Let $q \in L^2(\Omega)$ and consider the functional $\vartheta_q(w) = \langle q, L^*w \rangle$ for all $w \in \mathcal{D}(L^*)$. It is easy to see that $\vartheta_q \in \mathcal{D}'(L^*)$. Define the linear map $L_w \colon L^2(\Omega) \to \mathcal{D}'(L^*)$ as $L_w q = \vartheta_q$ for all $q \in L^2(\Omega)$. The operator L_w is the "weak version" of L, defined on the whole $L^2(\Omega)$. $\diamond$

Definition 4.2.2 The linear map $(L_w L^*)^{-1} \colon \mathcal{D}'(L^*) \to \mathcal{D}(L^*)$ is defined through the solution of the weak problem

$$\text{Find } z \in \mathcal{D}(L^*) \colon \langle L^*z, L^*w \rangle = \ell(w), \quad \forall w \in \mathcal{D}(L^*), \tag{4.2.1}$$

where $\ell \in \mathcal{D}'(L^*)$; that is, if $\hat{z} \in \mathcal{D}(L^*)$ solves (4.2.1), then $(L_w L^*)^{-1}\ell = \hat{z}$. $\diamond$

Owing to (ASM 1) and the Riesz representation theorem, (4.2.1) has a unique solution. Hence, $(L_w L^*)^{-1}$ is well-defined. It becomes clear in the next section that the notation $(L_w L^*)^{-1}$ is consistent and meaningful.

Remark 4.2.3 Assumption (ASM 2) is equivalent (see [77, Theorems 2.20 and 2.21]) to the assumption that

$$c_P\|u\| \leq \|Lu\|, \quad \forall u \in \mathcal{D}(L), \tag{ASM 3}$$

for some constant $c_P > 0$. This is the assumption that L satisfies a Poincaré-type inequality. Similar to above, (ASM 3) implies that $\mathcal{D}(L)$ is a Hilbert space with respect to the norm $\|\cdot\|_{\mathcal{D}(L)} = \|L(\cdot)\|$,

and this norm is equivalent to the respective graph norm on $\mathcal{D}(L)$; that is, there exists a constant $c_G > 0$ such that

$$c_G(\|u\|^2 + \|Lu\|^2) \leq \|Lu\|^2 \leq \|u\|^2 + \|Lu\|^2, \quad \forall u \in \mathcal{D}(L).$$

Similarly, (ASM 1) is equivalent to the assumption that $L\colon \mathcal{D}(L) \to L^2(\Omega)$ is surjective:

(ASM 4) $$L(\mathcal{D}(L)) = L^2(\Omega).$$ ◇

4.3 Properties of the operators

This section is devoted to the theoretical study of the properties of the operators introduced in Section 4.2. First, some abstract theory is presented. Then, the properties of $(L_w L^*)^{-1}$ relative to the $L^2(\Omega)$ inner product are shown. The main idea is to characterize the $L^2(\Omega)$ norm in terms of the norm in $\mathcal{D}'(L^*)$ and to properly represent the functional norm aiming at obtaining, in Section 4.4, an appropriate computable approximation of the $L^2(\Omega)$ minimization (4.1.3).

To aid precision and clarity below, note that $L^2(\Omega)$ can be embedded into $\mathcal{D}'(L^*)$. Indeed, for any $q \in L^2(\Omega)$, consider the functional $\ell_q(w) = \langle q, w\rangle$ for all $w \in \mathcal{D}(L^*)$. Using (ASM 1), it is easy to see that $\ell_q \in \mathcal{D}'(L^*)$. Then, the embedding operator $\mathcal{E}\colon L^2(\Omega) \to \mathcal{D}'(L^*)$ is defined as $\mathcal{E}q = \ell_q$ for all $q \in L^2(\Omega)$. Moreover, it is not difficult to show, using (ASM 1), that $\mathcal{E}\colon L^2(\Omega) \to \mathcal{D}'(L^*)$ is a bounded linear operator and, thus, it represents the continuous embedding of $L^2(\Omega)$ into $\mathcal{D}'(L^*)$.

The operator $(L_w L^*)^{-1}\mathcal{E}$ maps $L^2(\Omega)$ into $\mathcal{D}(L^*)$. Owing to (4.2.1) and the definition of $\mathcal{E}$, for $q \in L^2(\Omega)$, $(L_w L^*)^{-1}\mathcal{E}q$ equals the solution of the weak problem

(4.3.1) $$\text{Find } z \in \mathcal{D}(L^*)\colon \langle L^* z, L^* w\rangle = \langle q, w\rangle, \quad \forall w \in \mathcal{D}(L^*).$$

As customary, for simplicity of notation, we skip the embedding, $\mathcal{E}$, in the notation for the operator $(L_w L^*)^{-1}\mathcal{E}$ and consider $(L_w L^*)^{-1}\colon L^2(\Omega) \to \mathcal{D}(L^*)$ defined through the solution of the weak problem (4.3.1). This should lead to no confusion, since it should be clear from the context if $(L_w L^*)^{-1}$ denotes the operator $(L_w L^*)^{-1}\colon L^2(\Omega) \to \mathcal{D}(L^*)$ associated with the solution of the weak form (4.3.1) or $(L_w L^*)^{-1}\colon \mathcal{D}'(L^*) \to \mathcal{D}(L^*)$ introduced in Definition 4.2.2 and associated with the solution of the weak form (4.2.1); that is, the map $(L_w L^*)^{-1}\colon \mathcal{D}'(L^*) \to \mathcal{D}(L^*)$ can be considered as defined on $L^2(\Omega)$ via the embedding $\mathcal{E}$, allowing the consideration of the operator $(L_w L^*)^{-1}\colon L^2(\Omega) \to \mathcal{D}(L^*)$. The strict meaning behind this is provided by the weak form (4.3.1).

4.3.1 Abstract properties

First, a few basic results are collected in the following lemmas. The motivation behind the operator L_w is that it extends L (in fact, it extends $\mathcal{E}L$) on $L^2(\Omega)$, in the sense that L_w coincides with $\mathcal{E}L$ on $\mathcal{D}(L)$. This allows the general characterization below of the norm in $L^2(\Omega)$ over the entire space. The result is important in the formulation of the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method, since, as demonstrated in the next section, it essentially moves the infeasibility of (4.1.3), caused by the presence of the exact solution, $\hat{u}$, to the functional norm in $\mathcal{D}'(L^*)$.

Lemma 4.3.1 *The operator L_w coincides with $\mathcal{E}L$ on $\mathcal{D}(L)$.*

Proof. It is easy to see, from the definitions of L_w and $\mathcal{E}$, that L_w coincides with $\mathcal{E}L$ on $\mathcal{D}(L)$. ∎

Lemma 4.3.2 *The operator $L^*\colon \mathcal{D}(L^*) \to L^2(\Omega)$ is a bijective isometry.*

Proof. This property follows immediately from (ASM 1) and the surjectivity of L^* in (ASM 2), using that $\mathcal{D}(L^*)$ is endowed with the norm $\|\cdot\|_{\mathcal{D}(L^*)} = \|L^*(\cdot)\|$. ∎

Remark 4.3.3 Similarly, (ASM 3) and (ASM 4) (or, equivalently, as discussed in Remark 4.2.3, (ASM 1) and (ASM 2)) imply that $L\colon \mathcal{D}(L) \to L^2(\Omega)$ is a bijective isometry. ◇

It is not practical to work directly with a dual norm like $\|\cdot\|_{\mathcal{D}'(L^*)}$. Therefore, the operator $(L_wL^*)^{-1}\colon \mathcal{D}'(L^*) \to \mathcal{D}(L^*)$ is considered. As implied by the following lemma, it is the Riesz isomorphism between $\mathcal{D}(L^*)$ and $\mathcal{D}'(L^*)$, i.e., it is the isomorphism between a Hilbert space and its dual, mapping functionals to their representations with respect to the inner product in the Hilbert space, in accordance with the Riesz representation theorem. In essence, $(L_wL^*)^{-1}$ is the analog of the inverse Laplace operator in H^{-1}-type methods.

Lemma 4.3.4 *The operator $(L_wL^*)^{-1}\colon \mathcal{D}'(L^*) \to \mathcal{D}(L^*)$ is a bijective isometry. This, combined with* (4.2.1)*, means that*

$$\|\ell\|^2_{\mathcal{D}'(L^*)} = \langle L^*(L_wL^*)^{-1}\ell, L^*(L_wL^*)^{-1}\ell\rangle = \ell((L_wL^*)^{-1}\ell), \quad \forall \ell \in \mathcal{D}'(L^*).$$

In particular,

$$\|\mathcal{E}q\|^2_{\mathcal{D}'(L^*)} = \langle (L_wL^*)^{-1}\mathcal{E}q, q\rangle, \quad \forall q \in L^2(\Omega).$$

Proof. Owing to (4.2.1), it is an isometry since

$$\begin{aligned}\|\ell\|_{\mathcal{D}'(L^*)} &= \sup_{w\in\mathcal{D}(L^*)} \frac{|\ell(w)|}{\|L^*w\|} = \sup_{w\in\mathcal{D}(L^*)} \frac{|\langle L^*\hat{z}, L^*w\rangle|}{\|L^*w\|} \\ &= \|L^*\hat{z}\| = \|\hat{z}\|_{\mathcal{D}(L^*)} = \|(L_wL^*)^{-1}\ell\|_{\mathcal{D}(L^*)} = \|L^*(L_wL^*)^{-1}\ell\|,\end{aligned}$$

where $\hat{z} \in \mathcal{D}(L^*)$ is the solution of (4.2.1), i.e., $\hat{z} = (L_wL^*)^{-1}\ell$. The fact that $(L_wL^*)^{-1}$ is an isometry immediately implies that it is injective. Owing to (ASM 1), (4.2.1), and the Riesz representation theorem, it follows that $(L_wL^*)^{-1}$ is surjective. Indeed, for any $z \in \mathcal{D}(L^*)$, consider $\ell_z^\bullet \in \mathcal{D}'(L^*)$ defined as $\ell_z^\bullet(w) = \langle L^*z, L^*w\rangle$ for all $w \in \mathcal{D}(L^*)$. Then, clearly, $(L_wL^*)^{-1}\ell_z^\bullet = z$, showing that it is surjective.

Finally, using (4.2.1) and the definition of $\mathcal{E}$ (or, equivalently, using (4.3.1)), it holds that

$$\begin{aligned}\|\mathcal{E}q\|^2_{\mathcal{D}'(L^*)} &= \langle L^*(L_wL^*)^{-1}\mathcal{E}q, L^*(L_wL^*)^{-1}\mathcal{E}q\rangle \\ &= [\mathcal{E}q]((L_wL^*)^{-1}\mathcal{E}q) = \langle q, (L_wL^*)^{-1}\mathcal{E}q\rangle,\end{aligned}$$

where it is utilized that $(L_wL^*)^{-1}\mathcal{E}q$ solves (4.2.1) with $\ell = \mathcal{E}q$. ∎

The above lemmas together with the theorem below provide justification for the (symbolic) equality $(L_wL^*)^{-1} = (L^*)^{-1}L_w^{-1}$. More importantly, the following theorem essentially demonstrates that the (symbolic) map $L^*(L_wL^*)^{-1}L_w$ is the identity operator $I\colon L^2(\Omega) \to L^2(\Omega)$, which provides a characterization of the $L^2(\Omega)$ norm in terms of the functional norm on $\mathcal{D}'(L^*)$.

Theorem 4.3.5 (characterization of the L^2 norm) *The operator $L_w\colon L^2(\Omega) \to \mathcal{D}'(L^*)$ is a bijective isometry. In particular, this means that*

$$\|q\| = \|L_wq\|_{\mathcal{D}'(L^*)}, \quad \forall q \in L^2(\Omega).$$

Proof. Using the surjectivity of L^* in (ASM 2), L_w is an isometry since

$$\|L_wq\|_{\mathcal{D}'(L^*)} = \sup_{w\in\mathcal{D}(L^*)} \frac{|\langle q, L^*w\rangle|}{\|L^*w\|} = \|q\|.$$

The fact that L_w is an isometry immediately implies that it is injective. Consider arbitrary $\ell \in \mathcal{D}'(L^*)$ and let $\hat{z} \in \mathcal{D}(L^*)$ be the solution of (4.2.1), i.e., $\hat{z} = (L_wL^*)^{-1}\ell$. Then, by setting $\hat{q} = L^*\hat{z} \in L^2(\Omega)$, it follows from (4.2.1) that $L_w\hat{q} = \ell$. Thus, L_w is surjective. ∎

Corollary 4.3.6 *It holds that*

$$\|u\| = \|\mathcal{E}Lu\|_{\mathcal{D}'(L^*)}, \quad \forall u \in \mathcal{D}(L).$$

Proof. The equality follows from Theorem 4.3.5 and Lemma 4.3.1. ∎

4.3.2 Properties of $(L_w L^*)^{-1}$ in L^2

As discussed at the beginning of this section, $L^2(\Omega)$ is embedded into $\mathcal{D}'(L^*)$ and, hence, $(L_w L^*)^{-1}$ can be considered defined on $L^2(\Omega)$ via the embedding $\mathcal{E}$. This subsection focuses on the operator $(L_w L^*)^{-1}\mathcal{E}\colon L^2(\Omega) \to \mathcal{D}(L^*)$. The following lemmas establish that $(L_w L^*)^{-1}\mathcal{E}$ is continuous, self-adjoint, and positive definite with respect to the $L^2(\Omega)$ inner product.

Lemma 4.3.7 (L^2-continuity) *It holds that*

$$\|(L_w L^*)^{-1}\mathcal{E}q\| \leq \frac{1}{(c_P^*)^2}\|q\|, \quad \forall q \in L^2(\Omega).$$

Proof. Using (ASM 1) and (4.3.1), it follows

$$\begin{aligned}\|(L_w L^*)^{-1}\mathcal{E}q\| &\leq \frac{1}{c_P^*}\|L^*(L_w L^*)^{-1}\mathcal{E}q\| = \frac{1}{c_P^*}\sup_{w\in\mathcal{D}(L^*)}\frac{|\langle L^*(L_w L^*)^{-1}\mathcal{E}q, L^*w\rangle|}{\|L^*w\|} \\ &= \frac{1}{c_P^*}\sup_{w\in\mathcal{D}(L^*)}\frac{|\langle q, w\rangle|}{\|L^*w\|} \leq \frac{1}{(c_P^*)^2}\sup_{w\in\mathcal{D}(L^*)}\frac{|\langle q, w\rangle|}{\|w\|} \leq \frac{1}{(c_P^*)^2}\|q\|.\end{aligned}$$ ∎

Lemma 4.3.4 allows to characterize the inner product in $\mathcal{D}'(L^*)$ using the operator $(L_w L^*)^{-1}$. This is important for the considerations in Section 4.4, since by approximating $(L_w L^*)^{-1}$, the $\mathcal{D}'(L^*)$ norm is approximated, thus obtaining, in view of Theorem 4.3.5, computationally feasible approximations of the $L^2(\Omega)$ norm and the minimization (4.1.3). In practical finite element formulations, the $\mathcal{D}'(L^*)$ inner product characterization is needed for functions in $L^2(\Omega)$. This is the motivation behind the following result. It shows that $\langle (L_w L^*)^{-1}\mathcal{E}\cdot, \cdot\rangle$ defines an inner product in $L^2(\Omega)$, which, by Lemma 4.3.4, is precisely the inner product associated with $\|\cdot\|_{\mathcal{D}'(L^*)}$, but restricted, via the embedding $\mathcal{E}$, to $L^2(\Omega)$.

Lemma 4.3.8 *The operator $(L_w L^*)^{-1}\mathcal{E}\colon L^2(\Omega) \to \mathcal{D}(L^*)$ is self-adjoint and positive definite with respect to the $L^2(\Omega)$ inner product.*

Proof. Given $p, q \in L^2(\Omega)$, by (4.2.1) and the definition of $\mathcal{E}$ (or, equivalently, by (4.3.1)), it follows

$$\begin{aligned}\langle q, (L_w L^*)^{-1}\mathcal{E}p\rangle &= \langle L^*(L_w L^*)^{-1}\mathcal{E}q, L^*(L_w L^*)^{-1}\mathcal{E}p\rangle \\ &= \langle L^*(L_w L^*)^{-1}\mathcal{E}p, L^*(L_w L^*)^{-1}\mathcal{E}q\rangle = \langle p, (L_w L^*)^{-1}\mathcal{E}q\rangle,\end{aligned}$$

where $(L_w L^*)^{-1}\colon \mathcal{D}'(L^*) \to \mathcal{D}(L^*)$, as defined above.

Next, owing to Lemma 4.3.4,

$$\langle (L_w L^*)^{-1}\mathcal{E}q, q\rangle = \|\mathcal{E}q\|^2_{\mathcal{D}'(L^*)} \geq 0,$$

where the equality holds if and only if $q = 0$. ∎

Remark 4.3.9 Note that $(L_w L^*)^{-1}\mathcal{E}$ is not necessarily L^2-coercive (strictly positive definite), that is, $\langle (L_w L^*)^{-1}\mathcal{E}q, q\rangle \geq \alpha\|q\|^2$, for all $q \in L^2(\Omega)$, does not necessarily hold for any constant $\alpha > 0$. In view of Lemma 4.3.4, this reflects the fact that the $L^2(\Omega)$ norm is generally strictly stronger than the $\mathcal{D}'(L^*)$ norm on $L^2(\Omega)$. ◇

Remark 4.3.10 The assumption that L^* is L^2-coercive, (ASM 1), is important for the theory in this section, since basic results depend on it. Namely, it provides that the operator $(L_w L^*)^{-1}$ is well-defined, $\mathcal{D}(L^*)$ is a Hilbert space with respect to the norm $\|\cdot\|_{\mathcal{D}(L^*)} = \|L^*(\cdot)\|$, $\mathcal{D}'(L^*)$ can be endowed with the respective dual norm, and allows the above presented definition of the embedding $\mathcal{E}$. Also, the closedness of the operator L is important. Particularly, it provides that $(L^*)^* = L$. The assumption that L^* is surjective, (ASM 2), is used only in Lemma 4.3.2, Remark 4.3.3, Theorem 4.3.5, and Corollary 4.3.6, which are important results. ◇

4.4 The $(\mathcal{L}\mathcal{L}^*)^{-1}$ method

The $(\mathcal{L}\mathcal{L}^*)^{-1}$ method, which is a main focus of this chapter, is presented in this section. First, the method is formulated. Next, the corresponding linear algebra equations are discussed. Finally, the properties of the discrete formulation are studied.

4.4.1 Motivation and formulation

Let $\mathcal{U}^h$ be a finite element space and consider the operator equation, $Lu = f$, in (4.1.2). For simplicity of exposition, $\mathcal{U}^h$ is a subset of $\mathcal{D}(L)$ in this section. The extension of the formulation

to more general finite element spaces is discussed in Section 4.A. The purpose is to obtain $u^h \in \mathcal{U}^h$ that approximates the exact solution of (4.1.2) in the $L^2(\Omega)$ norm. Owing to Corollary 4.3.6 and Lemma 4.3.4, the minimization (4.1.3) can be equivalently expressed as

$$\begin{aligned} u^h &= \operatorname*{argmin}_{v^h \in \mathcal{U}^h} \|v^h - \hat{u}\|^2 = \operatorname*{argmin}_{v^h \in \mathcal{U}^h} \|\mathcal{E}L(v^h - \hat{u})\|^2_{\mathcal{D}'(L^*)} \\ &= \operatorname*{argmin}_{v^h \in \mathcal{U}^h} \|\mathcal{E}(Lv^h - f)\|^2_{\mathcal{D}'(L^*)} = \operatorname*{argmin}_{v^h \in \mathcal{U}^h} \langle (L_w L^*)^{-1}(Lv^h - f), Lv^h - f\rangle, \end{aligned} \tag{4.4.1}$$

where $\hat{u} \in \mathcal{D}(L)$ denotes the exact solution of (4.1.2). In view of the symmetry in Lemma 4.3.8, this leads to the weak problem

$$\text{Find } u^h \in \mathcal{U}^h \colon \langle (L_w L^*)^{-1} L u^h, Lv^h\rangle = \langle (L_w L^*)^{-1} f, Lv^h\rangle, \quad \forall v^h \in \mathcal{U}^h. \tag{4.4.2}$$

Observe that (4.4.1) and (4.4.2) are not computationally feasible, since the effect of $(L_w L^*)^{-1}$ cannot be computed in general. Therefore, a computable discrete version of $(L_w L^*)^{-1}$ is necessary. To this end, consider an additional (auxiliary) finite element space $\mathcal{Z}^\mathfrak{h} \subset \mathcal{D}(L^*)$. The discrete version of $(L_w L^*)^{-1}$ is obtained from the discrete version of (4.2.1), described as follows.

Definition 4.4.1 The linear map $(L_w L^*)^{-1}_\mathfrak{h} \colon \mathcal{D}'(L^*) \to \mathcal{Z}^\mathfrak{h}$ is defined through the solution of the discrete weak problem

$$\text{Find } z^\mathfrak{h} \in \mathcal{Z}^\mathfrak{h} \colon \langle L^* z^\mathfrak{h}, L^* w^\mathfrak{h}\rangle = \ell(w^\mathfrak{h}), \quad \forall w^\mathfrak{h} \in \mathcal{Z}^\mathfrak{h}, \tag{4.4.3}$$

where $\ell \in \mathcal{D}'(L^*)$.

As previously, when convenient, the operator $(L_w L^*)^{-1}_\mathfrak{h} \colon L^2(\Omega) \to \mathcal{Z}^\mathfrak{h}$ is considered (via the embedding $\mathcal{E}$), in which case, for $q \in L^2(\Omega)$, (4.4.3) takes the form

$$\text{Find } z^\mathfrak{h} \in \mathcal{Z}^\mathfrak{h} \colon \langle L^* z^\mathfrak{h}, L^* w^\mathfrak{h}\rangle = \langle q, w^\mathfrak{h}\rangle, \quad \forall w^\mathfrak{h} \in \mathcal{Z}^\mathfrak{h}. \tag{4.4.4}$$ ◇

Now, (4.4.1) and (4.4.2) can be approximated feasibly by replacing $(L_w L^*)^{-1}$ with $(L_w L^*)^{-1}_\mathfrak{h}$. This results in the following:

$$u^h = \operatorname*{argmin}_{v^h \in \mathcal{U}^h} \langle (L_w L^*)^{-1}_\mathfrak{h}(Lv^h - f), Lv^h - f\rangle, \tag{4.4.5}$$

$$\text{Find } u^h \in \mathcal{U}^h \colon \langle (L_w L^*)^{-1}_\mathfrak{h} L u^h, Lv^h\rangle = \langle (L_w L^*)^{-1}_\mathfrak{h} f, Lv^h\rangle, \quad \forall v^h \in \mathcal{U}^h, \tag{4.4.6}$$

which constitutes the discrete $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation. Alternatively, (4.4.4) and (4.4.6) can be combined into the system

$$\text{Find } (u^h, z^\mathfrak{h}) \in \mathcal{U}^h \times \mathcal{Z}^\mathfrak{h} \colon \begin{cases} \langle L^* z^\mathfrak{h}, L^* w^\mathfrak{h} \rangle + \langle u^h, L^* w^\mathfrak{h} \rangle = \langle f, w^\mathfrak{h} \rangle, & \forall w^\mathfrak{h} \in \mathcal{Z}^\mathfrak{h}, \\ \langle L^* z^\mathfrak{h}, v^h \rangle = 0, & \forall v^h \in \mathcal{U}^h. \end{cases} \tag{4.4.7}$$

In summary, exchanging $(L_w L^*)^{-1}$ for $(L_w L^*)_\mathfrak{h}^{-1}$ is practically trading the minimization of the $L^2(\Omega)$ norm of the error in (4.4.1) for computational feasibility. Namely, the resulting minimization problem (4.4.5) can be solved numerically but does not necessarily provide the L^2-orthogonal projection of the exact solution onto $\mathcal{U}^h$. In contrast, the standard $\mathcal{L}\mathcal{L}^*$ method introduced in [43] solves the L^2 minimization (4.4.1), but for the special choice $\mathcal{U}^h = L^*(\mathcal{Z}^\mathfrak{h})$, i.e., it trades the freedom of choosing a standard finite element space in the place of $\mathcal{U}^h$ for computational feasibility. Moreover, the $\mathcal{L}\mathcal{L}^*$ method uses the space $\mathcal{Z}^\mathfrak{h}$ (more precisely, the space $L^*(\mathcal{Z}^\mathfrak{h})$) to approximate the exact solution, $\hat{u}$, whereas, in the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method introduced above, $\mathcal{Z}^\mathfrak{h}$ serves as an auxiliary space to approximate the operator $(L_w L^*)^{-1} \colon \mathcal{D}'(L^*) \to \mathcal{D}(L^*)$ by the operator $(L_w L^*)_\mathfrak{h}^{-1} \colon \mathcal{D}'(L^*) \to \mathcal{Z}^\mathfrak{h} \subset \mathcal{D}(L^*)$. See Sections 4.5 and 4.8 for further and more detailed comparisons of the $(\mathcal{L}\mathcal{L}^*)^{-1}$ and other $\mathcal{L}\mathcal{L}^*$-type methods. The implications of approximating the minimization problem (4.4.1) by (4.4.5) are studied in Subsection 4.4.3.

4.4.2 Linear algebra equations

Here, the algebraic systems associated with (4.4.6) and (4.4.7) are formulated. Let $\{\, \phi_i^h \,\}_{i=1}^N$ and $\{\, \psi_i^\mathfrak{h} \,\}_{i=1}^M$ be the bases for $\mathcal{U}^h$ and $\mathcal{Z}^\mathfrak{h}$, respectively. Define the matrices $\boldsymbol{L} \in \mathbb{R}^{M \times N}$, $\boldsymbol{H} \in \mathbb{R}^{M \times M}$, $\boldsymbol{M} \in \mathbb{R}^{N \times N}$ (the $L^2(\Omega)$ mass matrix on $\mathcal{U}^h$), and the vector $\bar{\boldsymbol{f}} \in \mathbb{R}^M$ as

$$(\boldsymbol{L})_{ij} = \langle \phi_j^h, L^* \psi_i^\mathfrak{h} \rangle, \; (\boldsymbol{H})_{ij} = \langle L^* \psi_j^\mathfrak{h}, L^* \psi_i^\mathfrak{h} \rangle, \; (\boldsymbol{M})_{ij} = \langle \phi_j^h, \phi_i^h \rangle, \; (\bar{\boldsymbol{f}})_i = \langle f, \psi_i^\mathfrak{h} \rangle. \tag{4.4.8}$$

The functions in $\mathcal{U}^h$ and $\mathcal{Z}^\mathfrak{h}$ can be identified with their corresponding coefficient vectors with respect to the bases of the spaces. Namely, $u^h \in \mathcal{U}^h$, $\boldsymbol{u} \in \mathbb{R}^N$ and $z^\mathfrak{h} \in \mathcal{Z}^\mathfrak{h}$, $\boldsymbol{z} \in \mathbb{R}^M$ are identified with the expansions

$$u^h = \sum_{i=1}^N (\boldsymbol{u})_i \phi_i^h, \quad z^\mathfrak{h} = \sum_{i=1}^M (\boldsymbol{z})_i \psi_i^\mathfrak{h}.$$

Using this notation, the weak formulation (4.4.7) induces the following algebraic system of equations with a symmetric block matrix $\mathbb{A}$:

$$\mathbb{A}\begin{bmatrix}\boldsymbol{z}\\ \boldsymbol{u}\end{bmatrix}=\begin{bmatrix}\boldsymbol{H} & \boldsymbol{L}\\ \boldsymbol{L}^T & \end{bmatrix}\begin{bmatrix}\boldsymbol{z}\\ \boldsymbol{u}\end{bmatrix}=\begin{bmatrix}\bar{\boldsymbol{f}}\\ \boldsymbol{0}\end{bmatrix}. \tag{4.4.9}$$

Note that, owing to (ASM 1), $\boldsymbol{H}$ is a symmetric positive definite matrix. Eliminating $\boldsymbol{z}$ in (4.4.9) results in the following algebraic system for the respective Schur complement:

$$\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\boldsymbol{u}=\boldsymbol{L}^T\boldsymbol{H}^{-1}\bar{\boldsymbol{f}}. \tag{4.4.10}$$

Denote $\boldsymbol{A}=\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\in\mathbb{R}^{N\times N}$ and $\boldsymbol{f}=\boldsymbol{L}^T\boldsymbol{H}^{-1}\bar{\boldsymbol{f}}\in\mathbb{R}^N$. Then (4.4.10) becomes

$$\boldsymbol{A}\boldsymbol{u}=\boldsymbol{f}, \tag{4.4.11}$$

which is precisely the algebraic system induced by the weak form (4.4.6). Indeed, since the solution of (4.4.4) (i.e., the effect of $(L_wL^*)_{\mathfrak{h}}^{-1}$) is computed through the effect of $\boldsymbol{H}^{-1}$, the matrix $\boldsymbol{A}$ corresponds to the bilinear form and $\boldsymbol{f}$ corresponds to the right-hand side in (4.4.6); that is,

$$(\boldsymbol{A})_{ij}=\langle(L_wL^*)_{\mathfrak{h}}^{-1}L\phi_j^h,L\phi_i^h\rangle,\quad(\boldsymbol{f})_i=\langle(L_wL^*)_{\mathfrak{h}}^{-1}f,L\phi_i^h\rangle.$$

Clearly, $\boldsymbol{A}$ is nonsingular if and only if the matrix $\mathbb{A}$ in (4.4.9) is nonsingular.

4.4.3 Analysis

In this subsection, the properties of the operator $(L_wL^*)_{\mathfrak{h}}^{-1}$ and the discrete $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation are analyzed and studied in detail. The analysis of the discrete $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation and the properties of the matrix $\boldsymbol{A}$ is fundamentally founded upon the effect of replacing $(L_wL^*)^{-1}$ with $(L_wL^*)_{\mathfrak{h}}^{-1}$ on the characterization of the $L^2(\Omega)$ norm. The major result is the error estimate for the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method.

First, the discrete counterparts of Lemmas 4.3.7 and 4.3.8 are shown.

Lemma 4.4.2 (L^2-continuity) *It holds that*

$$\|(L_wL^*)_{\mathfrak{h}}^{-1}q\|\leq\frac{1}{(c_P^*)^2}\|q\|,\quad\forall q\in L^2(\Omega).$$

Proof. Using (ASM 1) and (4.4.4), it follows that

$$\begin{aligned}\|(L_w L^*)_\mathfrak{h}^{-1} q\| &\le \frac{1}{c_P^*}\|L^*(L_w L^*)_\mathfrak{h}^{-1} q\| = \frac{1}{c_P^*}\sup_{w^\mathfrak{h}\in\mathcal{Z}^\mathfrak{h}} \frac{|\langle L^*(L_w L^*)_\mathfrak{h}^{-1} q, L^* w^\mathfrak{h}\rangle|}{\|L^* w^\mathfrak{h}\|} \\ &= \frac{1}{c_P^*}\sup_{w^\mathfrak{h}\in\mathcal{Z}^\mathfrak{h}} \frac{|\langle q, w^\mathfrak{h}\rangle|}{\|L^* w^\mathfrak{h}\|} \le \frac{1}{c_P^*}\sup_{w\in\mathcal{D}(L^*)} \frac{|\langle q, w\rangle|}{\|L^* w\|} \\ &\le \frac{1}{(c_P^*)^2}\sup_{w\in\mathcal{D}(L^*)} \frac{|\langle q, w\rangle|}{\|w\|} \le \frac{1}{(c_P^*)^2}\|q\|.\end{aligned}$$

∎

Lemma 4.4.3 *The discrete operator $(L_w L^*)_\mathfrak{h}^{-1}\colon L^2(\Omega) \to \mathcal{Z}^\mathfrak{h}$ is self-adjoint and positive semidefinite with respect to the $L^2(\Omega)$ inner product.*

Proof. The proof is similar to that of Lemma 4.3.8. Note, however, that $(L_w L^*)_\mathfrak{h}^{-1}$ is not positive definite since it has a nontrivial (and infinite-dimensional) null space. This is to be expected since $(L_w L^*)_\mathfrak{h}^{-1}$ maps an infinite-dimensional space to a finite-dimensional one. Indeed, from (4.4.4), it follows that

$$\mathcal{N}((L_w L^*)_\mathfrak{h}^{-1}) = (\mathcal{Z}^\mathfrak{h})^\perp = \{\, q \in L^2(\Omega);\ \langle q, w^\mathfrak{h}\rangle = 0 \text{ for all } w^\mathfrak{h} \in \mathcal{Z}^\mathfrak{h} \,\} \tag{4.4.12}$$

is the null space of $(L_w L^*)_\mathfrak{h}^{-1}$. ∎

The corollary below is an immediate consequence of Lemma 4.4.3.

Corollary 4.4.4 *The matrix $\boldsymbol{A}$ in* (4.4.11) *is symmetric positive semidefinite, for all choices of $\mathcal{U}^h$ and $\mathcal{Z}^\mathfrak{h}$.*

Since $(L_w L^*)_\mathfrak{h}^{-1}$ is singular, the matrix $\boldsymbol{A}$ (or, equivalently, the matrix $\mathbb{A}$ in (4.4.9)) can be singular if the spaces $\mathcal{U}^h$ and $\mathcal{Z}^\mathfrak{h}$ are not selected carefully. The null space of $\boldsymbol{A}$ admits a simple but abstract characterization.

Lemma 4.4.5 *By identifying the vectors in $\mathbb{R}^N$ with the functions in $\mathcal{U}^h$, the null space of $\boldsymbol{A}$ is characterized as*

$$\mathcal{N}(\boldsymbol{A}) = \mathcal{U}^h \cap \left[L^*(\mathcal{Z}^\mathfrak{h})\right]^\perp.$$

Proof. Consider $v^h \in \mathcal{U}^h$ and its vector of coefficients $\boldsymbol{v} \in \mathbb{R}^N$. First, let $v^h \in [L^*(\mathcal{Z}^\mathfrak{h})]^\perp$. Then

$$(\boldsymbol{A}\boldsymbol{v})_i = \langle v^h, L^*(L_w L^*)_\mathfrak{h}^{-1} L\phi_i^h\rangle = 0 \quad (\text{since } L^*(L_w L^*)_\mathfrak{h}^{-1} L\phi_i^h \in L^*(\mathcal{Z}^\mathfrak{h})),$$

which implies that $\boldsymbol{v} \in \mathcal{N}(\boldsymbol{A})$ and, hence, $\mathcal{U}^h \cap [L^*(\mathcal{Z}^\mathfrak{h})]^\perp \subset \mathcal{N}(\boldsymbol{A})$. Conversely, let $\boldsymbol{v} \in \mathcal{N}(\boldsymbol{A})$ and denote $z^\mathfrak{h} = (L_w L^*)_\mathfrak{h}^{-1} L v^h \in \mathcal{Z}^\mathfrak{h}$. Then, by (4.4.4) and (ASM 1),

$$\begin{aligned} 0 = \boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} &= \langle L v^h, (L_w L^*)_\mathfrak{h}^{-1} L v^h \rangle = \langle L v^h, z^\mathfrak{h} \rangle \\ &= \langle L^* z^\mathfrak{h}, L^* z^\mathfrak{h} \rangle = \|L^* z^\mathfrak{h}\|^2 \geq (c_P^*)^2 \|z^\mathfrak{h}\|^2, \end{aligned}$$

implying that $L v^h \in \mathcal{N}((L_w L^*)_\mathfrak{h}^{-1})$. Combining this with (4.4.12) shows

$$0 = \langle L v^h, w^\mathfrak{h} \rangle = \langle v^h, L^* w^\mathfrak{h} \rangle, \quad \forall w^\mathfrak{h} \in \mathcal{Z}^\mathfrak{h}.$$

Thus, $v^h \in [L^*(\mathcal{Z}^\mathfrak{h})]^\perp$ and, hence, $\mathcal{N}(\boldsymbol{A}) \subset \mathcal{U}^h \cap [L^*(\mathcal{Z}^\mathfrak{h})]^\perp$. ∎

The following is a simple corollary of Lemma 4.4.5.

Corollary 4.4.6 *The null spaces of* $\boldsymbol{A}$ *and* $\boldsymbol{L}$ *coincide, that is,*

$$\mathcal{N}(\boldsymbol{A}) = \mathcal{N}(\boldsymbol{L}).$$

Proof. Let $v^h \in \mathcal{U}^h$ be a finite element function with a coefficient vector $\boldsymbol{v} \in \mathbb{R}^N$. Then

$$(\boldsymbol{L}\boldsymbol{v})_i = \langle v^h, L^* \psi_i^\mathfrak{h} \rangle, \quad \text{for } i = 1, \ldots, M,$$

implies that $\boldsymbol{v} \in \mathcal{N}(\boldsymbol{L})$ if and only if $v^h \in [L^*(\mathcal{Z}^\mathfrak{h})]^\perp$. Thus, owing to Lemma 4.4.5, $\mathcal{N}(\boldsymbol{A}) = \mathcal{N}(\boldsymbol{L})$.∎

Lemma 4.4.5 shows that $\boldsymbol{A}$ is always singular if $L^*(\mathcal{Z}^\mathfrak{h}) \subsetneq \mathcal{U}^h$. More generally, as shown below, $\boldsymbol{A}$ is guaranteed to be singular if $\mathcal{U}^h$ is of higher dimension than $\mathcal{Z}^\mathfrak{h}$.

Corollary 4.4.7 *If* $\dim(\mathcal{U}^h) > \dim(\mathcal{Z}^\mathfrak{h})$ *(i.e.,* $N > M$*), then* $\boldsymbol{A}$ *(as well as* $\mathbb{A}$*) is singular.*

Proof. If $N > M$, then there exists $\boldsymbol{v} \in \mathbb{R}^N \setminus \{\boldsymbol{0}\}$ such that $\boldsymbol{v} \in \mathcal{N}(\boldsymbol{L})$. Thus, by Corollary 4.4.6, $\boldsymbol{v} \in \mathcal{N}(\boldsymbol{A})$ and, hence, $\mathcal{N}(\boldsymbol{A}) \neq \{\boldsymbol{0}\}$. ∎

Theorem 4.3.5 and Lemma 4.3.4 show that $(L_w L^*)^{-1}$ together with L_w exactly recover the $L^2(\Omega)$ norm on the entire space, which is related to the mentioned equality $L^*(L_w L^*)^{-1} L_w = I$. However, replacing $(L_w L^*)^{-1}$ with $(L_w L^*)_\mathfrak{h}^{-1}$ cannot fully recover the $L^2(\Omega)$ norm. The following result shows that, instead, the $L^2(\Omega)$ norm is exactly recovered only on a subspace, $L^*(\mathcal{Z}^\mathfrak{h})$, and $L^*(L_w L^*)_\mathfrak{h}^{-1} L_w$ becomes a L^2-orthogonal projection. This is important for the coming considerations and results.

Lemma 4.4.8 (L^2-orthogonal projection) *Let $\Pi_*^{\mathfrak{h}}\colon L^2(\Omega) \to L^*(\mathcal{Z}^{\mathfrak{h}})$ be the L^2-orthogonal projection onto $L^*(\mathcal{Z}^{\mathfrak{h}})$. Then $\Pi_*^{\mathfrak{h}} = L^*(L_w L^*)_{\mathfrak{h}}^{-1} L_w$.*

Proof. Consider an arbitrary $q \in L^2(\Omega)$. Notice that $\Pi_*^{\mathfrak{h}} q \in L^*(\mathcal{Z}^{\mathfrak{h}})$ is characterized by the following weak form:

$$\langle \Pi_*^{\mathfrak{h}} q, L^* w^{\mathfrak{h}} \rangle = \langle q, L^* w^{\mathfrak{h}} \rangle, \quad \forall w^{\mathfrak{h}} \in \mathcal{Z}^{\mathfrak{h}}. \tag{4.4.13}$$

Let $\hat{z}^{\mathfrak{h}} = (L_w L^*)_{\mathfrak{h}}^{-1} L_w q \in \mathcal{Z}^{\mathfrak{h}}$. The definitions of $(L_w L^*)_{\mathfrak{h}}^{-1}$ and L_w imply

$$\langle L^* \hat{z}^{\mathfrak{h}}, L^* w^{\mathfrak{h}} \rangle = \langle q, L^* w^{\mathfrak{h}} \rangle, \quad \forall w^{\mathfrak{h}} \in \mathcal{Z}^{\mathfrak{h}}.$$

Thus, $\Pi_*^{\mathfrak{h}} q = L^* \hat{z}^{\mathfrak{h}}$ and, hence, $\Pi_*^{\mathfrak{h}} q = L^*(L_w L^*)_{\mathfrak{h}}^{-1} L_w q$. ∎

Corollary 4.4.9 *It holds that*

$$(\boldsymbol{A})_{ij} = \langle \Pi_*^{\mathfrak{h}} \phi_j^h, \phi_i^h \rangle, \quad (\boldsymbol{f})_i = \langle \Pi_*^{\mathfrak{h}} \hat{u}, \phi_i^h \rangle,$$

and the discrete $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation in (4.4.5) *and* (4.4.6) *can be equivalently expressed as*

$$u^h = \operatorname*{argmin}_{v^h \in \mathcal{U}^h} \|\Pi_*^{\mathfrak{h}}(v^h - \hat{u})\|^2, \tag{4.4.14}$$

$$\text{Find } u^h \in \mathcal{U}^h\colon \langle \Pi_*^{\mathfrak{h}} u^h, v^h \rangle = \langle \Pi_*^{\mathfrak{h}} \hat{u}, v^h \rangle, \quad \forall v^h \in \mathcal{U}^h. \tag{4.4.15}$$

Proof. This follows easily from Lemmas 4.3.1 and 4.4.8, using the obvious equality $f = L\hat{u}$. ∎

As shown in [43], discussed later in Section 4.5, and evident from Lemma 4.4.8, the result of $\Pi_*^{\mathfrak{h}} \hat{u}$ is computable through an application of $(L_w L^*)_{\mathfrak{h}}^{-1}$ (i.e., by solving (4.4.3)). This is a feature provided by the standard $\mathcal{L}\mathcal{L}^*$ method. In particular, the $\mathcal{L}\mathcal{L}^*$ method of [43] approximates the exact solution, $\hat{u}$, by $\Pi_*^{\mathfrak{h}} \hat{u}$. This justifies why formulations like (4.4.14) and (4.4.15) are computationally feasible. Corollary 4.4.9 is rather useful and interesting. It explains the effect on (4.1.3) when $(L_w L^*)^{-1}$ is replaced by $(L_w L^*)_{\mathfrak{h}}^{-1}$. Namely, the infeasible L^2-norm minimization of the error becomes a feasible, due to the standard $\mathcal{L}\mathcal{L}^*$ formulation, minimization of the projection of the error. This is, generally, a semi-norm minimization that only partially represents the $L^2(\Omega)$ norm, due to the necessary discretization of the operator $(L_w L^*)^{-1}$. Furthermore, Corollary 4.4.9 contributes to a considerable simplification of the proofs and considerations below.

Note that Corollary 4.4.7 establishes a necessary condition ($\dim(\mathcal{U}^h) \leq \dim(\mathcal{Z}^\mathfrak{h})$) for the invertibility of $\boldsymbol{A}$. A sufficient condition is more delicate. Lemma 4.4.5 suggests that the spaces $\mathcal{U}^h$ and $L^*(\mathcal{Z}^\mathfrak{h})$ should be "close" in a certain sense. This is made precise by the "inf-sup" condition below, which can be interpreted as a condition on the cosine of the abstract angle between the spaces $\mathcal{U}^h$ and $L^*(\mathcal{Z}^\mathfrak{h})$. Moreover, it implies a discrete (i.e., on $\mathcal{U}^h$) L^2-coercivity that is a stronger result than the nonsingularity of $\boldsymbol{A}$ and, in particular, provides information on the conditioning of $\boldsymbol{A}$; that is, even though the $L^2(\Omega)$ norm is only partially recovered, i.e., only on $L^*(\mathcal{Z}^\mathfrak{h})$, by the projection operator, the "proximity" of $\mathcal{U}^h$ and $L^*(\mathcal{Z}^\mathfrak{h})$ provided by the inf-sup condition implies a discrete "control" of the $L^2(\Omega)$ norm on $\mathcal{U}^h$.

Theorem 4.4.10 (inf-sup condition) *If there exists a constant $c_I > 0$ such that*

$$\inf_{v^h \in \mathcal{U}^h} \sup_{w^\mathfrak{h} \in \mathcal{Z}^\mathfrak{h}} \frac{|\langle v^h, L^* w^\mathfrak{h} \rangle|}{\|v^h\| \|L^* w^\mathfrak{h}\|} \geq c_I, \tag{4.4.16}$$

then the following spectral estimate holds:

$$c_I^2 \, \boldsymbol{v}^T \boldsymbol{M} \boldsymbol{v} \leq \boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v}, \quad \forall \boldsymbol{v} \in \mathbb{R}^N. \tag{4.4.17}$$

In particular, $\boldsymbol{A}$ and $\mathbb{A}$ are nonsingular.

Proof. Consider a finite element function $v^h \in \mathcal{U}^h$ and its corresponding coefficient vector $\boldsymbol{v} \in \mathbb{R}^N$. Then, owing to (4.4.16), (4.4.13), and Corollary 4.4.9, it follows that

$$\begin{aligned} c_I^2 \, \boldsymbol{v}^T \boldsymbol{M} \boldsymbol{v} = c_I^2 \|v^h\|^2 &\leq \left[\sup_{w^\mathfrak{h} \in \mathcal{Z}^\mathfrak{h}} \frac{|\langle v^h, L^* w^\mathfrak{h} \rangle|}{\|L^* w^\mathfrak{h}\|} \right]^2 = \left[\sup_{w^\mathfrak{h} \in \mathcal{Z}^\mathfrak{h}} \frac{|\langle \Pi_*^\mathfrak{h} v^h, L^* w^\mathfrak{h} \rangle|}{\|L^* w^\mathfrak{h}\|} \right]^2 \\ &= \|\Pi_*^\mathfrak{h} v^h\|^2 = \langle \Pi_*^\mathfrak{h} v^h, \Pi_*^\mathfrak{h} v^h \rangle = \langle \Pi_*^\mathfrak{h} v^h, v^h \rangle = \boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v}. \end{aligned}$$

□

Remark 4.4.11 Almost the same argument can be used to show that if $\lambda_{\min}$ is the smallest eigenvalue of the generalized eigenvalue problem $\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{M}\boldsymbol{v}$, then

$$\sqrt{\lambda_{\min}} = \inf_{v^h \in \mathcal{U}^h} \sup_{w^\mathfrak{h} \in \mathcal{Z}^\mathfrak{h}} \frac{|\langle v^h, L^* w^\mathfrak{h} \rangle|}{\|v^h\| \|L^* w^\mathfrak{h}\|}.$$

This implies that (4.4.17) holds if and only if the inf-sup condition (4.4.16) holds. Moreover, $\boldsymbol{A}$ and $\mathbb{A}$ are nonsingular if and only if

$$\inf_{v^h \in \mathcal{U}^h} \sup_{w^\mathfrak{h} \in \mathcal{Z}^\mathfrak{h}} \frac{|\langle v^h, L^* w^\mathfrak{h} \rangle|}{\|v^h\| \|L^* w^\mathfrak{h}\|} > 0.$$

◇

Remark 4.4.12 Obtaining inf-sup conditions of the form (4.4.16) for common finite element spaces is nontrivial. However, for the special choice of $\mathcal{U}^h = L^*(\mathcal{Z}^\mathfrak{h})$, it is easy to see that the inf-sup condition (4.4.16) holds with $c_I = 1$. In this case, $\boldsymbol{A} = \boldsymbol{M}$ and (4.4.14) reduces to $u^h = \Pi_*^\mathfrak{h} \hat{u}$, that is, the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method coincides with the standard $\mathcal{L}\mathcal{L}^*$ method when $\mathcal{U}^h = L^*(\mathcal{Z}^\mathfrak{h})$. See Section 4.5 for a further discussion on the relation of the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method to other $\mathcal{L}\mathcal{L}^*$-type methods. $\diamond$

The reverse spectral inequality can be easily shown without assuming the inf-sup condition (4.4.16).

Proposition 4.4.13 *The following spectral estimate holds:*

$$\boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} \leq \boldsymbol{v}^T \boldsymbol{M} \boldsymbol{v}, \quad \forall \boldsymbol{v} \in \mathbb{R}^N.$$

Proof. Let $v^h \in \mathcal{U}^h$ be a finite element function with a coefficient vector $\boldsymbol{v} \in \mathbb{R}^N$. From Corollary 4.4.9,

$$\boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} = \|\Pi_*^\mathfrak{h} v^h\|^2 \leq \|v^h\|^2 = \boldsymbol{v}^T \boldsymbol{M} \boldsymbol{v},$$

using the well known property of the orthogonal projection $\|\Pi_*^\mathfrak{h}\| = 1$. $\square$

The results above can be combined to obtain the spectral equivalence between $\boldsymbol{A}$ and $\boldsymbol{M}$:

$$c_I^2 \, \boldsymbol{v}^T \boldsymbol{M} \boldsymbol{v} \leq \boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} \leq \boldsymbol{v}^T \boldsymbol{M} \boldsymbol{v}, \quad \forall \boldsymbol{v} \in \mathbb{R}^N, \tag{4.4.18}$$

which can be equivalently expressed as

$$c_I^2 \|v^h\|^2 \leq \langle (L_w L^*)_\mathfrak{h}^{-1} L v^h, L v^h \rangle \leq \|v^h\|^2, \quad \forall v^h \in \mathcal{U}^h,$$

or

$$c_I^2 \|v^h\|^2 \leq \|\Pi_*^\mathfrak{h} v^h\|^2 \leq \|v^h\|^2, \quad \forall v^h \in \mathcal{U}^h. \tag{4.4.19}$$

As can be expected, (4.4.16) allows us to derive an important error estimate, which is the main result in this section. Indeed, while the operator $(L_w L^*)_\mathfrak{h}^{-1}$ recovers the $L^2(\Omega)$ norm only partially and it is clear from Corollary 4.4.9 that a global L^2-coercivity cannot hold, the discrete (on $\mathcal{U}^h$) control of the $L^2(\Omega)$ norm that is provided by (4.4.16) is sufficient for obtaining optimal convergence rates with respect to the $L^2(\Omega)$ norm. This is the content of the following abstract lemma, which

provides the analytical foundation for the error estimate below regarding the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method. It is a particular extension of Céa's lemma (see, e.g, [58, 61]) for formulations with symmetric bilinear forms. The result can be viewed as a specific adaptation of the general considerations in [86] and is easily shown by a standard argument from the finite element analysis of so-called "variational crimes" [58, Chapter 10]; see also [83, 41].

Lemma 4.4.14 *Consider a real Hilbert space $\mathcal{H}$ with norm $\|\cdot\|_{\mathcal{H}}$ and closed subspace $\mathcal{V} \subset \mathcal{H}$. Let $a(\cdot,\cdot)$ be a symmetric positive semidefinite bilinear form defined on $\mathcal{H} \times \mathcal{H}$ that is coercive on $\mathcal{V}$ and continuous on $\mathcal{H}$; that is,*

$$\alpha\|\chi\|_{\mathcal{H}} \leq a(\chi,\chi)^{1/2}, \quad a(w,w)^{1/2} \leq \beta\|w\|_{\mathcal{H}},$$

for all $\chi \in \mathcal{V}$, $w \in \mathcal{H}$, and some constants $\alpha, \beta > 0$. If $v \in \mathcal{V}$ and $\hat{v} \in \mathcal{H}$ are such that the "orthogonality" relation

$$a(v - \hat{v}, \chi) = 0, \quad \forall \chi \in \mathcal{V}, \tag{4.4.20}$$

is satisfied, then the following (quasi-)optimal error estimate holds:

$$\|v - \hat{v}\|_{\mathcal{H}} \leq \left(1 + \frac{\beta}{\alpha}\right) \inf_{\chi \in \mathcal{V}} \|\chi - \hat{v}\|_{\mathcal{H}}.$$

Proof. Observe that $a(\cdot,\cdot)$ is an inner product on $\mathcal{V}$ inducing a norm that is equivalent to $\|\cdot\|_{\mathcal{H}}$ on $\mathcal{V}$. It induces only a seminorm on $\mathcal{H}$, i.e., $a(\cdot,\cdot)$ is an "indefinite inner product" on $\mathcal{H}$. Note that the Cauchy-Schwarz inequality continues to hold in this case; see [72, Remark (1) on p. 176], i.e.,

$$|a(w,q)| \leq a(w,w)^{1/2} a(q,q)^{1/2}, \quad \forall w, q \in \mathcal{H}.$$

Let $\chi \in \mathcal{V}$. The proof proceeds as

$$\begin{aligned}
\|v - \hat{v}\|_{\mathcal{H}} &\leq \|\chi - \hat{v}\|_{\mathcal{H}} + \|v - \chi\|_{\mathcal{H}} \\
&\leq \|\chi - \hat{v}\|_{\mathcal{H}} + \frac{1}{\alpha} a(v - \chi, v - \chi)^{1/2} \qquad \text{(coercivity)} \\
&= \|\chi - \hat{v}\|_{\mathcal{H}} + \frac{1}{\alpha} \sup_{\xi \in \mathcal{V}} \frac{|a(v - \chi, \xi)|}{a(\xi,\xi)^{1/2}} \\
&= \|\chi - \hat{v}\|_{\mathcal{H}} + \frac{1}{\alpha} \sup_{\xi \in \mathcal{V}} \frac{|a(\hat{v} - \chi, \xi) + a(v - \hat{v}, \xi)|}{a(\xi,\xi)^{1/2}} \\
&= \|\chi - \hat{v}\|_{\mathcal{H}} + \frac{1}{\alpha} \sup_{\xi \in \mathcal{V}} \frac{|a(\chi - \hat{v}, \xi)|}{a(\xi,\xi)^{1/2}} \qquad \text{("orthogonality")}
\end{aligned}$$

$$\leq \|\chi - \hat{v}\|_{\mathcal{H}} + \frac{1}{\alpha} a(\chi - \hat{v}, \chi - \hat{v})^{1/2} \qquad \text{(Cauchy-Schwarz inequality)}$$
$$\leq \left(1 + \frac{\beta}{\alpha}\right) \|\chi - \hat{v}\|_{\mathcal{H}} \qquad \text{(continuity)}.$$

The final result follows by taking the infimum over $\chi \in \mathcal{V}$. ∎

The important L^2-norm error estimate for the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method can be derived now. The argument counts on the discrete L^2-coercivity of (4.4.14) given by (4.4.16), the natural L^2-continuity of (4.4.14), and Lemma 4.4.14 to show a (quasi-)optimal error estimate for the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method in the $L^2(\Omega)$ norm.

Theorem 4.4.15 (error estimate) *Assume that the inf-sup condition in* (4.4.16) *holds. If* $u^h \in \mathcal{U}^h$ *is the approximation obtained by the* $(\mathcal{L}\mathcal{L}^*)^{-1}$ *method (i.e., the solution to any of* (4.4.5), (4.4.6), (4.4.7), (4.4.14), *or* (4.4.15) *obtained, e.g., by solving any of the linear systems* (4.4.9) *or* (4.4.11)*) and* $\hat{u}$ *is the exact solution of* (4.1.2), *then*

$$\|u^h - \hat{u}\| \leq \left(1 + \frac{1}{c_I}\right) \inf_{v^h \in \mathcal{U}^h} \|v^h - \hat{u}\|.$$

Proof. Note that, in view of (4.4.6) and (4.4.15), the bilinear forms of interest here are $\langle \Pi_*^{\mathfrak{h}} \cdot, \cdot \rangle$ and $\langle (L_w L^*)_{\mathfrak{h}}^{-1} L \cdot, L \cdot \rangle$. Owing to Lemma 4.4.8, they coincide when they are both defined, i.e., on $\mathcal{D}(L)$. However, the bilinear form $\langle \Pi_*^{\mathfrak{h}} \cdot, \cdot \rangle$ is clearly well-defined on $L^2(\Omega) \times L^2(\Omega)$ and it is the one that is useful for this proof. Further information on extending the $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation is provided in Section 4.A.

First, since $\Pi_*^{\mathfrak{h}}$ is an orthogonal projection, it holds that

$$\langle \Pi_*^{\mathfrak{h}} q, q \rangle^{1/2} \leq \|q\|, \quad \forall q \in L^2(\Omega),$$

i.e., $\langle \Pi_*^{\mathfrak{h}} \cdot, \cdot \rangle$ is continuous on $L^2(\Omega)$. The left inequality in (4.4.19) can be written as

$$c_I \|v^h\| \leq \langle \Pi_*^{\mathfrak{h}} v^h, v^h \rangle^{1/2}, \quad \forall v^h \in \mathcal{U}^h,$$

which shows that $\langle \Pi_*^{\mathfrak{h}} \cdot, \cdot \rangle$ is coercive on the discrete space $\mathcal{U}^h$. Next, (4.4.15) implies the orthogonality property

$$\langle \Pi_*^{\mathfrak{h}} (u^h - \hat{u}), v^h \rangle = 0, \quad \forall v^h \in \mathcal{U}^h.$$

Thus, the error estimate follows from Lemma 4.4.14. ∎

Remark 4.4.16 Notice that the argument in Theorem 4.4.15 only needs the inf-sup condition (4.4.16) and $\mathcal{U}^h \subset L^2(\Omega)$. No other particular assumptions on $\mathcal{U}^h$ are necessary as long as a general $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation like (4.4.14) and (4.4.15) is used; see Section 4.A. $\diamond$

Remark 4.4.17 In general, all observations above also hold when c_I depends on the mesh parameter, h, instead of being a constant. In such a case, according to the estimate in Theorem 4.4.15, an h-dependence of c_I takes away from the convergence order that is implied by the approximation properties of $\mathcal{U}^h$. Also, this would affect the spectral equivalence estimate (4.4.18). $\diamond$

Remark 4.4.18 The estimate in Theorem 4.4.15 resembles results in the theory of mixed finite element methods [62, Section 2.3],[58, Chapter 12],[41, 87] and both cases utilize an argument from the analysis of the so-called "variational crimes"; see, e.g., [58, Chapter 10]. However, the approach here is different compared to more standard mixed finite element methods. Most notably, the solution of interest here is viewed as a minimizer of an unconstrained problem (4.4.14) and there is no apparent advantage for the theory of the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method to analyze a constrained minimization problem. $\diamond$

It is reasonable to expect that, for any fixed $\mathcal{U}^h$ (i.e., h is fixed), the corresponding approximation of $(L_w L^*)^{-1}$ by $(L_w L^*)_{\mathfrak{h}}^{-1}$ becomes better as $\mathfrak{h} \to 0$, in the sense that the representation of the $L^2(\Omega)$ norm on $\mathcal{U}^h$ improves. This is demonstrated below by showing, under mild assumptions on the approximation properties of $L^*(\mathcal{Z}^{\mathfrak{h}})$, that $c_I \to 1$ in (4.4.16) as $\mathfrak{h} \to 0$ and the $(\mathcal{L}\mathcal{L}^*)^{-1}$ solution approaches the L^2-orthogonal projection of $\hat{u}$ onto $\mathcal{U}^h$; that is, as $\mathfrak{h} \to 0$, the abstract angle between the spaces $\mathcal{U}^h$ and $L^*(\mathcal{Z}^{\mathfrak{h}})$ vanishes and the computational representation of the $L^2(\Omega)$ norm on $\mathcal{U}^h$ becomes closer to being exact, since it is exact on $L^*(\mathcal{Z}^{\mathfrak{h}})$. Furthermore, it is shown, under stronger assumptions on the approximation properties of $L^*(\mathcal{Z}^{\mathfrak{h}})$, that (4.4.16) can be maintained uniformly with c_I arbitrarily close to 1 by taking the ratio $h/\mathfrak{h}$ sufficiently large and keeping it fixed. This is a very basic study of how suitable approximation properties can provide inf-sup stability by appropriately selecting the configuration of spaces. These considerations need the following proposition. It shows that the inf-sup condition (4.4.16) can be equivalently expressed as a "sup-inf" condition. This can be interpreted as a condition on the sine of the abstract angle between the spaces $\mathcal{U}^h$ and $L^*(\mathcal{Z}^{\mathfrak{h}})$.

Proposition 4.4.19 *The inf-sup condition* (4.4.16) *is equivalent to*

$$\sup_{v^h \in \mathcal{U}^h} \frac{\|v^h - \Pi_*^{\mathfrak{h}} v^h\|}{\|v^h\|} \leq \sqrt{1 - c_I^2}. \tag{4.4.21}$$

Proof. Using (4.4.13) and the simple equality

$$\|q\|^2 = \|\Pi_*^{\mathfrak{h}} q\|^2 + \|q - \Pi_*^{\mathfrak{h}} q\|^2, \quad \forall q \in L^2(\Omega),$$

the equivalence follows from

$$\begin{aligned}
&\inf_{v^h \in \mathcal{U}^h} \left[\sup_{w^{\mathfrak{h}} \in \mathcal{Z}^{\mathfrak{h}}} \frac{|\langle v^h, L^* w^{\mathfrak{h}} \rangle|}{\|v^h\| \|L^* w^{\mathfrak{h}}\|} \right]^2 = \inf_{v^h \in \mathcal{U}^h} \frac{\|\Pi_*^{\mathfrak{h}} v^h\|^2}{\|v^h\|^2} = \inf_{v^h \in \mathcal{U}^h} \frac{\|v^h\|^2 - \|v^h - \Pi_*^{\mathfrak{h}} v^h\|^2}{\|v^h\|^2} \\
&= \inf_{v^h \in \mathcal{U}^h} \left[1 - \frac{\|v^h - \Pi_*^{\mathfrak{h}} v^h\|^2}{\|v^h\|^2} \right] = 1 - \sup_{v^h \in \mathcal{U}^h} \frac{\|v^h - \Pi_*^{\mathfrak{h}} v^h\|^2}{\|v^h\|^2}.
\end{aligned}$$

□

Generally, we do not have any explicit requirements on the approximation properties of $\mathcal{Z}^{\mathfrak{h}}$, as long as the inf-sup condition (4.4.16) holds. However, Proposition 4.4.19 suggests that the approximation properties of $L^*(\mathcal{Z}^{\mathfrak{h}})$ may not be completely neglected. In fact, if $L^*(\mathcal{Z}^{\mathfrak{h}})$ possesses approximation properties, the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method can always be made stable (in the sense that (4.4.16) can be enforced) as long as $\mathfrak{h}$ is taken sufficiently small for fixed $\mathcal{U}^h$. Indeed, let h (i.e., $\mathcal{U}^h$) be fixed and assume that $L^*(\mathcal{Z}^{\mathfrak{h}})$ satisfies an approximation bound like

$$\|\Pi_*^{\mathfrak{h}} v^h - v^h\| \leq C_{v^h} \mathfrak{h}^\gamma, \quad \forall v^h \in \mathcal{U}^h, \tag{4.4.22}$$

for $\gamma > 0$ and a constant $C_{v^h} > 0$ that generally depends on some Sobolev-space norm of v^h. Then, one can show, for any $v^h \in \mathcal{U}^h$, that

$$\|\Pi_*^{\mathfrak{h}} v^h - v^h\| \leq C_h \mathfrak{h}^\gamma \|v^h\|,$$

where the constant $C_h > 0$ can generally depend on the space $\mathcal{U}^h$. Therefore, (4.4.21) becomes arbitrary small, when $\mathfrak{h}$ is sufficiently close to zero. More precisely, $\sqrt{1 - c_I^2} = \mathcal{O}(\mathfrak{h}^\gamma)$ and $1 - c_I = \mathcal{O}(\mathfrak{h}^{2\gamma})$, using $c_I \in [0, 1]$ and the trivial $1 - c_I^2 = (1 - c_I)(1 + c_I)$; that is, the inf-sup condition can be enforced with a constant c_I arbitrary close to 1, as long as $\mathfrak{h}$ is taken sufficiently small, for fixed h.

Intuitively, this means that, as $\mathfrak{h} \to 0$, $(L_w L^*)_{\mathfrak{h}}^{-1}$ approaches $(L_w L^*)^{-1}$, the discrete $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation (4.4.5) approaches the L^2-norm minimization (4.4.1), and the $(\mathcal{L}\mathcal{L}^*)^{-1}$ approximation,

u^h, approaches the L^2-orthogonal projection of $\hat{u}$ onto $\mathcal{U}^h$. Indeed, consider the vector $\hat{\boldsymbol{f}} \in \mathbb{R}^N$ such that

$$(\hat{\boldsymbol{f}})_i = \langle \hat{u}, \phi_i^h \rangle.$$

Then the L^2-norm minimization (4.4.1) induces the linear system

$$\boldsymbol{M}\boldsymbol{u}_p = \hat{\boldsymbol{f}}, \tag{4.4.23}$$

where u_p^h denotes the L^2-orthogonal projection of $\hat{u}$ onto $\mathcal{U}^h$ and $\boldsymbol{u}_p \in \mathbb{R}^N$ is its respective coefficient vector. One can show that

$$|\boldsymbol{A} - \boldsymbol{M}| = \mathcal{O}(\mathfrak{h}^\gamma), \quad |\boldsymbol{f} - \hat{\boldsymbol{f}}| = \mathcal{O}(\mathfrak{h}^\gamma),$$

for any vector and its respective matrix norms $|\cdot|$. Thus, the $(\mathcal{L}\mathcal{L}^*)^{-1}$ linear system (4.4.11) approaches the L^2-orthogonal projection linear system (4.4.23), for fixed h, as $\mathfrak{h} \to 0$. A well known perturbation result from linear algebra (see, e.g., [88, Theorem 2.3.8]) implies that $\boldsymbol{u}$ also approaches $\boldsymbol{u}_p$. Namely,

$$|\boldsymbol{u} - \boldsymbol{u}_p| \leq \kappa(\boldsymbol{M})|\boldsymbol{u}_p| \left[\frac{|\boldsymbol{A} - \boldsymbol{M}|}{|\boldsymbol{M}|} + \frac{|\boldsymbol{f} - \hat{\boldsymbol{f}}|}{|\hat{\boldsymbol{f}}|} + \frac{|\boldsymbol{A} - \boldsymbol{M}|}{|\boldsymbol{M}|} \frac{|\boldsymbol{f} - \hat{\boldsymbol{f}}|}{|\hat{\boldsymbol{f}}|} \right],$$

where $\kappa(\boldsymbol{M})$ denotes the condition number of $\boldsymbol{M}$ with respect to the matrix norm $|\cdot|$. Thus,

$$|\boldsymbol{u} - \boldsymbol{u}_p| = \mathcal{O}(\mathfrak{h}^\gamma) \quad \text{and} \quad \|u^h - u_p^h\| = \mathcal{O}(\mathfrak{h}^\gamma).$$

Recall that, here, $u^h \in \mathcal{U}^h$ denotes the approximation obtained by the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method, $\boldsymbol{u} \in \mathbb{R}^N$ is its respective coefficient vector, and h is fixed as $\mathfrak{h}$ approaches zero.

The above argument does not exclude the possibility that, in general, the ratio $h/\mathfrak{h}$ may potentially need to grow to maintain the inf-sup condition (4.4.16) as $h \to 0$. However, assume that $\mathcal{U}^h$ is an H^1 (Lagrangian) finite element space on a quasi-uniform mesh, Ω is a polyhedral (or polygonal) domain, and

$$\|\Pi_*^{\mathfrak{h}} v^h - v^h\| \leq C\mathfrak{h}\|v^h\|_1, \quad \forall v^h \in \mathcal{U}^h,$$

where $\|\cdot\|_1$ is the norm on $H^1(\Omega)$ and the constant $C > 0$ does not depend on h, $\mathfrak{h}$, or v^h; that is, at least to a certain extent, the approximation properties of $L^*(\mathcal{Z}^{\mathfrak{h}})$ are on par with those of $\mathcal{U}^h$. Let $\mathfrak{h} = h/\tau$ for some constant $\tau \geq 1$. Then, using an inverse inequality [58, Theorem 4.5.11], we obtain

$$\frac{\|\Pi_*^{\mathfrak{h}} v^h - v^h\|}{\|v^h\|} \leq \frac{C}{\tau}.$$

Thus, if τ is taken sufficiently large (i.e., $\mathfrak{h}$ is sufficiently small relative to h), then (4.4.21) (and the respective (4.4.16)) can be enforced with c_I arbitrary close to 1 and the inf-sup condition is maintained as $h \to 0$ by keeping the ratio $h/\mathfrak{h} = \tau$ fixed. Similar to above, observe that $\sqrt{1-c_I^2} = \mathcal{O}(\tau^{-1})$ and $1 - c_I = \mathcal{O}(\tau^{-2})$.

In the discrete $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation (4.4.6), $(L_w L^*)^{-1}$ is replaced by $(L_w L^*)_{\mathfrak{h}}^{-1}$ (i.e., $\mathcal{D}(L^*)$ is replaced by $\mathcal{Z}^{\mathfrak{h}}$) leading to the loss of the L^2-orthogonal projection property of (4.4.2). However, Theorem 4.4.15 shows that when $\mathcal{Z}^{\mathfrak{h}}$ is appropriately chosen in relation to $\mathcal{U}^h$, so that the inf-sup condition (4.4.16) would hold, then the approximation $(L_w L^*)_{\mathfrak{h}}^{-1}$ of $(L_w L^*)^{-1}$ is of sufficient quality to guarantee (quasi-)optimal L^2-norm approximations on $\mathcal{U}^h$ of the exact solution. The above considerations show that under mild assumptions the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method can be made stable (i.e., the inf-sup condition (4.4.16) can be enforced) and under stronger assumptions this can be achieved with fixed ratio $h/\mathfrak{h}$. Deriving inf-sup conditions of type (4.4.16) for spaces $\mathcal{U}^h$, $\mathcal{Z}^{\mathfrak{h}}$ and operators L, L^* of interest is currently an open question, especially for $h/\mathfrak{h}$ being fixed and small so that the method is computationally efficient. It is not clear if this can be achieved with common finite element spaces serving as $\mathcal{Z}^{\mathfrak{h}}$ or special (ad-hoc) spaces are needed to guarantee the inf-sup stability (4.4.16). In Section 4.8, we investigate numerically the behavior of the method on model problems and using common finite element spaces as $\mathcal{Z}^{\mathfrak{h}}$ in which case the inf-sup condition (4.4.16) may not hold.

We close this section by stating the equivalence of the inf-sup condition (4.4.16) to the existence of a stable approximation operator $\Pi_z^{\mathfrak{h}} \colon \mathcal{D}(L^*) \to \mathcal{Z}^{\mathfrak{h}}$ that preserves, in a sense, a certain discrete (with respect to $\mathcal{U}^h$) version of the operator L^*.

Proposition 4.4.20 *The inf-sup condition* (4.4.16) *holds if and only if there exists a linear operator* $\Pi_z^{\mathfrak{h}} \colon \mathcal{D}(L^*) \to \mathcal{Z}^{\mathfrak{h}}$ *such that*

$$\|L^* \Pi_z^{\mathfrak{h}} w\| \leq \frac{1}{c_I} \|L^* w\|, \quad \langle v^h, L^* \Pi_z^{\mathfrak{h}} w \rangle = \langle v^h, L^* w \rangle, \qquad \forall w \in \mathcal{D}(L^*),\ \forall v^h \in \mathcal{U}^h.$$

This is a special case of the rather general considerations in [89], but it is not difficult to prove it directly for the setting here.

4.5 Other $\mathcal{L}\mathcal{L}^*$-type methods

This section is devoted to more standard $\mathcal{L}\mathcal{L}^*$-type approaches. All methods here and the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method of the previous section are related as they are founded upon the original $\mathcal{L}\mathcal{L}^*$ method introduced in [43]. Here, we consider all formulations on common terms to aid the comparison between them. They are further compared numerically in Section 4.8. Again, for simplicity of exposition, $\mathcal{U}^h$ is a subset of $\mathcal{D}(L)$ in this section and the extensions of the formulations to more general finite element spaces is discussed in Section 4.A.

First, consider the (standard) $\mathcal{L}\mathcal{L}^*$ formulation of [43]:

$$z_*^\mathfrak{h} = \underset{w^\mathfrak{h}\in\mathcal{Z}^\mathfrak{h}}{\operatorname{argmin}} \|L^* w^\mathfrak{h} - \hat{u}\|^2. \tag{4.5.1}$$

The resulting $\mathcal{L}\mathcal{L}^*$ approximation is $u_*^\mathfrak{h} = L^* z_*^\mathfrak{h} \in L^*(\mathcal{Z}^\mathfrak{h})$. The weak form corresponding to (4.5.1) is

$$\text{Find } z^\mathfrak{h} \in \mathcal{Z}^\mathfrak{h} \colon \langle L^* z^\mathfrak{h}, L^* w^\mathfrak{h}\rangle = \langle f, w^\mathfrak{h}\rangle, \quad \forall w^\mathfrak{h} \in \mathcal{Z}^\mathfrak{h};$$

that is, the weak form is precisely (4.4.4) with $q = f$, i.e., $z_*^\mathfrak{h} = (L_w L^*)_\mathfrak{h}^{-1} f$ and, clearly from (4.5.1) or Lemma 4.4.8, $u_*^\mathfrak{h} = \Pi_*^\mathfrak{h} \hat{u}$. In other words, the method provides the best L^2-norm approximation of $\hat{u}$ in $L^*(\mathcal{Z}^\mathfrak{h})$. The quality of the obtained solution depends on the approximation properties of $L^*(\mathcal{Z}^\mathfrak{h})$. Using the notation introduced in (4.4.8), the following is the linear system of equations that arises from (4.5.1):

$$\boldsymbol{H}\boldsymbol{z}_* = \bar{\boldsymbol{f}}. \tag{4.5.2}$$

To obtain an approximation on $\mathcal{U}^h$, the $\mathcal{L}\mathcal{L}^*$ solution, $u_*^\mathfrak{h}$, can be further projected onto $\mathcal{U}^h$:

$$u_{ts}^h = \underset{v^h\in\mathcal{U}^h}{\operatorname{argmin}} \|v^h - L^* z_*^\mathfrak{h}\|^2. \tag{4.5.3}$$

Computationally, this requires solving a linear system with the mass matrix $\boldsymbol{M}$. The minimizations (4.5.1) and (4.5.3) constitute the *"two-stage method"*. Alternatively, the minimizations in (4.5.1) and (4.5.3) can be combined resulting in the *"single-stage method"*:

$$(u_{ss}^h, z_\bullet^\mathfrak{h}) = \underset{(v^h,w^\mathfrak{h})\in\mathcal{U}^h\times\mathcal{Z}^\mathfrak{h}}{\operatorname{argmin}} \left[\omega\|L^* w^\mathfrak{h} - \hat{u}\|^2 + \|v^h - L^* w^\mathfrak{h}\|^2\right], \tag{4.5.4}$$

for a given constant weight $\omega > 0$. Note that $L^* z_\bullet^\mathfrak{h} \in L^*(\mathcal{Z}^\mathfrak{h})$ also approximates the exact solution, $\hat{u}$, but it is generally inferior, as an L^2-norm approximation, to the standard $\mathcal{L}\mathcal{L}^*$ solution, $u_*^\mathfrak{h}$,

since $u_*^{\mathfrak{h}}$ is the best approximation of $\hat{u}$ in $L^*(\mathcal{Z}^{\mathfrak{h}})$ in the $L^2(\Omega)$ norm. Also, the purpose here is to obtain approximations in $\mathcal{U}^h$. Therefore, we concentrate on u_{ss}^h. Formulation (4.5.4) resembles the "hybrid method" introduced in [44] with the difference that the first-order system least-squares (FOSLS) term is not present in (4.5.4).

As in Subsection 4.4.2, (4.5.3) and (4.5.4) induce the following block linear systems (cf., (4.4.9)), respectively:

$$\begin{bmatrix} \boldsymbol{H} & \\ -\boldsymbol{L}^T & \boldsymbol{M} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_* \\ \boldsymbol{u}_{ts} \end{bmatrix} = \begin{bmatrix} \bar{\boldsymbol{f}} \\ \boldsymbol{0} \end{bmatrix},$$

$$\mathbb{A}_{ss} \begin{bmatrix} \boldsymbol{z}_\bullet \\ \boldsymbol{u}_{ss} \end{bmatrix} = \begin{bmatrix} (\omega+1)\boldsymbol{H} & -\boldsymbol{L} \\ -\boldsymbol{L}^T & \boldsymbol{M} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_\bullet \\ \boldsymbol{u}_{ss} \end{bmatrix} = \begin{bmatrix} \omega\bar{\boldsymbol{f}} \\ \boldsymbol{0} \end{bmatrix}. \tag{4.5.5}$$

Similar to (4.4.10), $\boldsymbol{z}_*$ and $\boldsymbol{z}_\bullet$ can be eliminated in the above systems resulting in problems involving only $\boldsymbol{u}_{ts}$ and $\boldsymbol{u}_{ss}$. Namely, using the notation $\boldsymbol{f} = \boldsymbol{L}^T\boldsymbol{H}^{-1}\bar{\boldsymbol{f}} \in \mathbb{R}^N$ introduced above (4.4.11), the algebraic systems for the respective Schur complements corresponding to the methods in this chapter are the following:

$$\boldsymbol{A}\boldsymbol{u}_{inv} = \boldsymbol{f} \qquad ((\mathcal{L}\mathcal{L}^*)^{-1} \text{ method}), \tag{4.5.6}$$

$$\boldsymbol{M}\boldsymbol{u}_{ts} = \boldsymbol{f} \qquad \text{(two-stage method)}, \tag{4.5.7}$$

$$[(\omega+1)\boldsymbol{M} - \boldsymbol{A}]\boldsymbol{u}_{ss} = \omega\boldsymbol{f} \qquad \text{(single-stage method)}. \tag{4.5.8}$$

Corollary 4.4.9 demonstrates that the algebraic system (4.5.6) precisely corresponds to the least-squares problem (4.4.14) that minimizes the $L^*(\mathcal{Z}^{\mathfrak{h}})$ component of the error. It is possible to obtain similar minimization problems that characterize the solutions to (4.5.7) and (4.5.8) in relation to the exact solution, $\hat{u}$, aiding the comparison between the methods. Namely, the algebraic systems (4.5.6), (4.5.7), and (4.5.8) are associated with the following respective least-squares problems:

$$u_{inv}^h = \underset{v^h \in \mathcal{U}^h}{\operatorname{argmin}} \|\Pi_*^{\mathfrak{h}}(v^h - \hat{u})\|^2, \tag{4.5.9}$$

$$u_{ts}^h = \underset{v^h \in \mathcal{U}^h}{\operatorname{argmin}} \left[\|\Pi_*^{\mathfrak{h}}(v^h - \hat{u})\|^2 + \|v^h - \Pi_*^{\mathfrak{h}} v^h\|^2 \right], \tag{4.5.10}$$

$$u_{ss}^h = \underset{v^h \in \mathcal{U}^h}{\operatorname{argmin}} \left[\|\Pi_*^{\mathfrak{h}}(v^h - \hat{u})\|^2 + \frac{\omega+1}{\omega} \|v^h - \Pi_*^{\mathfrak{h}} v^h\|^2 \right]. \tag{4.5.11}$$

In comparison, the L^2-orthogonal projection is defined as

$$u_p^h = \underset{v^h \in \mathcal{U}^h}{\operatorname{argmin}} \|v^h - \hat{u}\|^2 = \underset{v^h \in \mathcal{U}^h}{\operatorname{argmin}} \left[\|\Pi_*^{\mathfrak{h}}(v^h - \hat{u})\|^2 + \|(I - \Pi_*^{\mathfrak{h}})(v^h - \hat{u})\|^2 \right],$$

but this formulation is generally infeasible because of the second term. Indeed, $(I-\Pi_*^{\mathfrak{h}})\hat{u}$ is generally not computationally obtainable, whereas $\Pi_*^{\mathfrak{h}}\hat{u}$ is available due to the $\mathcal{L}\mathcal{L}^*$ method (4.5.1). Thus, in a sense, all three methods (4.5.9), (4.5.10), and (4.5.11) trade the L^2-orthogonal projection for computational feasibility. The difference is that (4.5.9) drops the term for which there is no available information, whereas (4.5.10) and (4.5.11) replace it with "regularization" terms for the size of $(I-\Pi_*^{\mathfrak{h}})v^h$, i.e., they only drop $(I-\Pi_*^{\mathfrak{h}})\hat{u}$. Note that the second terms in (4.5.10) and (4.5.11) cannot be expected to contribute to the quality of approximation of the exact solution, since they do not contain any additional information on $\hat{u}$. However, those terms "stabilize" the methods and the resulting linear algebra systems (4.5.7) and (4.5.8) are always symmetric positive definite (hence, nonsingular).

Remark 4.5.1 It is not difficult to derive (4.5.10) from (4.5.3) by observing that

$$\|v^h-\Pi_*^{\mathfrak{h}}\hat{u}\|^2=\|\Pi_*^{\mathfrak{h}}(v^h-\Pi_*^{\mathfrak{h}}\hat{u})\|^2+\|(I-\Pi_*^{\mathfrak{h}})(v^h-\Pi_*^{\mathfrak{h}}\hat{u})\|^2=\|\Pi_*^{\mathfrak{h}}(v^h-\hat{u})\|^2+\|(I-\Pi_*^{\mathfrak{h}})v^h\|^2,$$

for any $v^h\in\mathcal{U}^h$. It is easy to see that the weak form corresponding to (4.5.10) induces the linear system (4.5.7).

Similarly, (4.5.11) can be derived from (4.5.4), but it is more challenging. Nevertheless, it is easy to verify that (4.5.8) can be associated with the weak formulation

$$\text{Find } u^h\in\mathcal{U}^h\colon \frac{\omega+1}{\omega}\langle(I-\Pi_*^{\mathfrak{h}})u^h,v^h\rangle+\langle\Pi_*^{\mathfrak{h}}u^h,v^h\rangle=\langle\Pi_*^{\mathfrak{h}}\hat{u},v^h\rangle,\quad\forall v^h\in\mathcal{U}^h,$$

which, clearly, corresponds to the minimization (4.5.11); see the proof of Theorem 4.5.3 below. $\diamond$

Remark 4.5.2 Observe that the formulation (4.5.11) approaches the one in (4.5.10) as $\omega\to\infty$. In fact, the two-stage method can be viewed as an extreme case of the single-stage method, when $\omega=\infty$. $\diamond$

Next, error estimates for the single- and two-stage methods are derived.

Theorem 4.5.3 (error estimate) *The following error estimate holds:*

$$\|u_\diamond^h-\hat{u}\|\le s\inf_{v^h\in\mathcal{U}^h}\|v^h-\hat{u}\|+(s+1)\inf_{w^{\mathfrak{h}}\in\mathcal{Z}^{\mathfrak{h}}}\|L^*w^{\mathfrak{h}}-\hat{u}\|,$$

where $u_\diamond^h=\{\,u_{ss}^h$ *or* $u_{ts}^h\,\}$ *and* $s=\{\,(\omega+1)/\omega$ *or* $1\,\}\ge 1$ *for the single- and two-stage methods, respectively.*

Proof. The weak form associated with (4.5.11) (or (4.5.10)), i.e., the one that induces (4.5.8) (or (4.5.7)), can be expressed as

$$\text{Find } u^h \in \mathcal{U}^h \colon a_s(u^h, v^h) = \langle \Pi_*^{\mathfrak{h}} \hat{u}, v^h \rangle, \quad \forall v^h \in \mathcal{U}^h,$$

where the symmetric bilinear form $a_s \colon L^2(\Omega) \times L^2(\Omega) \to \mathbb{R}$ is defined as

$$a_s(p, q) = s\langle (I - \Pi_*^{\mathfrak{h}})p, q \rangle + \langle \Pi_*^{\mathfrak{h}} p, q \rangle, \quad \forall p, q \in L^2(\Omega).$$

Clearly, a_s is L^2-equivalent, i.e.,

$$\|q\|^2 \leq a_s(q, q) \leq s\|q\|^2, \quad \forall q \in L^2(\Omega).$$

Also, the following "orthogonality" property holds:

$$a_s(u_\diamond^h - \Pi_*^{\mathfrak{h}} \hat{u}, v^h) = 0, \quad \forall v^h \in \mathcal{U}^h.$$

Thus, Céa's lemma implies that

$$\|u_\diamond^h - \Pi_*^{\mathfrak{h}} \hat{u}\| \leq s\|v^h - \Pi_*^{\mathfrak{h}} \hat{u}\|, \quad \forall v^h \in \mathcal{U}^h.$$

Using this, for any $v^h \in \mathcal{U}^h$, we obtain

$$\begin{aligned} \|u_\diamond^h - \hat{u}\| &\leq \|u_\diamond^h - \Pi_*^{\mathfrak{h}} \hat{u}\| + \|\Pi_*^{\mathfrak{h}} \hat{u} - \hat{u}\| \leq s\|v^h - \Pi_*^{\mathfrak{h}} \hat{u}\| + \|\Pi_*^{\mathfrak{h}} \hat{u} - \hat{u}\| \\ &\leq s\|v^h - \hat{u}\| + (s+1)\|\Pi_*^{\mathfrak{h}} \hat{u} - \hat{u}\|. \end{aligned}$$

□

Remark 4.5.4 Notice that, in general, the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method is the only one of the three (4.5.9), (4.5.10), and (4.5.11) that possesses an "orthogonality" property like (4.4.20) with respect to the exact solution, $\hat{u}$, whereas (4.5.10) and (4.5.11) satisfy such a property for the projection $\Pi_*^{\mathfrak{h}} \hat{u}$. Also, the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method is the only one of the three that does not, generally, have a uniform L^2-coercivity and depends on the inf-sup condition (4.4.16) to satisfy a discrete (i.e., on $\mathcal{U}^h$) L^2-coercivity. ◇

Theorem 4.5.3 suggests that the quality of the solutions in $\mathcal{U}^h$ obtained by the single- and two-stage methods can depend not only on the approximation properties of $\mathcal{U}^h$, but also on the approximation properties of $L^*(\mathcal{Z}^{\mathfrak{h}})$. In view of (4.5.3) and (4.5.4), this can be expected since

$L^*(\mathcal{Z}^\mathfrak{h})$ is the only "connection" between the resulting approximations on $\mathcal{U}^h$ and the exact solution, $\hat{u}$. According to the estimate in Theorem 4.5.3, optimal rates of convergence are obtainable when the approximation properties of $L^*(\mathcal{Z}^\mathfrak{h})$ are not worse than those of $\mathcal{U}^h$ and an optimal setting would be if they are on par. In particular, when $\mathcal{U}^h \subset L^*(\mathcal{Z}^\mathfrak{h})$, then all three methods (4.5.9), (4.5.10), and (4.5.11) coincide (in fact, $\boldsymbol{A} = \boldsymbol{M}$, the inf-sup condition (4.4.16) holds with $c_I = 1$, and the systems (4.5.6), (4.5.7), (4.5.8), and (4.4.23) coincide), and they all provide the L^2-orthogonal projection of $\hat{u}$ onto $\mathcal{U}^h$, i.e., $u^h_{inv} = u^h_{ts} = u^h_{ss} = u^h_p$. In general, for fixed h and assuming that the property (4.4.22) holds, a similar argument to the one following Proposition 4.4.19 shows that the linear systems (4.5.6), (4.5.7), and (4.5.8) approach (4.4.23) as $\mathfrak{h} \to 0$ and

$$\|u^h_\diamond - u^h_p\| = \mathcal{O}(\mathfrak{h}^\gamma),$$

where $u^h_\diamond = \{\, u^h_{inv}, u^h_{ts}, \text{ or } u^h_{ss} \,\}$. In the case $u^h_\diamond = u^h_{ss}$, the constant in the $\mathcal{O}$-notation depends on $1/\omega$, that is, for fixed h, the three approaches converge to the same method as $\mathfrak{h} \to 0$, which is the L^2-orthogonal projection (4.4.1).

4.6 Implementation and preconditioning

In this section, implementation and preconditioning of the linear systems introduced in the previous sections is discussed. In particular, we consider Krylov methods with block preconditioners.

The $(\mathcal{L}\mathcal{L}^*)^{-1}$ method can be implemented in a way similar to the H^{-1} method in [45]. Namely, in view of (4.4.18), the conjugate gradient method (CG) is potentially (depending on the inf-sup condition (4.4.16)) an adequate choice for solving the system (4.4.11). However, obtaining the residual and a matrix-vector product with $\boldsymbol{A}$ requires computing the effect of $(L_w L^*)^{-1}_\mathfrak{h}$, i.e., numerically inverting the matrix $\boldsymbol{H}$, which needs to be performed on each CG iteration. As in [45], $\boldsymbol{H}^{-1}$ can be replaced by an application of a symmetric positive definite preconditioner $\boldsymbol{B}^{-1}$. This is equivalent to replacing $(L_w L^*)^{-1}_\mathfrak{h} \colon \mathcal{D}'(L^*) \to \mathcal{Z}^\mathfrak{h}$ with a respective operator $B^{-1}_\mathfrak{h} \colon \mathcal{D}'(L^*) \to \mathcal{Z}^\mathfrak{h}$. It results in (4.4.5) being replaced by the modified (by a preconditioner) minimization

$$\tilde{u}^h_{inv} = \operatorname*{argmin}_{v^h \in \mathcal{U}^h} \langle B^{-1}_\mathfrak{h}(Lv^h - f), Lv^h - f\rangle. \tag{4.6.1}$$

More precisely, for $\ell \in \mathcal{D}'(L^*)$, $\tilde{z}^\mathfrak{h} = B^{-1}_\mathfrak{h}\ell$ with a coefficient vector $\tilde{\boldsymbol{z}} \in \mathbb{R}^M$ is defined as

$$\tilde{\boldsymbol{z}} = \boldsymbol{B}^{-1}\boldsymbol{\ell},$$

where $\boldsymbol{\ell} \in \mathbb{R}^M$ and $(\boldsymbol{\ell})_i = \ell(\psi_i^{\mathfrak{h}})$. In comparison, for $z^{\mathfrak{h}} = (L_w L^*)_{\mathfrak{h}}^{-1}\ell$ with a coefficient vector $\boldsymbol{z} \in \mathbb{R}^M$, it holds that

$$\boldsymbol{z} = \boldsymbol{H}^{-1}\boldsymbol{\ell}.$$

The weak form, associated with the minimization problem (4.6.1), induces the linear system

$$\tilde{\boldsymbol{A}}\tilde{\boldsymbol{u}}_{inv} = \tilde{\boldsymbol{f}},$$

where $\tilde{\boldsymbol{A}} \in \mathbb{R}^{N\times N}$, $\tilde{\boldsymbol{f}} \in \mathbb{R}^N$, and

$$(\tilde{\boldsymbol{A}})_{ij} = \langle B_{\mathfrak{h}}^{-1} L\phi_j^h, L\phi_i^h\rangle, \quad (\tilde{\boldsymbol{f}})_i = \langle B_{\mathfrak{h}}^{-1} f, L\phi_i^h\rangle.$$

In practice, the matrices $\boldsymbol{A}$ and $\tilde{\boldsymbol{A}}$ can be explicitly assembled only for small values of M and N, since they are generally dense. Even if $\boldsymbol{H}^{-1}$ and $\boldsymbol{B}^{-1}$ (i.e., $(L_w L^*)_{\mathfrak{h}}^{-1}$ and $B_{\mathfrak{h}}^{-1}$) can be applied in optimal time, i.e., in $\mathcal{O}(M)$ number of operations, the cost of assembling and storing $\boldsymbol{A}$ or $\tilde{\boldsymbol{A}}$ is prohibitive for large values of M and N. However, Krylov methods can clearly be used in a matrix-free way. Matrix-vector products with $\boldsymbol{A}$ and $\tilde{\boldsymbol{A}}$ can be computed without explicitly assembling the matrices. Indeed, let $\boldsymbol{v} \in \mathbb{R}^N$. Then, by (4.4.10),

$$\boldsymbol{A}\boldsymbol{v} = \boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\boldsymbol{v}.$$

Thus, computing the matrix-vector product $\boldsymbol{A}\boldsymbol{v}$ requires a single application of $\boldsymbol{H}^{-1}$ (i.e., of $(L_w L^*)_{\mathfrak{h}}^{-1}$) and matrix-vector products with $\boldsymbol{L}$ and $\boldsymbol{L}^T$, which can be efficiently assembled. Similarly, the right-hand side, $\boldsymbol{f}$, and the residual, $\boldsymbol{f} - \boldsymbol{A}\boldsymbol{v}$, can be computed. The same considerations apply to the computation of $\tilde{\boldsymbol{A}}\boldsymbol{v}$, $\tilde{\boldsymbol{f}}$, and $\tilde{\boldsymbol{f}} - \tilde{\boldsymbol{A}}\boldsymbol{v}$ by replacing $\boldsymbol{H}^{-1}$ with $\boldsymbol{B}^{-1}$ (i.e., replacing $(L_w L^*)_{\mathfrak{h}}^{-1}$ with $B_{\mathfrak{h}}^{-1}$).

If $\boldsymbol{B}$ is spectrally equivalent to $\boldsymbol{H}$, then, for some constants $c_s, C_s > 0$,

$$c_s\, \boldsymbol{z}^T\boldsymbol{H}^{-1}\boldsymbol{z} \le \boldsymbol{z}^T\boldsymbol{B}^{-1}\boldsymbol{z} \le C_s\, \boldsymbol{z}^T\boldsymbol{H}^{-1}\boldsymbol{z}, \quad \forall \boldsymbol{z} \in \mathbb{R}^M.$$

Thus,

$$c_s\, \boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} \le \boldsymbol{v}^T\tilde{\boldsymbol{A}}\boldsymbol{v} \le C_s\, \boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v}, \quad \forall \boldsymbol{v} \in \mathbb{R}^N,$$

or, equivalently,

$$c_s\|\Pi_*^{\mathfrak{h}} v^h\|^2 \le \langle B_{\mathfrak{h}}^{-1} Lv^h, Lv^h\rangle \le C_s\|\Pi_*^{\mathfrak{h}} v^h\|^2, \quad \forall v^h \in \mathcal{U}^h.$$

Moreover, consider $q \in L^2(\Omega)$ and the vector $\boldsymbol{l} \in \mathbb{R}^M$ such that $(\boldsymbol{l})_i = [L_w q](\psi_i^{\mathfrak{h}})$. Then, by the definition of L_w,

$$\begin{aligned}\langle L^* B_{\mathfrak{h}}^{-1} L_w q, q\rangle &= \boldsymbol{l}^T \boldsymbol{B}^{-1} \boldsymbol{l} \leq C_s\, \boldsymbol{l}^T \boldsymbol{H}^{-1} \boldsymbol{l} = C_s \langle L^* (L_w L^*)_{\mathfrak{h}}^{-1} L_w q, q\rangle \\ &= C_s \|\Pi_*^{\mathfrak{h}} q\|^2 \leq C_s \|q\|^2.\end{aligned}$$

Therefore, observing that the formulation (4.6.1) satisfies an "orthogonality" property like (4.4.20) and assuming the inf-sup condition (4.4.16), similar to Theorem 4.4.15, the following error estimate holds:

$$\|\tilde{u}_{inv}^h - \hat{u}\| \leq \left(1 + \frac{\sqrt{C_s}}{\sqrt{c_s} c_I}\right) \inf_{v^h \in \mathcal{U}^h} \|v^h - \hat{u}\|;$$

that is, the modified minimization (4.6.1) maintains the properties of the original $(\mathcal{L}\mathcal{L}^*)^{-1}$ method (4.4.5) when $\boldsymbol{B}$ is spectrally equivalent to $\boldsymbol{H}$.

Obtaining spectrally equivalent preconditioners of $\boldsymbol{H}$ for hyperbolic operators, L^*, is quite challenging. In the above approach, the quality of the preconditioner can affect not only the solver, but also the minimization formulation and the quality of the approximation $\tilde{u}_{inv}^h$. On the other hand, requiring a preconditioner of $\boldsymbol{H}$ is reasonable, since the efficient iterative solution of the standard $\mathcal{L}\mathcal{L}^*$ system (4.5.2) also needs such a preconditioner. Therefore, we propose a different path here, using the same tools (the preconditioner $\boldsymbol{B}$ and Krylov solvers) and solving the block system (4.4.9) directly, thus maintaining the original $(\mathcal{L}\mathcal{L}^*)^{-1}$ principle (4.4.5).

Based on well known block factorizations of 2×2 block matrices, we obtain the following symmetric preconditioner of the matrix $\mathbb{A}$ in (4.4.9) (see also [90]):

$$\begin{aligned}\mathbb{B}_{inv}^{-1} &= \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{B}^{-1}\boldsymbol{L} \\ & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}^{-1} & \\ & \boldsymbol{Z}_{inv}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \\ -\boldsymbol{L}^T\boldsymbol{B}^{-1} & \boldsymbol{I} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{B}^{-1}\boldsymbol{L} \\ & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}^{-1} & \\ -\boldsymbol{Z}_{inv}^{-1}\boldsymbol{L}^T\boldsymbol{B}^{-1} & \boldsymbol{Z}_{inv}^{-1} \end{bmatrix},\end{aligned} \tag{4.6.2}$$

where $\boldsymbol{Z}_{inv}$ is a symmetric preconditioner of the respective Schur complement of $\mathbb{A}$, $\boldsymbol{S}_{inv} = -\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L} = -\boldsymbol{A}$. Notice that the Schur complement of $\mathbb{A}$, $\boldsymbol{S}_{inv}$, that we are interested in, is symmetric negative semidefinite, by Corollary 4.4.4. Hence, $\mathbb{A}$ is generally a symmetric indefinite matrix. Also, the block preconditioner $\mathbb{B}_{inv}$ is positive definite when $\boldsymbol{Z}_{inv}$ is positive definite, and indefinite when $\boldsymbol{Z}_{inv}$ is negative definite. By (4.4.18), $\boldsymbol{S}_{inv}$ is spectrally equivalent to $-\boldsymbol{M}$ and this

equivalence depends on the inf-sup condition (4.4.16). Observe that applying $\mathbb{B}_{inv}^{-1}$ requires two applications of $\boldsymbol{B}^{-1}$ and two of $\boldsymbol{Z}_{inv}^{-1}$.

Similarly, the following symmetric preconditioner of the matrix $\mathbb{A}_{ss}$ in (4.5.5) can be formulated:

$$\begin{aligned}\mathbb{B}_{ss}^{-1} &= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{B}_{\omega}^{-1}\boldsymbol{L} \\ & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{\omega}^{-1} & \\ & \boldsymbol{Z}_{ss}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \\ \boldsymbol{L}^T\boldsymbol{B}_{\omega}^{-1} & \boldsymbol{I} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{B}_{\omega}^{-1}\boldsymbol{L} \\ & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{\omega}^{-1} & \\ \boldsymbol{Z}_{ss}^{-1}\boldsymbol{L}^T\boldsymbol{B}_{\omega}^{-1} & \boldsymbol{Z}_{ss}^{-1} \end{bmatrix},\end{aligned} \tag{4.6.3}$$

where $\boldsymbol{B}_{\omega}^{-1} = (\omega+1)^{-1}\boldsymbol{B}^{-1}$ and $\boldsymbol{Z}_{ss}$ is a symmetric preconditioner of the respective Schur complement of $\mathbb{A}_{ss}$, $\boldsymbol{S}_{ss} = \boldsymbol{M} - (\omega+1)^{-1}\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}$. Note that the block matrix $\mathbb{A}_{ss}$ is symmetric positive definite and the respective Schur complement of $\mathbb{A}_{ss}$, $\boldsymbol{S}_{ss}$, is spectrally equivalent to $\boldsymbol{M}$ without requiring the inf-sup condition (4.4.16) (i.e., even when $c_I = 0$) with the equivalence depending on ω.

In Subsection 4.8.3, we provide preliminary results with the above presented block preconditioners using $\boldsymbol{B}^{-1} = \boldsymbol{H}^{-1}$ and $\boldsymbol{Z}_{ss} = \boldsymbol{I}$, $\boldsymbol{Z}_{inv} = -\boldsymbol{I}$. We are interested in utilizing preconditioners based on algebraic multigrid methods [39, 91, 92] as $\boldsymbol{B}^{-1}$, to be investigated in follow-up work.

4.7 Application to linear hyperbolic problems

The considerations above are rather general. Here, we comment on certain particularities associated with the application of the methods to the hyperbolic problem (4.1.1).

The differential operator in (4.1.1) can be written as $L = \hat{L} + \sigma I$, where $\hat{L}u = \nabla \cdot \boldsymbol{b}u$. Then $L^* = \hat{L}^* + \sigma I$, where integration by parts (Green's formula) [40] implies that $\hat{L}^* w = -\boldsymbol{b} \cdot \nabla w$. Thus, the PDE adjoint to (4.1.1) is also hyperbolic of similar type to (4.1.1). In particular, when $\nabla \cdot \boldsymbol{b} = 0$, then $\hat{L}u = \boldsymbol{b} \cdot \nabla u$, i.e., $\hat{L}^* = -\hat{L}$.

Furthermore,

$$\mathcal{D}(L) = \mathcal{D}(\hat{L}) = \{\, u \in L^2(\Omega);\ \hat{L}u \in L^2(\Omega) \text{ and } u = 0 \text{ on } \Gamma_I \,\},$$
$$\mathcal{D}(L^*) = \mathcal{D}(\hat{L}^*) = \{\, w \in L^2(\Omega);\ \hat{L}^* w \in L^2(\Omega) \text{ and } w = 0 \text{ on } \Gamma_O \,\},$$

where Γ_O is the outflow portion of the boundary, i.e.,

$$\Gamma_O = \{\, \boldsymbol{x} \in \partial\Omega;\ \boldsymbol{n}(\boldsymbol{x}) \cdot \boldsymbol{b}(\boldsymbol{x}) > 0 \,\}.$$

Note that the boundary conditions in the definitions of $\mathcal{D}(L)$ and $\mathcal{D}(L^*)$ make sense in terms of traces; see the trace theorem in [33].

Under reasonable mild assumptions on $\boldsymbol{b}$, a Poincaré-type inequality for $\hat{L}^*$ is shown in [33, Lemma 2.4]. A similar argument shows the respective inequality for $\hat{L}$; cf., [32, Lemma 6.8]. This covers the assumptions (ASM 1), (ASM 3) (as well as (ASM 2), by Remark 4.2.3) for the case $\sigma \equiv 0$. The case of $\sigma \not\equiv 0$ is simpler and is studied in [35].

4.8 Numerical results

Numerical results are shown in this section that demonstrate the behavior of the methods presented and studied in this chapter. Also, experiments with the block preconditioners of Section 4.6 are provided. The software used for implementing and testing the methods is FEniCS, cbc.block [93], PETSc [94], and LEAP (a least-squares package based on FEniCS that is under development at University of Colorado, Boulder).

4.8.1 Experiment setting

The domain, boundaries, structure of the coefficient σ, and a typical unstructured quasi-uniform triangular mesh (the coarsest mesh used in our experiments here) are shown on Figure 4.1. Namely, the domain is $\Omega = (0,1)^2$. It is split in two subregions – $\Omega_{in} = (0.25, 0.75)^2$ and $\Omega_{out} = \Omega \setminus \Omega_{in}$. The coefficient σ is discontinuous – $\sigma = \sigma_{out}$ in Ω_{out} and $\sigma = \sigma_{in}$ in Ω_{in}. We choose σ_{out} small (i.e., Ω_{out} is a "thin" region) and σ_{in} relatively large (i.e., Ω_{in} is a "thick" region). In particular, the experiments here use $\sigma_{out} = 10^{-4}$ and $\sigma_{in} = \{\, 10^4 \text{ or } 10 \,\}$. The choice $\sigma_{in} = 10^4$ provides a case when very steep exponential layers form that are not well resolved by the meshes. In contrast, when $\sigma_{in} = 10$, the exponential layers are less steep and can be resolved by a reasonably fine mesh. In all test cases, $\boldsymbol{b} = [\cos\alpha, \sin\alpha]$, where $\alpha = 3\pi/16$. Also, we set $r \equiv 0$ and $g = 1$ on Γ_I, where,

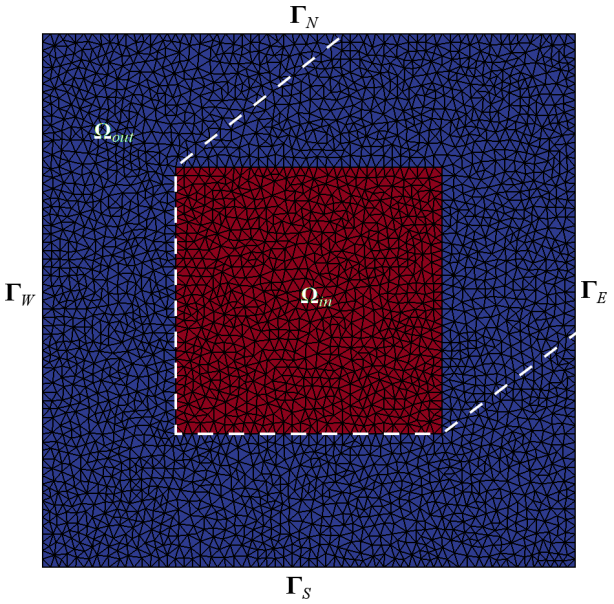


Figure 4.1: Experiment setting.

with the current choice of $\boldsymbol{b}$, $\Gamma_I = \Gamma_W \cup \Gamma_S$ and $\Gamma_O = \Gamma_E \cup \Gamma_N$. Thus, (4.1.1) becomes

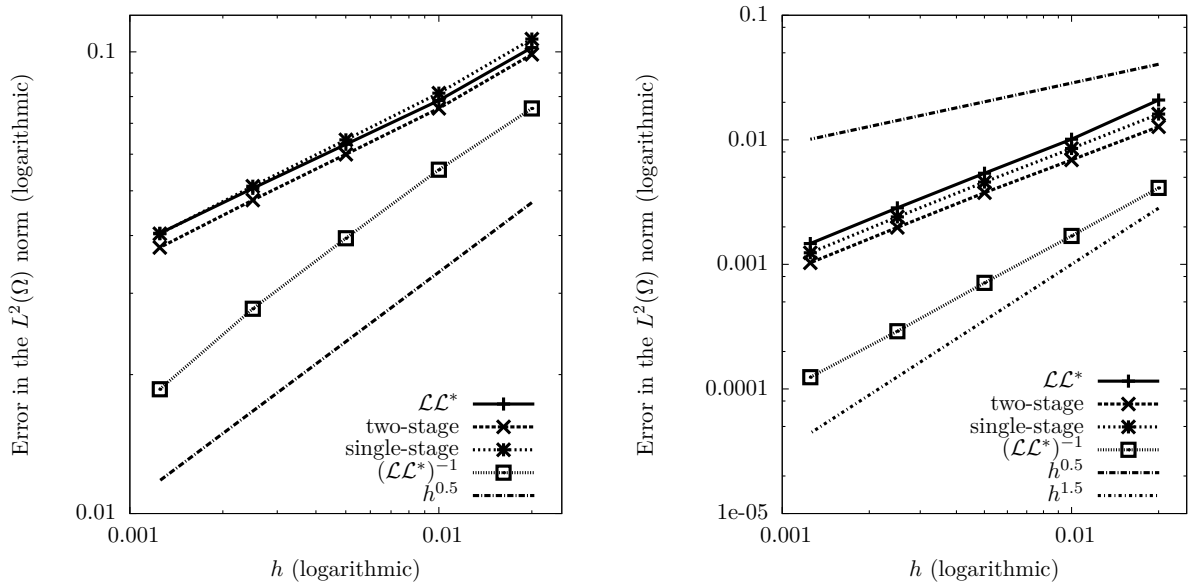
$$\begin{aligned} \boldsymbol{b} \cdot \nabla \psi + \sigma \psi &= 0 \quad \text{in } \Omega, \\ \psi &= 1 \quad \text{on } \Gamma_I. \end{aligned}$$

The dashed lines in Figure 4.1 show the locations of the exponential layers with the current choices of $\boldsymbol{b}$ and σ. The unstructured meshes do not follow the characteristics of the problem, but they resolve the coefficient σ (i.e., they take into account the subregions Ω_{in} and Ω_{out}). In all tests, we set $\omega = 1$ in (4.5.4).

4.8.2 Convergence experiments

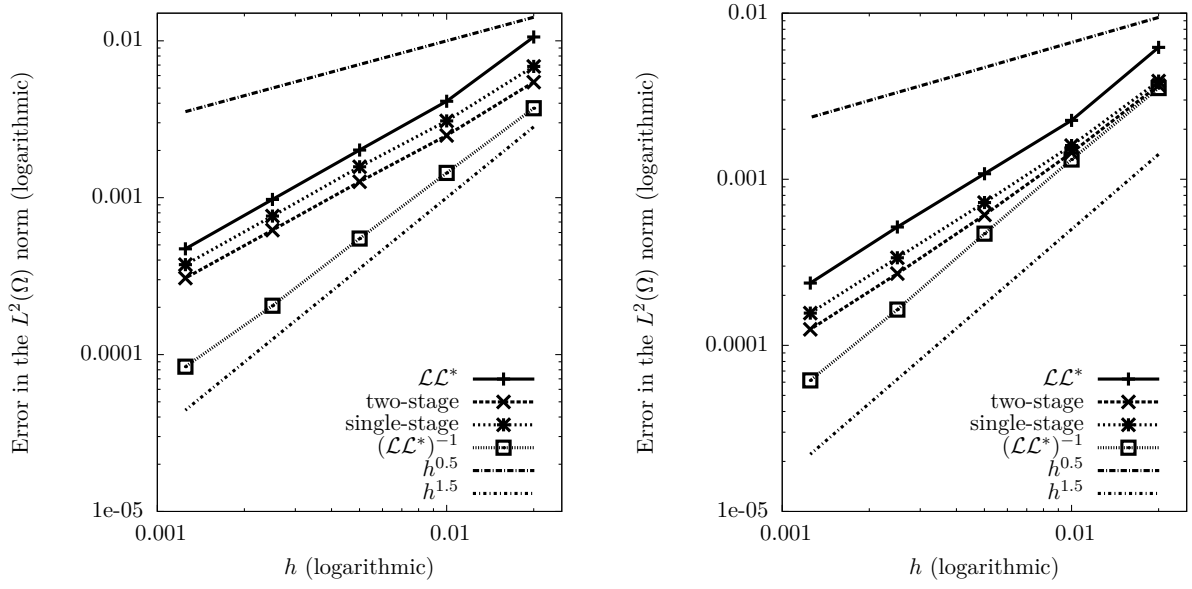
In this subsection, convergence with respect to the $L^2(\Omega)$ norm of the methods in this chapter is demonstrated. In all cases, standard Lagrangian ($\mathcal{C}^0$ piecewise polynomial) finite element spaces are utilized for $\mathcal{U}^h$ and $\mathcal{Z}^\mathfrak{h}$. Based on Corollary 4.4.7, it is always ensured that $\dim(\mathcal{Z}^\mathfrak{h}) > \dim(\mathcal{U}^h)$.

First, results for $\mathcal{U}^h$ – linear, $\mathcal{Z}^\mathfrak{h}$ – quadratic, both spaces on the same respective meshes, and $\sigma_{in} = 10^4$ are shown on Figure 4.2a. Note that, strictly speaking, the exact solution is in the Sobolev space $H^{3/2-\epsilon}(\Omega)$, for any $\epsilon > 0$. According to polynomial approximation theory [58], the optimal asymptotic rate of convergence of the L^2-norm approximations of the exact solution on $\mathcal{U}^h$ is $h^{3/2-\epsilon}$. However, the analytical solution possesses very steep exponential layers that on the scale of the



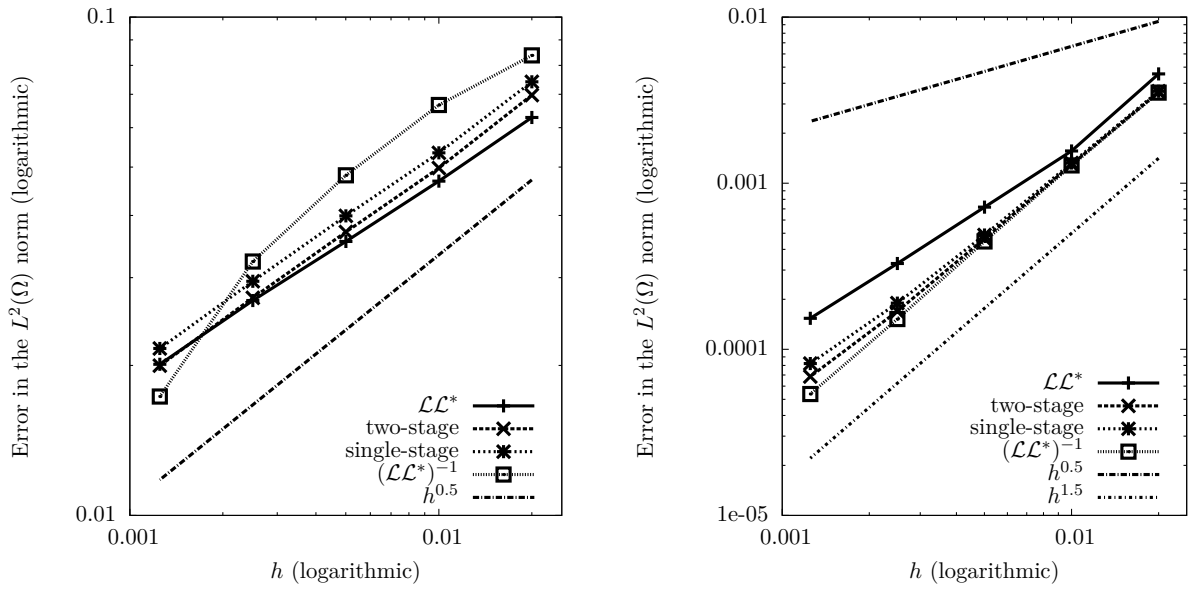
(a) $\sigma_{in} = 10^4$ (b) $\sigma_{in} = 10$

Figure 4.2: Convergence results. The spaces $\mathcal{U}^h$ and $\mathcal{Z}^\mathfrak{h}$ are on the same meshes, $\mathcal{U}^h$ – linear, $\mathcal{Z}^\mathfrak{h}$ – quadratic.



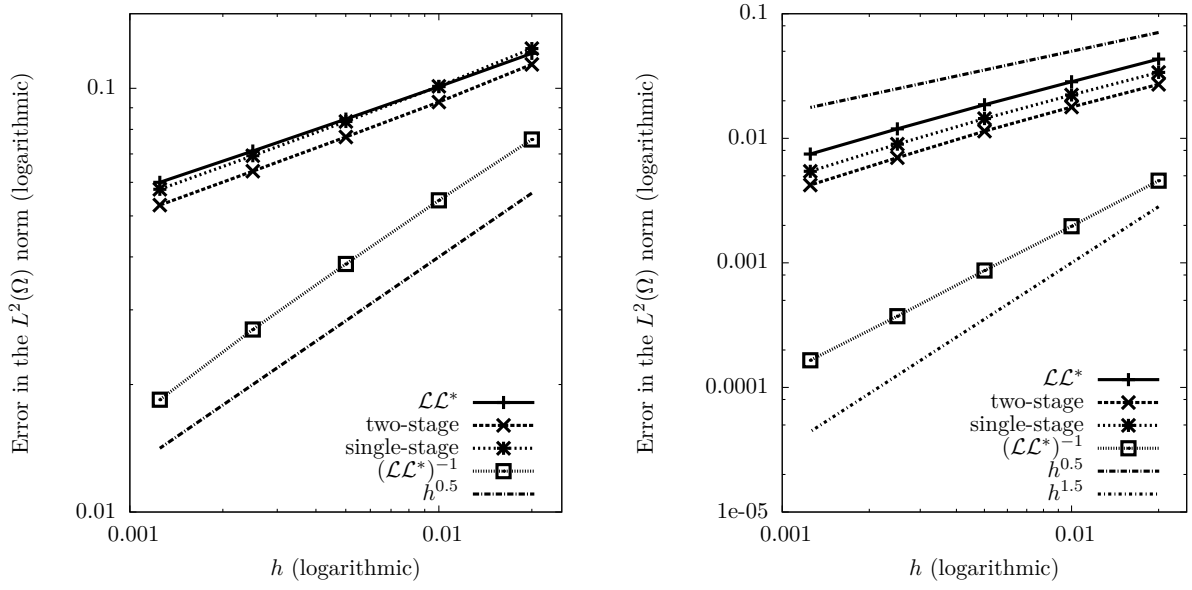
(a) $\mathcal{Z}^\mathfrak{h}$ – cubic (b) $\mathcal{Z}^\mathfrak{h}$ – quartic

Figure 4.3: Convergence results. The spaces $\mathcal{U}^h$ and $\mathcal{Z}^\mathfrak{h}$ are on the same meshes, $\mathcal{U}^h$ – linear, $\sigma_{in} = 10$.



(a) $\sigma_{in} = 10^4$ (b) $\sigma_{in} = 10$

Figure 4.4: Convergence results. The spaces $\mathcal{U}^h$ and $\mathcal{Z}^\mathfrak{h}$ are on the same meshes, $\mathcal{U}^h$ – linear, $\mathcal{Z}^\mathfrak{h}$ – quintic.



(a) $\sigma_{in} = 10^4$ (b) $\sigma_{in} = 10$

Figure 4.5: Convergence results. The spaces $\mathcal{U}^h$ and $\mathcal{Z}^\mathfrak{h}$ are both linear. Every mesh of $\mathcal{Z}^\mathfrak{h}$ is obtained by a single uniform refinement of the respective $\mathcal{U}^h$ mesh.

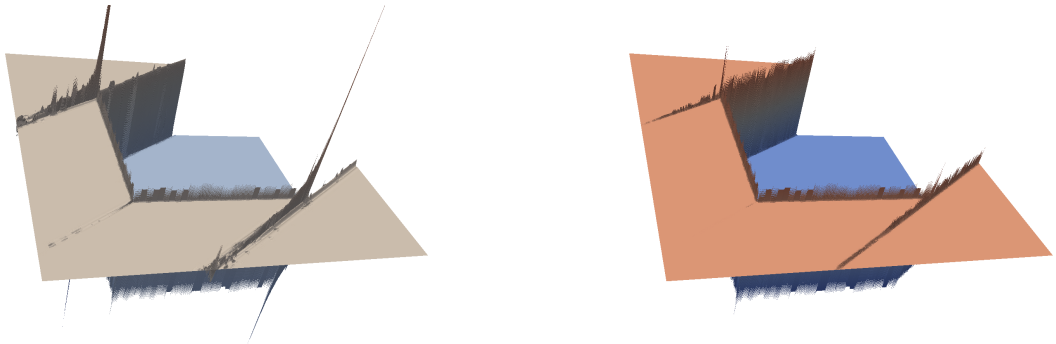
(coarser) meshes behave like discontinuities (which is a case of interest). Therefore, intuitively, until the mesh begins resolving the exponential layers (i.e., before the "asymptotic regime" starts settling), the exact solution can, in a sense, be seen as "discontinuous", i.e., nearly behaving as a function in $H^{1/2-\epsilon}(\Omega)$, and a rate of around $h^{1/2}$ can be considered as "optimal" initially with the potential of improving as the mesh is refined. More precisely, in view of the interpolation bounds of the polynomial approximation theory, the $H^{3/2-\epsilon}(\Omega)$ norm of the analytical solution is rather large and this is associated with a delayed "asymptotic regime" of convergence. Figure 4.2a demonstrates that the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method obtains an $h^{1/2}$ rate. In comparison, the $\mathcal{L}\mathcal{L}^*$, single-, and two-stage methods are slower to converge. Owing to Theorem 4.5.3, this can be explained with the approximation properties of $L^*(\mathcal{Z}^\mathfrak{h})$. It is interesting to notice that, in view of Theorem 4.5.3, the single- and two-stage methods demonstrate slightly "enhanced" convergence rates compared to the $\mathcal{L}\mathcal{L}^*$ method. The current theory cannot predict or explain such a behavior. It is unclear if this "enhanced" rate would be maintained once the "asymptotic regime" fully settles.

Note that the $h^{1/2}$ convergence of the $(\mathcal{L}\mathcal{L}^*)^{-1}$ approximations in Figure 4.2a does not necessarily mean that the inf-sup condition (4.4.16) holds with c_I independent of h. For example, observe Figure 4.2b, which shows the same experiment as above but with $\sigma_{in} = 10$, i.e., the exponential layers are now well resolved by the meshes and the optimal asymptotic rate $h^{3/2-\epsilon}$ is achievable. Notice that all methods, including the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method, demonstrate suboptimal rates. The $\mathcal{L}\mathcal{L}^*$, single-, and two-stage methods this time converge with equal rates, but slower than the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method and their respective errors are close to each other. The suboptimal convergence of the $(\mathcal{L}\mathcal{L}^*)^{-1}$ approximations indicates that the inf-sup condition (4.4.16) does not hold uniformly for this choice of spaces, i.e., c_I in (4.4.16) depends on h. Figures 4.2b, 4.3a, 4.3b, and 4.4b track the change (improvement) in the errors of the methods as the order of $\mathcal{Z}^\mathfrak{h}$ is increased, for $\sigma_{in} = 10$.

Next, Figure 4.4 (compare with Figure 4.2) shows an experiment with quintic $\mathcal{Z}^\mathfrak{h}$. Piecewise polynomial finite element spaces on triangles of order five (or higher) are special in the sense that they contain the space associated with the Argyris element; cf., [58]. In other words, $\mathcal{Z}^\mathfrak{h}$ has a $\mathcal{C}^1$ piecewise polynomial finite element subspace and $L^*(\mathcal{Z}^\mathfrak{h})$ contains a $\mathcal{C}^0$ piecewise polynomial finite element space. A counting argument shows that the $\mathcal{C}^0$ subspace of $L^*(\mathcal{Z}^\mathfrak{h})$ is smaller than $\mathcal{U}^h$. The results on Figure 4.4 suggest that this is not sufficient for the approximation properties of $L^*(\mathcal{Z}^\mathfrak{h})$ to be on par with those of $\mathcal{U}^h$, but the the inf-sup condition (4.4.16) may potentially hold. This

(a) $\mathcal{Z}^{\mathfrak{h}}$ – quadratic (b) $\mathcal{Z}^{\mathfrak{h}}$ – quintic

Figure 4.6: Plots of $(\mathcal{L}\mathcal{L}^*)^{-1}$ solutions in a linear $\mathcal{U}^h$, where the mesh in Figure 4.1 is refined 4 times.

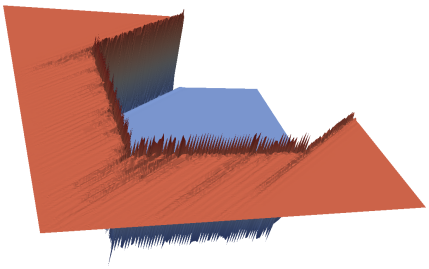


(a) $\mathcal{Z}^{\mathfrak{h}}$ – quadratic (b) $\mathcal{Z}^{\mathfrak{h}}$ – quintic

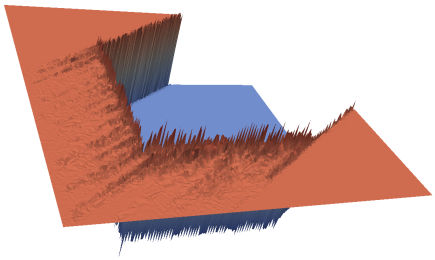
Figure 4.7: Plots of two-stage solutions in a linear $\mathcal{U}^h$, where the mesh in Figure 4.1 is refined 4 times.

is a subject of future investigation. Observe also that increasing the order of $\mathcal{Z}^{\mathfrak{h}}$ in Figure 4.4a, compared to Figure 4.2a, results in improved errors for the $\mathcal{L}\mathcal{L}^*$, single-, and two-stage methods, whereas this does not initially lead to an error improvement for the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method and only on finer meshes such an improvement can be observed. This creates the impression in Figure 4.4a that the $(\mathcal{L}\mathcal{L}^*)^{-1}$ solution converges with a rate higher than $h^{1/2}$. However, this is due to the sudden improvement in the size of the error, since the mesh is not sufficiently fine to resolve the steep layers and a rate around $h^{1/2}$ is to be expected, even from the actual best L^2-norm approximation on $\mathcal{U}^h$.

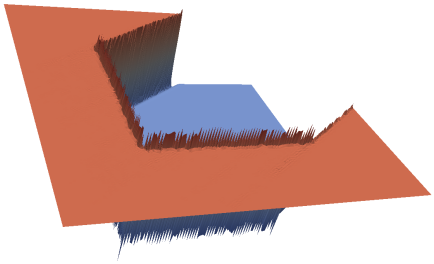
The spaces $\mathcal{U}^h$ and $\mathcal{Z}^h$ do not need to be on the same mesh. This is demonstrated on Figure 4.5 for the case when $\mathcal{Z}^h$ utilizes refined versions of the respective meshes of $\mathcal{U}^h$. The results are very similar, with slightly slower rates, to those on Figure 4.2.



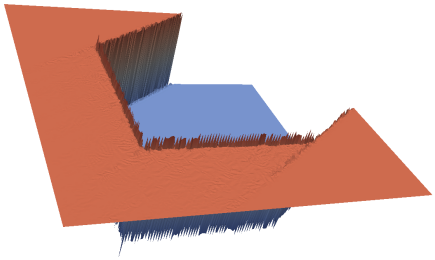
(a) $\mathcal{U}^h$ – linear, $\mathcal{Z}^\mathfrak{h}$ – quadratic



(b) $\mathcal{U}^h$ – cubic, $\mathcal{Z}^\mathfrak{h}$ – quartic



(c) $\mathcal{U}^h$ – linear, $\mathcal{Z}^\mathfrak{h}$ – quintic

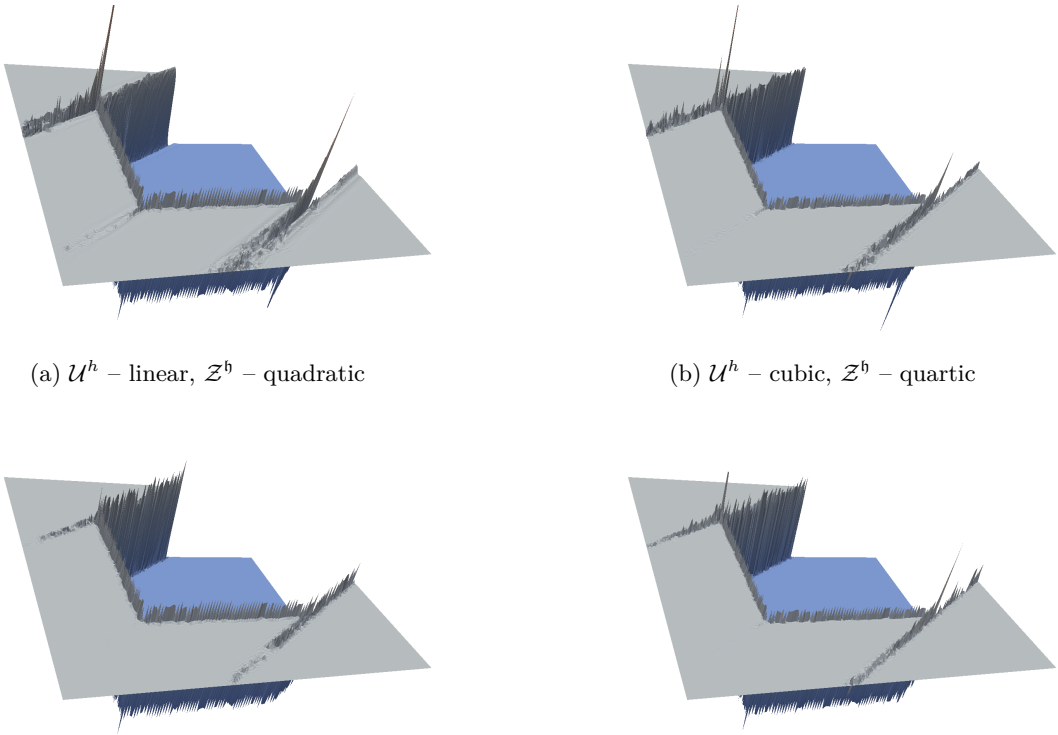


(d) $\mathcal{U}^h$ – cubic, $\mathcal{Z}^\mathfrak{h}$ – degree 7

Figure 4.8: Plots of $(\mathcal{L}\mathcal{L}^*)^{-1}$ solutions, where the mesh in Figure 4.1 is refined twice.

Interestingly, in view of Figures 4.2a, 4.4a, and 4.5a, the losses of optimal rate ($h^{3/2-\epsilon}$), caused by the unresolved exponential layers and the dependence of c_I on h, do not seem to add up in the results for the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method. It seems that the slowest non-optimality dominates, which here is mostly the non-optimality of the mesh, and we obtain a rate of around $h^{1/2}$ on coarser meshes.

The methods in this chapter target approximations in the $L^2(\Omega)$ norm and, as a result, provide a much better resolution of steep layers than a standard least-squares approach. However, more oscillations are now produced, which, due to the nature of the $L^2(\Omega)$ norm, do not prohibit convergence. Particularly, in view of Figures 4.6–4.9, the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method produces substantially less oscillations than the $\mathcal{L}\mathcal{L}^*$-type methods and, interestingly, the $(\mathcal{L}\mathcal{L}^*)^{-1}$ approach provides a slightly better resolution of steep layers, both contributing to smaller L^2-norm errors. This aligns with the observations that, in terms of solution quality, it is better to use $\mathcal{Z}^\mathfrak{h}$ to approximate $(L_w L^*)^{-1}$ than $L^*(\mathcal{Z}^\mathfrak{h})$ to approximate $\hat{u}$ or, similarly, it is better to relate $\mathcal{U}^h$ and $L^*(\mathcal{Z}^\mathfrak{h})$ via an inf-sup condition than via approximation properties. Particular plots of the solutions produced by the methods can



(a) $\mathcal{U}^h$ – linear, $\mathcal{Z}^\mathfrak{h}$ – quadratic

(b) $\mathcal{U}^h$ – cubic, $\mathcal{Z}^\mathfrak{h}$ – quartic

(c) $\mathcal{U}^h$ – linear, $\mathcal{Z}^\mathfrak{h}$ – quintic

(d) $\mathcal{U}^h$ – cubic, $\mathcal{Z}^\mathfrak{h}$ – degree 7

Figure 4.9: Plots of two-stage solutions, where the mesh in Figure 4.1 is refined twice.

be seen in Figures 4.6 and 4.7. The solutions provided by the single-stage method have a very similar appearance to the two-stage solutions. Further plots are shown on Figures 4.8 and 4.9, demonstrating the effect of using higher-order elements for $\mathcal{U}^h$. Namely, increasing the order of $\mathcal{U}^h$ can cause more oscillations, whereas increasing the order of $\mathcal{Z}^\mathfrak{h}$ reduces those oscillations.

In our experiments, we observe that the local L^2-norm error in subregions, where the solution is smooth, decreases with higher rates. Namely, in the case of a large contrast in σ (i.e., corresponding to Figures 4.2a and 4.4a), the rate of local convergence is around $\mathcal{O}(h)$ for all methods and both choices of $\mathcal{Z}^\mathfrak{h}$ (quadratic and quintic). This demonstrates that the "polluting" effect of the steep layers is limited to some extent and requires further investigation.

Finally, observe that in the majority of the results above the L^2-norm errors of the single- and two-stage methods are smaller than the respective errors of the $\mathcal{L}\mathcal{L}^*$ method. In all tests presented here, the two-stage method exhibits smaller errors compared to the single-stage method. In some cases, the convergence rates of the single- and two-stage approximations are "enhanced" (better) in comparison to the respective rates of the $\mathcal{L}\mathcal{L}^*$ method. It is not completely clear if this

"enhanced" error behavior is maintained asymptotically as $h \to 0$. Also, notice that as the order of $\mathcal{Z}^{\mathfrak{h}}$ is increased, the error graphs of the different methods become more grouped together. This is predicted by the theoretical considerations in the previous sections. Namely, as $\mathfrak{h} \to 0$, for fixed h, all methods approach the same formulation – the L^2-orthogonal projection onto $\mathcal{U}^h$.

4.8.3 Preconditioning experiments

Here, preliminary results with the block preconditioners, introduced in Section 4.6, are shown. In particular, we use $\boldsymbol{B}^{-1} = \boldsymbol{H}^{-1}$, $\boldsymbol{Z}_{ss} = \boldsymbol{I}$, and $\boldsymbol{Z}_{inv} = -\boldsymbol{I}$ to provide a basic idea about the behavior of the preconditioners. The effect of $\boldsymbol{B}^{-1}$ is computed using a sparse direct solver – MUMPS [95]. All numerical experiments in this subsection are for the case when $\mathcal{U}^h$ and $\mathcal{Z}^{\mathfrak{h}}$ are piecewise linear and $\mathcal{Z}^{\mathfrak{h}}$ utilizes meshes that are obtained from the respective $\mathcal{U}^h$ meshes by a single uniform refinement, i.e., $\mathfrak{h} = h/2$; that is, the results here correspond to Figure 4.5. In all tests, the iterative processes are stopped when the overall relative reduction of the norm of the preconditioned residual becomes less than 10^{-6}.

Table 4.1 shows the number of preconditioned GMRES(30) [96, 97] iterations for the $(\mathcal{L}\mathcal{L}^*)^{-1}$ system (4.4.9) using the block preconditioner $\mathbb{B}_{inv}^{-1}$ in (4.6.2). The preconditioner $\mathbb{B}_{inv}^{-1}$ is not scalable, as $h \to 0$ with the choice $\boldsymbol{Z}_{inv} = -\boldsymbol{I}$. As already discussed in Section 4.6, this can be associated with the dependence of c_I, in the inf-sup condition (4.4.16), on the mesh parameter, h, and, as a result, the spectral relation (4.4.18) does not hold uniformly (i.e., it depends on h). This suggests that further care is necessary in preconditioning the Schur complement $\boldsymbol{S}_{inv}$ and the simple choice $\boldsymbol{Z}_{inv} = -\boldsymbol{I}$ is insufficient in this case. We plan to further investigate this, together with the utilization of algebraic multigrid as $\boldsymbol{B}^{-1}$, in follow-up work.

In contrast, as discussed in Section 4.6, the block preconditioner $\mathbb{B}_{ss}^{-1}$ in (4.6.3) is scalable, as $h \to 0$, for the single-stage system (4.5.5) with the choice $\boldsymbol{Z}_{ss} = \boldsymbol{I}$, independently of the inf-sup condition (4.4.16). Also, preconditioned CG can be used in this case. Results are shown in Table 4.2.

Table 4.1: (Results for the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method) Number of preconditioned GMRES(30) iterations for the $(\mathcal{L}\mathcal{L}^*)^{-1}$ system (4.4.9) using relative tolerance 10^{-6} and the preconditioner $\mathbb{B}_{inv}^{-1}$ in (4.6.2) with $\boldsymbol{B}^{-1} = \boldsymbol{H}^{-1}$ and $\boldsymbol{Z}_{inv} = -\boldsymbol{I}$.

				$\sigma_{in} = 10^4$	$\sigma_{in} = 10$
h	$\mathfrak{h}$	$\dim(\mathcal{U}^h)$	$\dim(\mathcal{Z}^\mathfrak{h})$	iterations	iterations
0.02	0.01	3226	12645	73	51
0.01	0.005	12645	50065	105	67
0.005	0.0025	50065	199233	147	84
0.0025	0.00125	199233	794881	201	106
0.00125	0.000625	794881	3175425	255	119

Table 4.2: (Results for the single-stage method) Number of preconditioned CG iterations for the single-stage system (4.5.5) using relative tolerance 10^{-6} and the preconditioner $\mathbb{B}_{ss}^{-1}$ in (4.6.3) with $\boldsymbol{B}^{-1} = \boldsymbol{H}^{-1}$ and $\boldsymbol{Z}_{ss} = \boldsymbol{I}$.

				$\sigma_{in} = 10^4$	$\sigma_{in} = 10$
h	$\mathfrak{h}$	$\dim(\mathcal{U}^h)$	$\dim(\mathcal{Z}^\mathfrak{h})$	iterations	iterations
0.02	0.01	3226	12645	30	25
0.01	0.005	12645	50065	31	28
0.005	0.0025	50065	199233	32	31
0.0025	0.00125	199233	794881	34	34
0.00125	0.000625	794881	3175425	35	35

4.9 Regularizations

In this section, we discuss a certain type of regularization (stabilization) of the $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation (4.4.9), (4.5.9). Such approaches are known and used in the context of saddle-point problems; cf. [98]. In that context, stabilization is associated with adding terms to the $(2,2)$ block of the saddle-point matrix $\mathbb{A}$ in (4.4.9). Thus, in a way, the single-stage formulation (4.5.5) can be viewed as a special regularization of the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method. In fact, in view of (4.5.9)–(4.5.11), both the single- and two-stage methods can be seen as regularizations of the $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation.

As already mentioned, the stabilization terms in (4.5.10) and (4.5.11) do not contain any additional information on the exact solution. The idea here is to augment the $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation (4.5.9) with a FOSLS term, which certainly carries additional information on the particular PDE and its solution. This is similar to the hybrid method in [44], where the single-stage formulation (4.5.4) is combined with FOSLS, while here we look at combining the $(\mathcal{L}\mathcal{L}^*)^{-1}$ and FOSLS methods.

Consider the regularized (or "hybrid") formulation

$$
\begin{aligned}
u^h_{reg} &= \operatorname*{argmin}_{v^h \in \mathcal{U}^h} \left[\|\Pi^{\mathfrak{h}}_* (v^h - \hat{u})\|^2 + S \|L(v^h - \hat{u})\|^2 \right] \\
&= \operatorname*{argmin}_{v^h \in \mathcal{U}^h} \left[\|\Pi^{\mathfrak{h}}_* (v^h - \hat{u})\|^2 + S \|Lv^h - f\|^2 \right],
\end{aligned} \tag{4.9.1}
$$

for some scaling $S > 0$, which can generally depend on h. Here, (4.9.1) is expressed in this form for simplicity and it is easy to extend the considerations in this section for the general case when S varies between mesh elements.

In terms of the approach and notation in Section 4.4, (4.9.1) can be identified with the following minimization and weak forms:

$$
\begin{aligned}
&\text{Find } u^h_{reg} \in \mathcal{U}^h \colon \langle \Pi^{\mathfrak{h}}_* u^h_{reg}, v^h \rangle + S \langle L u^h_{reg}, L v^h \rangle \\
&\qquad\qquad = \langle \Pi^{\mathfrak{h}}_* \hat{u}, v^h \rangle + S \langle f, L v^h \rangle, \quad \forall v^h \in \mathcal{U}^h,
\end{aligned} \tag{4.9.2}
$$

$$
u^h_{reg} = \operatorname*{argmin}_{v^h \in \mathcal{U}^h} \langle [(L_w L^*)^{-1}_{\mathfrak{h}} + S\, I](Lv^h - f), Lv^h - f \rangle, \tag{4.9.3}
$$

$$
\begin{aligned}
&\text{Find } u^h_{reg} \in \mathcal{U}^h \colon \langle (L_w L^*)^{-1}_{\mathfrak{h}} L u^h_{reg}, L v^h \rangle + S \langle L u^h_{reg}, L v^h \rangle \\
&\qquad\qquad = \langle (L_w L^*)^{-1}_{\mathfrak{h}} f, L v^h \rangle + S \langle f, L v^h \rangle, \quad \forall v^h \in \mathcal{U}^h.
\end{aligned} \tag{4.9.4}
$$

Alternatively, the weak formulations (4.9.2) and (4.9.4), combined with the definition of $(L_w L^*)^{-1}_{\mathfrak{h}}$, can be expressed as the system

$$
\text{Find } (u^h_{reg}, z^{\mathfrak{h}}) \in \mathcal{U}^h \times \mathcal{Z}^{\mathfrak{h}} \colon \begin{cases} \langle L^* z^{\mathfrak{h}}, L^* w^{\mathfrak{h}} \rangle + \langle u^h_{reg}, L^* w^{\mathfrak{h}} \rangle = \langle f, w^{\mathfrak{h}} \rangle, & \forall w^{\mathfrak{h}} \in \mathcal{Z}^{\mathfrak{h}}, \\ \langle L^* z^{\mathfrak{h}}, v^h \rangle - S \langle L u^h_{reg}, L v^h \rangle = -S \langle f, L v^h \rangle, & \forall v^h \in \mathcal{U}^h. \end{cases} \tag{4.9.5}
$$

Clearly, utilizing the notation in Subsection 4.4.2, (4.9.5) induces the following saddle-point system of linear equations:

$$
\mathbb{A}_{reg} \begin{bmatrix} \boldsymbol{z} \\ \boldsymbol{u}_{reg} \end{bmatrix} = \begin{bmatrix} \boldsymbol{H} & \boldsymbol{L} \\ \boldsymbol{L}^T & -S\,\boldsymbol{V} \end{bmatrix} \begin{bmatrix} \boldsymbol{z} \\ \boldsymbol{u}_{reg} \end{bmatrix} = \begin{bmatrix} \bar{\boldsymbol{f}} \\ -S\boldsymbol{r} \end{bmatrix}, \tag{4.9.6}
$$

where $\boldsymbol{V} \in \mathbb{R}^{N \times N}$ and $\boldsymbol{r} \in \mathbb{R}^N$ are defined as:

$$
(\boldsymbol{V})_{ij} = \langle L\phi^h_j, L\phi^h_i \rangle, \quad (\boldsymbol{r})_i = \langle f, L\phi^h_i \rangle.
$$

Using the notation in (4.4.11), the resulting Schur complement from the elimination of $\boldsymbol{z}$ in (4.9.6) is $-\boldsymbol{A} - S\,\boldsymbol{V}$, which is negative definite and, hence, $\mathbb{A}_{reg}$ is an indefinite and nonsingular matrix.

Furthermore, the respective linear system for the Schur complement is the following:

$$(\boldsymbol{A} + S\,\boldsymbol{V})\boldsymbol{u}_{reg} = \boldsymbol{f} + S\boldsymbol{r},$$

which is induced by the weak formulations (4.9.2) and (4.9.4).

In view of (4.4.1) and (4.9.3), the regularized or hybrid formulation (4.9.1) can be seen as the result of approximating $(L_w L^*)^{-1}$ with $(L_w L^*)_{\mathfrak{h}}^{-1} + S\,I$. In particular, if elliptic problems are considered, this is related to the approach in [45]. Namely, considering the elliptic case with $L = \nabla\cdot$ and $L^* = -\nabla$ results in $(L_w L^*)^{-1} = (-\Delta)^{-1}$ and, for $S = h^2$, $(L_w L^*)_{\mathfrak{h}}^{-1} + S\,I = (-\Delta)_{\mathfrak{h}}^{-1} + h^2 I$, which essentially provides the H^{-1}-type least-squares term in [45]. The regularization $h^2 I$ in [45] is sufficient to recover the coercivity, originally rendered by $(-\Delta)^{-1}$, for the discrete operator $(-\Delta)_{\mathfrak{h}}^{-1}$.

However, the hyperbolic case seems more complicated and the theory in [45] does not carry over. Currently, it is not clear if and how S can be chosen so that a desired coercivity can be guaranteed for the operator $(L_w L^*)_{\mathfrak{h}}^{-1} + S\,I$. The choice $S = h^2$ seems intuitively reasonable here as well. Nevertheless, in our experiments, the regularized method (4.9.1), with $S = h^2$, provides solutions with noticeably worse $L^2(\Omega)$ errors in comparison to the original $(\mathcal{L}\mathcal{L}^*)^{-1}$ approach, but the errors are still a bit better in comparison to the $\mathcal{L}\mathcal{L}^*$, single-, and two-stage methods. As discussed in the introduction to this chapter, this is associated with FOSLS producing poor approximations to the solution of (4.1.1), when σ has large contrasts and the mesh size, h, is not sufficiently small to capture the solution features. Therefore, the minimization (4.9.1) is affected by the poor FOSLS behavior, since (4.9.1) combines the FOSLS and $(\mathcal{L}\mathcal{L}^*)^{-1}$ terms in a single functional; that is, there are challenges in both theoretical and practical sense. One possibility to address the practical side of the issue is to utilize a properly locally (i.e., element by element) scaled FOSLS method in (4.9.1), as the one in [81], instead of the currently used vanilla (i.e., unscaled) FOSLS formulation, which is known for its poor behavior. Another approach is to utilize discontinuous least-squares (DLS) methods [33]. This is beyond the scope of this dissertation, since the main focus here is on dual methods. Further investigations in that direction are a subject of future work. The main purpose of this section is to demonstrate that two rather natural regularization approaches (namely, adding FOSLS terms to obtain a hybrid method and stabilizing the saddle-point problem) can yield the same final formulation.

4.10 Conclusions and further development

We proposed the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method, together with the $\mathcal{L}\mathcal{L}^*$, single-, and two-stage methods, and studied their application to scalar linear hyperbolic PDEs, aiming at obtaining L^2-norm approximations on finite element spaces. Error estimates were shown, pointing to the factors that affect convergence and providing conditions that guarantee optimal rates. Also, numerical results were demonstrated. The methods clearly show L^2-norm convergence, often with acceptable rates. The $(\mathcal{L}\mathcal{L}^*)^{-1}$ method demonstrates the best convergence rates, but it induces the most difficult linear systems to solve.

The considerations in this chapter suggest further directions of research. A few of them are mentioned in the exposition and in Section 4.A. Some additional topics are the following: further investigation of the potential regularizations of the $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation including the utilization of first-order system least-squares (FOSLS) terms in a "hybrid" method; and the potential of using $\mathcal{Z}^\mathfrak{h}$ on different meshes (even if we have no freedom to choose the mesh of $\mathcal{U}^h$, we can select $\mathcal{Z}^\mathfrak{h}$ freely) that can be better tailored to the particular problem and, thus, obtain $(L_w L^*)^{-1}_\mathfrak{h}$ that better approximates $(L_w L^*)^{-1}$, in some sense. The inf-sup condition and its relation to the approximation properties of the finite element spaces is an interesting and very challenging topic. This would allow further comparison between the methods in terms of the derived error estimates. Currently, the numerical results and basic analysis suggest that the requirements on the approximation properties of $L^*(\mathcal{Z}^\mathfrak{h})$ may possibly be stronger than the inf-sup condition. At least, we observe that when $L^*(\mathcal{Z}^\mathfrak{h})$ provides neither appropriate approximation properties, nor a uniform inf-sup condition, then the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method seems less affected by the deficiencies of $L^*(\mathcal{Z}^\mathfrak{h})$ in terms of convergence rates, but it may suffer more in terms of the efficiency of the linear solver. Furthermore, it is intriguing to study the influence of the coefficient σ on the constant c_I in the inf-sup condition (4.4.16) and, thus, on the behavior of the method, as well as whether and how the Poincaré constants in (ASM 1), (ASM 3) affect (4.4.16).

The proposed block preconditioner is one approach to solving the $(\mathcal{L}\mathcal{L}^*)^{-1}$ and single-stage linear systems. It would be interesting to study the adaptation and utilization of other methods, developed for "saddle-point problems", towards solving the $(\mathcal{L}\mathcal{L}^*)^{-1}$ system. Preconditioning the matrix $\boldsymbol{H}$, coming from hyperbolic operators, L^*, also suggests further development, which is

applicable beyond the methods of this chapter.

4.A Generalizing the formulations

In this appendix, for completeness, we review possible generalizations and extensions of the formulations in this chapter. In particular, we discuss the "weak" treatment of the inflow boundary condition and the potential of utilizing general (possibly discontinuous) finite element spaces as $\mathcal{U}^h$. Considering $\mathcal{U}^h \subset \mathcal{D}(L)$, for the $(\mathcal{L}\mathcal{L}^*)^{-1}$, single-, and two-stage methods, and enforcing the boundary data by superposition corresponds to imposing the boundary condition "strongly". For the general case, when $\mathcal{U}^h \subset L^2(\Omega)$ is possibly piecewise discontinuous, it is necessary to impose the boundary condition in a "weak" sense (i.e., as a part of the variational formulation).

Recall that, for simplicity, the hyperbolic problem (4.1.1) was reformulated as the operator equation (4.1.2) using superposition to enforce the boundary data g on Γ_I. In particular, this simplifies the weak formulation associated with the $\mathcal{L}\mathcal{L}^*$ minimization (4.5.1), since the exact solution, $\hat{u}$, of (4.1.2) is in $\mathcal{D}(L)$ (i.e., $\hat{u} = 0$ on Γ_I). Alternatively, consider the original PDE (4.1.1) and let $\hat{\psi}$ denote its exact solution. The respective $\mathcal{L}\mathcal{L}^*$ minimization is

$$\xi_*^{\mathfrak{h}} = \operatorname*{argmin}_{w^{\mathfrak{h}} \in \mathcal{Z}^{\mathfrak{h}}} \|L^* w^{\mathfrak{h}} - \hat{\psi}\|^2. \tag{4.A.1}$$

The resulting $\mathcal{L}\mathcal{L}^*$ approximation is $\psi_*^{\mathfrak{h}} = L^* \xi_*^{\mathfrak{h}} = \Pi_*^{\mathfrak{h}} \hat{\psi} \in L^*(\mathcal{Z}^{\mathfrak{h}})$. Using integration by parts (Green's formula), the weak form corresponding to (4.A.1) is the following:

$$\text{Find } z^{\mathfrak{h}} \in \mathcal{Z}^{\mathfrak{h}} \colon \langle L^* z^{\mathfrak{h}}, L^* w^{\mathfrak{h}} \rangle = \langle r, w^{\mathfrak{h}} \rangle - \int_{\Gamma_I} (\boldsymbol{b} g) \cdot \boldsymbol{n}\, w^{\mathfrak{h}} \, \mathrm{d}\sigma, \quad \forall w^{\mathfrak{h}} \in \mathcal{Z}^{\mathfrak{h}}.$$

Note that only the right-hand side is different and it involves only given data.

Using the $\mathcal{L}\mathcal{L}^*$ principle (4.A.1), leads to minor changes in the single- and two-stage formulations. Indeed, it is sufficient to replace $\bar{\boldsymbol{f}}$ with $\bar{\boldsymbol{f}}_b$ and $\boldsymbol{f}$ with $\boldsymbol{f}_b = \boldsymbol{L}^T \boldsymbol{H}^{-1} \bar{\boldsymbol{f}}_b \in \mathbb{R}^N$ in the respective linear systems above, where $\bar{\boldsymbol{f}}_b \in \mathbb{R}^M$ is defined as

$$(\bar{\boldsymbol{f}}_b)_i = \langle r, \psi_i^{\mathfrak{h}} \rangle - \int_{\Gamma_I} (\boldsymbol{b} g) \cdot \boldsymbol{n}\, \psi_i^{\mathfrak{h}} \, \mathrm{d}\sigma.$$

Note that this can still be combined with a "strong" enforcement of the inflow boundary data on $\mathcal{U}^h$ by standard means of the finite element methods, which demonstrates the flexibility that least-squares methods often provide. The analysis in Section 4.5 remains valid. Furthermore, the

single- and two-stage formulations are clearly general enough and allow the utilization of general finite element spaces $\mathcal{U}^h \subset L^2(\Omega)$.

According to Lemma 4.3.1, L_w is, in a sense, an extension of the operator L (more precisely, of $\mathcal{E}L$) on $L^2(\Omega)$. Therefore, the $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation is extended to general spaces $\mathcal{U}^h \subset L^2(\Omega)$ by replacing the operator L with its "weak" version L_w. This idea is already applied, for theoretical purposes, in the proof of Theorem 4.4.15 and in Section 4.6, when considering the operator $L^* B_{\mathfrak{h}}^{-1} L_w$, since the bilinear forms need to be defined on the whole of $L^2(\Omega)$ to obtain error estimates with respect to the $L^2(\Omega)$ norm. In fact, the generalized $(\mathcal{L}\mathcal{L}^*)^{-1}$ method is precisely the one in (4.4.14) and (4.4.15). Note that this is not only a tool of analysis but results in feasible formulations. Indeed, the weak formulation (4.4.7) is already stated in such a general form with the exception of the right-hand side, which needs to be modified to accommodate the "weak" enforcement of the boundary condition. Again, it is sufficient to replace $\bar{\boldsymbol{f}}$ with $\bar{\boldsymbol{f}}_b$ and $\boldsymbol{f}$ with $\boldsymbol{f}_b$ in the respective linear systems. Clearly, the analysis in this chapter remains valid. Similarly, the modified (by a preconditioner) formulation (4.6.1) can be extended to general finite element spaces $\mathcal{U}^h \subset L^2(\Omega)$.

In summary, we observed that the $(\mathcal{L}\mathcal{L}^*)^{-1}$, single-, and two-stage formulations can be easily generalized to arbitrary finite element spaces $\mathcal{U}^h \subset L^2(\Omega)$. However, a piecewise discontinuous finite element space $\mathcal{U}^h$ (which is a case of high interest) is rather rich, whereas $L^*(\mathcal{Z}^{\mathfrak{h}})$ is constrained by the requirement $\mathcal{Z}^{\mathfrak{h}} \subset \mathcal{D}(L^*)$. This poses further difficulties in maintaining the inf-sup condition (4.4.16) or on par approximation properties of $L^*(\mathcal{Z}^{\mathfrak{h}})$, according to the estimate in Theorem 4.5.3, when $\mathcal{U}^h$ is discontinuous. Removing or reducing the constraint $\mathcal{Z}^{\mathfrak{h}} \subset \mathcal{D}(L^*)$ is a challenging topic and it is a subject of future work.

Finally, the considerations in this chapter are rather general. In the exposition above, for simplicity of notation and since the scalar PDE (4.1.1) is considered, only $L^2(\Omega)$ is used. Nevertheless, in general, L may come either from a scalar PDE or a first-order system of PDEs. In the latter case, the considerations in this chapter can be extended to systems as long as the occurrences of $L^2(\Omega)$ are replaced by the appropriate product L^2 spaces with their respective product L^2 norms and the assumptions are satisfied. This also suggests a subject of further investigations.

Chapter 5

Closing Remarks

This dissertation proposed and studied a few dual least-squares finite element methods for hyperbolic partial differential equations. Hard problems, both linear and nonlinear, were considered that yield solutions with sharp layers and discontinuities. Theory and numerical results were presented.

The method based on the Helmholtz decomposition was introduced and studied first as an extension to nonlinear balance laws of the ideas in [32, 34], that were introduced in the context of conservation laws. The formulation is naturally related to the notion of a weak solution and possesses important numerical conservation properties, as shown analytically, and accordingly demonstrates good shock-capturing capabilities. It also provides correct approximations to other nonlinear behaviors like rarefactions. The challenging convergence properties of the method were also discussed. Note that the ability of this approach to correctly and automatically detect and accurately approximate shocks is very important for its applicability. The numerical examples illustrate that it recovers accurately the speed of the shock. Also, it is a quite positive feature of the method that it possesses the potential to be extended to systems and higher-dimensional domains, while maintaining the natural connection to the notion of a weak solution. This would provide an interesting holistic approach to nonlinear first-order hyperbolic systems.

Next, the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method was proposed and studied for linear hyperbolic problems. This is a novel method that utilizes the ideas in the standard $\mathcal{L}\mathcal{L}^*$ approach in an intriguing way to obtain L^2-norm approximations on common finite element spaces. The properties of the method are studied rigorously. Also, the $\mathcal{L}\mathcal{L}^*$-type formulations, which are simplified versions of the so-called hybrid least-squares approach, are analyzed and studied. The particular technique in the

analysis of the convergence of the $\mathcal{L}\mathcal{L}^*$-type methods is interesting. Namely, the initial systems induced by the formulations are reduced, obtaining Schur complements and respective minimization principles regarding only the variable of interest. The error estimates are then obtained that in a consistent way compile the factors that affect the convergence. Moreover, the same auxiliary space that, in the $(\mathcal{L}\mathcal{L}^*)^{-1}$ approach, serves to approximate the action of an operator, in the context of the $\mathcal{L}\mathcal{L}^*$-type methods is used to actually approximate the solution to the particular PDE (partial differential equation). As a consequence, the $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation provides better approximations of the solution in the L^2 norm than the $\mathcal{L}\mathcal{L}^*$-type methods in similar configurations. This suggests that having a discrete approximation of good quality for the particular operator and the respective inf-sup stability are less demanding than having an auxiliary space with good approximation properties. The improved error in the $(\mathcal{L}\mathcal{L}^*)^{-1}$ method comes at the cost of more challenging linear systems of equations.

Overall, the least-squares approach provides an attractive framework for a variety of problems, including hyperbolic equations. This thesis contributes to the already known versatility and flexibility of least-squares methods, particularly concentrating on hyperbolic PDEs and, thus, continuing the effort in the field, presenting novel approaches, and adding to the already existing work in publications like [32, 33, 34]. In particular, in Chapter 3, the standard notion of a weak solution is integrated appropriately in a least-squares setting to obtain a method with the desired conservation property, while, in Chapter 4, the $(\mathcal{L}\mathcal{L}^*)^{-1}$ and $\mathcal{L}\mathcal{L}^*$-type formulations provide an interesting least-squares reformulations and approximations of the generally infeasible L^2-norm minimization. The analysis included in the dissertation is quite comprehensive.

There is a vast possibility of future work and further development. Many suggestions were already mentioned in the previous chapters. The following are some particular ideas:

- This thesis is mostly focused on methods for discretizing hyperbolic PDEs and linear solvers are not studied in detail. An important future work is on the development and the comprehensive study of linear solvers for the proposed methods. In particular, the main difficulty with the $(\mathcal{L}\mathcal{L}^*)^{-1}$ system is approximating the respective Schur complement in the cases when the inf-sup condition does not hold uniformly. While the Schur complement is not terribly ill-conditioned, it is global and dense. Hence, a preconditioner that requires access to the matrix

itself or local information about the matrix cannot be obtained directly. One interesting idea that deserves investigation is the utilization of the known static condensation (hybridization) technique [87] to "sparsify" the Schur complement and, thus, reduce the problem to solving linear systems with sparse and symmetric positive definite matrices – a setting that is considerably more manageable.

- Another important future study is the applicability of the approaches to hyperbolic systems and higher-dimensional domains. The $(\mathcal{L}\mathcal{L}^*)^{-1}$ and $\mathcal{L}\mathcal{L}^*$-type methods are developed in a quite general setting, which can potentially make them naturally applicable to systems of hyperbolic PDEs. On the other hand, the approach based on the Helmholtz decomposition may need more work to obtain practical implementations of the respective formulations. The basic idea is to utilize an appropriate and general Helmholtz decomposition to deduce the respective least-squares principle.

- It is interesting to address the entropy conditions in the context of the method in Chapter 3. The main questions are if they need to be imposed explicitly and how this can be achieved in a least-squares setting.

- Extending the methods by upwind techniques would be of practical value, since it can reduce some nonphysical oscillations in the solutions.

- Hyperbolic problems can be associated with wave propagation with finite speed. In this context, the dispersive and dissipative properties [9] of the least-squares methods are still unknown but important as they describe how waves of different frequencies are handled.

- Inf-sup stability of the $(\mathcal{L}\mathcal{L}^*)^{-1}$ formulation needs more investigation.

- The norm convergence of the method in Chapter 3 is still an open and challenging question. The considerations here and in [32] provide some foundation for a possible future study that can lead to obtaining a comprehensive convergence proof.

- Applying the $(\mathcal{L}\mathcal{L}^*)^{-1}$ and $\mathcal{L}\mathcal{L}^*$-type methods to the solution of nonlinear hyperbolic problems is also a topic of further investigation.

- Adaptive mesh refinement and the utilization of the ideas of nested iteration and full multigrid are an important topic of research as it would improve on the practical applicability of the methods.

In conclusion, this dissertation describes novel methods and provides ground for future development.

Bibliography

[1] Randall J LeVeque. *Numerical Methods for Conservation Laws*. Lectures in Mathematics: ETH Zürich. Birkhäuser, Basel, 2nd edition, 1992.

[2] Randall J LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press, New York, 2002.

[3] Edwige Godlewski and Pierre-Arnaud Raviart. *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, volume 118 of *Applied Mathematical Sciences*. Springer, New York, 1996.

[4] Eleuterio F Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics: A Practical Introduction*. Springer, Berlin, Heidelberg, 3rd edition, 2009.

[5] Marek Brandner, Jiří Egermaier, and Hana Kopincová. Numerical Schemes for Hyperbolic Balance Laws - Applications to Fluid Flow Problems. In Radostina Petrova, editor, *Finite Vol. Method - Powerful Means Eng. Des.*, chapter 2, pages 35–60. InTech, 2012.

[6] Elmer E Lewis and Warren F Miller. *Computational Methods of Neutron Transport*. American Nuclear Society, La Grange Park, IL, 1993.

[7] Thomas A Manteuffel, Klaus J Ressel, and Gerhard Starke. A Boundary Functional for the Least-Squares Finite-Element Solution of Neutron Transport Problems. *SIAM J. Numer. Anal.*, 37(2):556–586, 2000.

[8] C K Birdsall and A B Langdon. *Plasma Physics via Computer Simulation*. Series in Plasma Physics. Taylor & Francis, Boca Raton, 2004.

[9] G B Whitham. *Linear and Nonlinear Waves*. Pure and Applied Mathematics. Wiley, New York, 1974.

[10] Edwige Godlewski and Pierre-Arnaud Raviart. *Hyperbolic systems of conservation laws*, volume 3/4 of *Mathematiques & Applications*. Ellipses, Paris, 1991.

[11] Peter D Lax. *Hyperbolic Partial Differential Equations*, volume 14 of *Courant Lecture Notes in Mathematics*. American Mathematical Society, 2006.

[12] Peter Lax and Burton Wendroff. Systems of conservation laws. *Commun. Pure Appl. Math.*, 13(2):217–237, 1960.

[13] Daniele Antonio Di Pietro and Alexandre Ern. *Mathematical Aspects of Discontinuous Galerkin Methods*, volume 69 of *Mathématiques et Applications*. Springer, Berlin, Heidelberg, 2012.

[14] G J Le Beau, S E Ray, S K Aliabadi, and T E Tezduyar. SUPG finite element computation of compressible flows with the entropy and conservation variables formulations. *Comput. Methods Appl. Mech. Eng.*, 104(3):397–422, 1993.

[15] Pavel B Bochev and Jungmin Choi. A Comparative Study of Least-squares, SUPG and Galerkin Methods for Convection Problems. *Int. J. Comut. Fluid Dyn.*, 15(2):127–146, 2001.

[16] Dietmar Kröner. *Numerical Schemes for Conservation Laws*. Advances in Numerical Mathematics. Wiley, Teubner, Chichester, Stuttgart, 1997.

[17] Claes Johnson, Uno Nävert, and Juhani Pitkäranta. Finite element methods for linear hyperbolic problems. *Comput. Methods Appl. Mech. Eng.*, 45(1):285–312, 1984.

[18] F. Brezzi, L. D. Marini, and E. Süli. Discontinuous Galerkin Methods for First-Order Hyperbolic Problems. *Math. Model. Methods Appl. Sci.*, 14(12):1893–1903, 2004.

[19] Augusto Cesar Galeão and Eduardo Gomes Dutra do Carmo. A consistent approximate upwind Petrov-Galerkin method for convection-dominated problems. *Comput. Methods Appl. Mech. Eng.*, 68(1):83–95, 1988.

[20] Thomas J R Hughes, Leopoldo P Franca, and Gregory M Hulbert. A new finite element formulation for computational fluid dynamics: VIII. The galerkin/least-squares method for advective-diffusive equations. *Comput. Methods Appl. Mech. Eng.*, 73(2):173–189, 1989.

[21] Leopoldo P Franca, Sergio L Frey, and Thomas J R Hughes. Stabilized finite element methods: I. Application to the advective-diffusive model. *Comput. Methods Appl. Mech. Eng.*, 95(2):253–276, 1992.

[22] R Verfürth. Robust A Posteriori Error Estimates for Stationary Convection-Diffusion Equations. *SIAM J. Numer. Anal.*, 43(4):1766–1782, 2005.

[23] Blanca Ayuso and L Donatella Marini. Discontinuous Galerkin Methods for Advection-Diffusion-Reaction Problems. *SIAM J. Numer. Anal.*, 47(2):1391–1420, 2009.

[24] G J Le Beau and T E Tezduyar. Finite element computation of compressible flows with the SUPG formulation. In *Adv. Finite Elem. Anal. Fluid Dyn.*, volume 123 of *FED*, pages 21–27. ASME, 1991.

[25] BLAST: High-Order Finite Element Hydrodynamics. `http://www.llnl.gov/CASC/blast`.

[26] V A Dobrev, T E Ellis, Tz. V Kolev, and R N Rieben. Curvilinear finite elements for Lagrangian hydrodynamics. *Int. J. Numer. Methods Fluids*, 65(11-12):1295–1310, 2011.

[27] Veselin A Dobrev, Tzanio V Kolev, and Robert N Rieben. High-Order Curvilinear Finite Element Methods for Lagrangian Hydrodynamics. *SIAM J. Sci. Comput.*, 34(5):B606–B641, 2012.

[28] R Anderson, V Dobrev, Tz. Kolev, D Kuzmin, M Quezada de Luna, R Rieben, and V Tomov. High-order local maximum principle preserving (MPP) discontinuous Galerkin finite element method for the transport equation. *J. Comput. Phys.*, 334:102–124, 2017.

[29] Pavel B Bochev and Max D Gunzburger. *Least-Squares Finite Element Methods*, volume 166 of *Applied Mathematical Sciences*. Springer, New York, 2009.

[30] Bo-nan Jiang. *The Least-Squares Finite Element Method: Theory and Applications in Computational Fluid Dynamics and Electromagnetics*. Scientific Computation. Springer, Berlin, Heidelberg, 1998.

[31] Pavel B Bochev and Max D Gunzburger. Finite Element Methods of Least-Squares Type. *SIAM Rev.*, 40(4):789–837, 1998.

[32] Luke N Olson. *Multilevel Least-Squares Finite Element Methods for Hyperbolic Partial Differential Equations*. PhD thesis, University of Colorado at Boulder, Department of Applied Mathematics, 2003.

[33] H De Sterck, Thomas A Manteuffel, Stephen F McCormick, and Luke Olson. Least-Squares Finite Element Methods and Algebraic Multigrid Solvers for Linear Hyperbolic PDEs. *SIAM J. Sci. Comput.*, 26(1):31–54, 2004.

[34] H De Sterck, Thomas A Manteuffel, Stephen F McCormick, and Luke Olson. Numerical Conservation Properties of H(div)-Conforming Least-Squares Finite Element Methods for the Burgers Equation. *SIAM J. Sci. Comput.*, 26(5):1573–1597, 2005.

[35] P B Bochev and J Choi. Improved Least-squares Error Estimates for Scalar Hyperbolic Problems. *Comput. Methods Appl. Math.*, 1(2):115–124, 2001.

[36] Gerhard Starke. A First-Order System Least Squares Finite Element Method for the Shallow Water Equations. *SIAM J. Numer. Anal.*, 42(6):2387–2407, 2005.

[37] Graham F Carey and B N Jiang. Least-squares finite elements for first-order hyperbolic systems. *Int. J. Numer. Methods Eng.*, 26(1):81–93, 1988.

[38] Paul Houston, Max Jensen, and Endre Süli. hp-Discontinuous Galerkin Finite Element Methods with Least-Squares Stabilization. *J. Sci. Comput.*, 17(1):3–25, 2002.

[39] Ulrich Trottenberg, Cornelis W Oosterlee, and Anton Schüller. *Multigrid*. Academic Press, San Diego, 2001.

[40] Vivette Girault and Pierre-Arnaud Raviart. *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*, volume 5 of *Springer Series in Computational Mathematics*. Springer, Berlin, Heidelberg, 1986.

[41] Daniele Boffi, Franco Brezzi, and Michel Fortin. *Mixed Finite Element Methods and Applications*, volume 44 of *Springer Series in Computational Mathematics*. Springer, Berlin, Heidelberg, 2013.

[42] R Hiptmair. Finite elements in computational electromagnetism. *Acta Numer.*, 11:237–339, 2002.

[43] Z Cai, T A Manteuffel, S F McCormick, and J Ruge. First-Order System $\mathcal{LL}^*$ (FOSLL*): Scalar Elliptic Partial Differential Equations. *SIAM J. Numer. Anal.*, 39(4):1418–1445, 2001.

[44] K Liu, T A Manteuffel, S F McCormick, J W Ruge, and L Tang. Hybrid First-Order System Least Squares Finite Element Methods with Application to Stokes Equations. *SIAM J. Numer. Anal.*, 51(4):2214–2237, 2013.

[45] James H. Bramble, Raytcho D. Lazarov, and Joseph E. Pasciak. A least-squares approach based on a discrete minus one inner product for first order systems. *Math. Comput.*, 66(219):935–955, 1997.

[46] Z Cai, R Lazarov, T A Manteuffel, and S F McCormick. First-Order System Least Squares for Second-Order Partial Differential Equations: Part I. *SIAM J. Numer. Anal.*, 31(6):1785–1799, 1994.

[47] Zhiqiang Cai, Thomas A Manteuffel, and Stephen F McCormick. First-Order System Least Squares for Second-Order Partial Differential Equations: Part II. *SIAM J. Numer. Anal.*, 34(2):425–454, 1997.

[48] Zhiqiang Cai, Thomas A. Manteuffel, and Stephen F. McCormick. First-Order System Least Squares for Velocity-Vorticity-Pressure Form of the Stokes Equations, with Application to Linear Elasticity. *Electron. Trans. Numer. Anal.*, 3:150–159, 1995.

[49] Z Cai, T A Manteuffel, and S F McCormick. First-Order System Least Squares for the Stokes Equations, with Application to Linear Elasticity. *SIAM J. Numer. Anal.*, 34(5):1727–1741, 1997.

[50] P Bochev, Z Cai, T A Manteuffel, and S F McCormick. Analysis of Velocity-Flux First-Order System Least-Squares Principles for the Navier–Stokes Equations: Part I. *SIAM J. Numer. Anal.*, 35(3):990–1009, 1998.

[51] Pavel Bochev, Thomas A Manteuffel, and Stephen F McCormick. Analysis of Velocity-Flux Least-Squares Principles for the Navier–Stokes Equations: Part II. *SIAM J. Numer. Anal.*, 36(4):1125–1144, 1999.

[52] *First-order system least squares philosophy: Informal discussion of some advantages and disdavantages of First-Order System Least Squares (FOSLS).* manuscript by the FOSLS gang to provide basic understanding and encourage discussion.

[53] Sigeru Mizohata. *The Theory of Partial Differential Equations.* Cambridge University Press, 1973.

[54] Lawrence C Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics.* American Mathematical Society, 2nd edition, 2010.

[55] Michael Renardy and Robert C Rogers. *An Introduction to Partial Differential Equations*, volume 13 of *Texts in Applied Mathematics.* Springer, New York, 2nd edition, 2004.

[56] Fritz John. *Partial Differential Equations*, volume 1 of *Applied Mathematical Sciences.* Springer, New York, 4th edition, 1982.

[57] Robert C McOwen. *Partial Differential Equations: Methods and Applications.* Prentice Hall, Upper Saddle River, 2nd edition, 2003.

[58] Susanne C Brenner and L Ridgway Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics.* Springer, New York, 3rd edition, 2008.

[59] Philippe G Ciarlet. *The Finite Element Method for Elliptic Problems*, volume 40 of *Classics in Applied Mathematics.* Society for Industrial and Applied Mathematics, reprint edition, 2002.

[60] Dietrich Braess. *Finite Elements: Theory, fast solvers, and applications in solid mechanics*. Cambridge University Press, 3rd edition, 2007.

[61] Alexandre Ern and Jean-Luc Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences*. Springer, New York, 2004.

[62] Peter Monk. *Finite Element Methods for Maxwell's Equations*. Numerical Mathematics and Scientific Computation. Clarendon Press, Oxford, 2003.

[63] Markus Berndt, Thomas A Manteuffel, and Stephen F McCormick. Local error estimates and adaptive refinement for first-order system least squares (FOSLS). *Electron. Trans. Numer. Anal.*, 6:35–43, 1997.

[64] H De Sterck, T Manteuffel, S McCormick, J Nolting, J Ruge, and L Tang. Efficiency-based h- and hp-refinement strategies for finite element methods. *Numer. Linear Algebr. with Appl.*, 15(2-3):89–114, 2008.

[65] J H Adler, T A Manteuffel, S F McCormick, J W Nolting, J W Ruge, and L Tang. Efficiency Based Adaptive Local Refinement for First-Order System Least-Squares Formulations. *SIAM J. Sci. Comput.*, 33(1):1–24, 2011.

[66] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[67] J H Adler and P S Vassilevski. Improving Conservation for First-Order System Least-Squares Finite-Element Methods. In Oleg P Iliev, Svetozar D Margenov, Peter D Minev, Panayot S Vassilevski, and Ludmil T Zikatanov, editors, *Numerical Solution of Partial Differential Equations: Theory, Algorithms, and Their Applications: In Honor of Professor Raytcho Lazarov's 40 Years of Research in Computational Methods and Applied Mathematics*, volume 45 of *Springer Proceedings in Mathematics & Statistics*, pages 1–19, New York, 2013. Springer.

[68] J H Adler and P S Vassilevski. Error Analysis for Constrained First-Order System Least-Squares Finite-Element Methods. *SIAM J. Sci. Comput.*, 36(3):A1071–A1088, 2014.

[69] Peter D Lax. Weak solutions of nonlinear hyperbolic equations and their numerical computation. *Commun. Pure Appl. Math.*, 7(1):159–193, 1954.

[70] Thomas Y Hou and Philippe G Le Floch. Why Nonconservative Schemes Converge to Wrong Solutions: Error Analysis. *Math. Comput.*, 62(206):497–530, apr 1994.

[71] Jindřich Nečas. *Direct Methods in the Theory of Elliptic Equations*. Springer Monographs in Mathematics. Springer, Berlin, Heidelberg, 2012.

[72] Philippe G Ciarlet. *Linear and Nonlinear Functional Analysis with Applications*. Society for Industrial and Applied Mathematics, 2013.

[73] Jacques-Louis Lions and Enrico Magenes. *Non-Homogeneous Boundary Value Problems and Applications v. I*, volume 181 of *Die Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 1972.

[74] Robert Dautray and Jacques-Louis Lions. *Mathematical Analysis and Numerical Methods for Science and Technology: Volume 2 Functional and Variational Methods*. Springer, Berlin, Heidelberg, 2000.

[75] S Diehl. Scalar conservation laws with discontinuous flux function: I. The viscous profile condition. *Commun. Math. Phys.*, 176(1):23–44, feb 1996.

[76] Jean-Marie Emmanuel Bernard. Density results in Sobolev spaces whose elements vanish on a part of the boundary. *Chinese Ann. Math. Ser. B*, 32(6):823, 2011.

[77] Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer, New York, 2011.

[78] J E Dennis and Robert B Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, volume 16 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, reprint edition, 1996.

[79] Jorge Nocedal and Stephen J Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2nd edition, 2006.

[80] Beixiang Fang, Pingfan Tang, and Ya-Guang Wang. The Riemann problem of the Burgers equation with a discontinuous source term. *J. Math. Anal. Appl.*, 395(1):307–335, 2012.

[81] T. A. Manteuffel, S. Münzenmaier, and B. S. Southworth. Scaling and Solving the Self-Adjoint Form for Steady-State Transport, (in preparation).

[82] T A Manteuffel, S F McCormick, J Ruge, and J G Schmidt. First-Order System $\mathcal{LL}^*$ (FOSLL*) for General Scalar Elliptic Problems in the Plane. *SIAM J. Numer. Anal.*, 43(5):2098–2120, 2005.

[83] L Demkowicz and J Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. Part I: The transport equation. *Comput. Methods Appl. Mech. Eng.*, 199(23–24):1558–1572, 2010.

[84] L Demkowicz and J Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. II. Optimal test functions. *Numer. Methods Partial Differ. Equ.*, 27(1):70–105, 2011.

[85] Kôsaku Yosida. *Functional Analysis*. Classics in Mathematics. Springer, Berlin, Heidelberg, reprint edition, 1995.

[86] Ivo Babuška. Error-bounds for finite element method. *Numer. Math.*, 16(4):322–333, 1971.

[87] J. E. Roberts and J.-M. Thomas. Mixed and Hybrid Methods. In *Finite Element Methods*, Handbook of Numerical Analysis II, (Eds. P. Ciarlet and J. Lions), Elsevier/North Holland, Amsterdam, pages 523–639, 1991.

[88] David S Watkins. *Fundamentals of Matrix Computations*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs, and Tracts. Wiley, 3rd edition, 2010.

[89] Michel Fortin. An analysis of the convergence of mixed finite element methods. *RAIRO. Anal. numérique*, 11(4):341–354, may 1977.

[90] Kent-Andre Mardal and Ragnar Winther. Preconditioning discretizations of systems of partial differential equations. *Numer. Linear Algebr. with Appl.*, 18(1):1–40, 2011.

[91] Panayot S Vassilevski. *Multilevel Block Factorization Preconditioners: Matrix-based Analysis and Algorithms for Solving Finite Element Equations*. Springer, New York, 2008.

[92] Thomas A Manteuffel, Luke N Olson, Jacob B Schroder, and Ben S Southworth. A Root-Node–Based Algebraic Multigrid Method. *SIAM J. Sci. Comput.*, 39(5):S723–S756, 2017.

[93] Anders Logg, Kent-Andre Mardal, Garth N Wells, and Others. *Automated Solution of Differential Equations by the Finite Element Method: The FEniCS Book*, volume 84 of *Lecture Notes in Computational Science and Engineering*. Springer, Berlin, Heidelber, 2012.

[94] Satish Balay, Shrirang Abhyankar, Mark∼F. Adams, Jed Brown, Peter Brune, Kris Buschelman, Lisandro Dalcin, Victor Eijkhout, William∼D. Gropp, Dinesh Kaushik, Matthew∼G. Knepley, Lois Curfman McInnes, Karl Rupp, Barry∼F. Smith, Stefano Zampini, Hong Zhang, and Hong Zhang. PETSc Web page. `http://www.mcs.anl.gov/petsc`, 2016.

[95] P R Amestoy, I S Duff, J Koster, and J.-Y. L'Excellent. A Fully Asynchronous Multifrontal Solver Using Distributed Dynamic Scheduling. *SIAM J. Matrix Anal. Appl.*, 23(1):15–41, 2001.

[96] Yousef Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, 2nd edition, 2003.

[97] Hank A van der Vorst. *Iterative Krylov Methods for Large Linear Systems*, volume 13 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, 2003.

[98] Leopoldo P Franca, Thomas J R Hughes, and Rolf Stenberg. Stabilized Finite Element Methods. In Max D Gunzburger and Roy A Nicolaides, editors, *Incompressible Comput. Fluid Dyn. Trends Adv.*, pages 87–108. Cambridge University Press, 1993.