

A discourse on variational and geometric aspects of stability of discretizations

Pavel Bochev^{1,2}

*Sandia National Laboratories
Albuquerque, NM 87185, USA*

¹Computational Mathematics and Algorithms Department, Sandia National Laboratories, Albuquerque, NM 87185-1110 (pbboche@sandia.gov).

²This work was funded by the Applied Mathematical Sciences program, U.S. Department of Energy, Office of Energy Research and performed at Sandia National Labs, a multiprogram laboratory operated by Sandia Corporation, a Lockheed-Martin Company, for the U.S. DOE under contract number DE-AC-94AL85000.

Abstract

These lectures are devoted to variational and geometrical aspects of stable discretizations for Partial Differential Equations problems.

Variational principles have been in the arsenal of finite element methods since their inception in the early fifties. They are a powerful tool for stability and error analysis, and indeed, variational principles have remained unsurpassed in their ability to generate sharp error estimates. One of the main reasons for the tremendous success of variational methods lies in the fundamental connection that exists between variational principles, on one hand and optimization problems and the structure of PDE's on the other hand.

Differential complexes serve to provide another tool that can be used to encode PDE structure. Differential forms model global quantities rather than fields, and in many ways their formalism is closer to the first principles used to describe physical phenomena. Two decades ago Bossavit pointed out that differential forms can and should be used to analyze and develop discretizations that mimic the topological structure of the underlying PDE. This viewpoint proved to be the key to successful application of finite element methods in electromagnetics.

Since then there has been an increased interest in the use of such *geometrical* methods for the discretization of PDE's. The main goal of these lectures is to show how variational and geometric approaches can complement each other in the quest for accurate and stable discretizations.

We begin with a review of basic facts about variational methods and take time to consider three special cases of variational settings. Then we provide examples of finite element methods in each setting and discuss their stability. In particular, we show how variational principles associated with a given PDE are propagated to the discrete problems, thus forming the basis of the *variational approach* to stability of discretizations.

The lectures contain a necessarily brief introduction into the elements of differential forms calculus. After these preliminaries we embark on a mission to apply this formalism to describe topological structure of PDE problems. We choose the Kelvin principle as a model problem and derive factorization diagrams for the associated first-order optimality system. Using these diagrams we define *compatible* discretizations of the PDE equation and show how the structure encoded in the factorization diagram is being propagated through a wide range of different discretizations, thus forming the basis of the *geometrical approach* to stability of discretizations.

The last part of these lectures highlights the fundamental connection between geometrical structure of PDEs and variational characterizations of their stable discretizations. Our examples are based on the grid decomposition property and the commuting diagram property for the Kelvin principle.

We conclude with examples of alternative discretization methods that are capable of circumventing the rigid constraints imposed by geometrical structures upon finite dimensional representations of PDEs and explain why such methods are useful.

1 Introduction

Partial Differential Equations (PDE) are a fundamental modeling tool in science and engineering. Their applications range from design and modeling of semiconductor devices to global climate simulations. As a result, approximate numerical solution of PDE's is a task of tremendous practical importance.

A numerical solution of a PDE problem, which we will write symbolically as

$$\mathcal{L}U = F \tag{1}$$

involves two principal ingredients:

a *discretization* step, wherein the continuous problem is replaced by a finite dimensional algebraic equation

$$\mathcal{L}^h U^h = F^h \tag{2}$$

and

a *solution* step, wherein the algebraic problem (2) is solved either by a direct or an iterative solution method.

In (2) we follow the accepted custom of using h to denote a small positive parameter whose reciprocal h^{-1} is related to the dimension of the space where U^h is sought.

In these lecture notes our main focus will be on the first, discretization step and the three dominant paradigms that exist in the construction of (2):

- finite element methods where *projection* or *quasi-projection* principles are restricted to finite dimensional subspaces;
- finite difference methods where *differential operators* are approximated by algebraic operators;
- finite volume methods where *integral fluxes* are approximated by quadrature.

While both ingredients are vital to the success of computer simulations, it is the discretization step that ultimately holds the key to all fundamental properties, both desirable and undesirable, of any numerical method for PDE's. Our main goal will be to explain why some discretization choices work well, while some other perform poorly or lead to downright disasters. Such an undertaking cannot be accomplished without a keen appreciation of the mathematical structure of the PDE. This structure governs well-posedness of the PDE and reflects intrinsic properties of the physical phenomena that are being modeled, e.g., conservation laws, solution symmetries, positivity, and maximum principles. Importance of well-posedness has been noted long before the dawn of the computer age by Maxwell who in 1873 wrote³

³J. C. Maxwell. *Does the progress of Physical Science tend to give any advantage to the opinion of Necessity (or Determinism) over that of the Contingency of Events and the Freedom of the Will?* in [64, pp.434-463].



There are certain classes of phenomena, as I have said, in which a small error in the data only introduces a small error in the result. Such are, among others, the larger phenomena of the Solar System, and those in which the more elementary laws in Dynamics contribute the greater part of the result. The course of events in these cases is stable.

Of course, given a well-posed PDE problem we seek the same in its discrete counterpart, i.e., we would like small changes in the discrete data F^h to effect only small changes in the solution U^h so that the course of the discrete events also remains stable. In addition we also want to be able to make the difference between U^h and U as small as we wish by taking smaller and smaller values of h .

1.1 Importance of structure

Common sense dictates that a well-behaved discretization must somehow represent or reproduce the portion of the mathematical structure in (1) that is responsible for the well-posedness of this problem. For many years physical intuition served as a trusted guide in the development of successful discretization schemes, while mathematical rigor and understanding usually trailed behind. Numerical literature abounds with examples of methods that were discovered by imitating, albeit intuitively, the underlying physics, and which worked well even though initially the reasons for their superiority over other schemes were not fully understood. Classical examples are the Finite Element method in structural analysis, Yee's FDTD scheme [83], MacCormack's scheme [62], and the Box Integration method.

On the mathematical side, analytical techniques within each one of the discretization paradigms took on different routes and evolved into fairly self-contained disciplines with few if any overlaps. Each discipline relied upon its own set of analytical tools to identify the basic requirements for obtaining well-posed discrete problems in (2). For instance, finite element methods rely upon variational principles and their analysis draws upon the rich theory of Hilbert spaces and such powerful results as Riesz representation theorem. Compared to finite elements, finite difference and finite volume methods have been less amenable to functional analysis tools. However, regardless of the differences in analytic approaches in all cases we find that

$$\text{CONSISTENCY} + \text{STABILITY} = \text{CONVERGENCE}. \quad (3)$$

This is the celebrated *Lax equivalence theorem*; see [61, p.142] which is a result that transcends many of the boundaries in numerical analysis and remains valid in contexts other than numerical PDE's. *Convergence* signifies the fact that the approximate solution U^h can be made arbitrarily close to the exact solution U , provided h is small enough, which of course is the coveted prize in any numerical method. The recipe for this, according to (3) is to ensure that (2) is *stable* and *consistent*; i.e., that \mathcal{L}^h has a bounded inverse with respect to some norm and that \mathcal{L}^h is close to \mathcal{L} , again with respect to some norm. Then, if F^h is close to F , recovery of U^h from F^h will introduce only a small amount of error. Moreover, since $(\mathcal{L}^h)^{-1}$ is bounded, we can make this error as small as we wish by taking smaller h , i.e., the method is convergent.

It is important to recognize that, while not completely orthogonal, consistency is *not* synonymous with stability. It is wrong to assume that just by constructing an \mathcal{L}^h that is close to \mathcal{L} , we will automatically ensure that $(\mathcal{L}^h)^{-1}$ is well-behaved. In fact, consistency and stability capture different aspects of the discretization process and require each other to validate (3). This fact is universal and can be observed across all discretization platforms regardless of their individual differences. Another universal and equally remarkable fact is that "unstable" and "stable" discretizations constructed under different discretization paradigms are strikingly similar.

For instance, Galerkin method for scalar hyperbolic equations is a textbook example of a consistent but unstable finite element discretization. Its finite difference analogue is the central difference scheme. In both cases discrete solutions develop spurious oscillations unless the exact solutions is globally smooth. Also, in both cases a stable discretization can be obtained by introducing an upwind bias either in the finite element weight function or in the finite difference approximation of the derivatives.

Another example is the mixed finite element for the Stokes problem implemented with equal order interpolation for the velocity and the pressure variables. The finite difference analogue is a collocated scheme where pressure and velocity derivatives are approximated by the same stencils. In both cases the discrete problem is unstable and gives "strange" results. The fix in finite elements and finite difference contexts is essentially the same - pressure and velocity approximations are separated - what differs is the implementation of the separation which necessarily follows the discretization paradigms. In finite element methods it is effected by approximation of the variables by finite element spaces defined with respect to different triangulations, or having different polynomial degrees. In finite differences the separation process amounts to a use of a staggered scheme where pressure and velocity are placed at different mesh locations.

The striking similarity shared across different discretization platforms by troublesome discretizations on one hand, and their stable counterparts on the other hand, is not accidental. It highlights the complementary roles played by consistency and stability. Consistency is most closely related to the metric aspects of the discretization. A consistent method will only introduce a small error per time step or per spatial cell. However, this is not enough to capture completely structural or topological properties that may govern stability of the continuous PDE problem (and which are ultimately determined by the physics of the process that is being modeled by the PDE). The fact that this structure is relevant to the well-posedness of the discrete problem is clear by noting that in spite of the differences between FEM, FD and FV, stable discretizations in all cases are forced into what is essentially the same pattern of variable approximations. A close inspection of stable discrete models across a wide range of scientific and engineering applications will only confirm this fact!

1.2 Discovery of structure

Having agreed that structure is important the obvious questions are what components of this structure are relevant in the discrete world, and how to make our discretization compatible with them? In some cases the answer can be easily deduced directly from the physical process that is being modeled without resorting to more abstract mathematical

tools. For an example, consider the PDE

$$u_t(x, t) + F_x(u(x, t)) = 0 \quad \text{in } (0, 1). \quad (4)$$

This equation models, e.g., flow of a fluid through one dimensional pipe, and is an example of a *conservation law*. For any two points $0 < \underline{x} \leq \bar{x} < 1$ we have that

$$\int_{\underline{x}}^{\bar{x}} u_t dx = F(u(\bar{x}, t)) - F(u(\underline{x}, t)). \quad (5)$$

Equation (5) is merely a mathematical statement of the fact that the total change of fluid mass in any section of the pipe will only depend on the *fluxes* at the endpoints, i.e., the amount of mass leaving and entering the section. A finite difference discretization of (4) that imitates this property is given by

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x} (F_{i+1/2}^n - F_{i-1/2}^n), \quad (6)$$

where $F_{i\pm 1/2}^n$ are *numerical fluxes*. This scheme is *conservative* in the sense that the discrete mass will only change due to the fluxes at the endpoints 0 and 1, while inside the domain the mass flowing from one cell to another is conserved. Note that this is accomplished by *staggering* the locations of the numerical fluxes and the density function. If these variables are collocated, then the scheme will not be conservative, i.e., discrete conservation is a consequence of the grid topology. It is well-known that discrete conservation is critical for accurate calculation of shock waves; see [61, p.237]. However, conservation alone is not sufficient to provide stable and convergent approximations!

Indeed, conservation is a purely *topological* property of the scheme in (6) that is automatically satisfied as long as the variables remain staggered. Convergence on the other hand requires consistency which is a metric quality measured by the local truncation error in the flux calculation. To put it differently, in addition to *situating* the fluxes at the right places, we also need to *calculate* them accurately! Otherwise, (6) can be made arbitrarily inconsistent while remaining conservative.

The reason that (6) may not be stable is that so far we have captured just one of the relevant properties of (4), and have yet to account for other, equally important aspects of the physics behind conservation laws. If we track one of the particles in our one-dimensional pipe we will see that it moves with finite speed along a curve in the (x, t) plane. In mathematical terms we express this by saying that hyperbolic problems propagate information with finite speed along characteristics.

Consider now how these features affect stability of (6). Suppose we wish to compute $F_{1+1/2}^n$ using u_i^n and u_{i+1}^n . If the time step is too large, there will be enough time for the data from other cells to propagate to the flux location and make it dependent on more cell values. To prevent this from happening spatial and temporal discretization steps must satisfy the CFL (Courant, Friedrichs, Lewy) condition

$$\frac{\mathbf{v}\Delta t}{\Delta x} \leq 1 \quad (7)$$

where \mathbf{v} denotes the propagation speed. The CFL condition forces (6) to comply with the finite propagation speed by ensuring that numerical domain of dependence contains

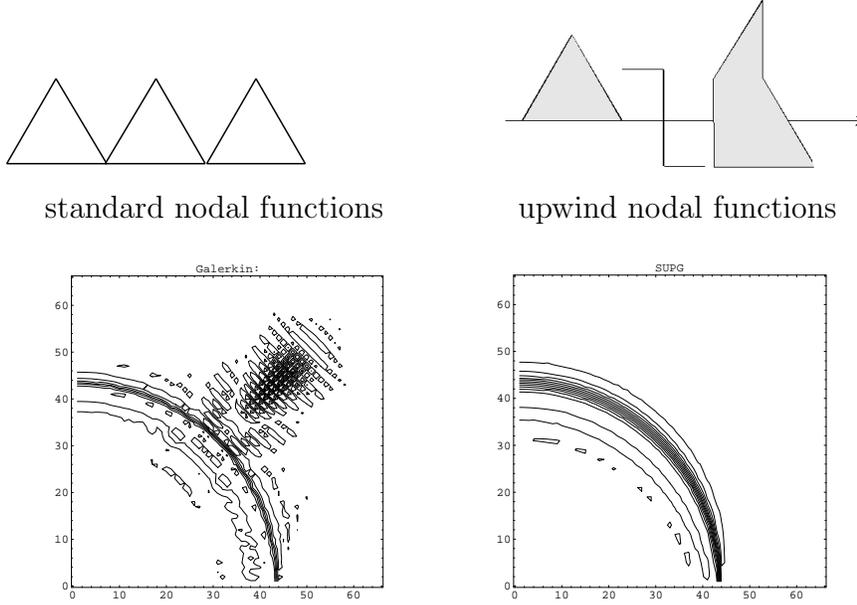


Figure 1: Galerkin vs. SUPG solution of circular advection problem

the true domain of dependence of (4); see [61, p.69]. Still, this is not enough to guarantee that (6) is stable, i.e., CFL is a necessary but not a sufficient stability condition.

The reason is simple: CFL does guarantee that the cells we use to compute the fluxes will contain all the necessary information, but it doesn't guarantee that this information will be used properly! To mimic the flow of information in (4) computation of fluxes in (6) must rely on data that comes from the "upwind" region, while information from the "downwind" regions is either neglected or given lesser importance. Failure to do so may lead to severe oscillations in the discrete solutions.

So far it appears that in a discretization of a problem like (4) mimicking its topology, i.e., the conservation, is less important for the stability than imitating the correct flow of information. For instance, a Galerkin method for the scalar equation

$$\nabla \cdot (\mathbf{b}\phi) = 0 \quad \text{in } \Omega; \quad \text{and} \quad \phi = \phi_- \quad \text{on } \Gamma_-,$$

where Γ_- is the *inflow* part of $\partial\Omega$, is given by the variational equation: seek $\phi \in \Phi^h$ such that $\phi = \phi_-$ on Γ_- and

$$\int_{\Omega} \nabla \cdot (\mathbf{b}\phi)\psi \, dx = 0 \tag{8}$$

for all $\psi \in \Phi^h$ such that $\psi = 0$ on Γ_- . If Φ^h is the usual "hat" function finite element space, solutions of (8) are not conservative⁴ and will develop spurious oscillations unless the solution is globally smooth. These oscillations are shown on the left side of Fig. 1.

⁴Continuous nodal Galerkin solutions are not conservative in the same sense as (6): there are no discrete fluxes that are conserved. However, Hughes et.al. [51] have shown that the weak equation (8) can be used to define *nodal* "fluxes" and that the sum of these fluxes is conserved on each element. Thus, continuous Galerkin method is locally conservative in the sense of nodal fluxes.

The Streamline Upwind Petrov-Galerkin (SUPG) formulation; see [47]

$$\int_{\Omega} \nabla \cdot (\mathbf{b}\phi)(\psi + h\mathbf{b} \cdot \nabla\psi) dx = 0 \quad (9)$$

is still not conservative but has improved stability, as can be seen on the right side of Fig. 1. The only difference between (8) and (9) is that the former uses weight functions that have upwind bias, see Fig. 1. We will return to this example in Section 3.3.

Ability to determine how different aspects of the mathematical structure affect discretizations is very important in practice. Besides being a prerequisite for well-posed discrete problems, this knowledge can help us solve problems more efficiently. For example, if we are in a setting where solutions of (4) will not develop shock waves, it may be more efficient to use a SUPG formulation implemented with high-order elements even though such a scheme is not conservative. However, if the goal is to compute correct shock positions and speeds, a better choice would be to use a conservative scheme.

Unfortunately, as the complexity of the mathematical structure of the PDE increases, the interplay between its components becomes more intricate and well-posedness of discrete models becomes more elusive.

1.3 Why do we need mathematicians?

So far we were able to imitate the physics of advection in our discrete model by using staggered grids to reproduce conservation, and by using upwinding to simulate the correct flow of information in the discrete problem. These solutions seem very logical, or at least easy to explain and justify to anyone who has observed, say the flow of water in a river.

However, not all PDE models divulge their structures so readily. As Maxwell pointed out, (see [64, pp.434-463])

There are other classes of phenomena which are more complicated, and in which cases of instability may occur, the number of such cases increasing, in an exceedingly rapid manner, as the number of variables increases.

Actually, even if the physics may seem simple enough, the PDE's that describe it may turn out to have a surprisingly rich structure. To illustrate this point consider the Poisson equation

$$-\Delta\phi = f \text{ in } \Omega \quad \text{and} \quad \phi = 0 \text{ on } \Gamma \quad (10)$$

and its equivalent first-order system form

$$\begin{cases} \nabla \cdot \mathbf{v} = f & \text{in } \Omega \\ \nabla\phi + \mathbf{v} = 0 & \text{in } \Omega \\ \phi = 0 & \text{on } \Gamma \end{cases} \quad (11)$$

A weak Galerkin form of (10) is to seek ϕ in some suitable function space Φ , such that

$$\int_{\Omega} \nabla\phi \cdot \nabla\psi dx = \int_{\Omega} f\psi dx \quad (12)$$

for all ψ in some other space Ψ . It is well-known, see [29], that a finite element solution of (12) that uses continuous nodal finite element spaces is stable and converges in the L^2 norm to all smooth exact solutions of (10) as $O(h^{r+1})$. Here r is the degree of the complete polynomials contained in the finite element space.

Situation with (11) is not so simple. One weak Galerkin form of (11) is to seek a pair (\mathbf{v}, ϕ) in some suitable function spaces such that

$$\begin{cases} \int_{\Omega} (\nabla \cdot \mathbf{v}) \xi \, dx = \int_{\Omega} f \xi \, dx \\ \int_{\Omega} \nabla \phi \cdot \mathbf{w} + \mathbf{v} \cdot \mathbf{w} \, dx = 0 \end{cases} \quad (13)$$

for all (\mathbf{w}, ξ) in some other function spaces. But (11) has other weak forms such as the div-div problem

$$\begin{cases} \int_{\Omega} (\nabla \cdot \mathbf{v}) \xi \, dx = \int_{\Omega} f \xi \, dx \\ \int_{\Omega} -\phi \nabla \cdot \mathbf{w} + \mathbf{v} \cdot \mathbf{w} \, dx = 0 \end{cases}, \quad (14)$$

the grad-div equations

$$\begin{cases} \int_{\Omega} -\mathbf{v} \cdot \nabla \xi \, dx = \int_{\Omega} f \xi \, dx \\ \int_{\Omega} -\phi \nabla \cdot \mathbf{w} + \mathbf{v} \cdot \mathbf{w} \, dx = 0 \end{cases}. \quad (15)$$

or the grad-grad formulation

$$\begin{cases} \int_{\Omega} -\mathbf{v} \cdot \nabla \xi \, dx = \int_{\Omega} f \xi \, dx \\ \int_{\Omega} \nabla \phi \cdot \mathbf{w} + \mathbf{v} \cdot \mathbf{w} \, dx = 0 \end{cases}. \quad (16)$$

It is clear that in seeking a finite element solution of (11) we face much more choices. First, we have to decide which one of the four weak forms (13)-(16) to use. Then we need to choose a pair of finite element spaces to approximate ϕ and \mathbf{v} . Here's a list of some possible choices:

1. use either one of (13)-(16) and continuous nodal elements for ϕ and \mathbf{v} ;
2. use (14), piecewise constants for ϕ and nodal elements for \mathbf{v} ;
3. use (14), nodal elements for ϕ and Raviart-Thomas elements [71] for \mathbf{v} ;
4. use (14), piecewise constants for ϕ and Raviart-Thomas elements for \mathbf{v} ;
5. use (16), nodal elements for ϕ and piecewise constant elements for \mathbf{v} .

It is clear that the list of discretization options can be continued indefinitely. However, it turns out that only few of the combinations between a variational problem and finite element spaces will lead to a stable finite element solution of (11)! Moreover, each stable and unstable finite element method will have its finite volume and finite difference twins. Thus, merely by switching from finite elements to, e.g., finite differences will not solve our difficulties.

What complicates the matters, compared to, e.g., (4), is that now well-posedness of (11) hinges on mathematical structures that are not so transparently connected to the physics that is being modeled by the PDE. As a result, the stable discretization choices are more obscure and "guessing" them right is not easy without first understanding what governs the well-posedness of (11), and then figuring out what does this mean for the discretization of this problem.

In these lectures we will investigate two different methodologies that can be used to address stability. The first one, discussed in Section 2, is based on *variational principles*. Variational principles are the mathematical foundation of finite element methods and rely upon results from functional analysis to give conditions for the well-posedness of the variational problems. In this approach, stability of discretizations is inferred from the type of the underlying variational principle and the fact that variational methods propagate this principle into the discrete equations.

The second methodology, presented in Section 4, has its roots in differential geometry and exterior calculus of differential forms. Differential forms are a powerful tool that can be used to separate topological and metric aspects of a PDE model. The result is a *factorization* diagram that expresses the PDE in terms of *equilibrium* equations and *constitutive* relations. These diagrams represent the foundations of geometrical modeling where they serve as templates for *compatible* discretizations. In this approach, stability of discretizations is assessed by measuring its conformity with a factorization diagram for the PDE.

In Section 7 we will talk about the fundamental links between geometrical structure of PDEs encoded in their factorization diagrams, and variational characterizations of their stability used in finite element methods. Then, in Section 8 we will explain why sometimes one is forced to consider discretizations that deviate from the stability rules laid down by geometrical or variational means. There we will briefly consider three popular classes of finite element methods that are designed to work in settings that violate these stability rules.

To save space and time we do not quote a number of results and definitions concerning Sobolev spaces and finite element approximation theory, and instead refer the reader to the monographs [1], [8], [29] and [41] for more detailed information on these subjects.

2 Variational approach to stability

This lecture will discuss variational methods for linear operator equations $Au = f$ in Hilbert spaces and conditions for their stable approximate solution. These abstract results are then applied to study stability of finite element methods.

2.1 Variational methods

A classical variational method attempts to solve $Au = f$ by turning it into a minimization problem. The idea is to construct, if possible, a functional whose minimizer coincides with the solution of the operator equation. Then, $Au = f$ is approximated by computing a minimizer out of a *finite dimensional* subspace. This minimizer is subject to a necessary condition (Euler-Lagrange equation) which is conveniently expressed in terms of bilinear forms and linear functionals. This setting is known as the classical Rayleigh-Ritz principle. Another optimization setting arises when the minimum is constrained by some linear constraint $\Lambda u = 0$. In this case, constrained minimization can be converted to the unconstrained problem of finding the saddle-points of a Lagrangian functional. Numerical approximation is obtained by restriction of the saddle-point optimality system (which is again given by a bilinear form and a linear functional) to finite dimensional subspaces for the state variable and the Lagrange multiplier.

However, the term “variational method” is routinely applied to methods that are not necessarily derived from optimization problems but still lead to equations expressed in terms of bilinear forms and linear functionals. A standard way to derive the variational equations in this case is to require that the inner product of the *residual* $Au - f$ and any function from a suitable *test space* vanishes. This formal “orthogonalization” approach and its modifications are known as Bubnov-Galerkin, Galerkin-Petrov, or Galerkin method of weighted residuals. To summarize, a variational method involves three steps:

- first, a variational problem is set up either by association of $Au = f$ with optimization, or by a formal Galerkin orthogonalization;
- next, an approximation is effected by restricting the variational equation to a finite dimensional subspace;
- last, a solution is computed by solving the ensuing finite dimensional equation.

The first step leads to a formal setting that can be described by a pair Hilbert spaces V_1 and V_2 , a linear functional $F : V_2 \mapsto \mathbb{R}$, and a bilinear form $Q(\cdot; \cdot) : V_1 \times V_2 \mapsto \mathbb{R}$. The variational problem consists of finding $u \in V_1$ such that

$$Q(u; v) = F(v) \quad \forall v \in V_2. \quad (17)$$

The second step introduces a pair of finite dimensional subspaces $V_1^h \subset V_1$ and $V_2^h \subset V_2$ and restricts (17) to these spaces, i.e., we seek $u^h \in V_1^h$ such that

$$Q(u^h; v^h) = F(v^h) \quad \forall v^h \in V_2^h. \quad (18)$$

Note that in (18) we have changed only the spaces but not the problem itself! This property is a hallmark of (conforming) variational methods, and it has some far reaching consequences for their stability and accuracy. Because a variational method computes solution of the original problem using a "smaller" space, we call such solutions *quasi-projections*. When (17) is associated with unconstrained minimization of a positive quadratic functional, then (18) defines a true inner product projection of the exact solution into V_1^h .

It is not hard to see that once bases for V_1^h and V_2^h are chosen, (18) is equivalent to a linear algebraic system

$$\mathbb{A}\mathbf{u} = \mathbf{f}. \tag{19}$$

where \mathbb{A} is a $\dim(V_2^h) \times \dim(V_1^h)$ matrix, and \mathbf{u} is a coefficient vector.

The space V_1 (resp. V_1^h) is called *trial* space and the space V_2 (resp. V_2^h) - test space. The choice of these spaces is very important because it governs the well-posedness of the variational problem and its discrete approximation. When V_1 and V_2 are finite dimensional their choice is limited to the selection of the dimensions - any two spaces of the same dimension are isomorphic. In infinite dimensions there is a much larger variety of spaces and choosing a "good" space is not trivial. For example,

- choosing V_1 to be a "large" space (e.g., functions with fewer derivatives) is good for proving existence because there is an abundant supply of candidate solutions. But choosing V_1 too large may lead to non-uniqueness and may admit nonphysical "spurious" modes that lead to instabilities.
- If V_1 is "small" (nice, smooth functions), it is easier to show uniqueness, but we may end up with a space that is too small to contain a solution. Also, a nice smooth space may be difficult to approximate conformingly with simple functions such as piecewise polynomials.

For the remainder of this section we will focus on conditions, expressed in terms of bilinear forms, that will quantify what it means to have spaces that are not "too small", not "too large" but "just right". Because we seek the most general set of conditions that ensure well-posedness of variational problems, initially we will not assume any connections between the variational equation (17) and an optimization problem. Then we will show how these conditions specialize to cases when variational equations can be associated with optimization problems.

2.2 Variational methods in finite dimensions

Many results concerning variational problems can be motivated and explained by examining them in standard Euclidean spaces. In this setting operator equations are simply linear systems of algebraic equations

$$\mathbb{A}\mathbf{u} = \mathbf{f} \tag{20}$$

where \mathbb{A} is $m \times n$ real matrix, $\mathbf{u} \in \mathbb{R}^n$, is a vector of unknowns and $\mathbf{f} \in \mathbb{R}^m$ is the right hand side vector. Our plan is to solve (20) using a variational method and then to express conditions for the well-posedness of (20) in the language of bilinear forms.

To this end, we make the identifications $V_1 \equiv \mathbb{R}^n$ and $V_2 \equiv \mathbb{R}^m$. A "weak" Galerkin variational formulation of (20) is to seek $\mathbf{u} \in V_1 = \mathbb{R}^n$ such that

$$\mathbf{v}^T \mathbb{A} \mathbf{u} = \mathbf{v}^T \mathbf{f} \quad \forall \mathbf{v} \in V_2 = \mathbb{R}^m. \quad (21)$$

The right hand side in (21) serves to define a linear functional $F : V_2 \mapsto \mathbb{R}$ given by

$$F(\mathbf{v}) = \mathbf{v}^T \mathbf{f},$$

the left hand side in (21) defines a bilinear form on $V_1 \times V_2$ given by

$$Q(\mathbf{u}; \mathbf{v}) = \mathbf{v}^T \mathbb{A} \mathbf{u}, \quad (22)$$

and so (21) takes the abstract form of (17): seek $\mathbf{u} \in V_1$ such that

$$Q(\mathbf{u}; \mathbf{v}) = F(\mathbf{v}) \quad \forall \mathbf{v} \in V_2. \quad (23)$$

We will consider three types of matrices \mathbb{A} for which (20) has a unique solution. In each case the matrix properties that guarantee uniqueness will be translated into a variational statement about the associated bilinear form in (21). Let us first assume that \mathbb{A} is invertible. Then \mathbb{A} is square and has full row (or column) rank. The following lemma gives a variational characterization of this property.

Lemma 1 *Suppose that*

$$\max_{\mathbf{v} \in \mathbb{R}^m} \frac{\mathbf{v}^T \mathbb{A} \mathbf{u}}{(\mathbf{v}^T \mathbf{v})^{1/2}} \geq C_1 (\mathbf{u}^T \mathbf{u})^{1/2}; \quad (24)$$

$$\max_{\mathbf{u} \in \mathbb{R}^n} \frac{\mathbf{v}^T \mathbb{A} \mathbf{u}}{(\mathbf{u}^T \mathbf{u})^{1/2}} \geq C_2 (\mathbf{v}^T \mathbf{v})^{1/2}. \quad (25)$$

Then \mathbb{A}^{-1} exists and (20) has a unique solution.

We leave the proof as an exercise. Using (22), conditions (24)-(25) can be written as

$$\max_{\mathbf{v} \in V_2} \frac{Q(\mathbf{u}; \mathbf{v})}{\|\mathbf{v}\|_{V_2}} \geq C_1 \|\mathbf{u}\|_{V_1}, \quad (26)$$

and

$$\max_{\mathbf{u} \in V_1} \frac{Q(\mathbf{u}; \mathbf{v})}{\|\mathbf{u}\|_{V_1}} \geq C_2 \|\mathbf{v}\|_{V_2}, \quad (27)$$

respectively. Forms that satisfy (26)-(27) are called *weakly coercive*. The main advantage of (26)-(27) over the algebraic form (24)-(25) is that it expresses unique solvability of (20) in terms of bilinear forms, norms and inner products, i.e., the fundamental objects of Hilbert spaces. As a result, conditions (26)-(27) can be extended to arbitrary Hilbert spaces.

Conditions (26)-(27) do not assume any special properties of \mathbb{A} and so they represent the most general criteria that guarantees the existence of \mathbb{A}^{-1} . If more is known about \mathbb{A} , these conditions can be specialized to reflect its structure. There are two important

classes of matrices that persistently arise in applications across many disciplines and we consider them next.

The first class contains matrices that are symmetric positive definite or real positive definite⁵. The reader is asked to verify that for such matrices there exists a positive constant C such that

$$\mathbf{u}^T \mathbb{A} \mathbf{u} \geq C \mathbf{u}^T \mathbf{u} \quad \forall \mathbf{u} \in \mathbb{R}^n. \quad (28)$$

This is precisely the variational quantification we are looking for, and which we immediately translate into the language of bilinear forms as

$$Q(\mathbf{u}; \mathbf{u}) \geq C \|\mathbf{u}\|_{V_1}^2, \quad (29)$$

where $Q(\cdot; \cdot)$ is the form defined in (22). Bilinear forms that satisfy (29) are called *coercive* or *V-elliptic*. Any coercive form trivially satisfies (26)-(27). Thus, coercivity always implies weak coercivity but not vice versa. In terms of matrices this is equivalent to the fact that every positive definite matrix is invertible, but not every invertible matrix is positive definite.

The second class contains symmetric indefinite matrices with the following 2×2 block structure

$$\begin{pmatrix} \mathbb{K} & \mathbb{B}^T \\ \mathbb{B} & \mathbf{0} \end{pmatrix}. \quad (30)$$

We assume that \mathbb{K} and \mathbb{B} are $n \times n$ and $m \times n$ matrices, respectively, and that $m \leq n$. The matrix in (30) is called *Karush-Kuhn-Tucker* (KKT) matrix and it arises in equality constrained quadratic programs (QP); see [70, p.443]. Let \mathbb{Z} denote a matrix whose columns are a basis for the nullspace of \mathbb{B} . The following lemma states sufficient conditions for the KKT matrix to be nonsingular; see [70, p.445].

Lemma 2 *Let \mathbb{B} has full row rank, and assume that $\mathbb{Z}^T \mathbb{K} \mathbb{Z}$ is positive definite. Then the KKT matrix is nonsingular.*

To translate this lemma into a variational statement let $V = \mathbb{R}^n$, $S = \mathbb{R}^m$. The matrices \mathbb{K} and \mathbb{B} serve to define the forms

$$a(\mathbf{u}, \mathbf{v}) = \mathbf{v}^T \mathbb{K} \mathbf{u} \quad \text{and} \quad b(\mathbf{u}, \mathbf{p}) = \mathbf{p}^T \mathbb{B} \mathbf{u}$$

on $V \times V$ and $S \times V$, respectively. Then, the linear system

$$\begin{pmatrix} \mathbb{K} & \mathbb{B}^T \\ \mathbb{B} & \mathbf{0} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{0} \end{pmatrix} \quad (31)$$

is equivalent to the variational problem: seek $(\mathbf{u}, \mathbf{p}) \in V \times S$ such that

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) + b(\mathbf{p}, \mathbf{v}) &= F(\mathbf{v}) \quad \forall \mathbf{v} \in V \\ b(\mathbf{r}, \mathbf{u}) &= 0 \quad \forall \mathbf{r} \in S. \end{aligned} \quad (32)$$

We leave it as an exercise for the reader to prove the following theorem.

⁵A matrix \mathbb{A} is real positive definite if $\mathbf{u}^T \mathbb{A} \mathbf{u} > 0$ for any real-valued vector \mathbf{u} . Equivalently, \mathbb{A} is real positive definite if $(\mathbb{A} + \mathbb{A}^T)/2$ is positive definite.

Theorem 1 Let $a(\cdot, \cdot)$ be coercive on $Z \times Z$, where

$$Z = \{\mathbf{u} \in V \mid b(\mathbf{r}, \mathbf{u}) = 0 \quad \forall \mathbf{r} \in S\}$$

and assume that

$$\max_{\mathbf{v} \in V} \frac{b(\mathbf{p}, \mathbf{v})}{\|\mathbf{v}\|_V} \geq C \|\mathbf{p}\|_S. \quad (33)$$

Then (32) has a unique solution.

With the identifications $V_1 = V_2 = V \times S$;

$$Q(\{\mathbf{v}, \mathbf{p}\}; \{\mathbf{v}, \mathbf{r}\}) = a(\mathbf{u}, \mathbf{v}) + b(\mathbf{p}, \mathbf{v}) + b(\mathbf{r}, \mathbf{u}), \quad (34)$$

and $F(\{\mathbf{v}, \mathbf{r}\}) \equiv F(\mathbf{v})$ problem (32) takes the form of (23). One can show that the assumptions on the forms $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ stated in Theorem 1 are sufficient to prove that the form defined in (34) is weakly coercive; see the Exercises.

Let us briefly talk about the roles of test and trial spaces in the linear algebra setting. The trial space V_1 is simply the space of unknowns. The test space V_2 has the same dimension as the right hand side and so it corresponds to the number of equations.

When V_2 is “larger” than V_1 ($\dim V_1 < \dim V_2$) there are more equations than unknowns, the system $\mathbb{A}\mathbf{u} = \mathbf{f}$ is overdetermined, and there may be no solution that will satisfy simultaneously all equations. If V_2 is smaller than V_1 ($\dim V_1 > \dim V_2$), the number of equations is less than the number of unknowns and so $\mathbb{A}\mathbf{u} = \mathbf{f}$ will have infinitely many solutions. As a result, a linear system is well-posed if and only if \mathbb{A} has a full column rank (implied by (26)) and $\dim V_1 = \dim V_2$ (implied by (27)). Thus, weak coercivity conditions (26)-(27) enforce simultaneously a condition on the rank of \mathbb{A} plus the correct “balance” between test and trial spaces.

Of course, in infinite dimensional Hilbert spaces we cannot speak of numbers of unknowns and equations. The relative “sizes” of the spaces are then given by the inclusion relations that may exist between them. When the Hilbert spaces are related to PDE’s their “size” can be measured in terms of the regularity of the functions they contain. A small space contains nice, smooth functions; a large space can include objects such as generalized functions (distributions like the Dirac delta function). Finding “perfectly balanced” test and trial spaces then requires a detailed knowledge of the PDE and its weak form, and is a major part of setting up a well-posed variational problem.

2.3 Variational problems in Hilbert spaces

We turn attention to variational equations in Hilbert spaces and sufficient conditions for their well-posedness, expressed in terms of bilinear forms.

2.3.1 Weakly coercive and coercive problems

The next definition extends (26)-(27) and (29) to forms in general Hilbert spaces.

Definition 1 A bilinear form $Q(\cdot; \cdot) : V_1 \times V_2 \mapsto \mathbb{R}$ is called *weakly coercive* if

$$\sup_{v \in V_2} \frac{Q(u; v)}{\|v\|_{V_2}} \geq C_2 \|u\|_{V_1} \quad \forall u \in V_1, \quad (35)$$

$$\sup_{u \in V_1} \frac{Q(u; v)}{\|u\|_{V_1}} \geq C_2 \|v\|_{V_2} \quad \forall v \in V_2. \quad (36)$$

Let $V_1 \equiv V_2$. A bilinear form $Q(\cdot; \cdot) : V \times V$ is called *coercive* if

$$Q(u; u) \geq C_2 \|u\|_V^2 \quad \forall u \in V. \quad (37)$$

The second condition (36) can be relaxed to

$$\sup_{u \in V_1} \frac{Q(u; v)}{\|u\|_{V_1}} > 0 \quad \forall 0 \neq v \in V_2. \quad (38)$$

We note for future reference that (35) is equivalent to the statement that for any $u \in V_1$ there exists $v \in V_2$ such that

$$Q(u; v) \geq C_2 \|u\|_{V_1} \|v\|_{V_2}. \quad (39)$$

Definition 2 A bilinear form $Q(\cdot; \cdot) : V_1 \times V_2 \mapsto \mathbb{R}$ is called *continuous*⁶ if

$$|Q(u; v)| \leq C_1 \|v\|_{V_2} \|u\|_{V_1}. \quad (40)$$

Theorem 2 (Necas [8]) Given two Hilbert spaces V_1 and V_2 , a linear functional $F : V_1 \mapsto \mathbb{R}$, and a bilinear form $Q(\cdot; \cdot) : V_1 \times V_2 \mapsto \mathbb{R}$ assume that

C.1 F is bounded;

C.2 $Q(\cdot; \cdot)$ is continuous;

C.3 $Q(\cdot; \cdot)$ is weakly coercive.

Then the variational problem: seek $u \in V_1$ such that

$$Q(u; v) = F(v) \quad \forall v \in V_2 \quad (41)$$

has a unique solution. Moreover,

$$\|u\|_{V_1} \leq \frac{1}{C_1} \|F\|. \quad (42)$$

As a corollary to this powerful result we have the celebrated Lax-Milgram Lemma.

Corollary 1 (Lax-Milgram) Let $V_1 = V_2 = V$ and assume C.1-C.2. If $Q(\cdot; \cdot)$ is coercive, then problem (41) has a unique solution and (42) holds.

In finite dimensions the types of variational problems solved by Theorem 2 and Corollary 1 correspond to linear systems with a nonsingular matrix and a positive definite matrix, respectively.

⁶In finite dimensions all bilinear forms are continuous.

Approximation of weakly coercive and coercive problems. To solve a variational problem we approximate the spaces where it is defined rather than the problem itself. Thus, a variational equation and its conforming discretization represent the *same* problem but posed on two different spaces. This has some very important consequences for the well-posedness of the approximate equation. Because this equation is defined in terms of exactly the same form as the problem it approximates, its well-posedness is governed by exactly the same rules as the well-posedness of the original problem.

This fact greatly "simplifies" the search for stable, uniquely solvable discrete problems because one only needs to find a pair of finite dimensional subspaces of V_1 and V_2 for which the form $Q(\cdot; \cdot)$ retains its continuity, coercivity or weak coercivity. Any pair of such spaces will automatically produce a well-posed discrete problem whose solution depends continuously on the data. Let us formalize these observations and then explain why we still had to place "simplify" in quotes.

Theorem 3 *Assume that all hypotheses of Theorem 2 hold. Let V_1^h and V_2^h be two closed subspaces of V_1 and V_2 , respectively, and assume that*

$$\sup_{v^h \in V_2^h} \frac{Q(u^h; v^h)}{\|v^h\|_{V_2}} \geq C_2^h \|u^h\|_{V_1} \quad \forall u^h \in V_1^h; \quad (43)$$

$$\sup_{u^h \in V_1^h} \frac{Q(u^h; v^h)}{\|u^h\|_{V_1}} \geq C_2^h \|v^h\|_{V_2} \quad \forall v^h \in V_2^h. \quad (44)$$

Then, the approximate problem: seek $u^h \in V_1^h$ such that

$$Q(u^h; v^h) = F(v^h) \quad \forall v^h \in V_2^h \quad (45)$$

has a unique solution and

$$\|u^h\|_{V_1} \leq \frac{1}{C_1} \|F\|. \quad (46)$$

Furthermore, the approximation error $u - u^h$ can be estimated by

$$\|u - u^h\|_{V_1} \leq \left(1 + \frac{C_1}{C_2^h}\right) \inf_{w^h \in V_1^h} \|u - w^h\|_{V_1}. \quad (47)$$

Proof.

1. Existence and uniqueness of u^h follows immediately from Theorem 2.
2. Error estimate. Let u and u^h denote the exact and the approximate solutions, respectively, and let w^h denote an arbitrary element of V_1^h . Then

$$\|u - u^h\|_{V_1} \leq \|u - w^h\|_{V_1} + \|u^h - w^h\|_{V_1} \quad \forall w^h \in V_1^h.$$

After taking an infimum over V_1^h the first term will produce the best approximation error, and so (47) will follow if we can show that the second term above is bounded by the same error. Because $V_2^h \subset V_2$, the exact solution u satisfies

$$Q(u; v^h) = F(v^h) \quad \forall v^h \in V_2^h,$$

while for u^h we have that

$$Q(u^h; v^h) = F(v^h) \quad \forall v^h \in V_2^h.$$

Subtracting these equations gives the fundamental *error orthogonality* relation⁷

$$Q(u - u^h; v^h) = 0 \quad \forall v^h \in V_2^h. \quad (48)$$

Adding and subtracting an arbitrary $w^h \in V_1^h$ in (48) yields the identity

$$Q(u - w^h + w^h - u^h; v^h) = Q(u - w^h; v^h) - Q(u^h - w^h; v^h) = 0$$

or,

$$Q(u^h - w^h; v^h) = Q(u - w^h; v^h).$$

From (40) (continuity of $Q(\cdot; \cdot)$)

$$Q(u^h - w^h; v^h) \leq C_1 \|u - w^h\|_{V_1} \|v^h\|_{V_2} \quad \forall w^h \in V_1^h,$$

or, assuming that $v^h \neq 0$

$$\frac{Q(u^h - w^h; v^h)}{\|v^h\|_{V_2}} \leq C_1 \|u - w^h\|_{V_1} \quad \forall w^h \in V_1^h.$$

Taking supremum over $v^h \neq 0$ gives

$$\sup_{v^h \in V_2^h} \frac{Q(u^h - w^h; v^h)}{\|v^h\|_{V_2}} \leq C_1 \|u - w^h\|_{V_1} \quad \forall w^h \in V_1^h,$$

while from (43) (weak coercivity)

$$\sup_{v^h \in V_2^h} \frac{Q(u^h - w^h; v^h)}{\|v^h\|_{V_2}} \geq C_2^h \|u^h - w^h\|_{V_1}.$$

After combining the lower and the upper bound we have that

$$\|u^h - w^h\|_{V_1} \leq \frac{C_1}{C_2^h} \|u - w^h\|_{V_1} \quad \forall w^h \in V_1^h.$$

As a result,

$$\|u - u^h\|_{V_1} \leq \left(1 + \frac{C_1}{C_2^h}\right) \|u - w^h\|_{V_1} \quad \forall w^h \in V_1^h.$$

The estimate follows by taking infimum over the space V_1^h . \square

For coercive forms we have the following result.

⁷This relation is another reason to call u^h a quasiprojection.

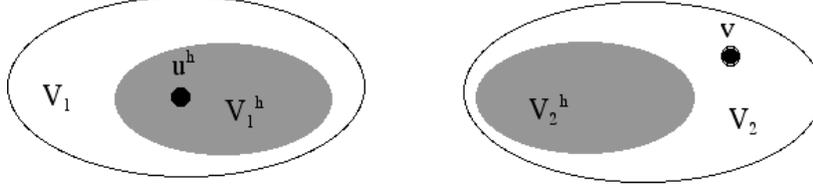


Figure 2: The space V_2^h may not contain v .

Theorem 4 *Assume all hypotheses of Lemma 1 (Lax-Milgram Lemma). Let V^h denote a closed subspace of V . Then the weak problem: seek $u^h \in V^h$ such that*

$$Q(u^h; v^h) = F(v^h) \quad \forall v^h \in V^h$$

has a unique solution,

$$\|u^h\|_V \leq C\|F\|$$

and

$$\|u - u^h\|_V \leq \frac{C_1}{C_2} \inf_{w^h \in V^h} \|u - w^h\|_V. \quad (49)$$

Some comments are now due regarding approximation of coercive and weakly coercive problems.

- The approximating spaces V_i^h , $i = 1, 2$, or V^h are only assumed to be closed subspaces of V_i or V ; they are not necessarily *finite dimensional*.
- The quantity $\inf_{w^h \in W^h} \|u - w^h\|_{V_1}$, where W^h stands for V_1^h or V^h , is the *best approximation error* out of V_1^h or V^h . Approximations for which the error is bounded by a constant times the best error are *quasi-optimal*. This is yet another reason to call u^h a *quasi-projection*. If $Q(\cdot; \cdot)$ is symmetric and coercive then it defines an equivalent inner product and u^h is a true projection of the exact solution.
- Continuity of $Q(\cdot; \cdot)$ is not included in the assumptions of Theorem 3 because it is inherited by all conforming subspaces of V_i .
- Weak coercivity of $Q(\cdot; \cdot)$ *must* be included in the assumptions of Theorem 3 because it is not inherited automatically by conforming subspaces of V_i . The inclusion $V_1^h \subset V_1$ implies that for any $u^h \in V_1^h$ there exists $v \in V_2$ such that (see (39))

$$Q(u^h; v) \geq C_2 \|u^h\|_{V_1} \|v\|_{V_2}.$$

However, existence of v is only guaranteed in the *larger* space V_2 and not in its proper subspace V_2^h ; see Fig.2. As a result, $Q(\cdot; \cdot)$ may fail to be weakly coercive on V_1^h and V_2^h , unless (43)-(44) have been explicitly imposed on these spaces. The fact that the generalized inf-sup conditions are necessary and sufficient for stable and accurate approximation of weakly coercive problems limits the choice of possible discretization spaces to pairs that satisfy (43)-(44).

- Coercivity of $Q(\cdot; \cdot)$ is, on the other hand, inherited on any closed subspace of V . Therefore, stable and accurate approximation of coercive problems only requires space conformity. This makes coercive problems much easier to approximate than weakly coercive equations.

2.3.2 Indefinite variational problems

In this section we consider variational problems that are extensions of the KKT linear system (31) to infinite dimensions. Given a pair of Hilbert spaces V and S , a bilinear form $a(\cdot, \cdot) : V \times V \mapsto \mathbb{R}$, a bilinear form $b(\cdot, \cdot) : V \times S \mapsto \mathbb{R}$, and a bounded linear functional $F(v) : V \mapsto \mathbb{R}$ we consider the problem: seek $u \in V$ and $p \in S$ such that

$$a(u, v) + b(v, p) = F(v) \quad \forall v \in V \quad (50)$$

$$b(u, q) = 0 \quad \forall q \in S. \quad (51)$$

In Section 3.2 we will see that this kind of variational problems is typical for saddle-point optimality conditions arising from the application of Lagrange multiplier techniques. Next theorem is an analogue of Theorem 1 and was established by F. Brezzi in his seminal paper [21].

Theorem 5 *Let*

$$Z = \{v \in V \mid b(v, q) = 0 \text{ for all } q \in S\}. \quad (52)$$

Assume that $a(\cdot, \cdot)$ is continuous on $V \times V$ and coercive on $Z \times Z$, i.e.,

$$|a(u, v)| \leq C_a \|u\|_V \|v\|_V \quad \forall u, v \in V \quad (53)$$

$$a(u, u) \geq \gamma_a \|u\|_V^2 \quad \forall u \in Z. \quad (54)$$

*Problem (50)-(51) has a unique solution if and only if*⁸

$$|b(u, p)| \leq C_b \|u\|_V \|p\|_S \quad \forall u \in V, p \in S \quad (55)$$

$$\sup_{v \in V} \frac{b(v, p)}{\|v\|_V} \geq \gamma_b \|p\|_S \quad \forall p \in S. \quad (56)$$

For a proof of this result we refer the reader to [41, p.57]. As an aside note, (56) is equivalent to the statement that for every $p \in S$ there exists $\mathbf{v} \in V$ such that

$$b(\mathbf{v}, p) \geq \gamma_b \|\mathbf{v}\|_V \|p\|_S. \quad (57)$$

With the identifications $V_1 = V_2 = V \times S$;

$$Q(\{v, p\}; \{v, q\}) = a(u, v) + b(p, v) + b(q, u), \quad (58)$$

⁸Condition (56) is often referred to as the inf-sup condition because of the equivalent form

$$\inf_{p \in S} \sup_{v \in V} \frac{b(v, p)}{\|p\|_S \|v\|_V} \geq \gamma_b,$$

see [23], [41] or [43].

and $F(\{v, q\}) \equiv F(v)$ problem (50)-(51) takes the form of (23) (compare with (32)). One can show that the assumptions on the forms $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ contained in Theorem 5 are sufficient to prove that the form defined in (58) is weakly coercive.

Problem (50)-(51) can be reduced to a coercive problem on $Z \times Z$ by noting that any solution u of (50)-(51) must be in Z . Therefore, any u that solves (50)-(51) will solve the variational equation: seek $u \in Z$ such that

$$a(u, v) = F(v) \quad \forall v \in Z. \quad (59)$$

If all assumptions of Theorem 5 hold for $a(\cdot, \cdot)$, then (59) will have a unique solution. In the next section we will see why (59) is not necessarily a better problem to solve than (50)-(51).

Approximation of indefinite problems. To approximate (50)-(51) we consider a pair of subspaces $V^h \subset V$ and $S^h \subset S$. We also have the discrete kernel subspace

$$Z^h = \{v^h \in V^h \mid b(v^h, q^h) = 0 \quad \forall q^h \in S^h\}. \quad (60)$$

The discrete problem is: seek $u^h \in V^h$ and $p^h \in S^h$ such that

$$a(u^h, v^h) + b(v^h, p^h) = F(v^h) \quad \forall v^h \in V^h \quad (61)$$

$$b(u^h, q^h) = 0 \quad \forall q^h \in S^h. \quad (62)$$

Problem (61)-(62) is a restriction of the weakly coercive equation (50)-(51). Since weak coercivity is not inherited automatically on V^h and S^h , well-posedness of (61)-(62) requires discrete spaces that verify the generalized inf-sup conditions of Theorem 3. These conditions are specialized to indefinite problems in the next theorem.

Theorem 6 *Let all assumptions of Theorem 5 hold for the forms $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$. Assume that $a(\cdot, \cdot)$ is coercive on $Z^h \times Z^h$:*

$$a(u^h, u^h) \geq \gamma_a^h \|u^h\|_V^2 \quad \forall u^h \in Z^h, \quad (63)$$

and that there exists $\gamma_b^h > 0$ such that

$$\sup_{v^h \in V^h} \frac{b(v^h, p^h)}{\|v^h\|_V} \geq \gamma_b^h \|p^h\|_S \quad \forall p^h \in S^h, \quad (64)$$

that is the inf-sup condition holds for the discrete pair (V^h, S^h) . Then,

1. Z^h is nonempty;
2. there exists a unique pair (u^h, p^h) that solves (61)-(62);
3. we have the error bound

$$\|u - u^h\|_V \leq K_{11} \inf_{v^h \in S^h} \|u - v^h\|_V + K_{12} \inf_{q^h \in S^h} \|p - q^h\|_S \quad (65)$$

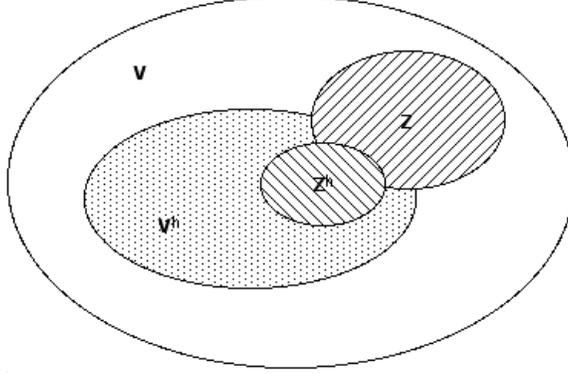


Figure 3: Z^h may not be a subspace of Z .

$$\|p - p^h\|_S \leq K_{21} \|u - u^h\|_V + K_{22} \inf_{q^h \in S^h} \|p - q^h\|_S \quad (66)$$

where

$$K_{11} = \left(1 + \frac{C_a}{\gamma_a^h} + \frac{C_a C_b}{\gamma_a^h \gamma_b^h}\right), \quad K_{12} = \frac{C_b}{\gamma_a^h} \delta(Z, Z^h),$$

$$K_{21} = \frac{C_b}{\gamma_b^h}, \quad K_{22} = \left(1 + \frac{C_b}{\gamma_b^h}\right),$$

and

$$\delta(Z, Z^h) = \sup_{z^h \in Z^h} \inf_{z \in Z} \|z - z^h\|_V.$$

Remark 1 If $Z^h \subset Z$ the distance $\delta(Z, Z^h)$ between these two spaces is zero and the estimate for u^h uncouples from the error of p^h

$$\|u - u^h\|_V \leq K_{11} \inf_{v^h \in S^h} \|u - v^h\|_V.$$

In this case, the error of u^h depends only on the approximation properties of V^h .

Remark 2 The error estimate for p^h cannot be uncoupled from that of u^h even if $Z^h \subset Z$. As a result, approximation of this variable always depends on the approximation properties of both V^h and S^h .

Remark 3 The constants in the estimate for u^h are proportional to $1/\gamma_b^h$, and the constants in the estimate for p^h - to $1/(\gamma_b^h)^2$. For some spaces the discrete inf-sup condition may hold with $\gamma_b^h \rightarrow 0$ as $h \rightarrow 0$. Such spaces lead to unstable discretizations and possible loss of convergence.

To conclude this section let us expound on the reasons why the coercive equation (59) may not be easier to approximate than the weakly coercive problem (50)-(51). To reap the benefits afforded by the coercivity of $a(\cdot, \cdot)$ requires a conforming subspace Z^h of Z . Then the problem: seek $u^h \in Z^h$ such that

$$a(u^h, v^h) = F(v^h) \quad \forall v^h \in Z^h \quad (67)$$

will be well-posed. The problem with this approach is that a conforming Z^h may be as difficult to construct as to find a pair (V^h, S^h) that satisfies the discrete inf-sup condition (64). To appreciate this, note that even if V^h and S^h are a stable pair, the space Z^h defined in (60) is not necessarily a subspace of Z ; see Fig.3.

2.4 Exercises

1. Show that if \mathbb{A} is symmetric and positive definite or real positive definite there exists a constant C that may depend on the space dimension, and such that (28) holds.
2. Prove that (21) is equivalent to (20).
3. Prove that (24)-(25) imply that \mathbb{A} is a square nonsingular matrix.
4. Assume that $Q(\cdot; \cdot)$ is coercive on $V \times V$ and let V^h denote a proper subspace of V . Prove that the form is also coercive on $V^h \times V^h$.
5. Let all assumptions of Lemma 1 hold. Prove that the form defined in (34) is weakly coercive.

3 Stability of finite element discretizations

Finite elements are variational methods that use piecewise polynomial subspaces defined with respect to some tessellation of the computational domain into simple geometrical shapes. This choice of approximating subspaces has been one of the main reasons for their practical appeal. First, finite element spaces are spanned by locally supported, piecewise polynomial functions and so they lead to the generation of sparse algebraic systems. Second, the finite element paradigm allows for almost automatic generation of high order methods in *arbitrary, unstructured* meshes.

Because finite elements are a variational method, all their fundamental properties, including well-posedness of the discrete equations, are governed by the variational principles embedded in their foundations. Brezzi and Fortin point out that (see [23, p.3])

a finite element method can only be considered in relation with a variational principle and a functional space. Changing the variational principle and the space in which it is posed leads to a different finite element approximation.

In this section we study the impact of different variational principles upon the finite element method.

3.1 FEM in unconstrained minimization setting

Consider the convex, quadratic functional

$$J(\phi; f) = \frac{1}{2} \int_{\Omega} |\nabla \phi|^2 d\Omega - \int_{\Omega} f \phi d\Omega \quad (68)$$

and the minimization principle

$$\min_{\phi \in H_0^1(\Omega)} J(\phi; f), \quad (69)$$

where f is a given function and $H_0^1(\Omega)$ denotes the space of functions that have square integrable first derivatives and that vanish on the boundary of the given domain Ω . The first-order necessary condition for the optimization problem (69) requires the first variation of the functional (68) to vanish. Therefore, the minimizer $\phi \in H_0^1(\Omega)$ of (68) satisfies the variational equation

$$Q_r(\phi; \psi) = F(\psi) \quad \forall \psi \in H_0^1(\Omega), \quad (70)$$

where

$$Q_r(\phi; \psi) = \int_{\Omega} \nabla \phi \cdot \nabla \psi \, d\Omega \quad \text{and} \quad F(\psi) = \int_{\Omega} f \psi \, d\Omega. \quad (71)$$

This equation has a connection with a familiar PDE problem. To see this connection, we integrate by parts in (70) to obtain

$$0 = \int_{\Omega} (\nabla \phi \cdot \nabla \psi - f \psi) \, d\Omega = - \int_{\Omega} \psi (\Delta \phi + f) \, d\Omega, \quad (72)$$

where we have assumed that ϕ is sufficiently smooth to justify the above integration. Since ψ is arbitrary, it follows that any sufficiently smooth⁹ minimizer of $J(\cdot; f)$ is a solution of the familiar Poisson problem

$$-\Delta \phi = f \quad \text{in } \Omega \quad \text{and} \quad \phi = 0 \quad \text{on } \Gamma, \quad (73)$$

The differential equation follows from (72) and the boundary condition (on the boundary Γ of Ω) from the fact that all admissible states are constrained to vanish on the boundary.

The correspondence between solutions of partial differential equations and unconstrained minimization problems is not unusual. Many physical phenomena are governed by energy principles which postulate that admissible states of the system are minimizers of some convex, quadratic energy functional. In this case, the optimization problem is the primary model of the physical process while the PDE problem is a strong (pointwise) expression of the first-order optimality condition.

A convex, quadratic energy functional defines an *energy norm* $|||\cdot|||$ and an *energy inner product* $((\cdot, \cdot))$ on its admissible space. This is the single, most important characteristic of unconstrained energy minimization. In our case, the admissible set is given by the Sobolev space $H_0^1(\Omega)$. The expression

$$J(\psi; 0) = \frac{1}{2} \int_{\Omega} |\nabla \psi|^2 \, d\Omega \equiv \frac{1}{2} |\psi|_1^2$$

⁹One appealing feature of the unconstrained energy minimization formulation is that every classical, i.e., twice continuously differentiable, solution of the Poisson equation is also a solution of the minimization problem (69) but the latter admits solutions which are not classical solutions of (73). These non-classical solutions of (69) are referred to as *weak solutions* of the Poisson problem.

and the form $Q_r(\cdot; \cdot)$ serve to define an equivalent norm and inner product on $H_0^1(\Omega)$, respectively, i.e., for the problem (69)

$$|||\phi||| \equiv J(\phi; 0)^{1/2} \quad \text{and} \quad ((\phi, \psi)) \equiv Q_r(\phi; \psi).$$

The norm-equivalence of (68) is a direct consequence of the Poincaré inequality

$$\lambda \|\psi\|_0 \leq |\psi|_1 \quad \forall \psi \in H_0^1(\Omega),$$

where λ is a constant whose value depends only on Ω and $\|\cdot\|_0$ denotes the standard norm for $L^2(\Omega)$. The inner product equivalence

$$(1 + \lambda^{-2})^{-1} \|\psi\|_1^2 \leq Q_r(\psi; \psi) \quad \text{and} \quad Q_r(\phi; \psi) \leq \|\phi\|_1 \|\psi\|_1 \quad (74)$$

follows from the identity $|\phi|_1^2 = Q_r(\phi; \phi)$ and the Cauchy inequality. Note that (74) implies coercivity of $Q_r(\cdot; \cdot)$ on $H_0^1(\Omega) \times H_0^1(\Omega)$.

Consider now the implications from the connection between (73) and (69) for the finite element solution of the Poisson equation. A finite element solution of (73) begins with rewriting this problem into a weak, variational form, essentially reversing the process through which we obtained (73) from (70). Then we choose a finite dimensional subspace X^h of $H_0^1(\Omega)$ and determine the finite element approximation from the weak problem (70) restricted to X^h , i.e.,

$$\text{seek } \phi^h \in X^h \text{ such that } Q_r(\phi^h; \psi^h) = F(\psi^h) \quad \forall \psi^h \in X^h. \quad (75)$$

Because (74) remains true for any subspace of $H_0^1(\Omega)$ the form $Q_r(\cdot; \cdot)$ is coercive on $X^h \times X^h$. Theorem 4 then asserts that (75) is a well-posed problem. Since $Q_r(\cdot; \cdot)$ is also an inner product, (75) defines an orthogonal projection of ϕ onto X^h with respect to the inner product $((\cdot, \cdot))$. To see this, note that

$$((\phi, \psi^h)) = F(\psi^h) \quad \forall \psi^h \in X^h$$

and

$$((\phi^h, \psi^h)) = F(\psi^h) \quad \forall \psi^h \in X^h$$

so that

$$((\phi - \phi^h, \psi^h)) = 0 \quad \forall \psi^h \in X^h.$$

Therefore, ϕ^h minimizes the energy norm of the error

$$|||\phi - \phi^h||| = \inf_{\psi^h \in X^h} |||\phi - \psi^h|||,$$

and, using the bounds in (74), we also have the quasi-optimal error estimate in the norm of $H_0^1(\Omega)$:

$$\|\phi - \phi^h\|_1 \leq C \inf_{\psi^h \in X^h} \|\phi - \psi^h\|_1.$$

As a result, a finite element for (73) is always guaranteed to compute the best possible approximation out of X^h . A different way to say this is that the finite element solution ϕ^h is the unique minimum of (68) out of X^h .

Recall that variational methods approximate the space and not the problem itself. Thus, the only difference between (70) and (75) is that the latter is posed over a finite dimensional subspace. But all bilinear forms in finite dimensional spaces are engendered by matrices and we can conclude that (75) is a linear algebraic system. Given a basis $\{\phi_i\}_{i=1}^N$ for X^h , this system has the form

$$\mathbb{A}\mathbf{x} = \mathbf{f}, \quad (76)$$

where $\mathbb{A}_{ij} = ((\phi_i, \phi_j)) = Q_r(\phi_i; \phi_j)$, $\mathbf{f}_i = F(\phi_i)$, and $\mathbf{x}_j = \xi_j$ with ξ_j denoting the coefficients in the expansion of ϕ^h in terms of the basis, i.e., $\phi^h = \sum_{i=1}^N \xi_i \phi_i$. From (71) and (74), it follows that \mathbb{A} is symmetric and positive definite matrix. In addition, the equivalence between the energy inner product $((\cdot, \cdot))$ and the standard inner product on $H_0^1(\Omega)$ implies the spectral equivalence between \mathbb{A} and the Gram matrix of $\{\phi_i\}_{i=1}^N$ with respect to the $H_0^1(\Omega)$ -inner product. This fact is useful for the design of efficient preconditioners for (76).

To summarize, when PDE problems are associated with unconstrained optimization of convex, quadratic energy functionals,

- the discrete problems have unique and stable solutions;
- the approximate solutions minimize an energy functional on the trial space so that they represent, in this sense, the best possible approximation;
- the linear systems used to determine the approximate solutions have symmetric and positive definite coefficient matrices;
- these matrices are spectrally equivalent to the Gram matrix of the trial space basis in the natural norm of $H_0^1(\Omega)$.

3.2 FEM in constrained optimization setting

Consider the quadratic functional

$$J(\mathbf{v}; \mathbf{f}) = \frac{1}{2} \int_{\Omega} |\nabla \mathbf{v}|^2 d\Omega - \int_{\Omega} \mathbf{f} \cdot \mathbf{v} d\Omega \quad (77)$$

and the minimization problem

$$\min_{\mathbf{v} \in \mathbf{H}^1(\Omega)} J(\mathbf{v}; \mathbf{f}) \quad \text{subject to} \quad \nabla \cdot \mathbf{v} = 0 \quad \text{in } \Omega, \quad (78)$$

where $\mathbf{H}_0^1(\Omega)$ is the vector analog of $H_0^1(\Omega)$. We proceed to introduce the Lagrange multiplier p , the Lagrangian functional

$$L(\mathbf{v}, q; \mathbf{f}) = J(\mathbf{v}; \mathbf{f}) - \int_{\Omega} q \nabla \cdot \mathbf{v} d\Omega, \quad (79)$$

and the unconstrained problem of determining saddle points of $L(\mathbf{v}, q; \mathbf{f})$. The first-order necessary conditions are equivalent to the weak problem:

seek $(\mathbf{u}, p) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ such that

$$\begin{aligned} \int_{\Omega} \nabla \mathbf{u} : \nabla \mathbf{v} \, d\Omega - \int_{\Omega} p \nabla \cdot \mathbf{v} \, d\Omega &= \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \, d\Omega \\ \int_{\Omega} q \nabla \cdot \mathbf{u} \, d\Omega &= 0 \end{aligned} \quad (80)$$

for all $(\mathbf{v}, q) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$.

If solutions to the constrained minimization problem (78) or, equivalently, of (80), are sufficiently smooth, then, using integration by parts, one obtains without much difficulty the Stokes equations

$$\begin{aligned} -\Delta \mathbf{u} + \nabla p &= \mathbf{f} \quad \text{and} \quad \nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega, \\ \mathbf{u} &= \mathbf{0} \quad \text{on } \Gamma, \end{aligned} \quad (81)$$

where \mathbf{u} is the *velocity* and p is the *pressure*. Thus, (80) is a weak formulation of the Stokes equations. Conversely, the Stokes equations give the strong optimality system for (79).

A second example of a constrained minimization problem is provided by the functional

$$J(\mathbf{v}) = \frac{1}{2} \int_{\Omega} |\mathbf{v}|^2 \, d\Omega,$$

the linear constraint

$$\nabla \cdot \mathbf{v} = f,$$

and the minimization problem

$$\min J(\mathbf{v}) \quad \text{subject to } \nabla \cdot \mathbf{v} = f, \quad (82)$$

where the minimization is effected over a suitable function space. Such problems arise in many applications; for example, in fluid mechanics, (82) is known as the *Kelvin principle* and, in structural mechanics (where \mathbf{u} is a tensor), as the *complimentary energy principle*; see [23, p.17]. Again, we use a Lagrange multiplier ψ to enforce the constraint and consider the Lagrangian functional

$$L(\mathbf{w}, \psi; f) = \frac{1}{2} \int_{\Omega} |\mathbf{w}|^2 \, d\Omega - \int_{\Omega} \psi (\nabla \cdot \mathbf{w} - f) \, d\Omega,$$

The optimality system obtained by setting the first variations of $L(\mathbf{v}, \psi; f)$ to zero is given by

seek $(\mathbf{v}, \phi) \in H(\Omega, \text{div}) \times L^2(\Omega)$ such that

$$\begin{aligned} \int_{\Omega} \mathbf{v} \cdot \mathbf{w} \, d\Omega - \int_{\Omega} \phi \nabla \cdot \mathbf{w} \, d\Omega &= 0 \\ \int_{\Omega} \xi \nabla \cdot \mathbf{v} \, d\Omega &= \int_{\Omega} f \xi \, d\Omega \end{aligned} \quad (83)$$

for all $(\mathbf{w}, \xi) \in H(\Omega, \text{div}) \times L^2(\Omega)$.

If solutions of (83) are sufficiently smooth we see that

$$\begin{aligned} \nabla \cdot \mathbf{v} &= f \quad \text{and} \quad \mathbf{v} + \nabla \phi = \mathbf{0} \quad \text{in } \Omega \\ \phi &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{84}$$

Problem (84) is the first-order Poisson equation that we first encountered in Section 1.3. Problem (83) is the second weak equation from the list of the four weak forms (13)-(16). If $V \equiv H(\Omega, \text{div})$, $S \equiv L^2(\Omega)$,

$$a(\mathbf{v}, \mathbf{w}) = \int_{\Omega} \mathbf{v} \cdot \mathbf{w} \, d\Omega \quad \text{and} \quad b(\phi, \mathbf{v}) = \int_{\Omega} \phi \nabla \cdot \mathbf{v} \, d\Omega,$$

then problem (83) takes the form of (50)-(51). The correspondence between solutions of partial differential equations and constrained optimization problems is also not unusual. In this case the PDE expresses the first-order optimality condition for the saddle-point optimization problem.

Let us now discuss the consequences from the association between a PDE and a saddle-point optimality system. Problems (80) and (83) have the abstract structure of (50)-(51). Therefore, their well-posedness is governed by the assumptions in Theorem 5 which, as we recall, specializes the more general Theorem 2 to saddle-point variational problems. This also means that (80) and (83) are only weakly coercive and that their finite element approximations are subject to the assumptions stated in Theorem 6.

Consider for example a mixed method for (83). To approximate this problem we choose a pair of subspaces $V^h \subset H(\Omega, \text{div})$ and $S^h \subset L^2(\Omega)$ and then restrict (83) to $V^h \times S^h$. The problem

seek $(\phi^h, \mathbf{w}^h) \in S^h \times V^h$ such that

$$\begin{aligned} \int_{\Omega} \mathbf{w}^h \cdot \mathbf{u}^h \, d\Omega + \int_{\Omega} \phi^h \nabla \cdot \mathbf{u}^h \, d\Omega &= 0 \\ \int_{\Omega} (\nabla \cdot \mathbf{w}^h) \xi^h \, d\Omega &= \int_{\Omega} f \xi^h \, d\Omega \end{aligned} \tag{85}$$

for all $(\xi^h, \mathbf{u}^h) \in S^h \times V^h$

is of the same type as the abstract problem (61)-(62). Consequently, it will not be a stable and accurate approximation of (83) unless

1. the form $a(\cdot, \cdot)$ is coercive on Z^h ;
2. the form $b(\cdot, \cdot)$ verifies the discrete inf-sup condition (64).

The single most important consequence from this fact is that we cannot choose the approximating spaces for the variables independently from each other. What is even more troublesome, the rather abstract nature of (64) doesn't provide any clues to as what might constitute a promising pair of spaces (V^h, S^h) . We will revisit these questions in Section

7. Another consequence is that the discrete problem now corresponds to the indefinite linear system (30). Such linear systems are more difficult to solve than the symmetric and positive definite systems arising from unconstrained optimization.

In summary, we see that whenever a partial differential equation problem is associated with constrained optimization problem,

- the discrete equations have unique and stable solutions only if the approximating spaces satisfy restrictive stability conditions;
- the approximate solutions are constrained projections rather than inner-product orthogonal projections;
- the linear systems used to determine the approximate solutions have symmetric but indefinite coefficient matrices.

3.3 Galerkin principles

In sections 3.1-3.2 we saw examples of PDE problems that can be associated with optimization problems. However, many PDE's are not related to optimization. The conservation law (4) from §1.1 is one such example. For this problem the differential equation is obtained from an integral conservation statement; see [61, p.17], and does not represent a strong form of an optimality condition.

In such cases finite element methods use formal residual orthogonalization to derive the variational equations. Each partial differential equation is multiplied by suitable testing function, integrated over the domain and then integrated by parts to equilibrate the derivatives. This procedure is often called a *weighted residual method*. When test and trial functions belong to the same spaces we speak of *Galerkin principles*, when they are drawn from different spaces the principle is called *Petrov-Galerkin*.

It is clear that any PDE equation, either related or unrelated to optimization, can be treated by a weighted residual method. Because of this universality, formal residual orthogonalization has been the natural choice in extending finite elements beyond differential equations problems associated with optimization principles. If, on the other hand the PDE is associated with optimization, we have already seen that Galerkin principles will recover the optimality system.

Let us consider some examples of PDE problems not associated with optimization. A simple example is provided by the Helmholtz equation problem

$$-\Delta\phi - k^2\phi = f \quad \text{in } \Omega \quad \text{and} \quad \phi = 0 \quad \text{on } \Gamma. \quad (86)$$

Using the formal Galerkin procedure we find the weak formulation of (86) to be

$$\int_{\Omega} (\nabla\phi \cdot \nabla\psi - k^2\phi\psi) d\Omega = \int_{\Omega} f\psi d\Omega \quad \forall \psi \in H_0^1(\Omega). \quad (87)$$

Note that the bilinear form on the left-hand side of (87) is symmetric but, if k^2 is larger than the smallest eigenvalue of $-\Delta$, it is not coercive, i.e., it does not define an inner

product on $H_0^1(\Omega) \times H_0^1(\Omega)$. As a result, proving the existence and uniqueness¹⁰ of weak solutions is not so simple a matter as it is for the Poisson equation case.

Another example is provided by the convection-diffusion-reaction problem

$$-\varepsilon\Delta\phi + \mathbf{b} \cdot \nabla\phi + c\phi = f \quad \text{in } \Omega \quad \text{and} \quad \phi = 0 \quad \text{on } \Gamma \quad (88)$$

and the companion *reduced problem*

$$\mathbf{b} \cdot \nabla\phi + c\phi = f \quad \text{in } \Omega \quad \text{and} \quad \phi = 0 \quad \text{on } \Gamma_-, \quad (89)$$

where \mathbf{b} is a given vector-valued function, c is a given scalar-valued function, and Γ_- denotes the *inflow* portion of Γ , i.e., the portion of Γ for which $\mathbf{b} \cdot \mathbf{n} < 0$, where \mathbf{n} denotes the outward unit normal vector.¹¹ Again, following a standard Galerkin procedure for (88) results in the Galerkin weak formulation

$$\int_{\Omega} \left(\varepsilon \nabla\phi \cdot \nabla\psi + \psi \mathbf{b} \cdot \nabla\phi + c\phi\psi \right) d\Omega = \int_{\Omega} f\psi d\Omega \quad \forall \psi \in H_0^1(\Omega). \quad (91)$$

For (89), the formal Galerkin process does not even require integration by parts and reduces to multiplication by a test function and integration:

$$\int_{\Omega} \left(\psi \mathbf{b} \cdot \nabla\phi + c\phi\psi \right) d\Omega = \int_{\Omega} f\psi d\Omega. \quad (92)$$

Now the bilinear forms on the left-hand side of (91) and (92) are neither symmetric nor coercive.

A *nonlinear* example of a problem without a minimization principle, but for which a weak formulation may be defined through a Galerkin method, is the Navier-Stokes system for incompressible, viscous flows given by

$$\begin{aligned} -\Delta\mathbf{u} + \mathbf{u} \cdot \nabla\mathbf{u} + \nabla p = \mathbf{f} \quad \text{and} \quad \nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega, \\ \mathbf{u} = \mathbf{0} \quad \text{on } \Gamma, \end{aligned} \quad (93)$$

where \mathbf{u} and p denote the velocity and pressure fields and the constant ν denotes the kinematic viscosity. A standard weak formulation analogous to (80) but containing an additional nonlinear term is given by

seek $(\mathbf{u}, p) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ such that

¹⁰In fact, solutions of (86) or (87) are not always unique.

¹¹Alternatively, (88) can be stated in conservative form as

$$-\nabla \cdot \sigma(\phi) + c\phi = f \quad \text{in } \Omega \quad \text{and} \quad \phi = 0 \quad \text{on } \Gamma \quad (90)$$

where $\sigma(\phi) = \sigma_D(\phi) + \sigma_C(\phi)$ is the total flux function and

$$\sigma_D(\phi) = \varepsilon\nabla\phi \quad \text{and} \quad \sigma_C(\phi) = -\mathbf{b}\phi$$

denote the diffusive and convective fluxes, respectively, and we have assumed that \mathbf{b} is solenoidal. The reduced problem is obtained when the diffusive flux is zero.

$$\nu \int_{\Omega} \nabla \mathbf{u} : \nabla \mathbf{v} \, d\Omega + \int_{\Omega} p \nabla \cdot \mathbf{v} \, d\Omega + \int_{\Omega} \mathbf{u} \cdot \nabla \mathbf{u} \cdot \mathbf{v} \, d\Omega = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \, d\Omega \quad (94)$$

$$\int_{\Omega} \mu \nabla \cdot \mathbf{u} \, d\Omega = 0 \quad (95)$$

for all $(\mathbf{v}, \mu) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$.

Despite the close resemblance between (80) and (94)–(95), these two problems are strikingly different in their variational origins. Specifically, the second problem does not represent an optimality system, i.e., there is no optimization problem associated with these weak equations. As a result, (94)–(95) cannot be derived in any other way but through the Galerkin procedure described above.

Problems (87), (91), (92) and (94)–(95) demonstrate how versatile Galerkin principles can be. However, this versatility exacts a toll on the requirements for the well-posedness of the weak problems and on the quality of the companion finite element algebraic equations.

Without an association with an optimization problem variational equations will default to the most general case of Theorem 2. This means that very little can be said about the structure of the ensuing finite element algebraic equations. It also means that well-posedness of the discrete equations is subject to all assumptions stated in Theorem 3, most notably, the discrete inf-sup conditions (43)–(44). The unpleasant consequence is that naively defined finite element methods that do not account for the fact that conformity is not enough to guarantee stability in the weakly coercive setting will fail.

The two problems (88) and (89) provide a textbook example of settings for which straightforward Galerkin methods lead to serious numerical difficulties for finite element methods. When ε is small compared to the grid spacing h (or zero as is the case for (89)), finite element solutions will develop spurious oscillations whenever the exact solution is not globally smooth or worse, may not even be well defined. This phenomena can be understood by inspecting the *stability* estimate

$$\sqrt{\varepsilon} \|\nabla \phi\|_0 + \|\phi\|_0 \leq C \|f\|_0$$

which shows that for ε small (or zero), control over the gradients exercised by the variational problem is very weak (or completely missing), see e.g., [58], or [59, Ch.9].

3.4 Summary

The examples given in §3.1–§3.3 show that theoretical and practical difficulties in defining a finite element method and solving the corresponding algebraic systems increase as their variational foundation becomes more and more estranged from that of a true inner product projection. This is the price one pays for the increasing generality of the applicability of the methods we considered as we moved from §3.1 to §3.3.

Constrained optimization problems lead to saddle point problems and restrictive stability conditions on the finite element spaces. If a problem possesses no associated optimization principle so that one is led to the formal Galerkin method, then things can get even worse.

	Rayleigh-Ritz	mixed Galerkin	Galerkin
associated optimization problem	unconstrained	constrained	none
properties of bilinear form	inner product equivalent	symmetric but indefinite	none in general
requirements for existence/uniqueness	none	inf-sup compatibility condition	two inf-sup conditions
requirements on discrete spaces	conformity	conformity and discrete inf-sup condition	conformity and two discrete inf-sup conditions
properties of discrete problems	symmetric, positive definite	symmetric but indefinite	indefinite, not symmetric

Table 1: Comparison of different settings for finite element methods in their most general sphere of applicability.

In both cases, i.e., for weakly coercive and saddle-point problems, we also face the task of finding stable pairs of test and trial spaces. The inf-sup conditions stated in Theorem 3 and Theorem 6 are not very helpful here. While these conditions completely characterize stable variational principles they refuse to provide any clues about the distinguishing characteristics of the spaces that may verify them.

A summary of the comparative features of finite element methods is given in Table 1, where the methods are classified according to the nature of the associated optimization problem, if there is one.

4 A geometric approach to stability of discretizations

PDE's offer convenient and powerful formalism to encode physical phenomena in terms of field functions. We saw that a PDE may be obtained from an optimization problem, or a conservation statement expressed in integral form. Because PDE's have been so widespread in science and engineering, we often tend to forget that besides the field equations formalism, there also exist global quantities that can be used equally well to describe the physics. In fact, many important models have been initially formulated in terms of such global quantities. For example, Maxwell stated the laws of electromagnetics using current, charge, electric and magnetic fluxes, electromotive force (EMF), and magnetomotive force, rather than field representations and differential equations.

In practice, what we can measure about a given field is not the field itself, but some global quantity such as the work of the field along a path, or the electromotive force along a curve, or the flux of a magnetic field through a surface. Thus, our information about the fields is expressed by numbers (the measurements) that are associated with geometrical objects in space (the lines, curves, surfaces and volumes where we measure). The field function then is a mathematical abstraction that represents a measurement over an infinitesimal region.

Mappings from oriented geometrical objects to real numbers are called *differential forms*. We can think of differential forms as field functions with attached measuring devices that return global physical quantities when swept over the appropriate regions in space. The process of the actual measurement can then be formalized through the notion of integration of differential forms.

Until recently importance of differential forms and geometrical viewpoints has not been fully appreciated in numerical PDE's. One notable exception is computational electromagnetics where, starting with the pioneering work of Bossavit [16], [17], and [18], there has been a tremendous interest in geometrical approaches to discretization. A comprehensive account of the recent work in this direction can be found in [74].

Perhaps one reason for the limited interest in geometrical aspects of discretization for PDE's is the fact that in most cases (electromagnetics being again the notable exception) PDE's can be studied, understood and approximated without invoking the language of differential forms. In contrast, Hamiltonian mechanics cannot be understood without differential forms and numerical methods there have exploited differential geometry ideas for quite a while; see e.g., [12] for a discussion of symplectic integration methods.

4.1 Exterior forms

We begin with a review of exterior forms and operations between them. Our presentation follows [7].

Definition 3 *An exterior form of degree k , $k \leq n$ or a k -form, is a mapping*

$$\omega : \underbrace{\mathbb{R}^n \times \dots \times \mathbb{R}^n}_k \mapsto \mathbb{R},$$

that is k -linear and antisymmetric:

$$\omega(\dots, \alpha \boldsymbol{\xi}' + \beta \boldsymbol{\xi}'', \dots) = \alpha \omega(\dots, \boldsymbol{\xi}', \dots) + \beta \omega(\dots, \boldsymbol{\xi}'', \dots);$$

$$\omega(\boldsymbol{\xi}_{i_1}, \dots, \boldsymbol{\xi}_{i_k}) = (-1)^\nu$$

where

$$\nu = \begin{cases} 0 & \text{if } (i_1, i_2, \dots, i_k) \text{ is even} \\ 1 & \text{if } (i_1, i_2, \dots, i_k) \text{ is odd} \end{cases}$$

The set of all k -forms is a linear space with dimension C_n^k . Given a k -form ω_k and l -form ω_l we define an operation between them as follows.

Definition 4 Exterior multiplication. *The exterior product $\omega_k \wedge \omega_l$ is the $(k+l)$ -form whose value on the $k+l$ vectors $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_k, \dots, \boldsymbol{\xi}_{k+l}$ is computed according to the formula*

$$\omega_k \wedge \omega_l = \sum (-1)^\nu \omega_k(\boldsymbol{\xi}_{i_1}, \dots, \boldsymbol{\xi}_{i_k}) \omega_l(\boldsymbol{\xi}_{j_1}, \dots, \boldsymbol{\xi}_{j_l})$$

where $i_1 < \dots < i_k; j_1 < \dots < j_l$ is a permutation of the indices $(1, 2, \dots, k+l)$ and

$$\nu = \begin{cases} 0 & \text{if permutation is even} \\ 1 & \text{if permutation is odd} \end{cases}.$$

To compute the exterior product we take all possible partitions of the $k+l$ vectors into groups of k and l vectors. For each partition we compute the values of ω_k and ω_l on each group of k and l vectors and multiply them. This gives one term in the sum above. This term enters with $+$ or $-$ sign depending on whether the order of the $k+l$ vectors represents an even or an odd permutation.

For example, the exterior product of k 1-forms is a k -form whose value on $(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_k)$ is computed as

$$(\omega_1^1 \wedge \omega_1^2 \wedge \dots \wedge \omega_1^k)(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_k) = \begin{vmatrix} \omega_1^1(\boldsymbol{\xi}_1) & \dots & \omega_1^k(\boldsymbol{\xi}_1) \\ \dots & \dots & \dots \\ \omega_1^1(\boldsymbol{\xi}_k) & \dots & \omega_1^k(\boldsymbol{\xi}_k) \end{vmatrix}.$$

Let us consider some examples of exterior forms and their multiplication in \mathbb{R}^3 . There we have 1, 2 and 3-forms and for completeness, we associate a 0-form with an arbitrary constant.

1-forms. A 1-form is a 1-linear mapping $\mathbb{R}^3 \mapsto \mathbb{R}$, i.e., it is a linear functional. The set of all 1-forms is a linear space that is dual to \mathbb{R}^3 , and is denoted by $(\mathbb{R}^3)^*$. The dimension of this space equals $C_3^1 = 3$. Let $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ be a basis in \mathbb{R}^3 . Given a vector $\boldsymbol{\xi}$ with coordinates $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3)$ relative to this basis, we can write $\boldsymbol{\xi}$ as

$$\boldsymbol{\xi} = \boldsymbol{\xi}_1 \mathbf{x}_1 + \boldsymbol{\xi}_2 \mathbf{x}_2 + \boldsymbol{\xi}_3 \mathbf{x}_3.$$

Then, we can make the association

$$\mathbf{x}_i \mapsto \boldsymbol{\xi}_i$$

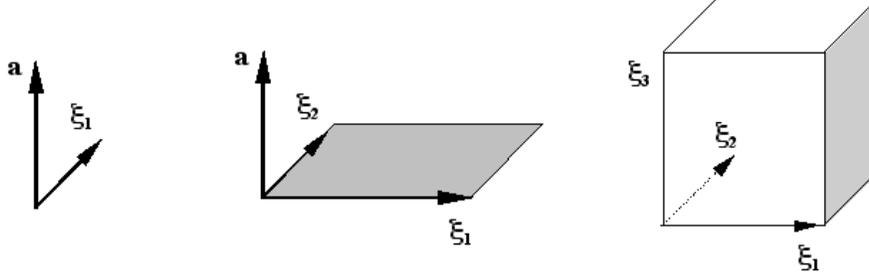


Figure 4: 1-, 2- and 3-forms in \mathbb{R}^3

between basis vectors and coordinates relative to these basis vectors. This association establishes \mathbf{x}_i as an 1-form that returns the i th coordinate of the vector to which it is applied:

$$\mathbf{x}_i(\boldsymbol{\xi}) = \xi_i .$$

The 1-forms \mathbf{x}_i are called *basic* because they span the dual space $(\mathbb{R}^3)^*$, i.e., every 1-form can be represented as

$$\omega_1(\boldsymbol{\xi}) = \alpha_1 \mathbf{x}_1(\boldsymbol{\xi}) + \alpha_2 \mathbf{x}_2(\boldsymbol{\xi}) + \alpha_3 \mathbf{x}_3(\boldsymbol{\xi}) .$$

This shows that every 1-form can be associated with a vector $\mathbf{a} = (\alpha_1, \alpha_2, \alpha_3)$ in \mathbb{R}^3 so that

$$\omega_1(\boldsymbol{\xi}) = \mathbf{a}^T \mathbf{x}(\boldsymbol{\xi}) ,$$

where $\mathbf{x}(\boldsymbol{\xi}) = (\mathbf{x}_1(\boldsymbol{\xi}), \mathbf{x}_2(\boldsymbol{\xi}), \mathbf{x}_3(\boldsymbol{\xi}))$. When \mathbf{x}_i is the canonical basis in \mathbb{R}^3 this is simply the dot product $\mathbf{a}^T \boldsymbol{\xi}$. If \mathbf{a} represents a force field, this dot product gives the work of this force field on the displacement $\boldsymbol{\xi}$, see Fig. 4.

The exterior product of two 1-forms ω_1^1, ω_1^2 is the 2-form

$$(\omega_1^1 \wedge \omega_1^2)(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = \begin{vmatrix} \omega_1^1(\boldsymbol{\xi}_1) & \omega_1^2(\boldsymbol{\xi}_1) \\ \omega_1^1(\boldsymbol{\xi}_2) & \omega_1^2(\boldsymbol{\xi}_2) \end{vmatrix} . \quad (96)$$

To find a geometrical interpretation of this formula consider the mapping $\omega : \mathbb{R}^3 \mapsto \mathbb{R} \times \mathbb{R}$ defined by

$$\boldsymbol{\xi} \mapsto \begin{pmatrix} \omega_1^1(\boldsymbol{\xi}) \\ \omega_1^2(\boldsymbol{\xi}) \end{pmatrix} .$$

This mapping transforms the parallelogram $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ into the parallelogram $(\omega(\boldsymbol{\xi}_1), \omega(\boldsymbol{\xi}_2))$. Formula (96) gives the area of this parallelogram, i.e., the value of the exterior product $\omega_1^1 \wedge \omega_1^2$ is the oriented area of the image of $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ in the ω_1^1, ω_1^2 coordinate frame.

2-forms. A 2-form is a bilinear and skew-symmetric mapping $\mathbb{R}^3 \times \mathbb{R}^3 \mapsto \mathbb{R}$:

$$\omega_2(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = -\omega_2(\boldsymbol{\xi}_2, \boldsymbol{\xi}_1) .$$

The set of all 2-forms is a real vector space with dimension $C_2^2 = 3$. We leave it as an exercise to show that every 2-form in \mathbb{R}^3 can be expressed as

$$\omega_2 = \alpha_1 \mathbf{x}_2 \wedge \mathbf{x}_3 + \alpha_2 \mathbf{x}_3 \wedge \mathbf{x}_1 + \alpha_3 \mathbf{x}_1 \wedge \mathbf{x}_2, \quad (97)$$

where \mathbf{x}_i are the basic 1-forms. Therefore,

$$(\mathbf{x}_1 \wedge \mathbf{x}_2, \mathbf{x}_2 \wedge \mathbf{x}_3, \mathbf{x}_3 \wedge \mathbf{x}_1)$$

is a basis in the space of 2-forms. From (96) we see that $\mathbf{x}_i \wedge \mathbf{x}_j$ is the area of the projection onto the $(\mathbf{x}_i, \mathbf{x}_j)$ coordinate plane.

From (97) we also see that every 2-form can be associated with a vector $\mathbf{a} = (\alpha_1, \alpha_2, \alpha_3)$ in \mathbb{R}^3 . Then, we can write 2-forms as formal triple vector products

$$\omega_2(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = \mathbf{a} \cdot \mathbf{x}(\boldsymbol{\xi}_1) \times \mathbf{x}(\boldsymbol{\xi}_2).$$

When \mathbf{x}_i is the canonical basis in \mathbb{R}^3 this is the usual triple product $\mathbf{a} \cdot \boldsymbol{\xi}_1 \times \boldsymbol{\xi}_2$. If \mathbf{a} represents a uniform velocity field of a fluid, this triple product gives the flux of the fluid through the parallelogram spanned by $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$, see Fig. 4.

3-forms. A 3-form in \mathbb{R}^3 is a trilinear, antisymmetric mapping $\mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3 \mapsto \mathbb{R}$. The set of all 3-forms is a linear space with dimension $C_3^3 = 1$. A basis for this space is given by the triple product

$$\mathbf{x}_1 \wedge \mathbf{x}_2 \wedge \mathbf{x}_3,$$

and so every 3-form can be expressed as $\omega_3 = \alpha \mathbf{x}_1 \wedge \mathbf{x}_2 \wedge \mathbf{x}_3$. Therefore

$$\omega_3(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3) = \alpha \begin{vmatrix} \mathbf{x}_1(\boldsymbol{\xi}_1) & \mathbf{x}_2(\boldsymbol{\xi}_1) & \mathbf{x}_3(\boldsymbol{\xi}_1) \\ \mathbf{x}_1(\boldsymbol{\xi}_2) & \mathbf{x}_2(\boldsymbol{\xi}_2) & \mathbf{x}_3(\boldsymbol{\xi}_2) \\ \mathbf{x}_1(\boldsymbol{\xi}_3) & \mathbf{x}_2(\boldsymbol{\xi}_3) & \mathbf{x}_3(\boldsymbol{\xi}_3) \end{vmatrix}.$$

If $\mathbf{x}_i = \mathbf{e}_i$ is the canonical basis in \mathbb{R}^3 , then

$$\omega_3(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3) = \alpha(\boldsymbol{\xi}_1 \cdot \boldsymbol{\xi}_2 \times \boldsymbol{\xi}_3)$$

is the oriented volume of the parallelepiped spanned by $\boldsymbol{\xi}_1$, $\boldsymbol{\xi}_2$ and $\boldsymbol{\xi}_3$, multiplied by α ; see Fig.4.

Exterior forms in \mathbb{R}^3 have some special properties that do not exist in other dimensions. Let \mathbf{x}_i denote an orthonormal basis in \mathbb{R}^3 . With every vector $\mathbf{a} \in \mathbb{R}^3$ we can associate the 1-form

$$\omega_1^{\mathbf{a}} = \mathbf{a}_1 \mathbf{x}_1 + \mathbf{a}_2 \mathbf{x}_2 + \mathbf{a}_3 \mathbf{x}_3, \quad (98)$$

and the 2-form

$$\omega_2^{\mathbf{a}} = \mathbf{a}_1 \mathbf{x}_2 \wedge \mathbf{x}_3 + \mathbf{a}_2 \mathbf{x}_3 \wedge \mathbf{x}_1 + \mathbf{a}_3 \mathbf{x}_1 \wedge \mathbf{x}_2. \quad (99)$$

Then, the exterior product of two 1-forms $\omega_1^{\mathbf{a}}$ and $\omega_1^{\mathbf{b}}$ is the vector product in \mathbb{R}^3 :

$$\omega_1^{\mathbf{a}} \wedge \omega_1^{\mathbf{b}} = \omega_2^{\mathbf{a} \times \mathbf{b}}, \quad (100)$$

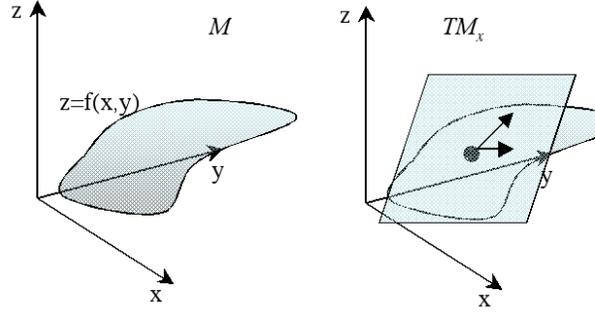


Figure 5: Manifold M and its tangent manifold at \mathbf{x}

and the exterior product of $\omega_1^{\mathbf{a}}$ and $\omega_2^{\mathbf{b}}$ is the dot product:

$$\omega_1^{\mathbf{a}} \wedge \omega_2^{\mathbf{b}} = (\mathbf{a}^T \mathbf{b}) \mathbf{x}_1 \wedge \mathbf{x}_2 \wedge \mathbf{x}_3 = \omega_3^{\mathbf{a}^T \mathbf{b}}.$$

These properties do not exist in other space dimensions. The reason is that only in \mathbb{R}^3 the space of all 2-forms is isomorphic to \mathbb{R}^3 itself, i.e., every 2-form on \mathbb{R}^3 can be associated with a vector in \mathbb{R}^3 . For $n \neq 3$, $C_n^2 \neq n$ and 2-forms are isomorphic to $\mathbb{R}^{C_n^2} \neq \mathbb{R}^n$.

To put it differently, vector calculus can only work in \mathbb{R}^3 . There's no natural way of defining vector products in higher dimensions. The reason we have a well-defined vector product in \mathbb{R}^3 is precisely because of the property (100), i.e., the fact that 2-forms can be identified with vectors in \mathbb{R}^3 .

4.2 Differential forms

Exterior forms operate on groups of vectors in the Euclidean space. Differential forms do the same thing, except that they draw their arguments from tangent spaces TM to differentiable manifolds M . Locally, a differentiable k -manifold embedded in \mathbb{R}^n is defined by the $n - k$ equations

$$f_i(\mathbf{x}) = 0 \quad i = 1, \dots, n - k,$$

where f_1, \dots, f_{n-k} are smooth functions $\mathbb{R}^n \mapsto \mathbb{R}$ such that ∇f_i are linearly independent. The tangent space at the point $\mathbf{x} \in M$ is defined by

$$TM_{\mathbf{x}} = (\text{span}\{\nabla f_1, \dots, \nabla f_{n-k}\})^\perp.$$

A differentiable manifold M and its tangent space at \mathbf{x} are shown on Fig.5. The union of all tangent spaces $TM = \cup_{\mathbf{x} \in M} TM_{\mathbf{x}}$ is called *tangent bundle* of M . In what follows we restrict attention to $M = \mathbb{R}^n$.

Definition 5 A differential k -form $\omega_k|_{\mathbf{x}}$ at a point $\mathbf{x} \in M$ is an exterior k -form on $TM_{\mathbf{x}}$. If such a form ω_k is given at every point $\mathbf{x} \in M$, and if it is differentiable, it is called k -form on M . A differential 0-form is a function on M . The set of all k -forms on M will be denoted by $W^k(M)$.

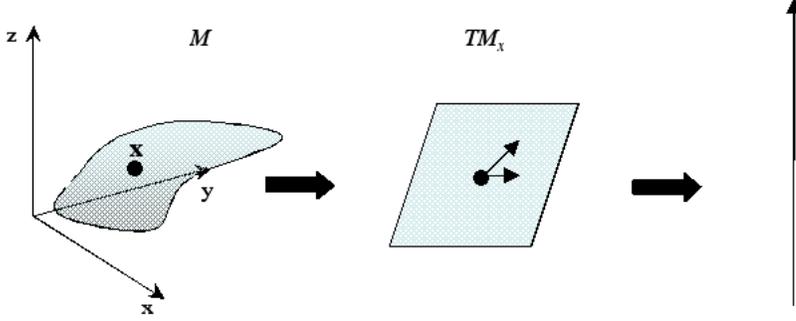


Figure 6: Differential k -form on M is an exterior k -form on $TM_{\mathbf{x}}$.

This definition is illustrated on Fig.6. We can think of differential forms as being composed from two separate parts. One is an exterior form that operates on vectors from $TM_{\mathbf{x}}$ just like any exterior form would. The second part is a function that modifies the value returned by the exterior form depending on the point \mathbf{x} . A 0-form only has a "functional" part because it doesn't accept vector arguments. This intuitive interpretation can be formalized, and one can prove; see [7, p.177], that any k -form on \mathbb{R}^n has the representation

$$\omega_k = \sum_{i_1 < \dots < i_k} a_{i_1, \dots, i_k}(\mathbf{x}) dx_{i_1} \wedge \dots \wedge dx_{i_k}, \quad (101)$$

where $a_{i_1, \dots, i_k}(\mathbf{x})$ are smooth functions and dx_i are the basic differential 1-forms on $TM_{\mathbf{x}}$. We will call the set of functions $\{a_{i_1, \dots, i_k}(\mathbf{x})\}$ a *proxy* of the k -form ω_k .

Exterior product is extended to differential forms in an obvious manner from the multilinear case. The reader is asked to verify that for $k + l \leq m = \dim M$ the product $\omega_k \wedge \omega_l$ defines a $(k + l)$ -form on M and that

$$\omega_k \wedge \omega_l = (-1)^{kl} \omega_l \wedge \omega_k. \quad (102)$$

Example 1 Evaluation of differential forms. Let $M = \mathbb{R}^2$,

$$\omega_1 = x_2 dx_1 - x_1 dx_2 \quad \text{and} \quad \omega_2 = (x_1 + x_2^2) dx_1 \wedge dx_2.$$

We compute ω_1 for ξ_2 shown on the left side in Fig.7 as follows. For this vector $\mathbf{x} = (1, 1)$ and¹² $dx_1(\xi) = dx_2(\xi) = 1$. Therefore,

$$\omega_1(\xi_2) = 2 - 2 = 0.$$

For ω_2 we take the pair (ξ_2, η_2) from the right side in Fig.7. Now $\mathbf{x} = (2, 1)$, $x_1 + x_2^2 = 3$, and

$$dx_1 \wedge dx_2(\xi_2, \eta_2) = \begin{vmatrix} dx_1(\xi_1) & dx_2(\xi_2) \\ dx_1(\eta_2) & dx_2(\eta_2) \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ -1 & 1 \end{vmatrix} = 2,$$

and so, $\omega_2(\xi_2, \eta_2) = 6$.

¹²Remember that dx_i act just like regular exterior basic forms, except that they operate on $TM_{\mathbf{x}}$. In this case $TM_{\mathbf{x}} = \mathbb{R}^2$. Therefore, $dx_i(\xi)$ simply returns the i th coordinate of ξ .

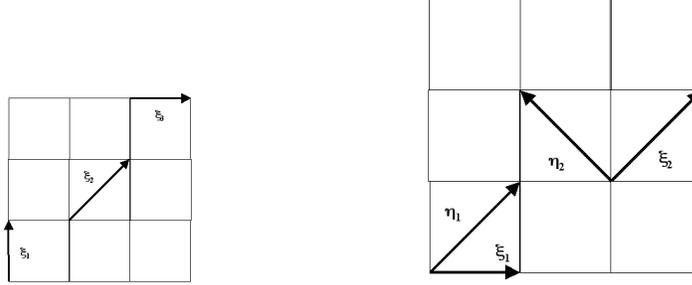


Figure 7: Vectors for Example 1.

4.2.1 Integration

Differential k -forms can be integrated over k -dimensional manifolds. The integral is defined in the usual manner by dividing the manifold into small pieces and taking the limit of the partial sums. The role of the pieces is played by oriented cells arranged in chains. The formal definition of an integral of a k -form over k -manifold is rather technical and will not be presented here. For details we refer the reader to [7, p.181]. However, the notions of a chain and a cell are important for approximation and discretization and we review them next.

Cells and chains Let M be an n -dimensional manifold. A k -cell K of M is the triple¹³ (\hat{K}, f, δ) where \hat{K} is a convex polyhedron in \mathbb{R}^k ; f is a differentiable map $f : \hat{K} \mapsto M$ and δ denotes orientation on \mathbb{R}^k .

Definition 6 A k -chain on a manifold M is a finite collection of k -cells K_1, \dots, K_r with multiplicities m_1, \dots, m_r . We write

$$C_k = \sum_{i=1}^r m_i K_i.$$

If K is a k -cell its boundary forms a $(k-1)$ -chain, denoted by ∂K . The $(k-1)$ -cells K_i of the boundary chain are the triples $(\hat{K}_i, f_i, \delta_i)$, where \hat{K}_i are the $(k-1)$ -dimensional faces of \hat{K} with orientations δ_i and f_i are maps $f_i : \hat{K}_i \mapsto \mathbb{R}^k$. The orientations δ_i are chosen so as to match the orientation of a given coordinate frame in \mathbb{R}^k . Each cell is taken with multiplicity one, i.e.,

$$\partial K = \sum K_i.$$

The boundary of a k -chain is

$$\partial C_k = \sum_{i=1}^r m_i \partial K_i.$$

¹³It is worth pointing out that the definition of a cell as the image of a standard convex polyhedron resembles a lot the definition of a finite element as the image of a standard reference element; see Ciarlet [29, p.78]

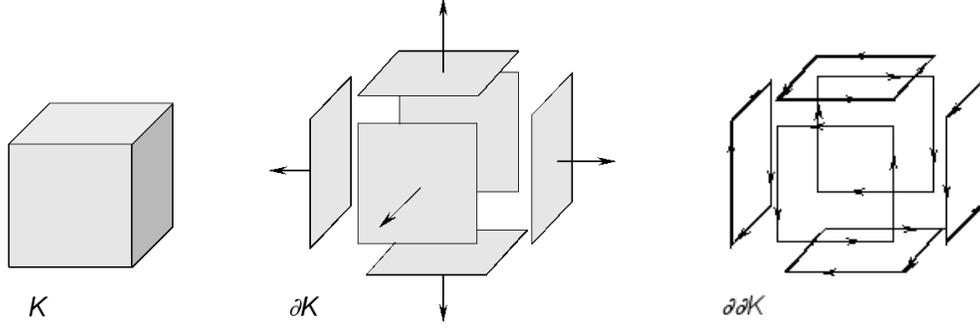


Figure 8: $\partial\partial K = 0$

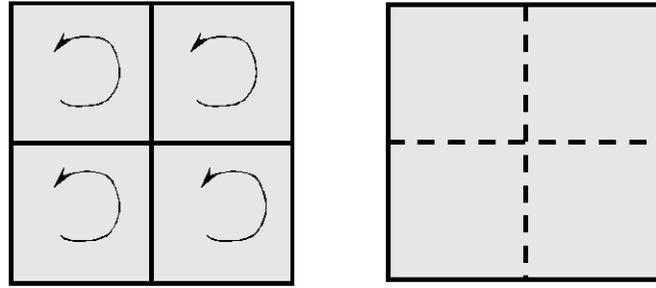


Figure 9: Boundary of a chain of four squares.

A fundamental property of the boundary operator is that

$$\partial\partial C_k = 0 \tag{103}$$

for any k -chain.

Example 2 Boundary of a boundary is zero. Consider a 3-chain consisting of one hexahedral 3-cell K . The boundary ∂K of this cell contains the six faces of K with orientation provided by the outward normal to each face. The boundary of each face consist of 4 edges. Therefore, $\partial\partial K$ is a chain of edges, where each edge of K enters twice with opposite multiplicities and so $\partial\partial K = 0$; see Fig.8.

Example 3 Boundary of a chain of quads. Consider the 2-chain C_2 of 4 quadrilaterals Q_i shown on Fig.9. The boundary ∂C_2 is sum of the boundaries ∂Q_i of each quadrilateral, i.e., their edges. If an edge is shared by two quadrilaterals it enters the sum twice with opposite multiplicities. The boundary thus contains only the edges that are not shared by two quads; see Fig.9.

Definition 7 Integral of a k -form over a k -chain. Let $C_k = \sum_{i=1}^r m_i K_i$.

$$\int_{C_k} \omega_k = \sum_{i=1}^r m_i \int_{K_i} \omega_k. \tag{104}$$

4.2.2 Exterior derivative

Exterior derivative takes a k -form into a $(k+1)$ -form on the same manifold. The definition we are about to give can be motivated by the following example.

Let \mathbf{a} be a smooth vector field in \mathbb{R}^3 . In vector calculus texts divergence of \mathbf{a} is normally defined as a formal dot product of the ∇ operator and \mathbf{a} , i.e.,

$$\operatorname{div} \mathbf{a} \equiv \nabla \cdot \mathbf{a} = \frac{\partial \mathbf{a}_1}{\partial x_1} + \frac{\partial \mathbf{a}_2}{\partial x_2} + \frac{\partial \mathbf{a}_3}{\partial x_3}.$$

Divergence can also be defined in a completely coordinate-independent way. Consider a point \mathbf{x} and a cube K that contains the point. Then¹⁴

$$\operatorname{div} \mathbf{a}(\mathbf{x}) = \lim_{\epsilon \rightarrow 0} \frac{\int_{\partial K} \mathbf{a} \cdot \mathbf{n} dS}{\operatorname{Vol}(\epsilon K)}. \quad (105)$$

Therefore, for small K

$$F(\partial K) \equiv \int_{\partial K} \mathbf{a} \cdot \mathbf{n} dS = \operatorname{div} \mathbf{a}(\mathbf{x})[\operatorname{Vol}(\epsilon K)] + o(\operatorname{Vol}(\epsilon K)),$$

i.e., divergence is the leading order term in the flux $F(\partial K)$. Exterior derivative of a k -differential form is defined following the same template: it is the leading multilinear term of the "flux", i.e., the integral of the form, computed on the boundary of a $(k+1)$ -cell. Note that this is consistent with the requirement that p -forms are integrated on p -chains.

Definition 8 *The exterior derivative $d\omega_k$ of ω_k is the principal multilinear part of the flux of ω_k over the boundary ∂K of the $(k+1)$ -cell K , i.e.,*

$$d\omega(\mathbf{x}) = \lim_{\epsilon \rightarrow 0} \frac{F_\omega(\partial K)}{\operatorname{Vol}(\epsilon K)} \quad (106)$$

where the flux is given by

$$F_\omega(\partial K) = \int_{\partial K} \omega_k.$$

It can be shown that $d\omega_k$ is indeed a $(k+1)$ -form and that (106) does not depend on the choice of K . If ω_k is given by (101), then (see [7, p.190])

$$d\omega_k = \sum_{i_1 < \dots < i_k} da_{i_1, \dots, i_k}(\mathbf{x}) \wedge dx_{i_1} \wedge \dots \wedge dx_{i_k}. \quad (107)$$

This formula gives a coordinate dependent representation of the exterior derivative. A fundamental property of d , that can be easily checked using (107), is that

$$dd\omega = 0 \quad (108)$$

¹⁴We will see later that such coordinate independent definitions of the gradient, curl and divergence form the basis for the *mimetic* finite difference operators introduced by M. Shashkov and M. Hyman [53]

for any differential form. Another important property, that can be easily verified using (107), is that

$$d(\omega_k \wedge \omega_l) = (d\omega_k) \wedge \omega_l + (-1)^k \omega_k \wedge (d\omega_l). \quad (109)$$

The unifying power of differential form abstraction is demonstrated in the next theorem. It merges the Newton-Leibniz Theorem, Stokes's Circulation Theorem, and Gauss's Divergence Theorem in one simple and elegant formula.

Theorem 7 For any ω

$$\int_{\partial C} \omega = \int_C d\omega. \quad (110)$$

As a corollary to this theorem and (109) we can prove the integration by parts formula

$$\int_{\partial C} \omega_k \wedge \omega_l = \int_C (d\omega_k) \wedge \omega_l + (-1)^k \int_C \omega_k \wedge (d\omega_l). \quad (111)$$

4.2.3 Duality

Consider a form ω_p and a chain $C_p = \sum_i m_i K_i^p$. We can "sample" ω on the cells of the chain by assigning to each cell the flux $F_\omega(K_i^p)$. The association

$$C_p \longrightarrow \left\{ \int_{K_i^p} \omega_p \right\} \quad (112)$$

between a p -form and its "sample" on a p -chain defines a mapping called *cochain* and denoted by C_p^* . The cochain can be thought of as an approximation or a representation of the differential form on a chain. It is approximate because the chain is finite and allows us to "sample" the form only at a finite number of configurations. The cochain is also dual to the chain because, using formula (104), it assigns to C_p a number, representing the global "flux". We express this duality relation by writing

$$(C_p, C_p^*) = F(\omega_p).$$

Consider now a form ω_p , its differential $d\omega_p$, a chain $C_{(p+1)}$ and its boundary chain

$$\partial C_{(p+1)} = C_p = \sum_i m_i K_p^i.$$

Let C_p^* and $C_{(p+1)}^*$ denote the co-chains of $\partial C_{(p+1)}$ and $C_{(p+1)}$, respectively. From the Stokes formula it follows that

$$(\partial C_{(p+1)}, C_p^*) = \int_{\partial C_{(p+1)}} \omega_p = \int_{C_{(p+1)}} d\omega_p = (C_{(p+1)}, C_{(p+1)}^*).$$

This identity serves to define an operator

$$d : C_p^* \mapsto C_{(p+1)}^* \quad (113)$$

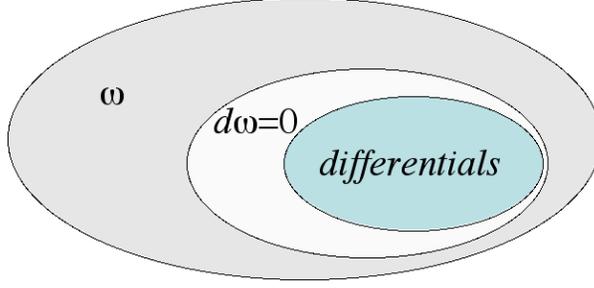


Figure 10: k th cohomology group.

called *coboundary*. In terms of this operator,

$$(\partial C_{(p+1)}, C_p^*) = (C_{(p+1)}, dC_p^*).$$

We can justify the use of the exterior derivative symbol by writing formally the Stokes theorem as

$$(\partial C, \omega) = (C, d\omega).$$

4.2.4 De Rham co-homology

A differential form ω is called *closed* if $d\omega = 0$. From (108) it is clear that every differential $d\omega$ of a k -form is a closed $(k + 1)$ -form. We call such forms *exact*. However, it turns out that there are closed forms that are not differentials. The "discrepancy" between closed forms and differentials can be measured by the dimension of the factor space

$$H_M^k = \ker\{d, W^k(M)\} / (dW^{k-1}(M)) \quad (114)$$

called *k-th cohomology group* of the manifold M ; see Fig.10. The dimension of this space is called *k-th Betti number* of M , b_k . The Betti numbers are topological invariants of M . If $M = \mathbb{R}^n$, then all closed forms are differentials. The same property is true for *contractible* and *star-shaped* regions¹⁵. This result is known as the *Poincare lemma*; see e.g., [73, p.69].

Theorem 8 Poincare lemma. *If M is star-shaped and $\omega_k \in W^k(M)$, $k > 0$ is closed, then $\omega_k = d\omega_{(k-1)}$ for some $\omega_{(k-1)} \in W^{(k-1)}(M)$.*

Structures consisting from spaces and an operator \mathcal{L} between them that has the property $\mathcal{L}\mathcal{L} \equiv 0$ are called homological complexes. The homological complex consisting of differential forms $W^k(M)$ and the operator d is called *De Rham complex*. In \mathbb{R}^3 De Rham's complex contains the forms $W^0(M)$, $W^1(M)$, $W^2(M)$, and $W^3(M)$. If M is contractible, or star-shaped, then the sequence

$$\mathbb{R} \hookrightarrow W^0(M) \xrightarrow{d} W^1(M) \xrightarrow{d} W^2(M) \xrightarrow{d} W^3(M) \longrightarrow 0 \quad (115)$$

¹⁵A region M is star-shaped if there exists a point $\mathbf{x}^* \in M$ such that if $\mathbf{x} \in M$ is arbitrary, the point $\mathbf{x}^* + \lambda(\mathbf{x} - \mathbf{x}^*)$ is in M for all $0 \leq \lambda \leq 1$. The formal definition of contractible M is more involved. Important example that will be sufficient for us are simply connected regular domains with connected boundary.

is *exact*, i.e., the closed forms in $W^0(M)$ are the constants and all closed forms in $W^k(M)$, $k = 1, 2, 3$ are differentials of $(k - 1)$ -forms. Therefore, for contractible and star-shaped regions $b_0 = 1$ and $b_1 = b_2 = 0$.

Note that for any 3-form in \mathbb{R}^3 we have that $d\omega_3 = 0$. Therefore, $W^3(M)$ contains only closed forms. The last link in (115), i.e., $W^3(M) \mapsto 0$, means that for regions without peculiarities all forms in $W^3(M)$ are in fact differentials, that is exact forms. But this also means that d is a surjective map from $W^2(M)$ into $W^3(M)$.

4.2.5 Hodge star operator

A manifold M can be endowed with a metric γ , that is a quadratic positive definite and symmetric form. The properties of differential forms discussed so far do not depend on the choice of this metric.

The Hodge $*_\gamma$ -operator is a mapping $W^k(M) \mapsto W^{m-k}(M)$ that depends on the metric selection on M . For a precise definition of this operator we refer to [73, p.356]. For the Euclidean metric the action of $*$ on the basic forms is given by (see [80, p.30])

$$\begin{aligned} *1 &= dx_1 \wedge dx_2 \wedge dx_3, \\ *dx_1 &= dx_2 \wedge dx_3; \quad *dx_2 = dx_3 \wedge dx_1; \quad *dx_3 = dx_1 \wedge dx_2, \\ *(dx_1 \wedge dx_2) &= dx_3; \quad *(dx_2 \wedge dx_3) = dx_1; \quad *(dx_3 \wedge dx_1) = dx_2, \end{aligned}$$

and

$$*(dx_1 \wedge dx_2 \wedge dx_3) = 1,$$

where 1 is the proxy of 0-forms. The Hodge operator defines an inner product on $W^k(M)$ according to

$$\langle \omega_k^1, \omega_k^2 \rangle = \int_M *_\gamma \omega_k^1 \wedge \omega_k^2.$$

Finally, we note that the Hodge operator can be used to connect two copies of an exact sequence into the following structure:

$$\begin{array}{ccccccc} W^0(\Omega) & \xrightarrow{d} & W^1(\Omega) & \xrightarrow{d} & W^2(\Omega) & \xrightarrow{d} & W^3(\Omega) \\ * \downarrow & & * \downarrow & & * \downarrow & & * \downarrow \\ W^3(\Omega) & \xleftarrow{d} & W^2(\Omega) & \xleftarrow{d} & W^1(\Omega) & \xleftarrow{d} & W^0(\Omega) \end{array} \quad (116)$$

4.3 Exercises

1. Prove that $\mathbf{x}_i \wedge \mathbf{x}_j = 0$ for $i \neq j$.
2. Show that $\omega_2^{ij} = \mathbf{x}_i \wedge \mathbf{x}_j$, $i < j$ is a basis for the space of all 2-forms.
3. Compute $\omega_1 = dx_1$ and $\omega_2 = dx_1 \wedge dx_2$ for the vectors shown on Fig.7.
4. Prove (108).
5. Prove (102).

5 Differential forms and PDE structure

In this section we use differential forms to describe the structure of second order PDE's. We will consider manifolds that are open and bounded regions Ω in \mathbb{R}^3 with smooth boundaries $\partial\Omega$. We note that in this case $T\Omega_{\mathbf{x}}$ is isomorphic to \mathbb{R}^3 . To simplify matters we fix the basis in \mathbb{R}^3 to be the Cartesian triple $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. Finally, we assume that Ω is contractible or star-shaped so that the De Rham complex

$$\mathbb{R} \hookrightarrow W^0(\Omega) \xrightarrow{d} W^1(\Omega) \xrightarrow{d} W^2(\Omega) \xrightarrow{d} W^3(\Omega) \longrightarrow 0 \quad (117)$$

is exact.

5.1 Differential forms in \mathbb{R}^3 and their proxies

Recall that in three dimensions 1 and 2 exterior forms were isomorphic to \mathbb{R}^3 and that every vector $\mathbf{a} \in \mathbb{R}^3$ engendered the forms $\omega_1^{\mathbf{a}}$ and $\omega_2^{\mathbf{a}}$ according to (98)-(99). A similar connection exists between vector fields in \mathbb{R}^3 and differential forms. With a smooth vector field

$$\mathbf{a}(\mathbf{x}) = \mathbf{a}_1(\mathbf{x})\mathbf{e}_1 + \mathbf{a}_2(\mathbf{x})\mathbf{e}_2 + \mathbf{a}_3(\mathbf{x})\mathbf{e}_3,$$

we can associate the 1-form

$$\omega_1^{\mathbf{a}} = \mathbf{a}_1(\mathbf{x})dx_1 + \mathbf{a}_2(\mathbf{x})dx_2 + \mathbf{a}_3(\mathbf{x})dx_3, \quad (118)$$

and the 2-form

$$\omega_2^{\mathbf{a}} = \mathbf{a}_1(\mathbf{x})dx_2 \wedge dx_3 + \mathbf{a}_2(\mathbf{x})dx_3 \wedge dx_1 + \mathbf{a}_3(\mathbf{x})dx_1 \wedge dx_2. \quad (119)$$

Therefore, vector fields in \mathbb{R}^3 serve as proxies for 1 and 2-forms. The proxies of 0 and 3-forms are scalar functions. If ϕ is a smooth function, then the 0-form ω_0^ϕ is the function ϕ itself, and the 3-form ω_3^ϕ is given by

$$\omega_3^\phi = \phi(\mathbf{x})dx_1 \wedge dx_2 \wedge dx_3. \quad (120)$$

5.1.1 Exterior derivative and grad, div and curl.

We show that exterior differentiation of a form in \mathbb{R}^3 corresponds to the application of grad, curl or div to its proxy.

Lemma 3 *Let ϕ and \mathbf{a} be a given function and a vector field defined on Ω . Then,*

$$d\omega_0^\phi = \omega_1^{\nabla\phi}, \quad (121)$$

$$d\omega_1^{\mathbf{a}} = \omega_2^{\nabla \times \mathbf{a}}, \quad (122)$$

and

$$d\omega_2^{\mathbf{a}} = \omega_3^{\nabla \cdot \mathbf{a}}. \quad (123)$$

Proof. Consider the 0-form ω_0^ϕ . From (107)

$$d\omega_0 = d\phi = \frac{\partial\phi}{\partial x_1}dx_1 + \frac{\partial\phi}{\partial x_2}dx_2 + \frac{\partial\phi}{\partial x_3}dx_3.$$

Using the association (118) between a 1-form and a vector field we obtain (121).

Next, consider $\omega_1^{\mathbf{a}}$. According to (107) its exterior derivative is

$$\begin{aligned} d\omega_1^{\mathbf{a}} &= d\mathbf{a}_1(\mathbf{x}) \wedge dx_1 + d\mathbf{a}_2(\mathbf{x}) \wedge dx_2 + d\mathbf{a}_3 \wedge (\mathbf{x})dx_3 \\ &= \left(\frac{\partial\mathbf{a}_1}{\partial x_2}dx_2 + \frac{\partial\mathbf{a}_1}{\partial x_3}dx_3 \right) \wedge dx_1 + \left(\frac{\partial\mathbf{a}_2}{\partial x_1}dx_1 + \frac{\partial\mathbf{a}_2}{\partial x_3}dx_3 \right) \wedge dx_2 \\ &\quad + \left(\frac{\partial\mathbf{a}_3}{\partial x_1}dx_1 + \frac{\partial\mathbf{a}_3}{\partial x_2}dx_2 \right) \wedge dx_3 \\ &= \left(\frac{\partial\mathbf{a}_3}{\partial x_2} - \frac{\partial\mathbf{a}_2}{\partial x_3} \right) dx_2 \wedge dx_3 + \left(\frac{\partial\mathbf{a}_1}{\partial x_3} - \frac{\partial\mathbf{a}_3}{\partial x_1} \right) dx_3 \wedge dx_1 \\ &\quad + \left(\frac{\partial\mathbf{a}_2}{\partial x_1} - \frac{\partial\mathbf{a}_1}{\partial x_2} \right) dx_1 \wedge dx_2. \end{aligned}$$

The coefficients multiplying the basic 2-forms are exactly the components of $\nabla \times \mathbf{a}$, which proves (122).

Finally, we use (107) to compute the exterior derivative of $\omega_2^{\mathbf{a}}$:

$$\begin{aligned} d\omega_2^{\mathbf{a}} &= d\mathbf{a}_1(\mathbf{x}) \wedge dx_2 \wedge dx_3 + d\mathbf{a}_2(\mathbf{x}) \wedge dx_3 \wedge dx_1 + d\mathbf{a}_3(\mathbf{x}) \wedge dx_1 \wedge dx_2 \\ &= \left(\frac{\partial\mathbf{a}_1}{\partial x_1} + \frac{\partial\mathbf{a}_2}{\partial x_2} + \frac{\partial\mathbf{a}_3}{\partial x_3} \right) dx_1 \wedge dx_2 \wedge dx_3 \\ &= (\nabla \cdot \mathbf{a}) dx_1 \wedge dx_2 \wedge dx_3. \end{aligned}$$

□

It is now clear that the familiar vector calculus identities

$$\nabla \times (\nabla\phi) = 0 \quad \text{and} \quad \nabla \cdot (\nabla \times \mathbf{a})$$

simply represent the fundamental identity $dd = 0$ in terms of the proxy fields. Also, specialized to proxies, the Stokes theorem (110) gives

$$\phi(\mathbf{p}) - \phi(\mathbf{q}) = \int_1 \nabla\phi \, dl \quad \text{where } \mathbf{p} - \mathbf{q} = \partial\mathbf{l}; \quad (124)$$

$$\int_1 \mathbf{a} \cdot \mathbf{t} \, dl = \int_S (\nabla \times \mathbf{a}) \cdot \mathbf{n} \, dS \quad \text{where } \mathbf{l} = \partial S \quad (125)$$

and

$$\int_S \mathbf{a} \cdot \mathbf{n} \, dS = \int_V \nabla \cdot \mathbf{a} \, dV \quad \text{where } S = \partial V. \quad (126)$$

We also have the analogue of Poincare's lemma for the proxies.

Theorem 9 Let \mathbf{a} , \mathbf{b} and ϕ denote two smooth vector fields and a smooth function in a star-shaped domain Ω . If

$$\nabla \times \mathbf{a} = 0 \quad \text{and} \quad \nabla \cdot \mathbf{b} = 0,$$

then there exist smooth vector fields \mathbf{c} , \mathbf{d} and a smooth function ψ , such that

$$\begin{aligned} \mathbf{a} &= \nabla \psi \\ \mathbf{b} &= \nabla \times \mathbf{c} \\ \phi &= \nabla \cdot \mathbf{d}. \end{aligned}$$

5.2 A differential De Rham complex in \mathbb{R}^3

From (117) we obtain the differential complex

$$\mathbb{R} \hookrightarrow C^\infty(\Omega) \xrightarrow{\nabla} \mathbf{C}^\infty(\Omega) \xrightarrow{\nabla \times} \mathbf{C}^\infty(\Omega) \xrightarrow{\nabla \cdot} C^\infty(\Omega) \longrightarrow 0, \quad (127)$$

by replacing forms with proxies. This complex is also exact (a fact that can be inferred directly from the vector calculus version of Poincaré's lemma (Theorem 9)).

Remark 4 The sequence (127) will not be exact when Ω has topological peculiarities such as holes and cycles. The values of the Betti numbers for such domains are related to the kind and type of peculiarities:

$$b_0 = \dim \left[\ker(\mathbf{grad}, C^\infty(\Omega)) \right]$$

gives the number of the connected components in Ω ;

$$b_1 = \dim \ker \left[(\mathbf{curl}, \mathbf{C}^\infty(\Omega)) / \mathbf{grad} (C^\infty(\Omega)) \right]$$

is the number of cycles (loops) in Ω , and

$$b_2 = \dim \ker \left[(\mathbf{div}, C^\infty(\Omega)) / \mathbf{curl} (C^\infty(\Omega)) \right]$$

gives the number of holes.

Remark 5 The number $\chi(\Omega) = b_0 - b_1 + b_2 - b_3$ is called Euler-Poincaré constant of Ω . It is a generalization of the famous Euler-Poincaré formula

$$\text{Nodes} - \text{Edges} + \text{Faces} - \text{Tetrahedrals} = \text{const}$$

for any simplicial mesh of a given domain $\Omega \subset \mathbb{R}^3$.

5.2.1 L^2 -based De Rham complex.

In PDE's one has to account for various boundary conditions. We will consider a simple situation wherein the boundary of Ω consists of a single smooth piece denoted by Γ . We will also switch to L^2 based proxies because of the importance of Sobolev spaces in PDE analysis. Let $H(\Omega, \mathbf{grad})$, $H(\Omega, \mathbf{curl})$, and $H(\Omega, \mathbf{div})$ denote spaces of square integrable functions whose gradients, curls and divergences are also square integrable. As usual, the space of all square integrable functions on Ω is denoted by $L^2(\Omega)$. The De Rham differential complex

$$\mathbb{R} \hookrightarrow H(\Omega, \mathbf{grad}) \xrightarrow{\nabla} H(\Omega, \mathbf{curl}) \xrightarrow{\nabla \times} H(\Omega, \mathbf{div}) \xrightarrow{\nabla \cdot} L^2(\Omega) \longrightarrow 0, \quad (128)$$

is an exact sequence of spaces. If we incorporate the boundary conditions in the spaces, the domains of the gradient, curl and divergence, relative to Γ are

$$H_0(\Omega, \mathbf{grad}) = \{\phi \in H(\Omega, \mathbf{grad}) \mid \phi = 0 \text{ on } \Gamma\}, \quad (129)$$

$$H_0(\Omega, \mathbf{curl}) = \{\mathbf{u} \in H(\Omega, \mathbf{curl}) \mid \mathbf{u} \times \mathbf{n} = 0 \text{ on } \Gamma\}, \quad (130)$$

$$H_0(\Omega, \mathbf{div}) = \{\mathbf{u} \in H(\Omega, \mathbf{div}) \mid \mathbf{u} \cdot \mathbf{n} = 0 \text{ on } \Gamma\}, \quad (131)$$

These spaces define a De Rham differential complex relative to Γ :

$$\mathbb{R} \hookrightarrow H_0(\Omega, \mathbf{grad}) \xrightarrow{\nabla} H_0(\Omega, \mathbf{curl}) \xrightarrow{\nabla \times} H_0(\Omega, \mathbf{div}) \xrightarrow{\nabla \cdot} L_0^2(\Omega) \longrightarrow 0. \quad (132)$$

This complex is also an exact sequence. The assumptions made about Ω imply that the last link in (128) and (132) is surjective.

5.3 Encoding the PDE structure: Tonti diagrams

Let (117) be an exact sequence that corresponds to the differential De Rham complex (128). We can connect two copies of this sequence by the Hodge operator as we already did in (116):

$$\begin{array}{ccccccc} W^0(\Omega) & \xrightarrow{d} & W^1(\Omega) & \xrightarrow{d} & W^2(\Omega) & \xrightarrow{d} & W^3(\Omega) \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ W^3(\Omega) & \xleftarrow{d} & W^2(\Omega) & \xleftarrow{d} & W^1(\Omega) & \xleftarrow{d} & W^0(\Omega) \end{array} \quad (133)$$

We will call the structure in (133) a *primal-dual* complex. It can be traversed along its horizontal and vertical links, which corresponds to application of the exterior derivative or the Hodge operator, respectively. In this manner we can obtain different diagrams relating primal and dual forms.

Consider the three diagrams obtained by taking the two columns on the left,

$$\begin{array}{ccc} W^0(\Omega) & \xrightarrow{d} & W^1(\Omega) \\ \downarrow & & \downarrow \\ W^3(\Omega) & \xleftarrow{d} & W^2(\Omega) \end{array} \quad (134)$$

the two columns in the middle

$$\begin{array}{ccc}
 W^1(\Omega) & \xrightarrow{d} & W^2(\Omega) \\
 \downarrow & & \downarrow \\
 W^2(\Omega) & \xleftarrow{d} & W^1(\Omega)
 \end{array} \tag{135}$$

and the two rightmost columns:

$$\begin{array}{ccc}
 W^2(\Omega) & \xrightarrow{d} & W^3(\Omega) \\
 \downarrow & & \downarrow \\
 W^1(\Omega) & \xleftarrow{d} & W^0(\Omega)
 \end{array} . \tag{136}$$

It turns out that these three diagrams can be used to encode the structure of all linear second order elliptic, parabolic and hyperbolic PDE's in three dimensions; see [46] for more details and examples with mixed boundary conditions. Throughout the literature such symbolic representations appear under a variety of names and graphical formats. For example, Mattiussi [65] calls them *factorization* diagrams. We will also use the term *Tonti diagrams* after E. Tonti who first used graphical representations not dissimilar to (134)-(136) in [76] (see also [77], [24]).

In what follows we will focus on the elliptic case and refer to Hiptmair [46], for factorization diagrams of parabolic and hyperbolic equations. The three elliptic equations that correspond to (134)-(136) can be represented by a single Tonti diagram as

$$\begin{array}{ccccc}
 \omega_{(k-1)} & & \xrightarrow{d} & & (-1)^k \omega_k \\
 *_{\alpha} & & \downarrow & & \downarrow & *_{\beta} \\
 \boxed{f} & -\omega_{n-(k-1)} & \xleftarrow{d} & & \omega_{(n-k)}
 \end{array} \tag{137}$$

where $k = 1, 2, 3$ and f is a given source term. The diagram in (137) is more than just an elegant way to represent equations. Using (137) and the properties of d we can infer many facts about the invariants of the equation. Another fundamental aspect of this diagram is that it separates topological from metric relationships in the problem. The primal and the dual equations

$$d\omega_{(k-1)} = (-1)^k \omega_k \quad \text{and} \quad d\omega_{n-k} = -\omega_{n-(k-1)} + f, \tag{138}$$

correspond to horizontal links in (133), i.e., they do not involve any metric relationships. These equations are called *topological field equations* in [75], and *equilibrium equations* in [46]. Note that the equilibrium equations are completely unrelated as they express

topological relations that exist independently of each other on the primal and the dual complexes. What connects them are the two equations

$$\omega_{n-(k-1)} = (*_{\alpha})\omega_{(k-1)} \quad \text{and} \quad \omega_{n-k} = (*_{\beta})\omega_k \quad (139)$$

that define "constitutive" relations between the primal and the dual forms using the Hodge operator.

Let us consider the case $k = 1$ and convince ourselves that (137) does indeed recover a familiar PDE problem. The primal equation involves the forms ω_0^{ϕ} and $\omega_1^{\mathbf{v}}$ with proxies ϕ and \mathbf{v} , respectively. The dual equation uses the forms $\omega_2^{\mathbf{w}}$ and ω_3^{ψ} with proxies \mathbf{w} and ψ , respectively. Finally, expressed in terms of the proxies, the action of the Hodge operators amounts to a scaling of the fields connected by a constitutive equation. As a result, translating (137) into proxy equations gives the diagram

$$\begin{array}{ccc} \phi & \xrightarrow{\nabla} & -\mathbf{v} \\ \psi = \alpha\phi & \downarrow & \downarrow \quad \mathbf{w} = \beta\mathbf{v} \\ \boxed{f} & -\psi \xleftarrow{\nabla \cdot} & \mathbf{w} \end{array} \quad (140)$$

where for simplicity we have assumed constant "material" properties. Therefore, in terms of proxies the horizontal links in (137) correspond to the differential equations¹⁶

$$\nabla\phi = -\mathbf{v} \quad \text{and} \quad \nabla \cdot \mathbf{w} = -\psi + f,$$

while the vertical links provide the constitutive relations

$$\psi = \alpha\phi \quad \text{and} \quad \mathbf{w} = \beta\mathbf{v}.$$

Using the constitutive equations we can eliminate either the dual or the primal proxy fields. In the former case we obtain the first-order system

$$\begin{aligned} \nabla \cdot \beta\mathbf{v} + \alpha\phi &= f \quad \text{and} \quad \mathbf{v} + \nabla\phi = \mathbf{0} \quad \text{in } \Omega \\ \phi &= 0 \quad \text{on } \Gamma. \end{aligned} \quad (141)$$

If we proceed to further eliminate \mathbf{v} from (141), the result is the second order problem

$$-\nabla \cdot \beta\nabla\phi + \alpha\phi = f. \quad (142)$$

For $\beta = 1$ and $\alpha = 0$ the problem (142) and the system (141) are exactly the Poisson equation (10) and its first-order form (11).

We leave it as an exercise for the reader to convert the cases $k = 2, 3$ to PDE's in terms of the proxy fields and to carry out the elimination. The result is a curl-curl and grad-div type second order PDE's.

¹⁶In [65] Mattiussi refers to the first equilibrium equation as the *kinematic* equation and calls the second one *balance* equation.

6 Application to discretization of PDE's

Tonti diagrams are a succinct representation tool for encoding PDE structure. If we agree that preserving this structure is important, or at least not harmful for the approximation of the PDE, then *compatible* discretization can be viewed as the process of creating discrete analogues of Tonti diagrams. Such a discretization paradigm requires two principal ingredients:

1. two sets $\{W_h^k(\omega)\}$ and $\{\widehat{W}_h^k(\Omega)\}$ of *primal* and *dual* discrete differential forms that are exact with respect to the operations d_h and \hat{d}_h ;
2. a discrete Hodge operator

$$*_k^h : W_h^k(\omega) \mapsto \widehat{W}_h^{(n-k)}(\Omega)$$

that connects the primal and the dual spaces.

We will see that stable and conforming discretizations of problems like (141) do adhere to this paradigm, regardless of the particular method chosen to define the algebraic equations. In other words,

strategies that are able to account for the structure of the PDE encoded by its Tonti diagram lead to stable FD, FV and FE methods.

In contrast, discretizations that violate one or more tenets of this paradigm are not inherently stable and require *stabilizing modifications* in order to work.

Bossavit was among the first to realize this in the context of finite element discretizations of the Maxwell's equations; see [16], [17] and [18]. In terms of our diagrams this is the middle section of (133), i.e, the case $k = 2$ in (137). This work was followed by Mattiussi who in [65] used "factorization" diagrams, essentially equivalent to the diagrams presented here, to study FE, FV and FD methods for thermostatics, i.e., a problem whose structure is encoded by (137) for $k = 1$. Further development and generalization of these ideas was provided by Hiptmair in his fundamental work on discrete Hodge operators [46].

Our presentation will draw upon these and other sources with the primary goal being to expose the common structural properties of compatible discretizations. Thus, at the moment we will intentionally leave the definition of $\{W_h^k(\omega)\}$, d_h , $\{\widehat{W}_h^k(\Omega)\}$, and \hat{d}_h vague and return to it whenever we discuss particular examples.

6.1 Compatible discretization

Let $n(k)$ and $\hat{n}(k)$ denote the dimensions of $W_h^k(\omega)$ and $\widehat{W}_h^k(\omega)$, respectively. Since $*_k^h$ is a mapping between two finite dimensional spaces it must have the form of a matrix relation. Hiptmair pointed out in [46], that a generic form of this operator is given by the equation

$$\mathbb{M}\omega_k^h = \mathbb{K}\hat{\omega}_{(n-k)}^h$$

where \mathbb{M} is $n(k) \times n(k)$ symmetric "mass" matrix and \mathbb{K} is a rectangular $n(k) \times \hat{n}(n - k)$ matrix. Using $\{W_h^k(\omega)\}$, $\{\widehat{W}_h^k(\Omega)\}$ and the discrete Hodge operator we define the discrete primal-dual complex

$$\begin{array}{ccccccc}
W_h^0(\Omega) & \xrightarrow{d_h} & W_h^1(\Omega) & \xrightarrow{d_h} & W_h^2(\Omega) & \xrightarrow{d_h} & W_h^3(\Omega) \\
*_0^h \downarrow & & *_1^h \downarrow & & *_2^h \downarrow & & *_3^h \downarrow \\
\widehat{W}_h^3(\Omega) & \xleftarrow{\hat{d}_h} & \widehat{W}_h^2(\Omega) & \xleftarrow{\hat{d}_h} & \widehat{W}_h^1(\Omega) & \xleftarrow{\hat{d}_h} & \widehat{W}_h^0(\Omega)
\end{array} \tag{143}$$

Each one of the primal and dual sequences in (143) is a model for a "discrete vector calculus" where many results such as the Stokes formula (110) hold true exactly. Then, a compatible discretization of a problem represented by the Tonti diagram (137) can be defined by a discrete version of this diagram built on the discrete complex (143):

$$\begin{array}{ccccc}
\omega_{(k-1)}^h & \xrightarrow{d_h} & (-1)^k \omega_k^h & & \\
*__{(k-1)}^h \downarrow & & \downarrow & & *_k^h \\
\boxed{f} & -\omega_{n-(k-1)}^h & \xleftarrow{\hat{d}_h} & \omega_{(n-k)}^h &
\end{array} \tag{144}$$

The diagram (144) represents a generic discretization template that maps the PDE structure onto discrete spaces. An important feature of this template is that the discrete equilibrium equations will be satisfied exactly as long as $W_h^k(\omega)$ and $\widehat{W}_h^k(\omega)$ are exact with respect to d_h and \hat{d}_h . This means that any discretization based on (144) will also imitate the conservation properties that are inherent to (137).

6.1.1 Discrete differential forms

We present examples of discrete differential complexes that commonly arise in the discretization of PDE's. Our main interest is in discretization techniques that involve partitioning of Ω into small subdomains. Thus, we will assume that Ω can be represented exactly by a family of 3-chains, parametrized by h :

$$\Omega = C_3^h = \sum m_i K_i.$$

The number h^{-1} is assumed to be proportional to the number of the 3-cells K_i in the chain, not counting their multiplicities.

Cochain complex. The first example of a discrete differential complex uses co-chains to approximate differential forms. It is representative of the type of discretization that one encounters in finite difference and finite volume methods. We call this type of discretizations *direct* because they arise from approximation of forms rather than their field proxies. As a result, the fields in direct discretizations are represented in a "cell-wise" sense by values assigned to the p -cells of C_3^h .

To fix ideas let Ω be the unit cube K^3 in \mathbb{R}^3 and let $C_3^h = K^3$. Therefore, our chain contains only one 3-cell, six 2-cells (the faces K_i^2), twelve 1-cells (the edges K_i^1) and eight 0-cells (the nodes K_i^0). We number and orient the 0,1,2-cells as shown on Fig.11, and assume "source" orientation for K^3 . The orientations are important for the correct calculation of the boundary and co-boundary operators. For example,

$$\partial K^3 = -K_1^2 + K_2^2 + K_3^2 - K_4^2 - K_5^2 + K_6^2,$$

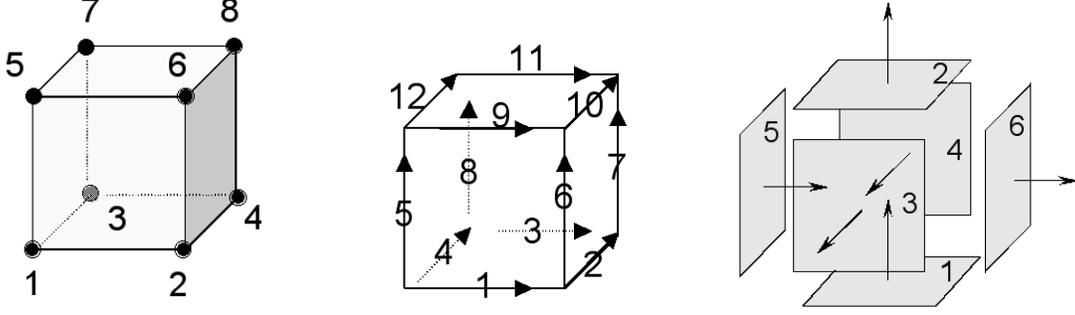


Figure 11: Numbering and orientation of 0,1 and 2-cells in a hexahedral.

the boundary of the bottom face is

$$\partial K_1^2 = K_1^1 + K_2^1 - K_3^1 - K_4^1,$$

the boundary of the top face is

$$\partial K_2^2 = K_9^1 + K_{10}^1 - K_{11}^1 - K_{12}^1,$$

the boundary of the sixth edge is

$$\partial K_6^1 = K_2^0 - K_6^0,$$

and similarly for the rest of the cells.

A form ω_p is represented on C_h^3 by its "fluxes"

$$F_\omega(K_i^p) = \int_{K_i^p} \omega_p$$

defined on the p -cells. These fluxes generate a finite dimensional set

$$W_h^p = \{F_\omega(K_i^p)\}$$

which we earlier identified with the p -cochain. Thus, we have four sets W_h^p that approximate differential forms of orders 0,1,2 and 3 on the chain C_h^3 . Now it is necessary to define an operator $d_h : W_h^p \mapsto W_h^{(p+1)}$ that will represent exterior differentiation on the cochain complex. To define this operator we apply the Stokes formula (110). Let $K_i^{(p+1)}$ be a p -cell, $p = 0, 1, 2$ and let

$$\partial K_i^{(p+1)} = \sum_j m_j K_j^p,$$

be the boundary of this cell. Then, by (110)

$$\int_{K_i^{(p+1)}} d\omega_p = \int_{\partial K_i^{(p+1)}} \omega_p. \quad (145)$$

Both the form and its derivative are represented on the chain by their fluxes:

$$\omega_p \mapsto \{F_\omega(K_i^p)\} \quad \text{and} \quad d\omega_p \mapsto \{F_{d\omega}(K_i^{(p+1)})\}.$$

Therefore, (145) can be restated as

$$F_{d\omega}(K_i^{(p+1)}) = \int_{K_i^{(p+1)}} d\omega_p = \int_{\partial K_i^{(p+1)}} \omega_p = \sum_j m_j \int_{K_j^p} \omega_p = \sum_j m_j F_\omega(K_j^p).$$

Therefore, the action of the exterior derivative on W_h^p is given by

$$d\{F_\omega(K_i^p)\} = \sum_j m_j F_\omega(K_j^p) \mapsto K_i^{(p+1)}. \quad (146)$$

where K_j^p are the boundary cells of $K_i^{(p+1)}$. This formula provides us with a well-defined operator $d_h : W_h^p \mapsto W_h^{(p+1)}$. This operator coincides with the coboundary operator. We also note that $d_h d_h = 0$, as desired from an exterior derivative.

Obviously, d_h can be expressed in matrix form. The rows of the matrix that represents $d_h : W_h^p \mapsto W_h^{(p+1)}$ will contain the multiplicities of the boundary pieces that form a $(p+1)$ -cell. For example, for the chain in Fig.11, $d_h : W_h^1 \mapsto W_h^2$ is given by the 6×12 matrix

$$\mathbb{C} = \begin{pmatrix} 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 \\ 1 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 0 & 0 \end{pmatrix}$$

This matrix corresponds to a curl. The mapping $d_h : W_h^2 \mapsto W_h^3$ is given by the 1×6 matrix

$$\mathbb{D} = (-1 \quad 1 \quad 1 \quad -1 \quad -1 \quad 1),$$

which corresponds to divergence. It is easy to see that $\mathbb{D}\mathbb{C} \equiv 0$. The matrix \mathbb{C} is the edge-face *mesh incidence* matrix, and \mathbb{D} is the face-cell incidence matrix. We also have the node-edge matrix \mathbb{G} with the property $\mathbb{C}\mathbb{G} = 0$. This matrix represents $d_h : W_h^0 \mapsto W_h^1$. \mathbb{G} , \mathbb{C} and \mathbb{D} depend only on the mesh connectivity and they will not change if the chain is deformed without changing its connectivity.

The strategy outlined above can be extended to chains of more than one 3-cell in an obvious manner. The W_h^p spaces are defined by the p -cochains, i.e., they represent forms by their fluxes on the p -cells. The coboundary operator then serves to define an exterior differentiation for W_h^p . It is also clear that this type of construction does not depend on the particular kind of cells used to form the chain.

Another important observation about the spaces and operations defined by cochains and coboundary operators is their complete metric independence. The definition of the exterior derivative depends only on the chain connectivity and not on the shapes of the cells in the chain. Thus, a deformation of the space that leaves the mesh connectivity intact will not change the definition of the exterior derivative. However, if a cell is broken

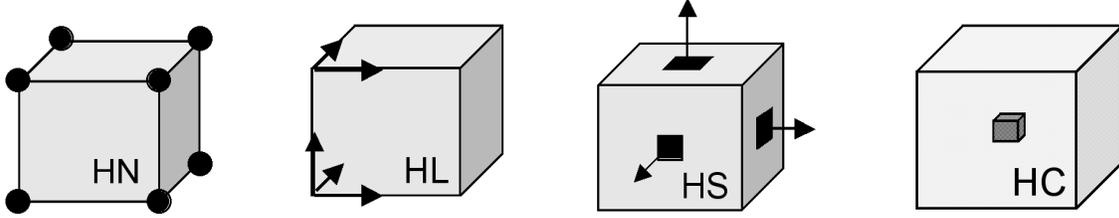


Figure 12: Mimetic complex.

into several smaller cells, or if new cells are added to the chain, d will change to a new operator.

For discretizations based on meshing, i.e., representation of the domain by finite chains, approximation of differential forms by cochains is almost an automatic choice. Its wide applicability to describe FE, FV and FD schemes was perhaps first realized and exploited by Mattiussi in [65]. In a recent work by Gross and Kotiuga [42], simplicial cochains and duality were used to develop a data structure for tetrahedral finite element meshes. Because of the universality of this principle it can be found either implicitly or explicitly in many existing discretization schemes. Examples of implicit applications are the direct discretization methods for electromagnetics [68], [69], mimetic methods [52], [53] and methods such as Yee's FDTD scheme [83] and FIT methods; see [79]. Methods that rely explicitly on differential forms formalism are usually dubbed "lattice vector calculus", or "lattice theories", see [72], [75], [74], and the references therein. Below we consider one popular example of implicit application of differential forms in discretizations of PDE's.

Mimetic complex. Mimetic finite differences (see [53], [54], [52], and [55]) provide a "discrete vector calculus" structure on logically rectangular grids or unstructured tetrahedral meshes. They can be considered as a particular example of the cochain complex discussed above.

A mimetic complex on a hexahedral mesh consists of four finite dimensional spaces, denoted by HN , HL , HS and HC , respectively, and operators

$$\text{GRAD} : HN \mapsto HL, \quad \text{CURL} : HL \mapsto HS, \quad \text{DIV} : HS \mapsto HC.$$

The spaces HN and HC are used to represent scalar functions by their nodal and cell center values respectively. HL and HS represent vector fields by their projections to the edges and to the normals to the faces of the mesh, respectively; see Fig.12. Thus, mimetic variables are "samples" of the proxy fields rather than fluxes.

Mimetic operators are defined using the coordinate-independent characterizations

$$\mathbf{l} \cdot \text{grad } \phi(\mathbf{x}) = \lim_{\text{Length}(\mathbf{l}) \rightarrow 0} \frac{\int_{\partial \mathbf{l}} \phi \, dl}{\text{Length}(\mathbf{l})},$$

$$\mathbf{n} \cdot \text{curl } \mathbf{a}(\mathbf{x}) = \lim_{\text{Area}(S) \rightarrow 0} \frac{\int_{\partial S} \mathbf{a} \cdot \mathbf{t} \, dl}{\text{Area}(S)},$$

$$\operatorname{div} \mathbf{a}(\mathbf{x}) = \lim_{\operatorname{Vol}(K) \rightarrow 0} \frac{\int_{\partial K} \mathbf{a} \cdot \mathbf{n} \, dS}{\operatorname{Vol}(K)},$$

of the gradient, curl and divergence. For example, on a hexahedral K^3 with faces numbered as in Fig.11,

$$\operatorname{DIV}|_{K^3} = \frac{1}{\operatorname{Vol}(K^3)} \left(-A_1 \mathbf{a}_1 + A_2 \mathbf{a}_2 + A_3 \mathbf{a}_3 - A_4 \mathbf{a}_4 - A_5 \mathbf{a}_5 + A_6 \mathbf{a}_6 \right)$$

where \mathbf{a}_i is the component of an HS vector on face K_i^2 and A_i is the area of this face. The connection between the mimetic divergence and the coboundary operator is clear. If

$$\mathbb{V} = \operatorname{Vol}(K^3) \quad \text{and} \quad \mathbb{A} = \operatorname{diag}(A_1, A_2, A_3, A_4, A_5, A_6)$$

we have that

$$\mathbb{V}(\operatorname{DIV}) = \mathbb{D} \mathbb{A},$$

where \mathbb{D} is the face-cell incidence matrix defined earlier. The matrices \mathbb{A} and \mathbb{V} represent the cell measures that are needed for the conversion of the mimetic variables to fluxes. The mimetic CURL is related to \mathbb{C} by a similar formula. Note that if mimetic variables are redefined to represent fluxes, then definition of mimetic operators will coincide with the coboundary. In this case GRAD, CURL and DIV will coincide with \mathbb{G} , \mathbb{C} and \mathbb{D} , respectively.

Finally, it is not hard to verify that

$$HN \quad \xrightarrow{\operatorname{GRAD}} \quad HL \quad \xrightarrow{\operatorname{CURL}} \quad HS \quad \xrightarrow{\operatorname{DIV}} \quad HC \quad (147)$$

is an exact sequence. This is equivalent to a validity of Poincare lemma ($\operatorname{CURL} \mathbf{a} = 0$ iff $\mathbf{a} = \operatorname{GRAD} \phi$, and $\operatorname{DIV} \mathbf{a} = 0$ iff $\mathbf{a} = \operatorname{CURL} \mathbf{b}$), and a Stokes theorem in the mimetic complex; see [55].

Whitney complex. Cochains and mimetic differences are examples of direct discretizations of differential forms. They replace a p -form by its p -cochain using either a sampling of its p -fluxes or its p -proxies on a p -chain. For direct discretizations proxy fields can be identified with "grid" functions that live on the p -cells of the mesh but are otherwise undefined. Also, in direct discretizations the action of the exterior derivative on p -forms is replaced by the action of coboundary on p -cochains.

A functional approach does not replace a p -form by a p -cochain. It uses the cochain to define a finite expansion of the proxy field in terms of some simple *basis* functions. Therefore, a functional approach does lead to an internal (conforming) approximation of $W^p(\Omega)$ in the sense that it creates a proper subspace $W_h^p(\Omega) \subset W^p(\Omega)$ of differential forms. A fundamental consequence from this fact is that the resulting discrete spaces inherit the exterior differentiation from $W^p(\Omega)$. This approach is representative of the type of discretization that arises from the application of finite element techniques.

As an example of a functional approximation we will consider the Whitney complex. This complex is defined on n -simplices and below we briefly review some of the relevant notions.

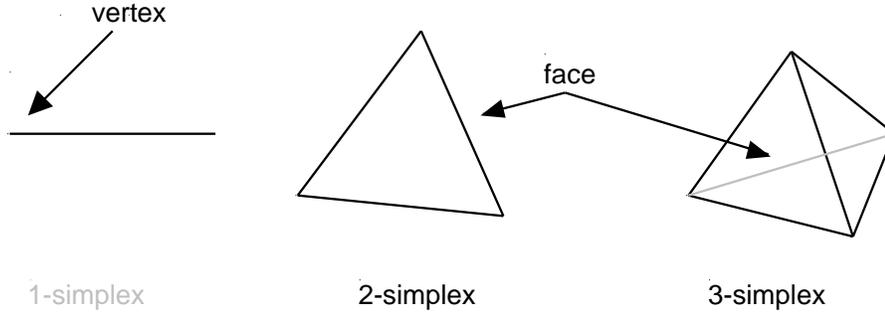


Figure 13: n -simplices in 1D, 2D and 3D

Definition 9 *n -simplex.* A set of $n + 1$ points $\{\mathbf{z}_i\}_{i=1}^{n+1}$ in \mathbb{R}^n is called n -simplex if the matrix

$$\begin{pmatrix} \mathbf{z}_{1,1} & \mathbf{z}_{1,2} & \mathbf{z}_{1,3} & \dots & \mathbf{z}_{1,n} & \mathbf{z}_{1,n+1} \\ \mathbf{z}_{2,1} & \mathbf{z}_{2,2} & \mathbf{z}_{2,3} & \dots & \mathbf{z}_{2,n} & \mathbf{z}_{2,n+1} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \mathbf{z}_{n,1} & \mathbf{z}_{n,2} & \mathbf{z}_{n,3} & \dots & \mathbf{z}_{n,n} & \mathbf{z}_{n,n+1} \\ 1 & 1 & 1 & \dots & 1 & 1 \end{pmatrix}$$

is non-singular.

The points \mathbf{z}_i are called *vertices* of the simplex. Fig.13 shows examples of 1,2 and 3-simplices. An n -simplex induces a local barycentric coordinate system.

Definition 10 Let $\mathbf{x} \in \mathbb{R}^n$ and let K^n be an n -simplex with vertices $\{\mathbf{z}_i\}_{i=1}^{n+1}$. The unique solution $\{\lambda_i(\mathbf{x})\}_{i=1}^{n+1}$ of the linear system

$$\begin{aligned} \lambda_1 \mathbf{z}_{1,1} + \dots + \lambda_{n+1} \mathbf{z}_{1,n+1} &= \mathbf{x}_1 \\ \lambda_1 \mathbf{z}_{2,1} + \dots + \lambda_{n+1} \mathbf{z}_{2,n+1} &= \mathbf{x}_2 \\ &\dots \dots \dots \\ \lambda_1 \mathbf{z}_{n,1} + \dots + \lambda_{n+1} \mathbf{z}_{n,n+1} &= \mathbf{x}_n \\ \lambda_1 + \dots + \lambda_{n+1} &= 1. \end{aligned} \tag{148}$$

is called *barycentric coordinates* of \mathbf{x} relative to K^n .

From (148) we see that

$$\sum_{i=1}^{n+1} \lambda_i(\mathbf{x}) = 1 \quad \text{and} \quad \mathbf{x} = \sum_{i=1}^{n+1} \lambda_i(\mathbf{x}) \mathbf{z}_i.$$

Other properties of barycentrics are that

- λ_i are linear functions of \mathbf{x} :

$$\lambda_i\left(\sum_{j=1}^m \alpha_j \mathbf{x}_j\right) = \sum_{j=1}^m \alpha_j \lambda_i(\mathbf{x}_j); \tag{149}$$

- each barycentric equals 1 at exactly one vertex and is zero at all other vertices:

$$\lambda_i(\mathbf{z}_j) = \delta_{ij}. \quad (150)$$

In [82] Whitney used barycentric coordinates to define differential forms on simplices. He associated 0-forms with λ_i . Then, for an arbitrary k -simplex $\{\mathbf{z}_i\}_{i=1}^{k+1}$ he defined the k -form

$$\omega_k = k! \sum_{j=1}^{k+1} (-1)^{j-1} (\lambda_m) d\lambda_1 \wedge \dots \wedge d\lambda_{m-1} \wedge d\lambda_{m+1} \wedge \dots \wedge d\lambda_k.$$

In three dimensions the Whitney forms associated with the nodes K_i^0 , edges K_{ij}^1 , faces K_{ijk}^2 and tetrahedrals K_{ijkl}^3 are

$$\begin{aligned} \omega_0 &= \lambda_i \\ \omega_1 &= \lambda_i d\lambda_j - \lambda_j d\lambda_i \\ \omega_2 &= 2(\lambda_i d\lambda_j \wedge d\lambda_k + \lambda_j d\lambda_k \wedge d\lambda_i + \lambda_k d\lambda_i \wedge d\lambda_j) \\ \omega_3 &= 6(\lambda_i d\lambda_j \wedge d\lambda_k \wedge d\lambda_l + \dots + \lambda_l d\lambda_i \wedge d\lambda_j \wedge d\lambda_k) \end{aligned}$$

The proxies associated with these forms are given by (see the Exercises)

$$\begin{aligned} w_0 &\longmapsto \lambda_i \\ w_1 &\longmapsto \lambda_i \nabla \lambda_j - \lambda_j \nabla \lambda_i \\ w_2 &\longmapsto 2(\lambda_i \nabla \lambda_j \times \nabla \lambda_k + \lambda_j \nabla \lambda_k \times \nabla \lambda_i + \lambda_k \nabla \lambda_i \times \nabla \lambda_j) \\ w_3 &\longmapsto 6 \text{Vol}(K_{ijkl}^3) \end{aligned}$$

These proxies span functional spaces $W_h^p(\Omega)$ which we will identify with the Whitney complex of differential forms. The space $W_h^0(\Omega)$ turns out to be exactly the *Lagrangian nodal finite element space* P^1 on tetrahedrals that has been in the arsenal of finite element methods since their inception. The space $W_h^3(\Omega)$ is also familiar - it contains functions that are constants on each tetrahedral. The remaining two members of the Whitney complex also correspond to finite element spaces, although they were introduced to the finite element community much later.

The Whitney space $W_h^2(\Omega)$ was rediscovered in 1977 by Raviart and Thomas [71] in \mathbb{R}^2 . The space $W_h^1(\Omega)$ was rediscovered in 1980 by Nedelec [66]. This explains the popular monikers *Raviart-Thomas (RT)* and *Nedelec* spaces that are widely used in the finite element community. Another popular term is *face* and *edge* elements that is derived from the locations of the degrees of freedom for these spaces. These locations are consistent with the fact that edge elements are proxies of 1-forms and face elements are proxies of 2-forms. Originally RT and Nedelec elements were given in coordinate-dependent forms as

$$\mathbf{a}\mathbf{x} + \mathbf{b} \quad \text{and} \quad \mathbf{a} \times \mathbf{x} + \mathbf{b},$$

respectively. From these formulas it is not very easy to make a connection between face and edge elements and 2 and 1-forms. This connection was made in 1984 by R. Kotiuga in his thesis [60].

Because each Whitney space is a proper subspace of some differential forms space, $W_h^p(\Omega)$ simply inherits the exterior derivative from the former. As a result, the Whitney complex

$$\mathbb{R} \hookrightarrow W_h^0(\Omega) \xrightarrow{\nabla} W_h^1(\Omega) \xrightarrow{\nabla \times} W_h^2(\Omega) \xrightarrow{\nabla \cdot} W_h^3(\Omega) \hookrightarrow 0, \quad (151)$$

is exact and provides conforming approximation of the L^2 -based differential De Rham complex (128). In finite elements transition from infinite dimensional spaces to their conforming subspaces is effected by *interpolation* operators $I_h^0 : H(\Omega, \mathbf{grad}) \mapsto W_h^0(\Omega)$, $I_h^1 : H(\Omega, \mathbf{curl}) \mapsto W_h^1(\Omega)$, $I_h^2 : H(\Omega, \mathbf{div}) \mapsto W_h^2(\Omega)$, and $I_h^3 : L^2(\Omega) \mapsto W_h^3(\Omega)$. These operators are defined by the degrees of freedom in each space. I_h^0 is the standard Lagrange nodal interpolant. I_h^3 is the L^2 orthogonal projection into the piecewise constant space. The other two interpolants project fields by using as degrees of freedom their circulations along the edges and fluxes across the faces of the simplicial mesh. These interpolants connect the Whitney complex and (128) in a structure

$$\begin{array}{ccccccc} H(\Omega, \mathbf{grad}) & \xrightarrow{\nabla} & H(\Omega, \mathbf{curl}) & \xrightarrow{\nabla \times} & H(\Omega, \mathbf{div}) & \xrightarrow{\nabla \cdot} & L^2(\Omega) \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ W_h^0(\Omega) & \xrightarrow{\nabla} & W_h^1(\Omega) & \xrightarrow{\nabla \times} & W_h^2(\Omega) & \xrightarrow{\nabla \cdot} & W_h^3(\Omega) \end{array} \quad (152)$$

that resembles the primal-dual complex (133). This structure forms a *commuting diagram* in the sense that in it applications of interpolation and differentiation commute. It turns out that this diagram embodies the stability conditions required by finite element methods based on stationary variational principles (mixed methods).

Analogues of the Whitney complex that have the same commuting diagram property exist for other types of elements, including hexahedrals, prisms, and pyramids; see [67], [78]; [23], and [13] for some examples. Exact sequences of *hp* finite elements were developed by Demkowicz et.al. [30], and in [45] Hiptmair formulated a general method for building exact sequences of affine finite element spaces.

One may wonder why it took so long for the finite element community to embrace shape functions other than the traditional Lagrangian nodal elements, when at the same time staggered discretizations were prevalent in electromagnetics, conservation laws and fluid dynamics. Perhaps the main cause is that finite elements were conceived in a continuum mechanics setting as a Rayleigh-Ritz approximation defined over piecewise polynomial spaces. There, nodal elements are the appropriate tool to approximate displacements and forces. However, this made finite elements "node"-centric and delayed their success in settings where nodal elements are not appropriate. On the positive side, the troubles experienced by finite elements in such settings provided the impetus for the explosive development of alternative formulations such as stabilized Galerkin methods, least-squares finite elements and Discontinuous Galerkin methods.

6.1.2 Exercises

1. Write the matrix \mathbb{G} that represents $d : W_0^h \mapsto W_1^h$ on the chain from Fig.11 and verify that $\mathbb{C}\mathbb{G} = 0$.
2. Find matrices \mathbb{G} , \mathbb{C} and \mathbb{D} for a chain consisting of two hexahedrals.
3. Find matrices \mathbb{G} , \mathbb{C} and \mathbb{D} for a chain consisting of one 3-simplex.
4. Show that

$$\lambda_i d\lambda_j - \lambda_j d\lambda_i = \omega_1^{(\lambda_i \nabla \lambda_j)} - \omega_1^{(\lambda_j \nabla \lambda_i)}$$

i.e., the proxy of ω_1 is indeed $\lambda_i \nabla \lambda_j - \lambda_j \nabla \lambda_i$.

5. Verify the formula for the proxy of ω_2 . Hint: show that

$$\lambda_i d\lambda_j \wedge \lambda_k = \omega_2^{(\lambda_i \nabla \lambda_j \times \nabla \lambda_k)}.$$

6. Draw $\lambda_i \nabla \lambda_j$ and $\lambda_i \cdot \nabla \lambda_j \times \nabla \lambda_k$.

6.2 Discretization patterns

In the realization of the compatible discretization paradigm (144) finite element, finite volume and finite difference methods rely upon different types of discrete differential complexes. Internal (conforming) approximations are typical for finite elements, while finite volume and finite difference methods use direct approximations related to the cochain complex. Notwithstanding the technical distinctions between these methods, realizations of (144) fall into two broad patterns which we call, following Hiptmair [46], *primal-dual* and *elimination*, respectively. These patterns are discussed next, using specific examples of discretization techniques to illustrate them. To highlight the perseverance of (144) across compatible discretizations we use the same model problem (84) throughout the discussion. The generic compatible discretization of this problem is given by the factorization diagram

$$\begin{array}{ccccc}
 \phi^h & \xrightarrow{\nabla^h} & -\mathbf{v}^h & & \\
 \psi^h = (*_{0,\alpha}^h)\phi^h & \downarrow & & \downarrow & \mathbf{w}^h = (*_{1,\beta}^h)\mathbf{v}^h \\
 \boxed{\mathbf{f}} & -\psi^h & \xleftarrow{(\nabla \cdot)^h} & & \mathbf{w}^h
 \end{array} \tag{153}$$

6.3 Primal-dual pattern

In a primal-dual method we implement the two equilibrium equations using two separate (but not necessarily distinct) differential complexes. Therefore, each horizontal link in (153) lives in its own complex and is connected to the other link by a discrete Hodge operator. This also means that we have two sets of variables to approximate.

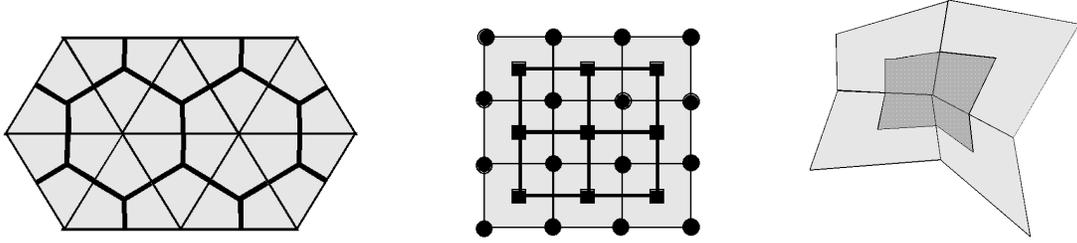


Figure 14: Primal-dual grids: Voronoi-Delauney (left), PEBI (center); Median Bisector (right).

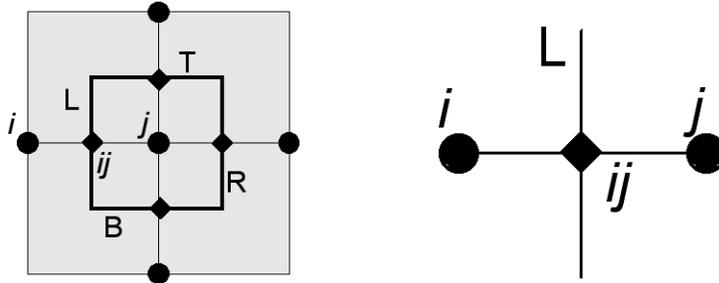


Figure 15: Discretization on topologically dual grids: grid fragment (left) and discrete Hodge operator (right)

6.3.1 Methods on dual grids

In principle, the two differential complexes can be defined on completely unrelated meshes. In practice, the two meshes cannot be too different because their approximation properties must remain close. One important example arises in methods that use topologically dual grids. In such grids each p -cell from the *primal* mesh corresponds to a $(n - p)$ -cell on the *dual* mesh; see [46] for a precise definition.

As a result, in topologically dual grids primal and dual variables enjoy a one-to-one relationship. Examples of such grids are Voronoi-Delauney triangulations, and perpendicular bisector grids on rectangles shown in Fig.14. Note that in the former case the primal grid is simplicial while the dual one is not. Not every primal-dual grid combination is topologically dual. The median bisector grid, shown in Fig.14, is one example.

To illustrate implementation of (153) on topologically dual grids consider the fragment of the grid complex shown in Fig.15. In this case the primal and the dual grids consist of the same type of rectangular cells and the one-to-one relationship is given by

$$\begin{aligned}
 \text{primal node} &\longleftrightarrow \text{dual cell center} \\
 \text{primal edge} &\longleftrightarrow \text{dual face}
 \end{aligned}
 \tag{154}$$

We implement (153) using this grid complex as follows. The primal equation is placed on the primal mesh where ϕ and \mathbf{v} are approximated by W_h^0 and W_h^1 spaces, respectively. There, for each primal edge K_{ij}^1 with nodes K_i^0 , K_j^0 and $\partial K_{ij}^1 = K_i^0 - K_j^0$, we have the discrete equation

$$-\mathbf{v}_{ij}^h = \phi_i^h - \phi_j^h.$$

The dual equation is placed on the dual mesh where \mathbf{w} and ψ are approximated by \widehat{W}_h^2 and \widehat{W}_h^3 , respectively. If the dual cell is oriented so that $\partial\hat{K} = K_R^2 - K_L^2 + K_T^2 - K_B^2$, then on this cell we have the equation

$$f^h - \psi_{\hat{K}}^h = \mathbf{w}_R^h - \mathbf{w}_L^h + \mathbf{w}_T^h - \mathbf{w}_B^h.$$

Thanks to (154) each primal variable corresponds to a dual one and the two equations are connected by the constitutive relations

$$\psi_{\hat{K}}^h = \alpha\phi_i^h \quad \text{and} \quad \mathbf{w}_{\hat{K}^2}^h = \beta\mathbf{v}_{K^1}^h.$$

i.e., the discrete Hodge operator amounts to simple scaling of the primal variables. Using this relations we could have eliminated, say the dual variables, to obtain a set of equations in terms of ϕ^h and \mathbf{v}^h only. These equations are representative of the type of equations one gets from the *box integration* co-volume method, or a *staggered* finite difference method. We see that in these methods the placement of the variables is implicitly guided by the structure of the first-order Poisson equation revealed in (153).

6.3.2 Primal-dual finite element method

A notable advantage of topologically dual meshes is the simplicity of the discrete Hodge operator. The fact that each primal variable has a corresponding dual variable provides for a clear-cut, unambiguous connection between the two equilibrium equations. However, topological duality exacts a toll on the generality of the mesh. For example, if the primal mesh consists of unstructured quadrilaterals or hexahedrals, the dual median bisector mesh is not *topologically dual*. In Voronoi-Delauney partitions one has also to impose conditions on the dihedral angles in order to ensure that each primal tetrahedral node lies in exactly one co-volume; see [68], [69].

As a result, when problem and domain features call for highly unstructured meshes, the use of topologically dual grids may become unfeasible. In such cases a viable alternative is to consider the use of the same differential complex for both equilibrium equations. Below we illustrate this idea using the Whitney complex (151).

To implement the primal equation in (153) we choose $W_h^0(\Omega)$ to approximate ϕ and $W_h^1(\Omega)$ to approximate \mathbf{v} . To implement the dual equation we use $W_h^2(\Omega)$ for \mathbf{w} and $W_h^3(\Omega)$ for ψ . In other words, ϕ and \mathbf{v} are approximated by nodal and edge (Nedelec) elements, while \mathbf{w} and ψ are approximated by face (RT) and discontinuous elements.

In contrast to a primal-dual method now all variables live in the same complex albeit in different parts. As a result, these variables cannot be connected by the same bijective relationship enjoyed by the topologically dual grids. The algebraic reason for this is clear - dimensions of Whitney spaces equal the numbers of nodes, edges, faces and cells in the mesh and these are different from each other. From a functional perspective the absence of a simple one-to-one relationship between different Whitney spaces is caused by the different continuity properties of these spaces. The nodal space $W_h^0(\Omega)$ contains continuous functions and serves as a domain for the gradient. The space $W_0^1(\Omega)$ is curl-conforming and contains functions that are continuous along the edges but can jump across the faces. In contrast, the face element space contains functions that are continuous across the faces but may jump along the edges of the mesh.

Thus, the price that we have pay for the ability to use arbitrary grids is reflected in the lack of a "free" discrete Hodge operator. To connect the primal and the dual components of (153) it is now necessary to define this operator in some way. We propose to do this by using a least-squares minimization process that penalizes the discrepancy between the primal and the dual fields. The equilibrium equations then become linear constraints that must be satisfied by the minimizers of this functional. Therefore, we are led to the constrained optimization problem:

seek $(\phi^h, \mathbf{v}^h, \mathbf{w}^h, \psi^h)$ in $W_h^0(\Omega) \times W_h^1(\Omega) \times W_h^2(\Omega) \times W_h^3(\Omega)$ such that

$$\mathcal{J}(\phi^h, \mathbf{v}^h, \mathbf{w}^h, \psi^h) = \frac{1}{2} \left(\|\psi^h - \alpha\phi^h\|_0^2 + \|\mathbf{w}^h - \beta\mathbf{v}^h\|_0^2 \right) \mapsto \min \quad (155)$$

subject to

$$\nabla\phi^h = -\mathbf{v}^h \quad \text{and} \quad \nabla \cdot \mathbf{w}^h = -\psi^h + f. \quad (156)$$

In this problem the discrete Hodge operators

$$(*_{0,\alpha}^h) : W_h^0(\Omega) \mapsto W_h^3(\Omega) \quad \text{and} \quad (*_{1,\beta}^h) : W_h^1(\Omega) \mapsto W_h^2(\Omega)$$

are defined implicitly via the optimization process. We can think of these operators as being least-squares projections from nodal to discontinuous and from edge to face elements, respectively.

It is possible to solve (155)-(156) by using Lagrange multipliers to enforce the constraints. However, a better strategy, that also reduces the number of variables, is to note that the constraint equations can be satisfied exactly in the spaces chosen for ϕ^h , \mathbf{v}^h , ψ^h and \mathbf{w}^h . Therefore, we can use the equilibrium equations to eliminate ψ^h and \mathbf{v}^h , and to obtain a least-squares minimization problem in terms of ϕ^h and \mathbf{w}^h only:

$$\min_{W_h^0 \times W_h^2} \frac{1}{2} \left(\|f - \nabla \cdot \mathbf{w}^h - \alpha\phi^h\|_0^2 + \|\mathbf{w}^h + \beta\nabla\phi^h\|_0^2 \right). \quad (157)$$

The least-squares problem (157) represents a realization of (153) with a particular choice for the discrete Hodge operator. An interesting feature of (157) is the cross-elimination pattern that leads to a problem expressed in terms of one primal and one dual variable. In the next section we will see examples of a different elimination strategy that results in a mixed finite element method. Problem (157) remains meaningful for $\alpha = 0$ as well.

Remark 6 *The minimization problem (157) was first proposed by Jespersen in [56] and subsequently used by many others in the least-squares community; see [27], [57], [25]. However, these references adopt a "node-centric" point of view and implement (157), or some augmented versions of it, using nodal elements, i.e., $W_h^0(\Omega)$, for the approximation of all unknowns. Our example shows that with a simple change in spaces the least-squares formulation can be turned into a compatible discretization. This least-squares method is conservative in the sense that the computed least-squares solution satisfies the equilibrium equations (156) exactly.*

6.4 Elimination pattern

In primal-dual methods both equilibrium equations are satisfied exactly because they occupy the correct locations on logically distinct discrete differential complexes. This is a sensible choice when we trust the equilibrium equations more than the constitutive relations and seek to preserve the qualitative features of the model. For instance, in a Darcy flow we would have the constitutive relation

$$\mathbf{w} = \frac{\kappa}{\mu} \mathbf{v},$$

where κ and μ are permeability and viscosity of the medium, respectively. Their values may be known only approximately and there's no reason to enforce such uncertain information firmly in the model. A method like (155) can maintain exactness on the equilibrium side while relaxing imprecise constitutive relation.

When we trust both the equilibrium equations and the constitutive relation, topologically dual meshes are the best choice because they allow to maintain all three relationships exactly. But if such a mesh is not feasible, and we still want to satisfy the constitutive relation, one of the equilibrium equations must be sacrificed. The reason is that a single mesh cannot support all three relations at the same time, a fact that we have already discussed in the context of Whitney spaces.

We can decide which equilibrium equation to relax based on the modeling goals. If a "kinematic" relationship like $\nabla\phi = -\mathbf{v}$ is considered less important than preserving the mass balance in $\nabla \cdot \mathbf{w} = f$, then we can proceed to eliminate the primal variables from (153). This type of methods is illustrated next.

6.4.1 Mixed finite element method

We want to implement (153) using the Whitney complex. The goal is to enforce exactly the balance equation in (140) (the bottom link in this diagram), and the constitutive relation (the two vertical links in (140)) The first task can be accomplished by using $W_h^2(\Omega)$ and $W_h^3(\Omega)$ to approximate \mathbf{w} and ψ , respectively. The two constitutive relations are enforced strongly by the elimination of the primal variables:

$$\phi = \frac{1}{\alpha} \psi \quad \text{and} \quad \mathbf{v} = \frac{1}{\beta} \mathbf{w}.$$

After elimination, the factorization diagram takes the form

$$\begin{array}{ccccc} \frac{\psi^h}{\alpha} & \xrightarrow{\nabla^h = ?} & -\frac{\mathbf{w}^h}{\beta} & & \\ \phi^h = \psi^h / \alpha & \uparrow & \uparrow & \mathbf{v}^h = \mathbf{w} / \beta & \\ \boxed{\mathbf{f}} & -\psi^h & \xleftarrow{(\nabla \cdot)} & \mathbf{w}^h & \end{array} \quad (158)$$

The absence of an h in the derivative designation along the bottom link indicates that this equation is exact. The dotted line and the question mark along the top link signify the fact this link has been broken in the sense that exterior differentiation acts in the "wrong"

direction relative to $W_h^3(\Omega)$ and $W_h^2(\Omega)$. Indeed, elimination of the primal variables has led to approximation of the 0-form ϕ by the 3-form ψ^h/α , and the 1-form \mathbf{v} by the 2-form \mathbf{w}^h/β . This combination of spaces is inconsistent with the action of the exterior derivative as a mapping from p -forms into $(p+1)$ forms. In finite elements we also speak of *non-conforming* discretizations. The space $W_h^3(\Omega)$, used for ϕ , is only a subspace of $L^2(\Omega)$ and not $H(\Omega, \mathbf{grad})$ and the action of the gradient is not determined there.

In finite elements this situation is circumvented by recasting the offending equation into a weak form, i.e., by shifting the derivative to the other variable using integration by parts. If we now write both equations in weak form, as is the custom in finite elements, we obtain the variational equation:

$$\begin{aligned} \text{seek } (\mathbf{w}^h, \psi^h) \in W_h^2(\Omega) \times W_h^3(\Omega) \text{ such that} \\ \int_{\Omega} \mathbf{w}^h \cdot \mathbf{u}^h d\Omega + \int_{\Omega} \psi^h \nabla \cdot \mathbf{u}^h d\Omega = 0 \\ \int_{\Omega} (\nabla \cdot \mathbf{w}^h) \xi^h d\Omega = \int_{\Omega} (f + \psi^h) \xi^h d\Omega \end{aligned} \tag{159}$$

for all $(\mathbf{u}^h, \xi^h) \in W_h^2(\Omega) \times W_h^3(\Omega)$.

It is easy to see that the second equation in (159) is a simple algebraic relation. A basis function for $W_h^3(\Omega)$ is a piecewise constant that equals one on one element and zero on all other elements. As a result, we can write the second equation as

$$\int_K \nabla \cdot \mathbf{w}^h d\Omega = \int_K f + \psi^h d\Omega, \quad \forall K \in \mathcal{T}_h.$$

This is equivalent to

$$\nabla \cdot \mathbf{w}^h = f^h + \psi^h,$$

where f^h is the L^2 projection of the source term into $W_h^3(\Omega)$.

6.4.2 Mixed mimetic method

It is instructive to compare the mixed finite element method (159) with a mimetic scheme based on the same elimination of the primal variables. Assuming that we have eliminated ϕ and \mathbf{v} a mimetic scheme that maintains the exact balance equation will use HS to approximate \mathbf{w} and HC to approximate ψ . This puts us in a situation much like the one depicted in (158), in fact we can draw almost the same factorization diagram but set in mimetic spaces:

$$\begin{array}{ccccc} \frac{\psi^h}{\alpha} & \dots \dots \dots & ? & \dots & -\frac{\mathbf{w}^h}{\beta} \\ \phi^h = \psi^h/\alpha & \uparrow & & \uparrow & \mathbf{v}^h = \mathbf{w}^h/\beta \\ \boxed{\mathbf{f}} & -\psi^h & \overleftarrow{\text{DIV}} & & \mathbf{w}^h \end{array} \tag{160}$$

The question mark again signifies the absence of a natural mimetic operator that will act as an exterior derivative from HC to HS . Mimetic methods solve this problem using a

trick that is very similar to the one employed by finite elements. The difference is that missing operator, denoted by $\overline{\text{GRAD}}$, is defined to be the adjoint of DIV , i.e., "integration by parts" is carried with respect to the mimetic inner product rather than the L^2 inner product. At first, this difference may seem superficial since finite element bases also induce a discrete inner product. However, the finite element inner product becomes undefined for degenerate cells because the mapping between computational and reference frames breaks down, while a mimetic inner product can still be defined for such cases; see [54]. This procedure can be repeated for GRAD and CURL so as to define their adjoint mimetic counterparts $\overline{\text{DIV}} : HL \mapsto HN$ and $\overline{\text{CURL}} : HS \mapsto HL$; see [54].

To summarize, we have the original mimetic complex

$$HN \xrightarrow{\text{GRAD}} HL \xrightarrow{\text{CURL}} HS \xrightarrow{\text{DIV}} HC$$

and its "adjoint" counterpart

$$HN \xleftarrow{\overline{\text{DIV}}} HL \xleftarrow{\overline{\text{CURL}}} HS \xleftarrow{\overline{\text{GRAD}}} HC.$$

Using these two complexes we obtain the mimetic "factorization" diagram

$$\begin{array}{ccccc} & \frac{\psi^h}{\alpha} & \xrightarrow{\overline{\text{GRAD}}} & -\frac{\mathbf{w}^h}{\beta} & \\ \phi^h = \psi^h/\alpha & \uparrow & & \uparrow & \mathbf{v}^h = \mathbf{w}/\beta \\ \boxed{\mathbf{f}} & -\psi^h & \xleftarrow{\text{DIV}} & \mathbf{w}^h & \end{array} \quad (161)$$

that defines a "mixed" mimetic method for (84). For more details about this method and its properties we refer to [39].

7 Application to stability of discretizations

Let us examine again the weak problem

seek $(\mathbf{v}, \phi) \in H(\Omega, \text{div}) \times L^2(\Omega)$ *such that*

$$\begin{aligned} \int_{\Omega} \mathbf{v} \cdot \mathbf{w} \, d\Omega - \int_{\Omega} \phi \nabla \cdot \mathbf{w} \, d\Omega &= 0 \\ \int_{\Omega} \xi \nabla \cdot \mathbf{v} \, d\Omega &= \int_{\Omega} f \xi \, d\Omega \end{aligned} \quad (162)$$

for all $(\mathbf{w}, \xi) \in H(\Omega, \text{div}) \times L^2(\Omega)$.

and its mixed finite element discretization

seek $(\mathbf{v}^h, \phi^h) \in V^h \times S^h$ *such that*

$$\int_{\Omega} \mathbf{v}^h \cdot \mathbf{w}^h d\Omega - \int_{\Omega} \phi^h \nabla \cdot \mathbf{w}^h d\Omega = 0$$

$$\int_{\Omega} \xi^h \nabla \cdot \mathbf{v}^h d\Omega = \int_{\Omega} f \xi^h d\Omega$$
(163)

for all $(\mathbf{w}^h, \xi^h) \in V^h \times S^h$.

We were led to (163) by three different paths:

- in Section 1.3 a Galerkin principle gave us (162) as one of the possible weak forms of (11) and (163) as a discretization of this form;
- in Section 3.2 we derived (162) and (163) from a saddle-point optimization problem and its first-order necessary condition;
- in Section 6.4.1 we obtained (163) from a factorization diagram that expressed (84) in terms of differential forms.

Stability of a Galerkin formulation is governed by the two generalized inf-sup conditions (43)-(44) in Theorem 3. The very general nature of these conditions makes them a poor guide for the identification of potential stable pairs of finite element spaces. Stability of a saddle-point approximation is governed by the conditions stated in Theorem 6. These conditions are more specific than those stated in Theorem 3. Thus, by accounting for an important variational aspect of (162) we were able to obtain a more precise characterization of the well-posedness of (163). Unfortunately, this doesn't make the task of finding stable spaces anymore easier because (64) remains rather vague about the traits that must be present in such spaces. As a result, for many years the search for stable elements was conducted by a trial and error approach, wherein a list of candidate spaces would be drafted and checked against the inf-sup condition. Perhaps, the best example that illustrates the limitations of this approach is linear elasticity where stable pairs eluded researchers for well over three decades; see [6].

It turns out that there exists a fundamental connection between the topological properties encoded in a DeRham complex and stability of mixed approximations to stationary problems. The single most important consequence from this fact is that knowledge of the complex's structure can be used to identify finite element spaces that will provide stable and accurate approximations of saddle-point problems. This connection was used by Arnold and Winther [5] to resolve the long standing problem of finding stable finite element spaces for elasticity.

Below we will use the Kelvin principle and the associated first-order Poisson problem to illustrate the correlation between numerical stability of stationary problems and differential complexes. We begin with the commuting diagram property of [31] and then proceed to discuss the grid decomposition property of Fix et. al. [36].

7.1 The commuting diagram property

Let us first show that (162) is a well-posed problem. We need to verify the assumptions of Theorem 5 for $V = H(\Omega, \text{div})$, $S = L^2(\Omega)$,

$$a(\mathbf{v}, \mathbf{w}) = \int_{\Omega} \mathbf{v} \cdot \mathbf{w} \, d\Omega \quad \text{and} \quad b(\phi, \mathbf{v}) = \int_{\Omega} \phi \nabla \cdot \mathbf{v} \, d\Omega.$$

In this case

$$Z = \{\mathbf{v} \in H(\Omega, \text{div}) \mid \nabla \cdot \mathbf{v} = 0 \text{ in } L^2(\Omega)\},$$

and for any $\mathbf{v} \in Z$ we have that

$$a(\mathbf{v}, \mathbf{v}) = \|\mathbf{v}\|_0^2 = \|\mathbf{v}\|_{H(\Omega, \text{div})}^2.$$

Therefore, $a(\cdot, \cdot)$ is coercive on $Z \times Z$ and (54) is verified. However, $a(\cdot, \cdot)$ is not coercive on all of V ! To check the inf-sup condition (56) note that divergence is surjective¹⁷ mapping $H(\Omega, \text{div}) \mapsto L^2(\Omega)$. Thus, given $\phi \in L^2(\Omega)$ there exists $\mathbf{v}_\phi \in H(\Omega, \text{div})$ such that

$$\nabla \cdot \mathbf{v}_\phi = \phi \quad \text{and} \quad \|\mathbf{v}_\phi\|_{H(\Omega, \text{div})} \leq C \|\phi\|_0.$$

This proves (56) because

$$\sup_{\mathbf{v} \in H(\Omega, \text{div})} \frac{b(\phi, \mathbf{v})}{\|\mathbf{v}\|_{H(\Omega, \text{div})}} \geq \frac{b(\phi, \mathbf{v}_\phi)}{\|\mathbf{v}_\phi\|_{H(\Omega, \text{div})}} = \frac{\|\phi\|_0^2}{\|\mathbf{v}_\phi\|_{H(\Omega, \text{div})}} \geq \frac{1}{C} \|\phi\|_0.$$

Next we turn attention to (163) with $V^h \equiv W_h^2(\Omega)$ and $S^h \equiv W_h^3(\Omega)$. The finite element interpolant $I_h^2 : H(\Omega, \text{div}) \mapsto W_h^2(\Omega)$ is given by

$$\int_{\partial K} I_h^2(\mathbf{v}) \cdot \mathbf{n} \, dS = \int_{\partial K} \mathbf{v} \cdot \mathbf{n} \, dS.$$

One can show; see [23, p.125], that this interpolant is well-defined and that

$$\|I_h^2 \mathbf{v}\|_{H(\Omega, \text{div})} \leq C \|\mathbf{v}\|_{H(\Omega, \text{div})} \quad \forall \mathbf{v} \in H(\Omega, \text{div}). \quad (164)$$

For $I_h^3 : L^2(\Omega) \mapsto W_h^3(\Omega)$ we consider the L^2 projection

$$\int_{\Omega} (I_h^3 \phi) \xi^h \, d\Omega = \int_{\Omega} \phi \xi^h \, d\Omega \quad \forall \xi^h \in W_h^3(\Omega).$$

Using these definitions it is possible to show that

$$\nabla \cdot (I_h^2 \mathbf{v}) = I_h^3(\nabla \cdot \mathbf{v}), \quad (165)$$

¹⁷This fact follows directly from the exactness of the De Rham complex where divergence corresponds to the last operator in the sequence. For domains without peculiarities we know that the last operator has to be a surjective mapping. For a direct proof see [23, p.136].

that is, the diagram

$$\begin{array}{ccc}
H(\Omega, \operatorname{div}) & \xrightarrow{\nabla \cdot} & L^2(\Omega) \\
I_h^2 \downarrow & & \downarrow I_h^3 \\
W_h^2(\Omega) & \xrightarrow{\nabla \cdot} & W_h^3(\Omega)
\end{array} \tag{166}$$

commutes. This fact has been first pointed out by Douglas and Roberts in [31].

The commuting diagram connects the factorization diagram (140) of the first-order Poisson equation with the factorization diagram (153) of its compatible discretization. Note that this connection is established between the two balance equations in these diagrams. The fundamental meaning of (165) is that the discrete balance equation is the same as the interpolated balance equation. This is not a coincidence because we obtained (163) by choosing the balance equation over the kinematic equation, i.e., this weak problem is *div-conforming* rather than *grad-conforming*.

We will now show that the relationship between the two balance equations encoded in (166) is *sufficient* to guarantee stability of the discrete problem (163). The following Lemma will prove essential for this purpose. It is a simplified version of Proposition 2.8, [23, p.58].

Lemma 4 *Assume that (V, S) is a pair of spaces such that the continuous inf-sup condition (56) is satisfied and let (V^h, S^h) denote a pair of their finite element subspaces. The discrete inf-sup condition (64) holds for the pair (V^h, S^h) if there exists a family of uniformly continuous operators*

$$\Pi^h : V \mapsto V^h$$

such that

$$\begin{cases} b(\Pi^h v - v, q^h) = 0 & \forall q^h \in S^h \\ \|\Pi^h v\|_V \leq C \|v\|_V & \forall v \in V \end{cases} . \tag{167}$$

Proof. Since (V, S) verifies (56) and $S^h \subset S$, from (57) it follows that for any given $q^h \in S^h$ there exists $v_q \in V$ such that

$$b(v_q, q^h) \geq \gamma_b \|v_q\|_V \|q^h\|_S.$$

Let

$$v_q^h := \Pi^h v_q.$$

Then, using the assumptions in (167) it is easy to see that

$$\begin{aligned} b(v_q^h, q^h) &= b(\Pi^h v_q, q^h) = b(v_q, q^h) \\ &\geq \gamma_b \|v_q\|_V \|q^h\|_S \geq \frac{\gamma_b}{C} \|\Pi^h v_q\|_V \|q^h\|_S = \frac{\gamma_b}{C} \|v_q^h\|_V \|q^h\|_S \end{aligned}$$

The last inequality verifies the inf-sup condition for (V^h, S^h) with $\gamma_b^h = \gamma_b/C$. \square

We are now ready to show that the commuting diagram property (166) implies that (163) is well-posed. The first assumption of Theorem 6 is that $a(\cdot, \cdot)$ is coercive on

$$Z^h = \{\mathbf{v}^h \in W_h^2(\Omega) \mid b(\mathbf{v}^h, \xi^h) = 0 \quad \forall \xi^h \in W_h^3(\Omega)\}.$$

$W_h^2(\Omega)$ and $W_h^3(\Omega)$ are the last two spaces in the Whitney sequence. Because this sequence is exact it follows that divergence is surjective mapping $W_h^2 \mapsto W_h^3$ and, thus

$$\nabla \cdot (W_h^2) = W_h^3.$$

In other words, the relationship between the last two spaces in the exact sequence (128) is reproduced by the last two members of the Whitney complex (151). This immediately implies the inclusion $Z^h \subset Z$, a property that is not valid for arbitrary choices of V^h and S^h . Remark 1 states that in this case the error estimate for \mathbf{v}^h uncouples from that for ϕ^h .

This inclusion also suffices to prove the first assumption in Theorem 6 because we already know that $a(\cdot, \cdot)$ is coercive on $Z \times Z$, and since coercivity is inherited on proper subspaces it follows that $a(\cdot, \cdot)$ is also coercive on $Z^h \times Z^h$.

To prove the inf-sup condition (64) note that the commuting diagram property (165) can be stated as

$$\int_{\Omega} \xi^h \nabla \cdot I_h^2 \mathbf{v} \, d\Omega = \int_{\Omega} \xi^h \nabla \cdot \mathbf{v} \, d\Omega \quad \forall \xi^h \in W_h^3.$$

This is equivalent to

$$\int_{\Omega} \xi^h \left(\nabla \cdot I_h^2 \mathbf{v} - \nabla \cdot \mathbf{v} \right) \, d\Omega = 0$$

and together with (164) it means that (167) holds for I_h^2 . The inf-sup condition now follows from Lemma 4.

7.2 The grid decomposition property

So far our experience indicates that a compatible (and stable!) discretization of the Kelvin principle seems to exclude nodal representations of the vector variable \mathbf{v} . This can be explained by the fact that \mathbf{v} is a proxy field of a 2-form and in both direct and functional approximations the order of the form determines where to store its discrete representation. For 0-forms this location is at the nodes, for 1-forms it is at the edges, and for 2-forms it is on the faces.

Thus, an interesting question is whether or not there exists a stable discretization of the Kelvin principle that uses nodal representation of \mathbf{v} . The answer to this question is affirmative and was given in a seminal paper by Fix et. al. in 1981, see [36].

In this paper they considered approximation of the Kelvin principle by a pair (V^h, S^h) where V^h is a subspace of $\mathbf{H}^1(\Omega)$, i.e., a nodal C^0 finite element space, and $S^h = \nabla \cdot (V^h)$. For example, if V^h is a piecewise linear space, then S^h is contained in the space of all piecewise constant functions. By construction this pair guarantees that divergence is a surjective map $V^h \mapsto S^h$. However, this by itself is clearly not enough to make (163) stable because it doesn't guarantee that the inf-sup condition will hold with a constant independent of the mesh size.

The main result of [36] was to prove that for such pairs of spaces the weak problem (163) will be stable if and only if the nodal space V^h satisfies the following property:

for every $\mathbf{v}^h \in V^h$ there exist $\mathbf{u}^h, \mathbf{w}^h \in V^h$ such that

$$\mathbf{v}^h = \mathbf{u}^h + \mathbf{w}^h$$

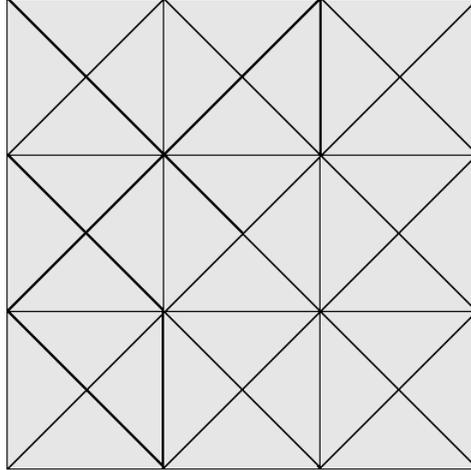


Figure 16: Criss-cross grid.

and

$$\nabla \cdot \mathbf{w}^h = 0, \quad \int_{\Omega} \mathbf{u}^h \cdot \mathbf{w}^h d\Omega = 0, \quad \|\mathbf{u}^h\|_0 \leq C \|\nabla \cdot \mathbf{v}^h\|_{-1}.$$

These three conditions constitute the *grid decomposition property*, or GDP. GDP implies that V^h can approximate well the nullspace of the divergence operator. In other words, V^h has a sufficiently large supply of vector fields that are exact curls. In [36] it was also shown that the criss-cross grid; see Fig. 16, has the grid decomposition property in 2D. Unfortunately, to this day there are no other known examples of grids that satisfy GDP.

Note that by virtue of its position in the Whitney complex the space $W_h^2(\Omega)$ has exactly the same properties as a space V^h that satisfies GDP: first, divergence maps it surjectively into $W_h^3(\Omega)$, and second, it contains a large subspace that consists of exact curls of $W_h^1(\Omega)$ fields. As a result, by assuming that GDP holds for V^h , and by enforcing surjection explicitly, we obtain a pair of spaces (V^h, S^h) that behaves just like a pair of spaces from a discrete exact sequence, which we know is sufficient to prove the inf-sup condition. The unique property of this pair is that V^h is a C^0 nodal finite element space.

This result is yet another example of how the properties of differential complexes inevitably surface in stable discretizations. Even though the authors of [36] were not aware of the connection between their grid decomposition property and the De Rham cohomology, they have formulated conditions that implicitly forced the finite element spaces into the correct topological relationship.

8 Finite element alternatives

Compatible discretizations imitate the intrinsic structure of PDE problems and, as a result, they lead to algebraic problems that are elegant and truthful representations of this structure in finite dimensions. The main appeal of such discretizations is in their ability to reproduce, in a finite dimensional setting, fundamental properties of the original PDE model that stem from that intrinsic structure.

There are however, some penalties associated with maintaining a high degree of structural fidelity in a discretization. Stability of compatible discretizations is contingent upon carefully balanced relationships between the discrete spaces on one hand and these spaces and the continuous problem on the other hand. Commuting diagrams, or grid decomposition properties that symbolize these relationships easily break down once the "wrong" spaces or grids are inserted.

The reasons to use the "wrong" spaces are first and foremost practical. Compatible discretizations impose rigid constraints on the spaces and operators between them. In a finite element setting this means that different fields will be approximated by different spaces, with different polynomial degrees and, perhaps defined with respect to different triangulations of the same region. This complicates finite element data structures and makes refinement more difficult because it is not easy to sustain the delicate relationships between the spaces when, e.g., hanging nodes are introduced. Refinement may also call for higher polynomial degrees on adjacent elements. Again, to maintain an exact sequence under such refinement policy is not simple; see for example [30]. Then, there's the node-centric viewpoint that dominated finite elements for quite a long time and that led to tremendous investments in simulation infrastructures based on Lagrangian elements. Reuse of these infrastructures has serious appeal and cannot be easily dismissed.

For these and perhaps many other reasons, methods that do not adhere to the structures imposed by the PDE, but achieve stability by some other means, have always attracted significant attention from researchers and practitioners alike. In this section we consider three examples of such *relaxed discretizations* that have become a focus of intensive research efforts in the last two decades.

The main difference between a compatible and a relaxed discretization is in the way they treat the structure of the given PDE problem. For a compatible discretization this structure is axiomatic and completely governs the design of the associated discrete configuration. In contrast, in a relaxed discretization the finite dimensional structure has been already predetermined (usually by factors not related to the problem on hand) and it is the problem that must be adapted to this structure. This adaptation usually takes the form of *stabilization*, wherein the original problem is modified to one that is stable, or *least-squares optimization*, wherein the problem is embedded into a completely new one.

We now embark on a brief excursion into the realm of these methods, using finite elements as a backdrop for our discussion. For a model problem we choose again the Kelvin principle and the associated first-order Poisson problem, thereby providing for a side by side comparison and contrast with the compatible methods discussed earlier.

8.1 Stabilization

The choice of $V^h = W_h^2(\Omega)$ and $S^h = W_h^3(\Omega)$ in (163) gives a compatible discretization of (162). This combination of spaces satisfies the inf-sup condition and leads to a well-posed discrete equation. But it provides a discontinuous approximation for ϕ and requires a face based finite element assembly. Suppose that instead, we insist on a continuous ϕ^h and prefer to use standard nodal assembly. A pair of spaces that meets these demands is given by $V^h = [W_h^0(\Omega)]^n$ and $S^h = W_h^0(\Omega)$. We call this pair *superconforming* because it replaces the third and the last members of the Whitney sequence by the first one which

is a proper subspace of the "stronger" space $H(\Omega, \mathbf{grad})$.

Consider now an alternative scenario where it is known in advance that solution of (162) will call for an extensive unstructured refinement of the mesh and highly anisotropic refinement in the polynomial degrees on each element. In this case numerical solution will be facilitated by discontinuous approximations in both \mathbf{v} and ϕ . A pair of spaces that meets this goal is $V^h = [W_h^3(\Omega)]^n$ and $S^h = W_h^3(\Omega)$. We call this pair of spaces *subconforming* because it replaces $W_h^2(\Omega)$ by the last member of the Whitney complex which is a proper subspace of the "weakest" space $L^2(\Omega)$.

Both the superconforming and the subconforming pairs do not satisfy the inf-sup condition and are unstable. There are however some important differences between these two pairs. The superconforming pair imposes more interelement continuity on the candidate solutions than required by the weak formulation of (162). In contrast, the subconforming pair of spaces completely relaxes interelement continuity of the candidate solutions so that (162) cannot be even restricted to these spaces without some additional preparation.

From the variational perspective, stability occurs when the sizes of test and trial spaces are matched so as to verify the necessary inf-sup conditions. From this vantage point loss of stability in superconforming approximation is caused by the "shrinking" of the spaces, while for subconforming methods the reason is the excessive "enlargement". From the geometrical perspective the reason for the loss of stability in both cases is in the breakup of the factorization diagram.

These two scenarios describe the settings that have led to stabilized Galerkin methods and Discontinuous Galerkin methods, respectively. Next we turn attention to these two classes of methods and show how each one of them responds to the loss of stability due to the space selection.

8.1.1 Stabilized Galerkin methods

Stabilized Galerkin methods were pioneered by Hughes et. al. in [47] for convection dominated problems (see also [58] for analysis and [48] for modifications of the original SUPG method). Subsequently, these methods were extended to stationary problems with particular emphasis on incompressible fluid flows; see [49], [50], [32], [22], [38], [10] and the references therein. The method we are about to discuss appeared very recently in [63] in the context of a Darcy flow problem. We will present this method in a simplified setting that matches our model problem (162).

There exists an important and fundamental distinction in the way stabilization works for advection dominated problems and stationary problems, such as (162), even though the terms that effect the stabilization appear to be of the same kind. In the former case, stabilization adds artificial dissipation. For stationary problems it amounts to *penalization* of a Lagrangian functional by a norm of the Lagrange multiplier. The main thrust in the application of this idea has always been to enable stable approximations of saddle-point problems by standard continuous nodal finite element spaces, including equal order interpolation, *without incurring the penalty error*. This places stabilized methods on the *superconforming* side of our taxonomy.

Let us now discuss how to stabilize the Lagrangian functional

$$L(\mathbf{w}, \psi; f) = \frac{1}{2} \int_{\Omega} |\mathbf{w}|^2 d\Omega - \int_{\Omega} \psi(\nabla \cdot \mathbf{w} - f) d\Omega, \quad (168)$$

that was associated with the Kelvin principle (82), so that it will work for superconforming pairs of spaces. First, we will try to guess the appropriate penalty term and then we'll deal with the penalty error. An ad hoc argument to aid our guess is as follows. A superconforming approximation of the Lagrange multiplier ψ has more regularity than (168) can handle. Indeed, if ψ is approximated by a subspace of $H(\Omega, \mathbf{grad})$, we must be able to control its gradient. A formulation based on (168) is too weak to do that because the Galerkin bilinear form associated with (162)

$$Q_G(\{\phi, \mathbf{v}\}; \{\psi, \mathbf{w}\}) = \int_{\Omega} \mathbf{v} \cdot \mathbf{w} d\Omega - \int_{\Omega} \phi \nabla \cdot \mathbf{w} d\Omega + \int_{\Omega} \psi \nabla \cdot \mathbf{v} d\Omega,$$

does not provide any control over $\nabla \phi$.

To strengthen (168) we penalize by $\nabla \psi$:

$$L_{\delta}(\mathbf{w}, \psi; f) = \frac{1}{2} \int_{\Omega} |\mathbf{w}|^2 d\Omega - \int_{\Omega} \psi(\nabla \cdot \mathbf{w} - f) d\Omega - \delta \int_{\Omega} |\nabla \psi|^2 d\Omega. \quad (169)$$

The penalty term is subtracted because (168) is maximized with respect to ψ . The first-order optimality system for (169) is

seek $(\mathbf{v}, \phi) \in H(\Omega, \text{div}) \times H(\Omega, \mathbf{grad})$ such that

$$\begin{aligned} \int_{\Omega} \mathbf{v} \cdot \mathbf{w} d\Omega - \int_{\Omega} \phi \nabla \cdot \mathbf{w} d\Omega &= 0 \\ \int_{\Omega} \psi \nabla \cdot \mathbf{v} d\Omega + \delta \int_{\Omega} \nabla \psi \cdot \nabla \phi d\Omega &= \int_{\Omega} f \psi d\Omega \end{aligned} \quad (170)$$

for all $(\mathbf{w}, \psi) \in H(\Omega, \text{div}) \times H(\Omega, \mathbf{grad})$.

With (170) we associate the *penalized Galerkin* bilinear form

$$Q_{PG}(\{\phi, \mathbf{v}\}; \{\psi, \mathbf{w}\}) = \int_{\Omega} \mathbf{v} \cdot \mathbf{w} d\Omega - \int_{\Omega} \phi \nabla \cdot \mathbf{w} d\Omega + \int_{\Omega} \psi \nabla \cdot \mathbf{v} d\Omega + \delta \int_{\Omega} \nabla \psi \cdot \nabla \phi d\Omega.$$

It is an easy matter to show that

$$Q_{PG}(\{\phi, \mathbf{v}\}; \{\phi, \mathbf{v}\}) = \|\mathbf{v}\|_0^2 + \delta \|\nabla \phi\|_0^2,$$

i.e., the penalized form is coercive and controls the H^1 seminorm of ϕ . Note that coercivity norms are consistent with the fact that \mathbf{v} represents the gradient of ϕ . Accordingly, (170) is not subject to an inf-sup stability condition and can be approximated by any pair of finite element subspaces. The price that (170) pays for this flexibility is the penalty error that will be incurred by its solutions.

A key observation that became the hallmark of stabilized Galerkin methods is that penalty errors can be avoided by using *consistent* terms to effect the regularization in

(169). A consistent term is one that vanishes on all sufficiently smooth solutions. In the present case the relevant term is provided by the L^2 norm of the "kinematic" equation $\nabla\phi + \mathbf{v} = 0$. Subtracting this term from (168) with $\delta = 1/4$ gives the *consistently* regularized Lagrangian functional

$$L_\delta(\mathbf{w}, \psi; f) = \frac{1}{2} \int_{\Omega} |\mathbf{w}|^2 d\Omega - \int_{\Omega} \psi(\nabla \cdot \mathbf{w} - f) d\Omega - \frac{1}{4} \int_{\Omega} |\nabla\psi + \mathbf{v}|^2 d\Omega. \quad (171)$$

The optimality system for (171) is given by

seek $(\mathbf{v}, \phi) \in H(\Omega, \text{div}) \times H(\Omega, \mathbf{grad})$ such that

$$\begin{aligned} \int_{\Omega} \mathbf{v} \cdot \mathbf{w} d\Omega - \int_{\Omega} \phi \nabla \cdot \mathbf{w} d\Omega - \frac{1}{2} \int_{\Omega} (\nabla\phi + \mathbf{v}) \mathbf{w} d\Omega &= 0 \\ \int_{\Omega} \psi \nabla \cdot \mathbf{v} d\Omega + \frac{1}{2} \int_{\Omega} (\nabla\phi + \mathbf{v}) \cdot \nabla\psi d\Omega &= \int_{\Omega} f\psi d\Omega \end{aligned} \quad (172)$$

for all $(\mathbf{w}, \psi) \in H(\Omega, \text{div}) \times H(\Omega, \mathbf{grad})$.

This problem is associated with the *Galerkin-Least-Squares* bilinear form

$$\begin{aligned} Q_{GLS}(\{\phi, \mathbf{v}\}; \{\psi, \mathbf{w}\}) \\ = \int_{\Omega} \mathbf{v} \cdot \mathbf{w} d\Omega - \int_{\Omega} \phi \nabla \cdot \mathbf{w} d\Omega + \int_{\Omega} \psi \nabla \cdot \mathbf{v} d\Omega + \frac{1}{2} \int_{\Omega} (\nabla\phi + \mathbf{v}) \cdot (\nabla\psi + \mathbf{w}) d\Omega. \end{aligned}$$

It is easy to see that

$$Q_{GLS}(\{\phi, \mathbf{v}\}; \{\phi, \mathbf{v}\}) = \frac{1}{2} \left(\|\mathbf{v}\|_0^2 + \|\nabla\phi\|_0^2 \right)$$

and so the Galerkin-Least-Squares form has the same stability properties as the penalized Galerkin form. As a result, (172) can be approximated by any pair of conforming subspaces. The main difference between (170) and (172) is that the latter is a *consistent* formulation and its solutions will not incur penalty error.

A stabilized method based on (172) was introduced in [63] for a Darcy flow problem. The term *least-squares* that we used in reference to this method comes from the form of the penalty term used in (171).

8.1.2 Discontinuous Galerkin methods

Discontinuous Galerkin (DG) methods have had a long and distinguished career in non-linear hyperbolic problems and to a lesser extent in elliptic problems such as the Kelvin principle that we use as our primary example. For a comprehensive historical survey, the reasons why DG development for elliptic problems lagged behind, and a taxonomy of DG methods for (84) we refer to [4].

Conceptually, the origins of DG in elliptic problems can be traced back to the classical interior penalty method of M. Wheeler [81] and the work of D. Arnold in [3] (see also [9] for related ideas).

Interior penalty methods are based on the observation that interelement continuity can be removed from the finite element spaces and imposed weakly by a modification of the variational problem. The same idea serves as a basis for numerous domain decomposition methods, including the FETI (Finite Element Tearing and Interconnecting) family of algorithms; see [33].

For an illustration let us consider a DG formulation for the first-order Poisson equation (84). First we observe that neither one of the weak formulations (13)-(16), or (162), discussed so far is meaningful for the subconforming pair of spaces $V^h = [W_h^3(\Omega)]^n$ and $S^h = W_h^3(\Omega)$. Therefore, to use these spaces we'll have to build the variational formulation from scratch and in a manner that takes into account *explicitly* the finite element partition of the domain Ω into finite elements K . This is very different from the case of superconforming approximation where any one of the weak formulations (13)-(16), or (162) would have worked.

To this end we assume that $\mathcal{T}_h = \mathcal{C}_3^h$ is a given triangulation of Ω . We multiply each equation in (84) by a test function and then integrate these equations by parts over each element $K \in \mathcal{T}_h$. The result is

$$\begin{aligned} \int_K \mathbf{v} \cdot \mathbf{w} \, dV - \int_K \phi \nabla \cdot \mathbf{w} \, dV &= \int_{\partial K} \phi \mathbf{w} \cdot \mathbf{n}_K \, dS \\ - \int_K \nabla \xi \cdot \mathbf{v} \, dV &= \int_K f \xi \, d\Omega + \int_{\partial K} \xi \mathbf{v} \cdot \mathbf{n}_K \, dS \end{aligned} \quad (173)$$

This weak problem serves as a foundation for a DG method introduced by Cockburn and Shu in [28]. In this method we

seek $(\mathbf{v}^h, \phi^h) \in [W_h^3(\Omega)]^n \times W_h^3(\Omega)$ such that

$$\begin{aligned} \int_K \mathbf{v}^h \cdot \mathbf{w}^h \, dV - \int_K \phi^h \nabla \cdot \mathbf{w}^h \, dV &= \int_{\partial K} \hat{\phi}_K^h \mathbf{w} \cdot \mathbf{n}_K \, dS \\ - \int_K \nabla \xi^h \cdot \mathbf{v}^h \, dV &= \int_K f \xi^h \, d\Omega + \int_{\partial K} \xi^h \hat{\mathbf{v}}_K^h \cdot \mathbf{n}_K \, dS \end{aligned} \quad (174)$$

for all $K \in \mathcal{T}_h$ and $(\mathbf{w}^h, \xi^h) \in [W_h^3(\Omega)]^n \times W_h^3(\Omega)$.

To close the definition of the DG method, the fluxes $\hat{\mathbf{v}}_K^h$ and $\hat{\phi}_K^h$ must be specified in terms of the variables \mathbf{v}^h, ϕ^h and the boundary conditions.

The choice of the fluxes is a very delicate issue in DG methods because they are responsible for gluing together the solution from the disjoint pieces on each element. It is clear that without the fluxes (174) is neither stable nor consistent and so, their role is to both stabilize and provide consistency. From this point of view (174) is also a *stabilized* formulation, albeit quite different from the stabilized Galerkin methods considered earlier. In addition, flux definitions can affect the sparsity and the symmetry of the stiffness matrix. For a detailed list of flux definitions and the ensuing DG methods we refer to [4].

8.2 Least-squares optimization

We have already encountered an example of a least-squares optimization problem in Section 6.3.2. There the least-squares functional (155) served to define a weak discrete Hodge

operator. The final least-squares formulation (157), implemented with the appropriate members of the Whitney sequence, represented a compatible discretization which preserved the equilibrium and kinematic equations and relaxed the constitutive relation.

Least-squares methods were however, conceived in a completely different manner that had nothing to do with factorization diagrams or differential complexes. For a historical perspective and review of recent developments we refer to [14] while here we will focus only on least-squares for (84).

Least-squares finite element methods were motivated by the same considerations as stabilized Galerkin methods. Their primary goal was to circumvent inf-sup stability conditions in saddle-point problems and to enable the use of arbitrary combinations of standard C^0 Lagrangian finite element spaces. Thus, (157) would be considered by many least-squares purists as an aberration because it uses the very same face elements a least-squares formulation is supposed to avoid. We will return to this in a moment. The principal difference between stabilized and least-squares methods is in the way they approach this task. A stabilized method retains the original variational principle but modifies it by "strengthening" the variational form. As we saw in (171), very often this stabilization is effected by using least-squares type terms. Thus, many stabilized Galerkin methods can be viewed as a weighted average of a standard Galerkin principle and a least-squares optimization principle.

A least-squares method completely discards the original variational principle and constructs a new one by using the idea of *residual minimization*. This idea is in a sense orthogonal to the formal *residual orthogonalization* of Galerkin principles and has the same universal applicability. As a result, least-squares principles can be applied to virtually any PDE equation problem, with or without an associated variational principle.

A least-squares method for (84) is given by:

$$\begin{aligned} \text{seek } (\mathbf{v}, \phi) \in H(\Omega, \text{div}) \times H_0(\Omega, \mathbf{grad}) \text{ such that} \\ J(\mathbf{v}, \phi) \leq J(\mathbf{w}, \psi) \end{aligned} \quad (175)$$

$$\text{for all } (\mathbf{w}, \psi) \in H(\Omega, \text{div}) \times H_0(\Omega, \mathbf{grad})$$

and where

$$J(\mathbf{w}, \psi) = \frac{1}{2} \left(\|f - \nabla \cdot \mathbf{w}\|_0^2 + \|\mathbf{w} + \beta \nabla \psi\|_0^2 \right). \quad (176)$$

The minimizers of (176) are subject to a necessary minimum condition that can be expressed as

$$\begin{aligned} \text{seek } (\mathbf{v}, \phi) \in H(\Omega, \text{div}) \times H_0(\Omega, \mathbf{grad}) \text{ such that} \\ Q_{LS}(\{\mathbf{v}, \phi\}; \{\mathbf{w}, \psi\}) = F(\{\mathbf{w}, \psi\}) \end{aligned} \quad (177)$$

$$\text{for all } (\mathbf{w}, \psi) \in H(\Omega, \text{div}) \times H_0(\Omega, \mathbf{grad})$$

where

$$Q_{LS}(\{\mathbf{v}, \phi\}; \{\mathbf{w}, \psi\}) = \int_{\Omega} (\nabla \cdot \mathbf{v})(\nabla \cdot \mathbf{w}) d\Omega + \int_{\Omega} (\nabla \phi + \mathbf{v}) \cdot (\nabla \psi + \mathbf{w}) d\Omega$$

and

$$F(\{\mathbf{w}, \psi\}) = \int_{\Omega} f \nabla \cdot \mathbf{w} \, d\Omega.$$

This setting was first used in a least-squares method by Jespersen in 1977; see [56]. For some of the founding work in this area we refer to Fix et. al. [34], [35], and [37]; Bramble et. al. [19] and [20]; more recent references are [27], [57], [25] and [26].

One can prove that (see, e.g., [25])

$$Q_{LS}(\{\mathbf{v}, \phi\}; \{\mathbf{v}, \phi\}) \geq C \left(\|\mathbf{v}\|_{H(\Omega, \text{div})}^2 + \|\phi\|_{H(\Omega, \mathbf{grad})}^2 \right)$$

and that $F(\cdot)$ is a continuous functional $H(\Omega, \text{div}) \times H_0(\Omega, \mathbf{grad}) \mapsto \mathbb{R}$. As a result, Lax-Milgram lemma can be used to show that (177) is a well-posed variational problem that has a unique solution. This is where the main appeal of least-squares principles lies: we were able to formulate a Rayleigh-Ritz principle for a problem that was originally associated with saddle-point optimization.

Once the Rayleigh-Ritz framework is established, discretization can proceed by using any combination of conforming finite element subspaces $V^h \subset H(\Omega, \text{div})$ and $S^h \subset H_0(\Omega, \mathbf{grad})$. The form $Q_{LS}(\cdot; \cdot)$ is guaranteed to be coercive on any such pair of spaces and, as a result, the discrete problem

seek $(\mathbf{v}^h, \phi^h) \in V^h \times S^h$ *such that*

$$Q_{LS}(\{\mathbf{v}^h, \phi^h\}; \{\mathbf{w}^h, \xi^h\}) = F(\{\mathbf{w}^h, \xi^h\}) \quad (178)$$

for all $(\mathbf{w}^h, \xi^h) \in V^h \times S^h$

will have a unique solution. One can show that this solution converges to the exact solution of (84) at the best possible rate.

An important remark that we wish to make at this point is that (178) is well-posed for *any* pair of conforming subspaces of $H(\Omega, \text{div}) \times H_0(\Omega, \mathbf{grad})$, including the the pair $(W_h^2(\Omega), W_h^0(\Omega))$, i.e., the pair used in (157). This means that while not in the mainstream of least-squares methods, the compatible version (157) is a bona fide least-squares method.

9 Closing remarks

One of the fundamental reasons for the tremendous success of finite element methods has been their reliance upon variational principles and the existing connection between variational principles and quasi-projections.

Variational principles are a powerful tool for stability and error analysis, and indeed, they have remained unsurpassed in their ability to generate sharp error estimates. Other discretization methods have freely borrowed ideas and techniques from finite elements in order to carry out convergence and error analysis; see e.g., [11] for an example from mimetic methods. The power of variational tools have been also noted in adaptivity where recasting, e.g., a finite volume scheme as a Petrov-Galerkin method allows to take advantage of adjoint formulations for the error estimation.

Nevertheless, there are certain limitations as to how much one can accomplish by relying on variational principles alone. The aforementioned example from linear elasticity, where for over three decades stable finite element spaces eluded researchers and practitioners, is one good example. Also, variational methods remain taciturn about the causes for the remarkable similarities between stable and unstable discretizations across virtually all discretization platforms.

This is where geometrical modeling ideas based on differential forms, differential complexes and factorization diagrams become an indispensable tool. Geometrical modeling possesses an extraordinary ability to codify and express fundamental topological properties of PDE's in a simple and unified manner. This in turn provides the springboard for an unified stability and error analysis of seemingly distinct discretization methods. Recently, there has been a significant effort to bring these tools either explicitly or implicitly to the mainstream of numerical PDE's, most notably by Bossavit [18], Hiptmair [46], Shahskov and Hyman [53],[54], Teixeira [74] and many others. Nicolaidis [68] and [69] has also demonstrated that error analysis can be developed directly in the setting of the exact sequence by using only the orthogonal decomposition property. Hiptmair have followed in [46] by a formal analysis entirely based on differential form formalism.

It is safe to say that in the coming years geometrical modeling and compatible discretizations will be an increasingly important tool in numerical simulations. Differential complexes are expected to play key role in the identification of stable discretizations of the Einstein equations of relativity and simulations of black hole collisions; see [2]. These complexes also turn out to be of prime importance in the design of algebraic solvers because compatible discretizations propagate the PDE structure to the associated algebraic problem; see [15] and [40]. Finally, by revealing the structure of PDE problems, differential complexes also serve to provide us with better understanding of the design principles that must be incorporated in alternative stabilized Galerkin, discontinuous Galerkin and least-squares types of methods.

10 Acknowledgements

The material presented in these lecture notes is the outgrowth from numerous formal and informal discussions that I had over the last eighteen months with my colleagues and collaborators. Special thanks go to Martin Berggren, David Day, Jonathan Hu, Rich Lehoucq, Allen Robinson, John Shadid, Ray Tuminaro (all from Sandia National Laboratories), Misha Shashkov (Los Alamos National Laboratory), Tom Manteuffel and Steve McCormick (University of Colorado, Boulder), Leszek Demkowicz (UT Austin), and Roy Nicolaidis (Carnegie Mellon University). Finally, a big "thank you" to Max Gunzburger (Florida State) who taught me about finite elements and is always ready to extend a helping hand.

References

- [1] R. Adams, *Sobolev Spaces*, Academic, New York, 1975.
- [2] A.M. Alekseenko and D. Arnold, A new symmetric hyperbolic formulation for the Einstein equations; submitted to *Phys. Rev. D*
- [3] D. Arnold, An interior penalty finite element method with discontinuous elements, *SIAM J. Num. Anal.* 19, pp.742-760, 1982.
- [4] D. Arnold, F. Brezzi, B. Cockburn, and L. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems, *SIAM Journal on Numerical Analysis*, 39, pp. 1749-1779, 2002.
- [5] D. Arnold and R. Winther, Mixed finite elements for elasticity, *Numer. Math.*, 42, pp. 401-419, 2002.
- [6] D. Arnold, Differential complexes and numerical stability, Proceedings of the International Congress of Mathematicians, Beijing 2002, Volume I: Plenary Lectures.
- [7] V. I. Arnold, *Mathematical Methods of Classical mechanics*, Springer Graduate Texts in Mathematics, Springer, 1989.
- [8] A. Aziz (editor), *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, Academic Press, 1972.
- [9] I. Babuška, The finite element method with Lagrange multipliers, *Numer. Math.*, 20, pp. 179–192, 1973.
- [10] M. A. Behr, L. P. Franca, T. E. Tezduyar; Stabilized finite element methods for the velocity-pressure-stress formulation of incompressible flows, *Computer Methods in Appl. Mech. Engrg.*, 104, pp. 31-48, 1993.
- [11] M. Berndt, K. Lipnikov, J. Moulton, M. Shashkov, Convergence of mimetic finite difference discretizations of the diffusion equation. *East-West Journal on Numerical Mathematics*, Vol. 9/4, pp. 253-316, 2001.
- [12] P. Bochev and C. Scovel, On quadratic invariants and symplectic structure, *BIT* 34, pp. 337-345, 1994.
- [13] P. Bochev and A. Robinson, *Matching algorithms with physics: exact sequences of finite element spaces*, in: Collected Lectures on the Preservation of Stability Under Discretization, edited by D. Estep and S. Tavener, SIAM, Philadelphia, 2001.
- [14] P. Bochev and M. Gunzburger, Least-squares finite element methods for elliptic equations, *SIAM Review*, 40/4, pp. 789–837, 1998.
- [15] P. Bochev, J. Hu, A. Robinson and R. Tuminaro; Towards robust 3D Z-pinch simulations: discretization and fast solvers for magnetic diffusion in heterogeneous conductors, to appear in *ETNA*, 2002.

- [16] A. Bossavit and J. Verite, A mixed fem-biem method to solve 3-d eddy current problems, *IEEE Trans. Magnetism*, 18, pp. 431-435, 1982.
- [17] A. Bossavit, A rationale for “edge-elements” in 3-D fields computations, *IEEE Trans. Magnetism*, 24, pp. 74-79, 1988.
- [18] A. Bossavit, *Computational Electromagnetism*, Academic Press, 1998.
- [19] J. Bramble and A. Schatz, Least-squares methods for $2m$ th order elliptic boundary value problems, *Math. Comp.*, 25, pp. 1-32, 1971.
- [20] J. Bramble and J. Nitsche, A generalized Ritz-least-squares method for Dirichlet problems, *SIAM J. Numer. Anal.*, 10, pp. 81-93, 1973.
- [21] F. Brezzi, On existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers, *RAIRO Model. Math. Anal. Numer.*, 21, pp. 129-151, 1974.
- [22] F. Brezzi and J. Douglas, Stabilized mixed methods for the Stokes problem, *Numer. Math.*, 53, pp. 225-235, 1988.
- [23] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element methods*, Springer-Verlag, 1991.
- [24] K. Bowden, On general physical systems theories, *Int. J. General Systems*, 18, pp. 61-79, 1990.
- [25] Z. Cai, R. Lazarov, T. Manteuffel, and S. McCormick, First order system least-squares for second order partial differential equations: Part I, *SIAM Numer. Anal.*, 31, pp. 1785-1799, 1994.
- [26] G. Carey, A. Pehlivanov, and R. Lazarov, Least-squares mixed finite element methods for second order elliptic problems, *SIAM J. Numer. Anal.*, 31, pp. 1368-1377, 1994.
- [27] C. Chang, Finite element approximation for grad-div type systems in the plane, *SIAM J. Numer. Anal.*, 29, pp. 452-461, 1992.
- [28] B. Cockburn and C.-W. Shu, The Local Discontinuous Galerkin method for convection-diffusion systems, *SIAM J. Num. Anal.* 35, pp. 2440-2463, 1998.
- [29] P. Ciarlet, *Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
- [30] L. Demkowicz, P. Monk, L. Vardapetyan, and W. Rachowicz, De Rham Diagram for hp -finite element spaces, TICAM Report 99-06, TICAM, UT Austin, 1999.
- [31] J. Douglas and J. Roberts, Mixed finite element methods for second order elliptic problems, *Math. Appl. Comp.* 1, pp. 91-103, 1982.

- [32] J. Douglas and J. Wang, An absolutely stabilized finite element method for the Stokes problem. *Math. Comp.*, 52, pp. 495-508, 1989.
- [33] C. Farhat and F.-X. Roux. A Method of Finite Element Tearing and Interconnecting and its Parallel Solution Algorithm. *Int. J. Numer. Meth. Engrg.*, 32, pp. 1205-1227, 1991.
- [34] G. Fix and M. Gunzburger, On least squares approximations to indefinite problems of the mixed type, *Inter. J. Numer. Meth. Engng.*, 12, pp. 453-469, 1978.
- [35] G. Fix, M. Gunzburger and R. Nicolaides, On finite element methods of the least-squares type, *Comput. Math. Appl.*, 5, pp. 87-98, 1979.
- [36] G. Fix, M. Gunzburger and R. Nicolaides, On mixed finite element methods for first-order elliptic systems, *Numer. Math.*, 37, pp. 29-48, 1981.
- [37] G. J. Fix, E. Stephan, On the finite element least squares approximation to higher order elliptic systems, *Arch. Rat. Mech. Anal.*, 91/2, pp. 137-151, 1986.
- [38] L. P. Franca and R. Stenberg; Error analysis of some Galerkin least-squares methods for the elasticity equations, *SIAM J. Numer. Anal.*, 28/6, pp. 1680-1697, 1991.
- [39] V. Ganzha, R. Liska, M. Shashkov and C. Zenger, Mimetic finite difference methods for elliptic equations on unstructured triangular grid; Technische Univ. Muenchen, Report TUM-I0108, December, 2001.
- [40] C. Garasi, J. Hu, A. Robinson, R. Tuminaro and P. Bochev, An improved algebraic multigrid method for solving Maxwell's equations, to appear and *SIAM J. Sci. Comp.*, 2003
- [41] V. Girault and P. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer, Berlin, 1986.
- [42] P.W Gross and P.R. Kotiuga, Data structures for geometric and topological aspects of finite element algorithms. In: F.L. Teixeira (editor) *Geometric methods for computational electromagnetics*, PIER32, EMW Publishing, Cambridge, MA, 2001.
- [43] M. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows*, Academic Press, Boston, 1989.
- [44] R. Hiptmair, Multigrid method for Maxwell's equations, *SIAM J. Numer. Anal.*, 36, pp. 204-225, 1998.
- [45] R. Hiptmair, Canonical construction of finite element spaces, *Math. Comp.*, 68, pp. 1325-1346, 1999.
- [46] R. Hiptmair, Discrete Hodge operators, *Numer. Math.*, 90, pp. 265-289, 2001.

- [47] T.J.R. Hughes and A. Brooks, A theoretical framework for Petrov-Galerkin methods with discontinuous weighting functions: Application to the streamline-upwind procedure, *Finite Elements in Fluids*, 4, edited by R.H. Gallagher et al, J. Willey & Sons pp. 47-65, 1982.
- [48] T.J.R. Hughes, M. Mallet and A. Mizukami, A new finite element formulation for computational fluid dynamics: II. Beyond SUPG, *Comput. Meth. Appl. Mech. Engrg.* 54, pp. 341-355, 1986.
- [49] T.J.R. Hughes, L. Franca, and M. Balestra, A new finite element formulation for computational fluid dynamics: V. Circumventing the Babuska-Brezzi condition: A stable Petrov-Galerkin formulation of the Stokes problem accommodating equal-order interpolations, *Comput. Meth. Appl. Mech. Engrg.*, 59, pp. 85-99, 1986.
- [50] T.J.R. Hughes and L. Franca, A new finite element formulation for computational fluid dynamics: VII. The Stokes problem with various well-posed boundary conditions: symmetric formulations that converge for all velocity pressure spaces, *Comput. Meth. Appl. Mech. Engrg.*, 65, pp. 85-96, 1987.
- [51] T.J.R. Hughes, G. Engel, L. Mazzei, M. Larson; The continuous Galerkin method is locally conservative. *J. Comp. Phys.*, 163/2, pp. 467-488, 2000.
- [52] J. M. Hyman, and M. Shashkov, Mimetic discretizations for Maxwell's equations, *J. Comput. Phys.*, 151, pp. 881-909, 1999.
- [53] J. M. Hyman, and M. Shashkov, Natural discretizations for the divergence, gradient and curl on logically rectangular grids, *Int. J. Comp. & Math with Applic.*, 33, pp. 88-104, 1997.
- [54] J. M. Hyman, and M. Shashkov, Adjoint operators for the natural discretizations of the divergence, gradient and curl on logically rectangular grids, *Applied Numerical Mathematics*, 25, pp. 413-442, 1997.
- [55] J. M. Hyman, and M. Shashkov, The orthogonal decomposition theorems for mimetic finite difference schemes. *SIAM J. Num. Anal.*, 36, pp. 788-818, 1999.
- [56] D. Jespersen, A least-squares decomposition method for solving elliptic equations, *Math. Comp.*, 31, pp. 873-880, 1977.
- [57] B. N. Jiang and L. Povinelli, Optimal least-squares finite element methods for elliptic problems, *Comp. Meth. Appl. Mech. Engrg.*, 102, pp. 199-212, 1993.
- [58] C. Johnson, U. Navert and J. Pitkaranta, Finite element methods for linear hyperbolic problems, *Comput. Meth. Appl. Mech. Engrg.* 45, pp. 285-312, 1984.
- [59] C. Johnson, *Numerical Solution of PDE's by the Finite Element Method*, Cambridge University Press, 1992.

- [60] P.R. Kotiuga, *Hodge Decomposition and Computational Electromagnetics*, Thesis, Department of electrical engineering, McGill University, Montreal, 1984.
- [61] R. Leveque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge Texts in Applied Mathematics, Cambridge University Press, 2002.
- [62] MacCormack R.W. The effect of viscosity in Hypervelocity Impact Cratering, *AIAA* paper 69-354, 1969.
- [63] A. Masud and T.J.R. Hughes, A stabilized mixed finite element method for Darcy flow, *Comp. Meth. Appl. Mech. Engrg.* 191, pp. 4341-4370, 2002.
- [64] J. C. Maxwell, Does the progress of Physical Science tend to give any advantage to the opinion of Necessity (or Determinism) over that of the Contingency of Events and the Freedom of the Will? in Chapter 14, L. Campbell and W. Garnett *The life of James Clark Maxwell*, Macmillan, London, 1882.
- [65] C. Mattiussi, An analysis of finite volume, finite element and finite difference methods using some concepts from algebraic topology, *Journal of Computational Physics*, 133, pp. 289-309, 1997.
- [66] J. Nedelec, Mixed finite elements in \mathbb{R}^3 , *Numer. Math.*, 35, pp. 315-341, 1980.
- [67] J. Nedelec, A new family of finite element methods in \mathbb{R}^3 , *Numer. Math.*, 50, pp. 57-81, 1986.
- [68] R. Nicolaides, Direct discretization of planar div-curl problems, *SIAM J. Numer. Anal.*, 29, pp. 32-56, 1992.
- [69] R. Nicolaides and X. Wu, Covolume solutions of three-dimensional div-curl equations, *SIAM J. Num. Anal.*, 34/6, pp. 2195-2203, 1997.
- [70] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Series in Operations Research, Springer, 1999.
- [71] Raviart P.A., Thomas J.M., A mixed finite element method for second order elliptic problems, *Mathematical aspects of the finite element method*, I. Galligani, E. Magenes, eds. Lecture Notes in Math. 606, Springer-Verlag, New York 1977.
- [72] W. Schwalm, B. Moritz, M. Giona and M. Schwalm, Vector difference calculus for physical lattice models, *Physical Review E*, 59/1, pp. 1217-1233, 1999.
- [73] M.E. Taylor, *Partial Differential Equations. Basic Theory.*, Springer Verlag, 1999.
- [74] F.L. Teixeira (editor) *Geometric Methods for Computational Electromagnetics*, PIER32, EMW Publishing, Cambridge, MA, 2001.
- [75] F.L. Teixeira and W. C. Chew, Lattice electromagnetic theory from a topological viewpoint. *J. Math. Phys.*, 40/1, pp. 169-187, 1999.

- [76] E. Tonti, On the mathematical structure of a large class of physical theories, *Lincei, Rend. Sc. Fis. Mat. e Nat.*, 52/1, pp. 51-56, 1972.
- [77] E. Tonti, The algebraic-topological structure of physical theories, *Proc. Conf. on Symmetry, Similarity and Group Theoretic Meth. in Mech*, pp. 441-467, Calgary, Canada, 1974.
- [78] J. S. Van Welij, Calculation of eddy currents in terms of H on hexahedra, *IEEE Trans. Magnetism*, 21, pp. 2239-2241, 1985.
- [79] T. Weiland, Numerical solution of Maxwell's equations for static, resonant and transient problems, in *Studies in Electrical and Electronic Engineering 28B*, T. Berceci, ed., URSI International Symposium on Electromagnetic Theory Part B, Elsevier, New York, pp. 537-542, 1986.
- [80] S. Weintraub, *Differential Forms. A Complement to Vector Calculus*, Academic Press, 1997
- [81] M. Wheeler, An elliptic collocation-finite element method with interior penalties, *SIAM, J. Numer. Anal.* 15, pp. 152-161, 1978.
- [82] H. Whitney, *Geometric Integration Theory*, Princeton University Press, 1957.
- [83] K. S. Yee, Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media, *IEEE Trans. Antennas and Propagation*, 14, pp. 302-307, 1966.