

Finite Element Methods
Based on Least-Squares and Modified Variational
Principles

Pavel Bochev¹
University of Texas at Arlington
Department of mathematics
`bochev@uta.edu`

April 14, 2004

¹This work is partially supported by Com²MaC-KOSEF and the National Science Foundation under grant number DMS-0073698.

Contents

Preface	viii
Acknowledgments	viii
1 Introduction	1
1.1 Notation	5
2 Review of variational principles	7
2.1 Unconstrained energy minimization	7
2.2 Saddle-point optimization problems	10
2.3 Galerkin methods	16
3 Modified variational principles	21
3.1 Modification of constrained problems	22
3.1.1 The penalty method	24
3.1.2 Penalized and Augmented Lagrangian formulations . .	25
3.1.3 Consistent stabilization	26
3.2 Problems without optimization principles	30
3.2.1 Artificial diffusion and SUPG	31
3.3 Modified variational principles: concluding remarks	32
4 Least-squares methods: first examples	33
4.1 Poisson equation	34
4.2 Stokes equations	35
4.3 PDE's without optimization principles	36
4.4 A critical look	37
4.4.1 Some questions and answers	41

5	CLSP and DLSP	43
5.1	The abstract problem	45
5.2	Continuous least-squares principles	47
5.3	Discrete least-squares principles	50
6	ADN systems	55
6.1	ADN differential operators	56
6.2	CLSP for ADN operators	60
6.3	First-order ADN systems	62
6.4	CLSP for first order systems	64
6.5	DLSP for first-order systems	66
6.5.1	Least-squares for Petrovski systems	66
6.5.2	Least-squares for first-order ADN systems	69
6.6	Concluding remarks	78
7	Least-squares for incompressible flows	81
7.1	First-order equations	82
7.1.1	The velocity-vorticity-pressure equations	82
7.1.2	The velocity-pressure-stress equations	87
7.1.3	Velocity gradient-based transformations	90
7.1.4	First-order formulations: concluding remarks	93
7.2	Inhomogeneous boundary conditions	94
7.3	Least-squares methods	95
7.3.1	Non-equivalent least-squares	97
7.3.2	Weighted least-squares methods	99
7.3.3	H^{-1} least-squares methods	102
7.4	Navier-Stokes equations	103
8	Least squares for $-\Delta u = f$	107
8.1	First-order systems	108
8.1.1	Inhomogeneous boundary conditions	109
8.2	Continuous Least Squares Principles	109
8.2.1	Error estimates	110
8.2.2	Conditioning and preconditioning of discrete systems	111
9	Least-squares methods that stand apart	113
9.1	Least-squares collocation methods	113
9.2	Restricted least-squares methods	115
9.3	Least-squares optimization methods	116

A	The Complementing Condition	119
A.1	Velocity-Vorticity-Pressure Equations	120
A.2	Velocity-Pressure-Stress Equations	124

List of Tables

- 3.1 Comparison of different settings for finite element methods in their most general sphere of applicability. 22

- 7.1 Classification of boundary conditions for the Stokes and Navier-Stokes equations: velocity-vorticity-pressure formulation. . . . 88
- 7.2 Rates of convergence with and without the weights. Velocity-vorticity-pressure formulation with (7.4) and (7.17). 100
- 7.3 Convergence rates with and without the weights. Velocity-pressure-stress formulation. 101

Preface

These lecture notes contain an expanded version of the short course *Finite element methods based on least-squares and modified variational principles* presented at POSTECH on July 5-6, 2001. While this topic is broad enough to include such diverse methods as mixed Galerkin finite elements (where a quadratic positive functional is modified via Lagrange multipliers) to bona-fide least-squares finite elements, we have tried to keep the focus of the presentation on methods which involve, explicitly, or implicitly, application of least-squares principles. Our choice is largely motivated by the recent popularity of such finite element methods and the ever increasing number of practical applications where they have become a viable alternative to the more conventional Galerkin methods.

Space and time limitations have necessarily led to some restrictions on the range of topics covered in the lectures. Besides personal preferences and tastes, which are responsible for the definite “least-squares” bias of these notes, the material selection was also shaped by the existing level of mathematical maturity of the methods. As a result, the bulk of the notes is devoted to the development of least-squares methods for first-order ADN elliptic systems with particular emphasis on the Stokes equations. This choice allows us to draw upon the powerful elliptic regularity theory of Agmon, Douglis and Nirenberg [11] in the analysis of least-squares principles. At the same time, it is general enough so as to expose universal principles occurring in the design of least-squares methods.

For the reader who decides to pursue the subject beyond these notes we recommend the review article [59] and the book [6]. A good summary of early developments, especially in the engineering field can be found in [119]. Least-squares methods for hyperbolic problems and conservation laws remain much less developed which is the reason why we have not included this topic here. The reader interested in such problems is referred to the existing literature, namely [94], [95], [96], [97], [118], and [117] for applications to the Euler equations and hyperbolic systems; [113], [114] for studies of least-squares for scalar hyperbolic problems; and [115] and [116] for convection-diffusion problems.

Acknowledgements

I owe my understanding and appreciation of finite element methods to Max Gunzburger. Max introduced me to least-squares methods and over the years his constant help and encouragement were instrumental for my research.

My work also greatly benefited from the numerous contacts that I had over the years with George Fix, Raytcho Lazarov, Tom Manteuffel, Steve McCormick, and many other excellent researchers.

I always had the support and understanding of my wife Biliana; her patience and care kept me on track while preparing these lecture notes. My deepest thanks to her.

I am grateful for the support of my research provided by the National Science Foundation through grants DMS-0073698 and DMS-9705793.

A special gratitude is reserved for the people in *Com²Mac* center at POSTECH, KOSEF, and in particular to Professors J. H. Kwak, K. I. Kim, and H. S. Oh for their hospitality and making these lecture notes possible.

Chapter 1

Introduction

Importance of variational principles in finite element methods stems from the fact that a finite element method is first and foremost a *quasi-projection* scheme. The paradigm that describes and defines quasi-projections is a synthesis of two components: *a variational principle* and *a closed subspace*. And indeed, a finite element method is completely determined by specifying the variational principle (usually given in terms of a *weak equation* derived from the PDE) and the closed, in fact, finite dimensional subspace. The approximate solutions are then characterized as

quasi-projections of the exact weak solutions onto the closed subspace.

From mathematical viewpoint, the success of this scheme stems from the intrinsic link between variational principles and partial differential equations. From a practical viewpoint, the great appeal of finite element methods (and their wide acceptance in the engineering community) is rooted in the choice of approximating spaces spanned by locally supported, piecewise polynomial functions defined on simple geometrical shapes. The combination of these two ingredients has spawned a truly remarkable class of numerical methods which is unsurpassed in terms of its mathematical maturity and practical utility.

While both the choice of the finite element space and the variational principle play critical role in the finite element method, it is the variational principle that determines the fundamental properties of finite elements, both the favorable ones and the negative ones. Let us recall that there are three different kinds of variational principles that lead to three fundamentally different types of quasi-projections and finite element methods. The first

one stems from *unconstrained* minimization of a positive, convex functional in a Hilbert space and seeks a *global minimum* point. The second variational principle seeks an *equilibrium point*, while the third one is not related to optimization problems at all. In Chapter 2 we will consider examples of finite element methods defined in each one of these three variational settings.

Global minimization of convex functionals, i.e., the first variational setting, offers by far the most favorable environment for a finite element method. In this case the finite element solution is characterized as a true projection with respect to a *problem dependent* inner product in some Hilbert space, i.e., the finite element method is essentially a variant of the classical Rayleigh-Ritz projection with a specific (piecewise polynomial!) choice of the closed subspace. For instance, in linear elasticity, which is among the first successful applications of finite elements, the state \mathbf{u} of an elastic body under given body force \mathbf{f} , surface displacement \mathbf{g} and surface traction \mathbf{t} is characterized as one having the minimum potential energy¹

$$\mathcal{E} = \frac{1}{2} \int_{\Omega} \boldsymbol{\sigma}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{u}) dx - \int_{\Omega} \mathbf{f} \cdot \mathbf{u} dx + \int_{\Gamma_T} \mathbf{t} \cdot \mathbf{u} dS.$$

This connection was not immediately recognized as the principal reason behind the success of the method and some early attempts to extend finite elements beyond problems whose solutions can be characterized as global minimizers encountered serious difficulties.

To understand the cause for these difficulties it suffices to note that mathematical and computational properties of inner product projections on one hand and saddle-point or formal Galerkin principles, on the other hand are strikingly different. Numerical approximation of saddle points, which is the defining paradigm of mixed Galerkin methods, requires strict adherence of the discrete space to restrictive compatibility conditions. Orthogonalization of residuals in the formal Galerkin method can lead to occurrence of spurious oscillations. In both cases we are confronted with the task of solving much less structured algebraic problems than those arising from inner product projections.

Combination of all these factors makes saddle-point and formal Galerkin quasi-projections much more sensitive to variational crimes. Nevertheless, the fact that such difficulties exists does not by any means diminish the overall appeal of the finite element method. It is merely an attestation to the fact that problems without natural energy principles are much harder

¹Here $\boldsymbol{\sigma}(\mathbf{u}) = 2\mu\boldsymbol{\varepsilon}(\mathbf{u}) + \lambda\text{tr}(\boldsymbol{\varepsilon}(\mathbf{u}))\mathbf{I}$ is the stress, $\boldsymbol{\varepsilon}(\mathbf{u}) = \frac{1}{2}(\nabla\mathbf{u} + (\nabla\mathbf{u})^T)$ is the strain, \mathbf{u} is the displacement and λ and μ are the Lamé moduli.

to solve to begin with. In fact, any discretization method that works well for problems with energy principles will inevitably experience similar complications for problems without such principles. However, within the finite element paradigm we can approach these problems in a very systematic and consistent manner by focusing on the variational principle as the main culprit, while in other methods one is confined to a set of remedies defined in an ad hoc manner.

More precisely, the key role of the quasi-projection in the finite element method naturally points towards the exploration of

alternative, externally defined variational principles

in lieu of the naturally occurring quasi-projections². This brings us to the two principal and philosophically different approaches that exist today and whose aim is to obtain better projections (or quasi-projections). The first approach retains the principal role of the naturally occurring quasi-projection but modifies it with terms that make it resemble more a true inner product projection. Some methods that belong to this category are Galerkin-Least-Squares [33]; stabilized Galerkin [26], [34], [32]; the SUPG class of methods [39], [40], [41], [42], [24], [30] and [31]; augmented Lagrangian [21], and penalty [20], [23], [38] formulations, among others. Chapter 3 offers a sampling of several popular finite element methods that belong to these categories.

In contrast, the second approach abandons completely the natural quasi-projection and proceeds to define an artificial, externally defined energy-type principle for the PDE. Typically, the “energy” principle is defined by virtue of residual minimization in some Hilbert spaces, thus the terms “least-squares principles” and “least squares finite elements” are used to describe the ensuing variational equations and finite element methods. In Chapter 4 we take a first look at these methods which will remain in the focus of all subsequent chapters.

Residual minimization is as universal as the residual orthogonalization of Galerkin methods. Thus, it is applicable to virtually any PDE. However, residual minimization differs fundamentally from formal residual orthogonalization in having the potential to recover the attractive features of Rayleigh-Ritz principles. For the same reason least-squares residual minimization differs from methods based on modified variational principles be-

²Another possibility is to modify the finite element spaces by “enriching” them with, e.g., bubble functions. This enrichment is related, and in many cases equivalent, to modification of the variational principle; see e.g., [27], [36] and [35]. Thus, we do not pursue this topic here.

cause such methods are not capable of recovering all of the advantages of the Rayleigh-Ritz setting.

Finite element methods based on least-squares variational principles have been the subject of extensive research efforts over the last two decades. While these efforts have paid off in turning least-squares into a viable alternative to standard and modified Galerkin methods, formulation of a good least-squares method requires careful analysis. Since such methods are based on inner product projections they tend to be exceptionally robust and stable. As a result, one is often tempted to forego analyses and proceed with the seemingly most natural least-squares formulation. As we shall see such “shortcuts” do not necessarily lead to methods that can fully exploit the advantages of least-squares principles.

Among the factors responsible for this renaissance of least-squares after a somewhat disappointing start in the early seventies³ a key role was played by the idea of transformations to equivalent first-order systems. This helped to circumvent the need to work with impractical C^1 finite element spaces and led to a widespread use of least-squares in fluid flow computations; see [48]–[58], [98]–[101], [108]–[111] and [104], among others. From the mathematical standpoint another idea, namely the notion of *norm-equivalence* of least-squares functionals emerged as a universal prerequisite for recovering fully the Rayleigh-Ritz setting. However, it was soon realized that norm-equivalence is often in conflict with practicality, even for first-order systems (see [48], [56] and [58]); and because practicality is usually the rigid constraint in the algorithmic development, norm equivalence was often sacrificed.

This brings us to the main theme of these notes which is to establish the reconciliation between practicality, as driven by algorithmic development, and norm-equivalence, as motivated by mathematical analyses, as the defining paradigm of least-squares finite element methods. The key components of this paradigm are introduced in Chapter 5 and include a *continuous least-squares principle* (CLSP) which describes a mathematically well-posed, but perhaps impractical, variational setting, and an associated *discrete least-squares principle* (DLSP) which describes an algorithmically feasible setting. The association between a CLSP and a DLSP follows four

³Early examples of least-squares methods suffered from serious disadvantages that seriously limited their appeal. For instance, such methods often demanded higher (compared with Galerkin methods) solution regularity to establish convergence. Similarly, in many cases discretization required impractical C^1 or better finite element spaces and led to algebraic problems with higher than usual condition numbers; see e.g., [46], [60]–[61]. Furthermore, in most cases it wasn’t clear how to precondition these problems efficiently.

universal patterns which lead to four classes of least-squares finite elements with distinctly different properties.

In Chapter 6 we develop this paradigm for the important class of first-order systems that are elliptic in the sense of Agmon-Douglis-Nirenberg [11]. In particular, we show that degradation of fundamental properties of least-squares method such as condition numbers, asymptotic convergence rates, and existence of spectrally equivalent preconditioners occurs when DLSP deviates from the mathematical setting induced by a given CLSP.

Then, in Chapters 7–8 the least-squares approach is further specialized to the Stokes equations and the Poisson problem, respectively. The discussion is rounded up in Chapter 9 with a brief summary of least-squares methods that do not fit into the mold of Chapter 6.

For the convenience of the reader we have decided to include some of the details that accompany the application of ADN theory for the development of the methods in Chapter 6. Most of this material is collected in Appendix A where the Complementing Condition of [11] is verified for two first-order forms of the Stokes equations.

1.1 Notation

Throughout these notes we try to adhere to standard notations and symbols. Ω will denote an open bounded domain in \mathcal{R}^n , $n = 2$ or 3 , having a sufficiently smooth boundary Γ . Throughout, vectors will be denoted by bold face letters, e.g., \mathbf{u} , tensors by underlined bold faced capitals, e.g., $\underline{\mathbf{T}}$, and C will denote a generic positive constant whose meaning and value changes with context. For $s \geq 0$, we use the standard notation and definition for the Sobolev spaces $H^s(\Omega)$ and $H^s(\Gamma)$ with corresponding inner products denoted by $(\cdot, \cdot)_{s,\Omega}$ and $(\cdot, \cdot)_{s,\Gamma}$ and norms by $\|\cdot\|_{s,\Omega}$ and $\|\cdot\|_{s,\Gamma}$, respectively. Whenever there is no chance for ambiguity, the measures Ω and Γ will be omitted from inner product and norm designations. We will simply denote the $L^2(\Omega)$ and $L^2(\Gamma)$ inner products by (\cdot, \cdot) and $(\cdot, \cdot)_\Gamma$, respectively. We recall the space $H_0^1(\Omega)$ consisting of all $H^1(\Omega)$ functions that vanish on the boundary and the space $L_0^2(\Omega)$ consisting of all square integrable functions with zero mean with respect to Ω . Also, for negative values of s , we recall the dual spaces $H^s(\Omega)$.

By $(\cdot, \cdot)_{\mathbf{X}}$ and $\|\cdot\|_{\mathbf{X}}$ we denote inner products and norms, respectively, on the product spaces $\mathbf{X} = H^{s_1}(\Omega) \times \cdots \times H^{s_n}(\Omega)$; whenever all the indices s_i are equal we shall denote the resulting space by $[H^{s_1}(\Omega)]^n$ or by $\mathbf{H}^s(\Omega)$ and simply write $(\cdot, \cdot)_{s,\Omega}$ and $\|\cdot\|_{s,\Omega}$ for the inner product and norm, respectively.

Due to the limited space we do not quote a number of relevant results concerning Sobolev spaces and finite element approximation theory, instead we refer the reader to the monographs [1], [2], [3] and [4] for more detailed information on these subjects.

Chapter 2

Review of variational principles

In this chapter we present three well-known examples of finite element methods. Each example highlights one of the three naturally occurring variational principles. The purpose of this review is to expose the key role played by the different types of quasi-projections for the analytical and computational properties of the ensuing finite element methods.

2.1 Unconstrained energy minimization

Consider the convex, quadratic functional

$$J(\phi; f) = \frac{1}{2} \int_{\Omega} |\nabla \phi|^2 d\Omega - \int_{\Omega} f \phi d\Omega \quad (2.1)$$

and the minimization principle

$$\min_{\phi \in H_0^1(\Omega)} J(\phi; f), \quad (2.2)$$

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 . Setting the first variation of (2.1) to zero gives the first-order necessary condition for (2.2). Therefore, we find that the minimizer $\phi \in H_0^1(\Omega)$ of the functional (2.1) satisfies the variational equation

$$Q_r(\phi; \psi) = F(\psi) \quad \forall \psi \in H_0^1(\Omega), \quad (2.3)$$

where

$$Q_r(\phi; \psi) = \int_{\Omega} \nabla \phi \cdot \nabla \psi \, d\Omega \quad \text{and} \quad F(\psi) = \int_{\Omega} f \psi \, d\Omega. \quad (2.4)$$

To see the connection between the minimization principle (2.2) and partial differential equations, we integrate by parts¹ in (2.3) to obtain

$$0 = \int_{\Omega} (\nabla \phi \cdot \nabla \psi - f \psi) \, d\Omega = - \int_{\Omega} \psi (\Delta \phi + f) \, d\Omega. \quad (2.5)$$

Since ψ is arbitrary, it follows that every sufficiently smooth minimizer of $J(\cdot; f)$ is a solution of the familiar Poisson problem

$$-\Delta \phi = f \quad \text{in } \Omega \quad \text{and} \quad \phi = 0 \quad \text{on } \Gamma. \quad (2.6)$$

The boundary condition follows from the fact that all admissible states were required to vanish on Γ .

We note that (2.3) makes sense for functions ϕ that vanish on Γ and that have merely square integrable first derivatives. On the other hand, (2.6) requires ϕ to have two continuous derivatives. Thus, one 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 (2.2) but the latter admits solutions which are not classical solutions of (2.6). These non-classical solutions of (2.2) are referred to as *weak solutions* of the Poisson problem.

The correspondence between minimizers of (2.2) and solutions of (2.6) is not a rare coincidence. A large number of physical processes is governed by energy minimization principles similar to the one considered above. The first-order optimality systems of these principles can be transformed into differential equations, provided the minimizer is smooth enough.

The analytic and computational advantages of the energy minimization setting stem from the fact that the expression

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

defines an *equivalent norm* on the space $H_0^1(\Omega)$. As a result, $Q_r(\cdot; \cdot)$ defines an *equivalent inner product* on $H_0^1(\Omega)$. The *norm-equivalence* of the functional (2.1) is a direct consequence of the Poincaré inequality

$$\lambda \|\psi\|_0 \leq |\psi|_1 \quad \forall \psi \in H_0^1(\Omega), \quad (2.7)$$

¹Assuming that the minimizer ϕ of $J(\cdot; f)$ is sufficiently smooth to justify the above integration.

where λ is a constant whose value depends only on Ω . The inner product equivalence

$$(1 + \lambda^{-2})^{-1} \|\psi\|_1^2 \leq Q_r(\psi; \psi) \quad \text{and} \quad Q_r(\phi; \psi) \leq \|\phi\|_1 \|\psi\|_1, \quad (2.8)$$

follows from the identity $|\phi|_1^2 = Q_r(\phi; \phi)$ and the Cauchy inequality. Thus, the energy principle (2.2) gives rise to the the equivalent *energy norm*

$$|||\phi||| \equiv J(\phi; 0)^{1/2}$$

and the equivalent *energy inner product*

$$((\phi, \psi)) \equiv Q_r(\phi; \psi).$$

Let us now investigate the computational advantages of this setting in the finite element method. We consider a weak solution ϕ and its finite element approximation ϕ^h . This approximation is determined by solving the variational problem

$$\text{seek } \phi^h \in X^h \text{ such that } Q_r(\phi^h; \psi^h) = F(\psi^h) \quad \forall \psi^h \in X^h, \quad (2.9)$$

where X^h is a finite dimensional subspace of $H_0^1(\Omega)$. Note that (2.9) is simply (2.3), restricted to X^h .

First, we observe that the *conformity*² of X^h and the fact that (2.8) holds for all functions belonging to $H_0^1(\Omega)$ imply that (2.9) defines an *orthogonal projection* of ϕ onto X^h with respect to the inner product $((\cdot, \cdot))$. From the fact that the exact solution satisfies the discrete problem and (2.9) it follows 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

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

As a result, ϕ^h minimizes the energy norm of the error, i.e.,

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

In conjunction with the continuity and coercivity bounds of (2.8) this bound gives an error estimate in the norm of $H_0^1(\Omega)$:

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

²In the sense that the inclusion $X^h \subset H_0^1(\Omega)$ holds for all h

Second, we observe that the norm-equivalence of the energy functional also implies *stability* in the norm of $H_0^1(\Omega)$. This follows from the coercivity bound in (2.8) which shows that the energy norm controls the gradient of the weak solution.

Lastly, let us examine the linear algebraic system that corresponds to the weak equation (2.9). Given a basis $\{\phi_i\}_{i=1}^N$ of X^h this system has the form

$$A\Phi^h = F, \quad (2.10)$$

where $A_{ij} = ((\phi_i, \phi_j)) = Q_r(\phi_i; \phi_j)$, $F_i = F(\phi_i)$, and $(\Phi^h)_j = c_j$ are the unknown coefficients of ϕ^h . From (2.4) and (2.8) it follows that A is symmetric and positive definite matrix. Moreover, the equivalence between the energy inner product defined by $Q_r(\cdot; \cdot)$ and the standard inner product on $H_0^1(\Omega) \times H_0^1(\Omega)$ implies spectral equivalence between A and the Gram matrix of $\{\phi_i\}_{i=1}^N$ in $H_0^1(\Omega)$ -inner product. This fact is useful for the design of efficient preconditioners for (2.10).

All attractive features described so far stem from exactly two factors: characterization of all weak solutions as minimizers of unconstrained energy functional and the fact that X^h is a subspace of $H_0^1(\Omega)$. As a result, the finite element solution ϕ^h is an orthogonal projection of the exact solution ϕ onto X^h . Moreover, as long as the inclusion $X^h \subset H_0^1(\Omega)$ holds,

- the discrete problems will have unique solutions;
- the approximate solutions will 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 will have symmetric and positive definite coefficient matrices;
- these matrices will be spectrally equivalent to the Gram matrix of the trial space basis in the natural norm of $H_0^1(\Omega)$.

2.2 Saddle-point optimization problems

We consider a setting in which weak solutions of PDE's are characterized via constrained minimization of convex, quadratic functionals. We note that a constrained optimization problem can be formally recast into an unconstrained one by simply restricting the admissible space by the constraint. The two settings are equivalent and, in theory, finite element methods may be based on either setting.

In practice, the choice of settings will depend on the ease with which the constraint can be imposed on a finite element space. Some constraints are trivial to impose, while other constraints require complicated construction of finite element spaces. In such a case one may choose to use Lagrange multipliers instead. This results in weak problems of the *saddle-point* type and finite element methods which lack many of the attractions of the Rayleigh-Ritz setting.

To illustrate how different constraints affect the choice of variational formulations for the finite element method consider again the weak Poisson problem (2.6). This variational equation gives the first-order necessary condition for the *unconstrained minimization* of (2.1). In actuality this problem is constrained in the sense that all admissible states are required to vanish on the boundary Γ . However, we avoided dealing explicitly with this constraint by minimizing (2.1) over $H_0^1(\Omega)$. Of course, now it is necessary to approximate $H_0^1(\Omega)$, but we have avoided Lagrange multipliers³. Moreover, finite element subspaces of $H_0^1(\Omega)$ are not at all hard to find; see, e.g., [3].

Now let us consider the quadratic functional

$$J(\mathbf{u}; \mathbf{f}) = \frac{1}{2} \int_{\Omega} |\nabla \mathbf{u}|^2 d\Omega - \int_{\Omega} \mathbf{f} \cdot \mathbf{u} d\Omega \quad (2.11)$$

and the minimization problem

$$\min_{\mathbf{u} \in \mathbf{H}^1(\Omega)} J(\mathbf{u}; \mathbf{f}) \quad \text{subject to } \nabla \cdot \mathbf{u} = 0 \text{ and } \mathbf{u}|_{\Gamma} = 0, \quad (2.12)$$

where $\mathbf{H}^1(\Omega)$ is the vector analog of $H^1(\Omega)$. To avoid Lagrange multipliers this problem can be converted to unconstrained minimization of (2.11) on the space

$$\mathbf{Z} = \{\mathbf{v} \in \mathbf{H}^1(\Omega) \mid \nabla \cdot \mathbf{v} = 0; \mathbf{u}|_{\Gamma} = 0\} \equiv \{\mathbf{v} \in \mathbf{H}_0^1(\Omega) \mid \nabla \cdot \mathbf{v} = 0\}$$

of solenoidal functions belonging to $\mathbf{H}_0^1(\Omega)$. We then pose the unconstrained minimization problem

$$\min_{\mathbf{u} \in \mathbf{Z}} J(\mathbf{u}; \mathbf{f}). \quad (2.13)$$

The first-order necessary condition for (2.13) is

seek $\mathbf{u} \in \mathbf{Z}$ such that

³There are instances when this approach is useful, especially for inhomogeneous boundary conditions posed on complicated regions; see, e.g., [17].

$$\int_{\Omega} \nabla \mathbf{u} : \nabla \mathbf{v} \, d\Omega = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \, d\Omega \quad \forall \mathbf{v} \in \mathbf{Z}. \quad (2.14)$$

It is easy to see that $\int_{\Omega} \nabla \mathbf{u} : \nabla \mathbf{v} \, d\Omega$ is coercive and continuous on $\mathbf{Z} \times \mathbf{Z}$ so that (2.13) has a unique solution. Therefore, (2.13) provides a Rayleigh-Ritz setting for (2.12). The problem is that in order to use this setting to define a finite element method we must construct a conforming subspace of \mathbf{Z} . This is not trivial⁴ at all, at least compared with satisfying the constraint $\mathbf{u} = \mathbf{0}$ and so we introduce the Lagrange multiplier function p , the Lagrangian functional

$$L(\mathbf{u}, p; \mathbf{f}) = J(\mathbf{u}; \mathbf{f}) - \int_{\Omega} p \nabla \cdot \mathbf{u} \, d\Omega, \quad (2.15)$$

and the unconstrained problem of determining saddle points of $L(\mathbf{u}, p; \mathbf{f})$. The first-order necessary conditions for (2.15) are equivalent to the weak problem:

seek (\mathbf{u}, p) in an appropriate function space such that $\mathbf{u} = \mathbf{0}$ on Γ and

$$\begin{aligned} \int_{\Omega} \nabla \mathbf{u} : \nabla \boldsymbol{\xi} \, d\Omega - \int_{\Omega} p \nabla \cdot \boldsymbol{\xi} \, d\Omega &= \int_{\Omega} \mathbf{f} \cdot \boldsymbol{\xi} \, d\Omega \\ \int_{\Omega} \mu \nabla \cdot \mathbf{u} \, d\Omega &= 0 \end{aligned} \quad (2.16)$$

for all $(\boldsymbol{\xi}, \mu)$ in the corresponding function space.

If solutions to the constrained minimization problem (2.12) or, equivalently, of (2.16), are sufficiently smooth, then, using integration by parts, one obtains without much difficulty the Stokes equations

$$\begin{aligned} -\Delta \mathbf{u} + \nabla p &= \mathbf{f} \quad \text{and} \quad \nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega, \\ \mathbf{u} &= \mathbf{0} \quad \text{on } \Gamma, \end{aligned} \quad (2.17)$$

where \mathbf{u} is the *velocity* and p is the *pressure*. Thus, (2.16) is a weak formulation of the Stokes equations. Solutions of (2.17) are determined up to a hydrostatic pressure mode. This mode can be eliminated by imposing an additional constraint on the pressure variable. A standard method of doing this is to require that

$$\int_{\Omega} p \, dx = 0. \quad (2.18)$$

⁴It is much easier to construct a *non-conforming* solenoidal space. One example are Raviart-Thomas spaces; see [22].

A second example of a constrained minimization problem is

$$\min J(\mathbf{u}) \quad \text{subject to } \nabla \cdot \mathbf{u} = f, \quad (2.19)$$

where the “energy” functional is given by

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

In fluid mechanics, (2.19) is known as the *Kelvin principle* and, in structural mechanics (where \mathbf{u} is a tensor), as the *complimentary energy principle*. The constraint in (2.19) defines an *affine subspace* which makes it even harder to satisfy! Therefore, we are forced again to consider a Lagrange multiplier p to enforce the constraint and the Lagrangian functional

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

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

seek (\mathbf{u}, p) belonging to some appropriate function space such that

$$\begin{aligned} \int_{\Omega} \mathbf{u} \cdot \mathbf{v} d\Omega - \int_{\Omega} p \nabla \cdot \mathbf{v} d\Omega &= 0 \\ \int_{\Omega} q \nabla \cdot \mathbf{u} d\Omega &= \int_{\Omega} f q d\Omega \end{aligned} \quad (2.20)$$

for all (\mathbf{v}, q) belonging to the corresponding function space.

If solutions to the constrained minimization problem (2.19) or, equivalently, of (2.20), are sufficiently smooth, then integration by parts can be used to show that

$$\begin{aligned} \nabla \cdot \mathbf{u} &= f \quad \text{and} \quad \mathbf{u} + \nabla p = \mathbf{0} \quad \text{in } \Omega \\ p &= 0 \quad \text{on } \Gamma. \end{aligned} \quad (2.21)$$

If \mathbf{u} is eliminated from this system, we obtain the Poisson problem (2.6) for p . Thus, (2.20) is another weak formulation⁵ of the Poisson problem (2.6).

⁵One reason why one would want to solve (2.21) instead of dealing directly with the Poisson equation (2.6) is that in many applications $\mathbf{u} = -\nabla\phi$ may be of greater interest than ϕ , e.g., heat fluxes vs. temperatures, or velocities vs. pressures, or stresses vs. displacements. Thus, since differentiation of an approximation ϕ^h could lead to a loss of precision, the direct approximation of $\nabla\phi$ becomes a matter of considerable interest.

Both examples of saddle-point optimization problems can be cast into the abstract form

$$a(u, v) + b(v, p) = \mathcal{F}(v) \quad \forall v \in V \quad (2.22)$$

$$b(u, q) = \mathcal{G}(q) \quad \forall q \in S, \quad (2.23)$$

where V and S are appropriate function spaces, $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ are bilinear forms on $V \times V$ and $V \times S$, respectively, and $\mathcal{F}(\cdot)$ and $\mathcal{G}(\cdot)$ are linear functionals on V and S , respectively. The system (2.22)–(2.23) is a typical optimality system for constrained minimization problems in which the bilinear form $a(\cdot, \cdot)$ is symmetric and is related to a convex, quadratic functional and (2.23) is a weak form of the constraint.

Well-posedness of (2.22)–(2.23) requires, among other things the following two conditions; see, e.g., [17], [19]:

$$\sup_{u \in Z} \frac{a(u, v)}{\|u\|_V} \geq \alpha \|v\|_V \quad \forall v \in Z \quad (2.24)$$

and

$$\sup_{v \in V} \frac{b(v, q)}{\|v\|_V} \geq \beta \|q\|_S \quad \forall q \in S, \quad (2.25)$$

where the subspace Z is defined by

$$Z = \{z \in V \mid b(z, q) = 0 \quad \forall q \in S\}.$$

The first bound is almost always satisfied because $a(\cdot, \cdot)$ is defined by a quadratic functional. The second bound (2.25), represents a *compatibility* condition between the space V and the Lagrange multiplier space S . It is more difficult to verify but is still satisfied for all problems of practical interest. Thus, from theoretical viewpoint the use of Lagrange multipliers did not introduce some serious difficulties. As we shall see in a moment, the use of multipliers will, however, considerably complicate the finite element method.

Suppose that $V^h \subset V$ and $S^h \subset S$ are two finite element subspaces of the “correct” function spaces. We restrict (2.22)–(2.23) to these spaces to obtain the discrete problem

$$a(u^h, v^h) + b(v^h, p^h) = \mathcal{F}(v^h) \quad \forall v^h \in V^h \quad (2.26)$$

$$b(u^h, q^h) = \mathcal{G}(q^h) \quad \forall q^h \in S^h, \quad (2.27)$$

which is a linear algebraic system of the form

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} U^h \\ P^h \end{pmatrix} = \begin{pmatrix} F^h \\ G^h \end{pmatrix}. \quad (2.28)$$

The vectors U^h and P^h contain the coefficients of the unknown functions u^h and p^h , and A and B are blocks generated by the forms in (2.22)–(2.23). The matrix in (2.28) is symmetric and *indefinite*; in contrast, the system (2.10) for the Rayleigh-Ritz method was symmetric and *positive definite*. Thus, (2.28) is more difficult to solve.

Still, solving (2.28) is not the main problem, making sure that this system is nonsingular and gives meaningful approximations is! Indeed, equations (2.26)–(2.27) are a discrete saddle-point problem. Therefore, unique, stable solvability of these equations is subject to the same conditions as were necessary for (2.22)–(2.23). In particular, it can be shown that (2.26)–(2.27) is well posed if and only if V^h and S^h satisfy the well-known inf-sup⁶, or Ladyzhenskaya-Babuska-Brezzi (LBB),⁷ or div-stability condition⁸

there exists $\beta > 0$, independent of h , such that

$$\sup_{v \in V^h} \frac{b(v, q)}{\|v\|_V} \geq \beta \|q\|_S \quad \forall q \in S^h \quad (2.29)$$

and the bilinear form $a(\cdot, \cdot)$ is coercive on $Z^h \times Z^h$, where $Z^h \subset V^h$ denotes the subspace of function satisfying the discrete constraint equations, i.e.,

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

The difficulty here is that

the inf-sup condition does not follow from the inclusions $V^h \subset V$ and $S^h \subset S$,

which is in sharp contrast with Rayleigh-Ritz setting where conformity was sufficient to provide well-posed discrete problems.

Note that the solution $(u^h, p^h) \in V^h \times S^h$ of (2.26)–(2.27) is not a projection of the solution $(u, p) \in V \times S$ of (2.22)–(2.23). To see this, note that (2.22)–(2.23) may be expressed in the equivalent form: seek $(u, p) \in V \times S$ such that

$$Q_s(u, p; v, q) = \mathcal{H}(v, q) \quad \forall (v, q) \in V \times S,$$

⁶The terminology “inf-sup” originates from the equivalent form $\inf_{q \in S^h} \sup_{v \in V^h} \frac{b(q, v)}{\|q\|_S \|v\|_V} \geq \beta$ of this condition.

⁷The terminology “LBB” originates from the facts that this condition was first explicitly discussed in the finite element setting by Brezzi [19] and that is a special case of the general weak-coercivity condition given by Babuska [16] for finite element methods and that, in the continuous setting of the Stokes equation, this condition was first proved by Ladyzhenskaya [7].

⁸The terminology “div-stability” arises from the application of this condition to the Stokes problem in which the constraint equation is $\nabla \cdot \mathbf{u} = 0$.

where $Q_s(u, p; v, q) \equiv a(u, v) + b(v, p) + b(u, q)$ and $\mathcal{H}(v, q) \equiv \mathcal{F}(v) + \mathcal{G}(q)$. Likewise, (2.26)–(2.27) is equivalent to seeking $(u^h, p^h) \in V^h \times S^h$ such that

$$Q_s(u^h, p^h; v^h, q^h) = \mathcal{H}(v^h, q^h) \quad \forall (v^h, q^h) \in V^h \times S^h.$$

These relations easily imply the usual finite element “orthogonality relation”

$$Q_s(u - u^h, p - p^h; v^h, q^h) = 0 \quad \forall (v^h, q^h) \in V^h \times S^h.$$

However, this does *not* by itself imply, even though $V^h \subset V$ and $S^h \subset S$, that (u^h, p^h) is an orthogonal projection onto $V^h \times S^h$ of the exact solution $(u, p) \in V \times S$ nor does it imply that the errors $u - u^h$ and $p - p^h$ are quasi-optimally accurate. This follows from the fact that $Q_s(\cdot; \cdot)$ does not define an inner product on $V \times S$.

2.3 Galerkin methods

Galerkin methods represent a formal (and very general) methodology that can be used to derive variational formulations directly from PDE’s. The paradigm of a Galerkin method is the *residual orthogonalization*. This principle can be applied to any PDE, even if there’s no underlying optimization problem. On the other hand, as we shall see, if such an optimization problem exists, then Galerkin methods do recover the associated optimality system. Because of this universality, Galerkin method has been a natural choice for extending finite elements beyond differential equations problems associated with minimization principles.

Let us first show that a Galerkin method can recover the optimality system if the PDE is associated with an optimization problem. For the model Poisson problem (2.6), the standard Galerkin approach is to multiply the differential equation by a *test* function ψ that vanishes on Γ , then integrate the result over the domain Ω , and then apply Green’s formula to equilibrate the order the highest derivatives applied to the unknown ϕ and the test function ψ ; the result is exactly (2.3). For the Stokes problem (2.17), we multiply the first equation by a test function \mathbf{v} that vanishes on the boundary Γ , integrate the result over Ω , and then integrate by parts in both terms to move one derivative onto the test function. We also multiply the second equation by a test function q and integrate the result over Ω . This process results in exactly (2.16). Thus, we were able to derive exactly the same weak formulations as before, working directly from the differential equation and without appealing to any calculus of variations ideas. However, it is clear

that there is some ambiguity associated with Galerkin methods, i.e., there are some choices faced in the process. A given differential equation problem can give rise to more than one weak formulation; we already saw this for the Poisson problem for which we obtained the weak formulations (2.3) and (2.20).

Let us now apply Galerkin method to a problem for which no corresponding minimization principle exists. A simple example is provided by the Helmholtz equation problem

$$-\Delta\phi - k^2\phi = f \quad \text{in } \Omega \quad \text{and} \quad \phi = 0 \quad \text{on } \Gamma. \quad (2.30)$$

Using the same procedure as for the Poisson equation we find the weak formulation of (2.30) to be

$$\int_{\Omega} (\nabla\phi \cdot \nabla\psi - k^2\phi\psi) d\Omega = \int_{\Omega} f\psi d\Omega \quad \forall \psi \in H_0^1(\Omega). \quad (2.31)$$

Note that the bilinear form on the left-hand side of (2.31) 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 uniqueness⁹ of weak solutions is not so simple a matter as it is for the Poisson equation case.

Another example of a problem without an associated optimization principle is the convection-diffusion-reaction equation

$$-\varepsilon\Delta\phi + \mathbf{b} \cdot \nabla\phi + c\phi = f \quad \text{in } \Omega \quad \text{and} \quad \phi = 0 \quad \text{on } \Gamma. \quad (2.32)$$

Following the familiar Galerkin procedure for (2.32) results in the weak formulation

$$\int_{\Omega} (\varepsilon\nabla\phi \cdot \nabla\psi + \psi\mathbf{b} \cdot \nabla\phi + c\phi\psi) d\Omega = \int_{\Omega} f\psi d\Omega \quad \forall \psi \in H_0^1(\Omega). \quad (2.33)$$

Now the bilinear form on the left-hand side of (2.33) is neither symmetric or coercive.

The weak formulations (2.31) and (2.33) are examples of the abstract problem: seek $u \in V$ such that

$$Q_g(u; v) = \mathcal{F}(v) \quad \forall v \in V, \quad (2.34)$$

where $Q_g(\cdot; \cdot)$ is a bilinear form and $\mathcal{F}(\cdot)$ a linear functional. Conforming finite element approximations are defined in the usual manner. One chooses

⁹In fact, solutions of (2.30) or (2.31) are not always unique.

a finite element subspace $V^h \subset V$ and then poses (2.34) on the subspace, i.e., one seeks $u^h \in V^h$ such that

$$Q_g(u^h; v^h) = \mathcal{F}(v^h) \quad \forall v^h \in V^h. \quad (2.35)$$

In general, the bilinear form $Q_g(\cdot; \cdot)$ is not coercive and/or symmetric and thus does not define an equivalent inner product on V . As a result, unlike the Rayleigh-Ritz setting, the conformity of approximating space is not sufficient to insure that the discretized problem (2.35) is well posed nor that the approximate solution is quasi-optimally accurate.¹⁰ To insure that it is indeed well posed, one must have that at least the weak coercivity or (general) inf-sup conditions

$$\inf_{u^h \in V^h} \sup_{v^h \in V^h} \frac{Q_g(u^h; v^h)}{\|u^h\| \|v^h\|} \geq C \quad \text{and} \quad \sup_{u^h \in V^h} \frac{Q_g(u^h; v^h)}{\|u^h\|} \geq 0$$

hold. We also note that the standard finite element ‘‘orthogonality’’ relation

$$Q_g(u - u^h; v^h) = 0 \quad \forall v^h \in V^h \quad (2.36)$$

is easily derived from (2.34) and (2.35). Since the bilinear form $Q_g(\cdot; \cdot)$ does not define an equivalent inner product on V , (2.36) does *not* imply that u^h is a projection onto V^h of the exact solution $u \in V$, even though $V^h \subset V$. For the same reason and equivalently, (2.36) does not truly state that the error $u - u^h$ is orthogonal to the approximating subspace V^h .

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

$$\begin{aligned} -\nu \Delta \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} + \nabla p &= \mathbf{f} & \text{in } \Omega \\ \nabla \cdot \mathbf{u} &= 0 & \text{in } \Omega \\ \mathbf{u} &= 0 & \text{on } \Gamma, \end{aligned} \quad (2.37)$$

where the constant ν denotes the kinematic viscosity. A standard weak formulation analogous to (2.16) but containing an additional nonlinear term is given by

$$\begin{aligned} \nu \int_{\Omega} \nabla \mathbf{u} : \nabla \mathbf{v} \, d\Omega + \int_{\Omega} p \nabla \cdot \mathbf{v} \, d\Omega \\ + \int_{\Omega} \mathbf{u} \cdot \nabla \mathbf{u} \cdot \mathbf{v} \, d\Omega = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \, d\Omega \quad \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega), \end{aligned} \quad (2.38)$$

¹⁰The discretized weak formulation (2.35) is equivalent to a linear algebraic system of the type (2.10), but unlike the Rayleigh-Ritz setting, the coefficient matrix A is now not symmetric for the weak formulation (2.33) and may not be positive definite for this problem and for (2.31); in fact, it may even be singular.

$$\int_{\Omega} q \nabla \cdot \mathbf{u} \, d\Omega = 0 \quad \forall q \in L_0^2(\Omega). \quad (2.39)$$

Despite the close resemblance between (2.16) and (2.38)–(2.39), 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 attached to these weak equations. As a result, (2.38)–(2.39) cannot be derived in any other way but through the Galerkin procedure described above.

All these examples show the ease with which one can obtain weak problems for virtually any partial differential equation by following the Galerkin recipe. The process used to derive the weak equations always leads to a variational problem and did not require any prior knowledge of whether or not there is a naturally existing minimization principle. However, the versatility of the Galerkin method comes at a price. The limited expectations the method has with respect to an available mathematical structure for the differential equation also makes its analysis and implementation a more difficult matter than that for methods rooted in energy minimization principles.

Chapter 3

Modified variational principles

The examples given in §2.1–§2.3 show that the further the variational framework for a finite element method deviates from the Rayleigh-Ritz setting, the greater are the levels of theoretical and practical complications associated with the method. These observations are summarized in Table 3.1. Given the advantages of the Rayleigh-Ritz setting it is not surprising that much effort has been spent in trying to recover or at least restore some of its attractive properties to situations where it does not occur naturally. Historically, these efforts have developed in two distinct directions, one based on

modifications of naturally occurring variational principles

and the other on the use of

externally defined, artificial energy functionals.

The second approach ultimately leads to *bona fide least-squares* variational principles and finite element methods which are potentially capable of recovering the advantages of the Rayleigh-Ritz setting.

This chapter will focus on the first class of finite element methods. Even though these methods do not recover all of the advantages of the Rayleigh-Ritz setting they lead to important examples of finite element methods that are used in practice. This class of methods also provides an illustration of another useful application of least-squares as stabilization tool.

	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	general inf-sup condition
requirements on discrete spaces	conformity	conformity and discrete inf-sup condition	conformity and general discrete inf-sup condition
properties of discrete problems	symmetric, positive definite	symmetric but indefinite	indefinite, not symmetric

Table 3.1: Comparison of different settings for finite element methods in their most general sphere of applicability.

3.1 Modification of constrained problems

The focus of this section will be on problems that are associated with constrained optimization of some convex, quadratic functional, i.e., we consider the problem

$$\min_{u \in V} J(u) \quad \text{subject to } \Lambda(u) = 0. \quad (3.1)$$

In (3.1) $J(\cdot)$ is a given energy functional, V a suitable function space, and $\Lambda(\cdot)$ a given constraint operator. We assume that the constraint $\Lambda(U) = 0$ is not a benign constraint, i.e., it is not easy to enforce on functions belonging to V . In §2.2, the Lagrange multiplier method was used to enforce the constraint. This led to the *Lagrangian functional*

$$L(u, \mu) = J(u) + \langle \mu, \Lambda(u) \rangle \quad (3.2)$$

and the associated *mixed Galerkin method*. Note that (3.2) may be viewed as a modification of the naturally occurring functional $J(\cdot)$ associated with the given problem.

An alternate way to treat the constraint is through penalization; one sets up an unconstrained minimization problem for the *penalized functional*

$$J_\rho(u) = J(u) + \rho \|\Lambda(u)\|^2, \quad (3.3)$$

where ρ is a parameter and $\|\cdot\|$ is a norm that the user has to choose. The use of penalty functionals in lieu of Lagrange functionals is one possibility for developing better variational principles; however, the penalty approach does not necessarily lead to better approximations.

One can combine Lagrange multipliers with penalty terms leading to the *augmented Lagrangian functional*

$$L_a(u, \mu) = J(u) + \langle \mu, \Lambda(u) \rangle + \rho \|\Lambda(u)\|^2 \quad (3.4)$$

and the associated augmented Lagrangian method which result from its unconstrained minimization. One can also penalize the Lagrangian functional with a term involving the Lagrange multiplier instead of the constraint, leading to the *penalized Lagrangian functional*

$$L_p(u, \mu) = J(u) + \langle \mu, \Lambda(u) \rangle + \rho \|\mu\|^2 \quad (3.5)$$

and the associated penalized Lagrangian method.

Solutions of optimization problems connected with any of the functionals (3.3)–(3.5) are not, in general, solutions of (3.1).¹ This potential disadvantage associated with the use of these functionals can be overcome by penalizing with respect to the residuals of the Euler-Lagrange equations of (3.1), leading to the *consistently modified Lagrangian functional*

$$L_m(u, \mu) = J(u) + \langle \mu, \Lambda(u) \rangle + \rho \|\delta J(u)\|^2 \quad (3.6)$$

and a *Galerkin least-squares* method. In (3.6), $\delta J(\cdot)$ denotes the first variation of the functional $J(\cdot)$. Another possibility is to use both $\delta J(\cdot)$ and its adjoint $\delta J(\cdot)^*$. Then we have the consistent modification

$$L_m(u, \mu) = J(u) + \langle \mu, \Lambda(u) \rangle + \rho (\delta J(u), \delta J(u)^*) \quad (3.7)$$

Alternatively, one can add the residuals to the Lagrange multiplier term, leading to another consistently modified Lagrangian functional

$$L_c(u, \mu) = J(u) + \langle \mu, \Lambda(u) + \delta J(u) \rangle \quad (3.8)$$

and a *stabilized Galerkin method*. Both (3.6) and (3.8) are *consistent* modification of the functional $J(u)$, i.e., optimization with respect these functionals yield solutions of the given problem (3.1).

¹On the other hand, at least formally, optimization with respect to the functional (3.2) does yields a solution of (3.1).

In the next few sections we examine several examples of modified variational principles and their associated finite element methods. As a model problem we use the familiar Stokes equations (2.17) and the optimization problem (2.12). After a brief discussion of the classical penalty formulation we turn attention to several examples of *consistently* modified variational principles. The interested reader can find more details about the methods and other related issues in [18, 28, 29, 20, 38] for penalty methods; [41, 26, 34, 32, 33, 25] for Galerkin least-squares and stabilized Galerkin methods; and in [21] for augmented Lagrangian methods.

3.1.1 The penalty method

The *penalty* method for the Stokes equations (see [38]) is to minimize the *penalized* energy functional

$$J_\varepsilon(\mathbf{u}, \mathbf{f}) = \int_\Omega \frac{1}{2} |\nabla \mathbf{u}|^2 - \mathbf{f} \cdot \mathbf{u} d\Omega + \frac{1}{\varepsilon} \|\nabla \cdot \mathbf{u}\|_0^2 \quad (3.9)$$

over $\mathbf{H}_0^1(\Omega)$. Note that this unconstrained optimization problem has the form (3.3). The Euler-Lagrange equations are given by (compare with the problem (2.14)!):

seek $\mathbf{u} \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla \mathbf{u} : \nabla \mathbf{v} d\Omega + \frac{1}{\varepsilon} \int_\Omega \nabla \cdot \mathbf{u} \nabla \cdot \mathbf{v} d\Omega = \int_\Omega \mathbf{f} \cdot \mathbf{v} d\Omega \quad \forall \mathbf{v} \in H_0^1(\Omega).$$

Alternatively, we could have obtained the same weak problem starting from the regularized Stokes problem

$$\begin{aligned} \Delta \mathbf{u} + \nabla p &= \mathbf{f} & \text{in } \Omega \\ \nabla \cdot \mathbf{u} &= -\varepsilon p & \text{in } \Omega, \end{aligned} \quad (3.10)$$

eliminating p using the second equation, and applying a formal Galerkin process. In the next section we will see that the same regularized problem can also be obtained starting from a penalized Lagrangian formulation!

It may come as a surprise to the reader, but the penalty formulation based on (3.9) does not really avoid the inf-sup condition (2.29) completely! Early on it has been noticed that exact integration leads to a locking effect²

²In the sense that the approximate solution starts to converge to zero as $h \mapsto 0$ even when the exact solution is different from zero.

and that the use of reduced integration can circumvent this problem. Further studies of this phenomena have revealed that (see e.g., [45], [37]) penalty formulation can be always related to a mixed formulation by virtue of an *implicitly induced* “pressure” space. The exact form of this space depends on the treatment of the penalty term. For instance, if exact integration is used this space can be identified with divergencies of functions in V^h , i.e.,

$$P^h = \{q^h = \nabla \cdot \mathbf{v}^h \mid \mathbf{v}^h \in V^h\}.$$

In any case, the pair (V^h, P^h) still must satisfy the inf-sup condition even though the pressure space is not explicitly present in the formulation.

3.1.2 Penalized and Augmented Lagrangian formulations

Instead of penalizing the original Stokes energy functional in these methods one penalizes the associated Lagrangian functional according to (3.4) and (3.5). We will see in a moment that in some cases this leads to the same regularized Stokes problem as in the previous section.

The *penalized* Lagrangian method for the Stokes problem is defined by adding the penalty term $(\varepsilon/2)\|p\|_0^2$ to (2.15). This produces the *penalized Lagrangian*

$$L_\varepsilon(\mathbf{u}, p; \mathbf{f}) = L(\mathbf{u}, p; \mathbf{f}) + \frac{\varepsilon}{2}\|p\|_0^2.$$

This functional has the form of (3.5). If we write the optimality system for the new functional, taking variation with respect to the Lagrange multiplier p gives the penalized equation

$$\int_\Omega q \nabla \cdot \mathbf{u} d\Omega + \varepsilon \int_\Omega q p d\Omega = 0 \quad \forall q \in L_0^2(\Omega).$$

This equation is weak form of the modified continuity equation in (3.10). Because it holds for all q we can conclude that

$$\nabla \cdot \mathbf{u} = -\varepsilon p \quad \text{in } \Omega.$$

Therefore, using the penalized Lagrangian leads to essentially the same formulation (3.10) as direct penalization of the Stokes energy functional by the incompressibility constraint.

Another variation of the penalized Lagrangian method is to penalize (2.15) by the gradient of the pressure leading to the penalized Lagrangian

$$L_\varepsilon(\mathbf{u}, p; \mathbf{f}) = L(\mathbf{u}, p; \mathbf{f}) + \frac{\varepsilon}{2}\|\nabla p\|_0^2.$$

This variation of the penalized Lagrangian method is equivalent to regularization of the Stokes problem by $\varepsilon\Delta p$. As in (3.10) the regularization is effected by modification of the continuity equation, leading to the *regularized Stokes* problem

$$\begin{aligned}\Delta \mathbf{u} + \nabla p &= \mathbf{f} && \text{in } \Omega \\ \nabla \cdot \mathbf{u} &= \varepsilon \Delta p && \text{in } \Omega,\end{aligned}\tag{3.11}$$

in which case it is also necessary to close the equations by adding a Neumann boundary condition on the pressure. Because the weak form of (3.11) will include ∇p , the pressure space must be continuous. This formulation cannot be directly related to a penalty method based on penalization of the Stokes energy functional.

Regularization of the Stokes problem according to (3.10) or (3.11) improves the quasi-projection associated with the saddle-point problem for (2.15) by changing the zero block in the algebraic system (2.28) to a positive definite block. The new algebraic system has the form

$$\begin{pmatrix} A & B \\ B^T & \varepsilon B \end{pmatrix} \begin{pmatrix} U^h \\ P^h \end{pmatrix} = \begin{pmatrix} F^h \\ G^h \end{pmatrix}.\tag{3.12}$$

For (3.10) B is the mass matrix of the pressure basis, while for (3.11) B is the Dirichlet matrix of this basis (this matrix is positive definite provided the zero mean constraint (2.18) is satisfied by the pressure.) Therefore, the advantage of (3.12) over (2.28) is that now we have to solve a symmetric and positive definite algebraic system instead of an indefinite problem.

The *augmented* Lagrangian method results from changing (2.15) according to (3.4). In other words, instead of penalizing $L(\mathbf{u}, p; \mathbf{f})$ by the norm of the Lagrange multiplier p we now penalize this functional by the norm of the constraint. The augmented Lagrangian for the Stokes problem is, therefore, given by

$$L_\varepsilon(\mathbf{u}, p; \mathbf{f}) = L(\mathbf{u}, p; \mathbf{f}) + \frac{\varepsilon}{2} \|\nabla \cdot \mathbf{u}\|_0^2.$$

For further details regarding these methods we refer to [21] and [5].

3.1.3 Consistent stabilization

The idea of *consistent* stabilization is to effect the stabilization by means of terms that vanish on the exact solution. The modification is carried in a manner which introduces the desired terms to the variational equation. As a result, consistency is achieved thanks to the fact that the modified variational equation is always satisfied by the exact solution. These methods, widely known as Galerkin-Least-squares, or stabilized Galerkin were introduced in [41], and studied in [26], [33]-[34], among others.

The method of Hughes, Franca and Balestra

From (3.12) we saw that regularization of the Stokes problem improves the quasi-projection by adding a positive-definite term to the mixed algebraic problem (2.28). Because regularization directly adds the desired pressure term to the equations it is always accompanied by a penalty error proportional to ε . The idea of consistent stabilization is to add the pressure term by including it in an expression that always vanishes on the exact solution.

An obvious candidate for this task is the residual of the momentum equation which contains the desired term ∇p . However, this residual also contains the second order term $-\Delta \mathbf{u}$ which is not meaningful for standard, C^0 finite element spaces. The solution is to introduce the stabilizing term separately on each element (unless of course one is willing to consider continuously differentiable velocity approximations). Thus, one possibility, considered in [41], is to change the discrete continuity equation (2.27) to

$$b(\mathbf{u}^h, q^h) + \alpha \sum_{\mathcal{K}} h_{\mathcal{K}}^2 (-\Delta \mathbf{u}^h + \nabla p^h - f, \nabla q^h)_0 = 0. \quad (3.13)$$

This modification introduces the stabilizing term $(\nabla p^h, \nabla q^h)$ which gives the same block in the linear system as the penalty method based on (3.11), but without the penalty error. However, as with (3.11), the pressure space must contain at least first degree polynomials because otherwise the stabilizing term will not give any contribution to the matrix. A more subtle issue is the space for the velocity: if \mathbf{u} is approximated by piecewise linear finite elements the term $\Delta \mathbf{u}^h$ does not contribute to the matrix and consistency is lost! This problem can be avoided either by using higher degree polynomials for the velocity, or by using a projection of the second order term; see [43].

Let us now show rigorously that (3.13) does indeed give a coercive bilinear form. Although it is possible to look for a suitable interpretation of (3.13) in terms of bilinear forms in Sobolev spaces, it is easier to work directly with the discrete equations. For this purpose we introduce a mesh dependent norm

$$|||(\mathbf{u}^h, p^h)||| = \left(\|\mathbf{u}^h\|_1^2 + \sum_{\mathcal{K} \in \mathcal{T}_h} h_{\mathcal{K}}^2 \|\nabla p\|_{0,\mathcal{K}}^2 \right)^{1/2} \quad (3.14)$$

and a mesh dependent bilinear form

$$\begin{aligned} Q(\{\mathbf{u}^h, p^h\}; \{\mathbf{v}^h, q^h\}) &= a(\mathbf{u}^h, \mathbf{v}^h) + b(p^h, \mathbf{v}^h) - b(q^h, \mathbf{u}^h) \\ &+ \alpha \sum_{\mathcal{K} \in \mathcal{T}_h} h_{\mathcal{K}}^2 (-\Delta \mathbf{u}^h + \nabla p^h, \nabla q^h)_{0,\mathcal{K}}. \end{aligned} \quad (3.15)$$

We will show that form (3.15) is coercive in (3.14) on $V^h \times S^h$. Indeed, using Poincaré's inequality (2.7) for $a(\mathbf{u}, \mathbf{u}) = |\mathbf{u}|_1^2$ and the inverse inequality (see [3]) for the second order term

$$\begin{aligned}
Q(\{\mathbf{u}^h, p^h\}; \{\mathbf{u}^h, p^h\}) &= a(\mathbf{u}^h, \mathbf{u}^h) + \\
&\quad \alpha \sum_{\mathcal{K} \in \mathcal{T}_h} h_{\mathcal{K}}^2 \left((-\Delta \mathbf{u}^h, \nabla p^h)_{0, \mathcal{K}} + (\nabla p^h, \nabla p^h)_{0, \mathcal{K}} \right) \\
&\geq C_P \|\mathbf{u}^h\|_1^2 + \alpha \sum_{\mathcal{K} \in \mathcal{T}_h} h_{\mathcal{K}}^2 \left(\|\nabla p^h\|_{0, \mathcal{K}}^2 - \|\Delta \mathbf{u}^h\|_{0, \mathcal{K}} \|\nabla p^h\|_{0, \mathcal{K}} \right) \\
&\geq C_P \|\mathbf{u}^h\|_1^2 + \alpha \sum_{\mathcal{K} \in \mathcal{T}_h} h_{\mathcal{K}}^2 \left(\|\nabla p^h\|_{0, \mathcal{K}}^2 - C_i h^{-1} \|\nabla \mathbf{u}^h\|_{0, \mathcal{K}} \|\nabla p^h\|_{0, \mathcal{K}} \right).
\end{aligned}$$

From the ϵ -inequality

$$C_i h^{-1} \|\nabla \mathbf{u}^h\|_{0, \mathcal{K}} \|\nabla p^h\|_{0, \mathcal{K}} \leq 2C_i h^{-2} \|\nabla \mathbf{u}^h\|_{0, \mathcal{K}}^2 + \frac{1}{2} \|\nabla p^h\|_{0, \mathcal{K}}^2$$

which gives bound for the mesh-dependent term:

$$\begin{aligned}
&\alpha \sum_{\mathcal{K} \in \mathcal{T}_h} h_{\mathcal{K}}^2 \left(\|\nabla p^h\|_{0, \mathcal{K}}^2 - C_i h^{-1} \|\nabla \mathbf{u}^h\|_{0, \mathcal{K}} \|\nabla p^h\|_{0, \mathcal{K}} \right) \\
&\geq \frac{\alpha}{2} \sum_{\mathcal{K} \in \mathcal{T}_h} \left(\frac{h^2}{2} \|\nabla p^h\|_{0, \mathcal{K}}^2 - 2C_i \|\nabla \mathbf{u}^h\|_{0, \mathcal{K}}^2 \right) \\
&= \frac{\alpha}{2} \sum_{\mathcal{K} \in \mathcal{T}_h} h^2 \|\nabla p^h\|_{0, \mathcal{K}}^2 - 2\alpha C_i \|\nabla \mathbf{u}^h\|_0^2.
\end{aligned}$$

As a result,

$$Q(\{\mathbf{u}^h, p^h\}; \{\mathbf{u}^h, p^h\}) \geq (C_P - 2\alpha C_i) \|\mathbf{u}^h\|_1^2 + \frac{\alpha}{2} \sum_{\mathcal{K} \in \mathcal{T}_h} h^2 \|\nabla p^h\|_{0, \mathcal{K}}^2.$$

The choice of the parameter α is very important for proper stabilization. First, note that a very small α will effectively reduce the stabilized formulation to the usual mixed Galerkin method. At the same time α cannot be chosen too large because then the term $(C_P - 2\alpha C_i)$ will become negative! In fact, even such “innocent” looking value as $\alpha = 1$ has been found to be “destabilizing” for some regions. Looking back at the coefficient of the velocity norm it seems reasonable to choose α so that

$$\frac{C_P}{2C_i} > \alpha > 0.$$

The problem is that both C_P (the Poincare constant) and C_i (the inverse inequality constant) are hard to find in general. This is especially true when triangulations are unstructured and involve elements of different sizes and aspect ratios. One case when C_i is known is for square elements and Q_2 spaces. Then its value equals $270/11$; see [41].

Galerkin-Least Squares method of Franca and Frey

Galerkin-Least squares (GLS) stabilization is the next logical step from the consistent stabilization method of [41]. It is based again on adding a properly weighted term which contains the residual of the momentum equation in (2.17), but now this term is of *least-squares type*; see [34]. The second order velocity derivative in the momentum equation makes it necessary again to add stabilizing terms on an element by element basis and the the modified discrete continuity equation now takes the form

$$b(\mathbf{u}^h, q^h) + \alpha \sum_{\mathcal{K}} h_{\mathcal{K}}^2 (-\Delta \mathbf{u}^h + \nabla p^h - f, -\Delta \mathbf{v}^h + \nabla q^h)_0 = 0. \quad (3.16)$$

The name “least-squares” can be explained as follows. If the Lagrange functional for the Stokes problem is penalized by the square of the L^2 -norm residual of the momentum equation

$$\frac{\alpha}{2} \| -\Delta \mathbf{u} + \nabla p - f \|_0^2$$

then the first variation of the penalized functional will include the terms

$$(-\Delta \mathbf{u} + \nabla p - f, -\Delta \mathbf{v} + \nabla q)_0.$$

This is precisely the situation described by the abstract setting of (3.6). The coercivity bound for GLS can be established using the same techniques as in the previous method, and it again depends on the choice of α :

$$Q(\{\mathbf{u}^h, p^h\}; \{\mathbf{u}^h, p^h\}) \geq (C_P - 2\alpha C_i) \|\mathbf{u}^h\|_1^2 + \frac{\alpha}{2} \sum_{\mathcal{K} \in \mathcal{T}_h} h^2 \|\nabla p^h\|_{0,\mathcal{K}}^2.$$

Thus, effecting stabilization through GLS encounters the same difficulties as the method of [41] - parameter α depends on the values of Poincare and inverse inequality constants. To see why this also happens in the Galerkin Least-Squares setting, consider the mesh dependent term

$$\alpha \sum_{\mathcal{K}} h_{\mathcal{K}}^2 \| -\Delta \mathbf{u}^h + \nabla p^h \|_{0,\mathcal{K}}^2$$

that appears in GLS form $Q(\{\mathbf{u}^h, p^h\}; \{\mathbf{u}^h, p^h\})$. To show coercivity this term is bounded from below by

$$\alpha \sum_{\mathcal{K}} h_{\mathcal{K}}^2 \left(\|\nabla p^h\|_{0,\mathcal{K}}^2 - \|\Delta \mathbf{u}^h\|_{0,\mathcal{K}}^2 \right)$$

and $\|\Delta \mathbf{u}^h\|_{0,\mathcal{K}}^2$ is converted to a first-order term using the inverse inequality. This necessarily introduces the constant C_i into the coercivity bound.

The method of Douglas and Wang

This method, introduced in [32], is very similar to the GLS method of [34], but it cannot be linked directly to addition of a least-squares type term to the Lagrangian functional (2.15). The modified discrete continuity equation for Douglas-Wang stabilization is

$$b(\mathbf{u}^h, q^h) + \alpha \sum_{\mathcal{K}} h_{\mathcal{K}}^2 (-\Delta \mathbf{u}^h + \nabla p^h - f, \Delta \mathbf{v}^h + \nabla q^h)_0 = 0. \quad (3.17)$$

The seemingly minor change of the sign in front of the second order term for the test function allows to derive coercivity bound which is independent of the parameter α :

$$Q(\{\mathbf{u}^h, p^h\}; \{\mathbf{u}^h, p^h\}) \geq C_P \|\mathbf{u}^h\|_1^2 + \alpha C \sum_{\mathcal{K} \in \mathcal{T}_h} h^2 \|\nabla p^h\|_{0,\mathcal{K}}^2.$$

As a result, this method is stable for any positive value of α . This method can be interpreted as using the adjoint operator \mathcal{L}^* to effect the stabilization, i.e., it has the form (3.7). Again, the actual implementation depends on the order of the finite element space used for the velocity.

3.2 Modification of problems without optimization principles

For differential equation problems not related to minimization principles such as (3.1), the weak formulation

$$Q_g(u; v) = F(v) \quad \forall v \in V \quad (3.18)$$

is not an optimality system; instead, it is a formal statement of residual orthogonalization. Modifications are now effected directly to (3.18). Adding a small dissipative term yields the modified weak problem

$$Q_g(u; v) + \varepsilon(D(u), D(v)) = F(v) \quad \forall v \in V \quad (3.19)$$

and *artificial diffusion methods*. In (3.19), ε denotes an artificial diffusivity coefficient and $D(\cdot)$ denotes a differential operator. Similar to penalty methods, (3.19) leads to inconsistencies in the sense that its solutions are not, in general, solutions of (3.18). Consistency errors can be avoided if one uses equations residuals $R(u)$ in the modified problem

$$Q_g(u; v) + (R(u), W(v)) = F(v) \quad \forall v \in V.$$

If the test function $W(\cdot)$ is the same as $R(\cdot)$, one is led to *Galerkin least-squares methods*; if $W(\cdot)$ is different, one can be led to a class of *upwinding methods*. Modification of the test function in (3.18)

$$Q_g(u; R(v)) = F(v) \quad \forall v \in V$$

lead to *Petrov-Galerkin methods* which are another class of *upwinding methods*.

In many cases, exactly the same methods can be derived by direct modification of the differential equations or direct modification of a corresponding Galerkin weak form (3.18). If an optimization principle such as (3.1) is available, the same methods can often be also derived through modification of the functional $J(\cdot)$. The first approach is the least revealing and the last the most with respect to the fundamental role played by variational principles. One should also note that two modifications that appear different may lead to the same method and a single modification can give rise to different methods depending on the choices made for the function spaces, norms, etc.

3.2.1 Artificial diffusion and SUPG

Below we consider two examples of modified formulations for the reduced problem

$$\mathbf{b} \cdot \nabla \phi + c\phi = f \quad \text{in } \Omega \quad \text{and} \quad \phi = 0 \quad \text{on } \Gamma_-. \quad (3.20)$$

In (3.20) the symbol Γ_- is used to denote the inflow portion of the boundary. We refer the reader to [39, 40, 24, 30, 44, 31] for more details about the resulting *upwind* schemes.

Application of the Galerkin method to (3.20) gives the weak equation

$$\int_{\Omega} (\psi \mathbf{b} \cdot \nabla \phi + c\phi \psi) d\Omega = \int_{\Omega} f \psi d\Omega \quad \forall \psi \in H^1(\Omega); \quad \psi = 0 \quad \text{on } \Gamma_-. \quad (3.21)$$

The artificial diffusion method for (3.20) modifies (3.21) to

$$\varepsilon \int_{\Omega} \nabla \phi \cdot \nabla \psi d\Omega + \int_{\Omega} (\psi \mathbf{b} \cdot \nabla \phi + c\phi \psi) d\Omega = \int_{\Omega} f \psi d\Omega \quad (3.22)$$

while the consistent SUPG method (see [39, 44]) employs the weak problem

$$\int_{\Omega} h(\mathbf{b} \cdot \nabla \phi + c\phi - f)(\mathbf{b} \nabla \cdot \psi) d\Omega + \int_{\Omega} (\psi \mathbf{b} \cdot \nabla \phi + c\phi \psi) d\Omega = \int_{\Omega} f\psi d\Omega. \quad (3.23)$$

3.3 Modified variational principles: concluding remarks

Each of the mixed-Galerkin, stabilized Galerkin, penalty, and augmented Lagrangian class of methods have their adherents and are used in practice; none, however, have gained universal popularity. Part of the problem is that the success of these methods often critically depends on various mesh-dependent calibration parameters that must be fine tuned from application to application. The purpose of these parameters is to adjust the relative importance between the original variational principle and the modification term. Often, the best possible value of the parameter cannot be determined in a constructive manner, leading to under/over stabilization or even loss of stabilization; see, e.g., [34]. The analysis of many of these methods also remains an open problem for important nonlinear equations such as the Navier-Stokes equations.

Chapter 4

Least-squares methods: first examples

In this chapter we take a first look at some possible answers to the following question:

for any given partial differential equation problem, is it possible to define a sensible convex, unconstrained minimization principle if one is not already available, so that a finite element method can be developed in a Rayleigh-Ritz-like setting?

Given the attractive computational and analytic advantages of true inner product projections, this question seems very logical. Obviously, to answer this question we cannot use the methods discussed in §2.2, §2.3, and Chapter 3. In §2.2, a saddle-point variational principle was introduced from the very beginning as a way of dealing with the constraints. In §2.3, it was demonstrated that the formal Galerkin method leads to weak problems whose features are always inextricably tied to those of the partial differential equation problem. In Chapter 3, we saw that modifications of the natural variational principle can recover some but not all of the desirable features of the Rayleigh-Ritz setting.

Modern least-squares finite element methods are a methodology that answers this question in a positive way through a variational framework based on the idea of *residual minimization*. This idea is as universal as the idea of *residual orthogonalization* which is the basis of the Galerkin method and so it can be applied to virtually any PDE problem. However, unlike the residual orthogonalization, when properly executed, *residual minimization* has the potential to define inner product projections even if the original problem is not at all associated with optimization.

The central premise underlying least-squares principles is the interpretation of a selected measure of the residual as an “energy” that must be minimized, with the exact solution being the one having zero energy. From this perspective, an appropriate least-squares “energy” functional can be set up immediately by summing up the squares of the equation residuals, each one measured in some suitable norm. The resulting energy functional more often than not has no physical meaning, but it offers the advantage of transforming the partial differential problem into an equivalent convex, unconstrained minimization problem.

In order to fully emulate the Rayleigh-Ritz setting it is critical to define a least-squares functional that is also *norm-equivalent* in some Hilbert space. Then, least-squares variational principles fit into the attractive category of orthogonal projections in Hilbert spaces with respect to problem-dependent inner products. Once the partial differential equation problem is recast into such a variational framework, stability prerequisites such as inf-sup conditions are no longer needed for the well-posedness of the weak problem. Let us now try to apply these ideas to some of the examples from §2.1–§2.3.

4.1 Poisson equation

Let us begin with the Poisson problem (2.6) and ignore the fact that for this problem there already exist a convex energy functional (2.1) and unconstrained optimization problem (2.2). We will proceed directly with the PDE (2.6). In order to point out another advantage of least-squares methods, we will generalize (2.6) to include the inhomogeneous boundary condition $\phi = g$ on Γ . Thus, there are two residuals: the differential equation residual

$$-\Delta\phi - f$$

and the boundary condition residual

$$\phi - g.$$

To define an “energy” functional based on these two residuals. we choose the simplest L^2 -norm:

$$J(\phi; f, g) = \|\Delta\phi + f\|_0^2 + \|\phi - g\|_{0,\Gamma}^2. \quad (4.1)$$

This convex, quadratic functional is minimized by the exact solution,¹ i.e., by ϕ such that $-\Delta\phi = f$ in Ω and $\phi = g$ on Γ . Then, we set up a least-squares minimization principle

¹To be precise, the exact solution must be sufficiently smooth because otherwise the term $\Delta\phi$ will not be square integrable.

seek ϕ in a suitable space X such that $J(\phi; f, g) \leq J(\psi; f, g)$ for all $\psi \in X$.

Next, using standard techniques from the calculus of variations, it is easy to see that all minimizers of (4.1) must satisfy the optimality system

seek $\phi \in X$ such that

$$\begin{aligned} \int_{\Omega} \Delta\phi \Delta\psi \, d\Omega + \int_{\Gamma} \phi\psi \, d\Gamma \\ = - \int_{\Omega} f \Delta\psi \, d\Omega + \int_{\Gamma} g\psi \, d\Gamma \quad \forall \psi \in X. \end{aligned} \quad (4.2)$$

The final steps are to choose a trial space $X^h \subset X$ and then restrict (4.2) to X^h to obtain²

seek $\phi^h \in X^h$ such that

$$\begin{aligned} \int_{\Omega} \Delta\phi^h \Delta\psi^h \, d\Omega + \int_{\Gamma} \phi^h\psi^h \, d\Gamma \\ = - \int_{\Omega} f \Delta\psi^h \, d\Omega + \int_{\Gamma} g\psi^h \, d\Gamma \quad \forall \psi^h \in X^h. \end{aligned} \quad (4.3)$$

This is simply a linear algebraic system.

Using integration by parts, it is easy to see that smooth solutions of (4.2) satisfy the biharmonic boundary value problem

$$-\Delta\Delta\phi = \Delta f \quad \text{in } \Omega \quad (4.4)$$

and

$$-\Delta\phi = f \quad \text{and} \quad \frac{\partial(\Delta\phi + f)}{\partial n} - (\phi - g) = 0 \quad \text{on } \Gamma. \quad (4.5)$$

Therefore, smooth solutions of (4.2) satisfy a differentiated form of that problem. Equivalently, the minimization of the least-squares functional (4.1) corresponds to the solving the biharmonic problem (4.4) and (4.5). Of course, solutions of the latter are solutions of the Poisson problem.

4.2 Stokes equations

Consider now the Stokes equations (2.17). For this problem there's no "natural" unconstrained, convex, quadratic minimization problem; we only have

²The system (4.3) can also be derived by directly minimizing the functional (4.1) over the finite element subspace X^h .

the constrained optimization problem (2.12). However, we can define an “artificial” energy functional by minimizing the sum of the squares of the L^2 -norms of the equation residuals, i.e.,

$$J(\mathbf{u}, p; \mathbf{f}, \mathbf{g}) = \| -\Delta \mathbf{u} + \nabla p - \mathbf{f} \|_0^2 + \| \nabla \cdot \mathbf{u} \|_0^2 + \| \mathbf{u} - \mathbf{g} \|_{0,\Gamma}^2. \quad (4.6)$$

Then, the optimality system corresponding to the minimization of this functional is given by

$$\begin{aligned} \int_{\Omega} (-\Delta \mathbf{u} + \nabla p) \cdot (-\Delta \mathbf{v} + \nabla q) d\Omega + \int_{\Omega} (\nabla \cdot \mathbf{u})(\nabla \cdot \mathbf{v}) d\Omega \\ + \int_{\Gamma} \mathbf{u} \cdot \mathbf{v} d\Gamma = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} d\Omega + \int_{\Gamma} \mathbf{g} \cdot \mathbf{v} d\Gamma, \end{aligned} \quad (4.7)$$

where \mathbf{u} and p belong to appropriate (unconstrained) function spaces and where \mathbf{v} and q are arbitrary in those function spaces. We can then define a discrete problem by either restricting (4.7) to appropriate finite element subspaces for the velocity and pressure or, equivalently, by minimizing the functional (4.6) with respect to those approximating spaces. Note that smooth solutions of (4.7), or equivalently, smooth minimizers of (4.6), are not directly solutions of the Stokes equations, but instead are solutions of an equivalent system of partial differential equations that may be determined from the Stokes equations through differentiations and linear combinations. The order of that system is higher than that for the Stokes equations, e.g., the equations include terms such as $\Delta \Delta \mathbf{u}$ and Δp .

4.3 PDE's without optimization principles

Least-squares principle can be applied to problems for which no natural minimization principle, either constrained or unconstrained, exists. For example, for the Helmholtz problem (2.30), we can define the functional

$$J(\phi; f, g) = \| \Delta \phi + k^2 \phi + f \|_0^2 + \| \phi - g \|_{0,\Gamma}^2 \quad (4.8)$$

and then proceed as in the Poisson case to derive, instead of (4.2), the weak formulation

seek $\phi \in X$ such that

$$\begin{aligned} \int_{\Omega} (\Delta \phi + k^2 \phi)(\Delta \psi + k^2 \psi) d\Omega + \int_{\Gamma} \phi \psi d\Gamma \\ = - \int_{\Omega} f(\Delta \psi + k^2 \psi) d\Omega + \int_{\Gamma} g \psi d\Gamma \quad \forall \psi \in X. \end{aligned} \quad (4.9)$$

Another example is provided by the convection-diffusion problem (2.32) for which we can define the functional

$$J(\phi; f, g) = \| -\Delta\phi + \mathbf{b} \cdot \nabla\phi + f \|_0^2 + \|\phi - g\|_{0,\Gamma}^2 \quad (4.10)$$

and then derive the weak formulation

seek $\phi \in X$ such that

$$\begin{aligned} & \int_{\Omega} (-\Delta\phi + \mathbf{b} \cdot \nabla\phi)(-\Delta\psi + \mathbf{b} \cdot \nabla\psi) d\Omega + \int_{\Gamma} \phi\psi d\Gamma \\ & = - \int_{\Omega} f(-\Delta\psi + \mathbf{b} \cdot \nabla\psi) d\Omega + \int_{\Gamma} g\psi d\Gamma \quad \forall \psi \in X. \end{aligned} \quad (4.11)$$

4.4 A critical look

The variational equations, i.e., weak formulations, derived from least-squares principles all have the form

seek U in some suitable function space X such that

$$Q(U; V) = F(V) \quad \forall V \in X, \quad (4.12)$$

where U denotes the relevant set of dependent variables, $Q(\cdot; \cdot)$ is a symmetric bilinear form, and $F(\cdot)$ is a linear functional. In contrast to the weak problems of §2.1–§2.3:

- the bilinear forms in the least-squares weak formulations are all symmetric;
- in all cases the bilinear forms may possibly be coercive;
- it is now possible to obtain positive definite discrete algebraic systems in all cases.

In general, positive definiteness³ is a consequence of the *norm-equivalence* of the least-squares functional and here we have not yet established that any of the functionals introduced in this section are norm equivalent, i.e., that the expressions

$$J(\phi; 0, 0) = \|\Delta\phi\|_0^2 + \|\phi\|_{0,\Gamma}^2$$

for the Poisson equation,

$$J(\mathbf{u}, p; \mathbf{0}, \mathbf{0}) = \| -\Delta\mathbf{u} + \nabla p \|_0^2 + \|\nabla \cdot \mathbf{u}\|_0^2 + \|\mathbf{u}\|_{0,\Gamma}^2$$

³Positive semi-definiteness is obvious.

for the Stokes equations,

$$J(\phi; 0, 0) = \|\Delta\phi + k^2\phi\|_0^2 + \|\phi\|_{0,\Gamma}^2$$

for the Helmholtz equation, and

$$J(\phi; 0, 0) = \|\Delta\phi + \mathbf{b} \cdot \nabla\phi\|_0^2 + \|\phi\|_{0,\Gamma}^2$$

for the convection-diffusion equation define equivalent norms on the Hilbert spaces over which the respective least-squares functionals are minimized. It turns out that this issue is essentially equivalent to the well-posedness of the boundary value problem in some function spaces.

While mathematical well-posedness is important we should not forget that the ultimate goal is to devise a good computational algorithm. Therefore, the methods must also be practical. This is a rather subjective characteristic, but if we want to be competitive with existing methods it is desirable that

- the matrices and right-hand sides of the discrete problem should be “easily” computable,
- discretization should be accomplished using standard, “easy to use” finite element spaces
- discrete problem should have a “manageable” condition number.

Let us see if the methods devised so far meet our criteria for practicality. First, all four variational equations include terms such as either

$$\int_{\Omega} \Delta\phi\Delta\psi \, d\Omega \quad \text{or} \quad \int_{\Omega} \Delta\mathbf{u} \cdot \Delta\mathbf{v} \, d\Omega.$$

and the corresponding discrete equations include terms such as either

$$\int_{\Omega} \Delta\phi^h\Delta\psi^h \, d\Omega \quad \text{or} \quad \int_{\Omega} \Delta\mathbf{u}^h \cdot \Delta\mathbf{v}^h \, d\Omega.$$

Recall that finite element spaces consist of piecewise polynomial functions. Therefore, each term is well-defined within an element. The problem is that these terms will not be well-defined across element boundaries unless the finite element spaces are continuously differentiable. In more than one dimension such spaces are hardly practical. As a result, any method that uses such terms, including the methods introduced here, is impractical. A further observation is that the condition numbers of the discrete problems

associated with these methods, even if we use smooth finite element spaces, are $O(h^{-4})$. This should be contrasted with, e.g., the Rayleigh-Ritz finite element method for the Poisson equation for which the condition number of the discrete problem is $O(h^{-2})$. Therefore, the least-squares finite element methods discussed so far fail the third practicality criterion as well. Another observation is that weak solutions are now required to possess two square integrable derivatives as opposed to only one in Galerkin methods. Early examples of least-squares finite element methods shared these practical disadvantages and for these reasons they did not, at first, gain popularity.

These observations indicate that development of a practical and mathematically solid least-squares method requires more than merely choosing the most obvious least-squares functional. This should not come as a surprise if we recall that

least-squares functionals are not necessarily physical quantities, i.e., unlike an energy minimization principle derived from physical laws, a least-squares principle can be set up in many different ways!

In particular, some of these ways may turn out to be less than useful. We will see that this ambiguity is in actuality an asset as it allows us to better “fine tune” the least-squares method to the problem in hand.

Let us now introduce some of the techniques that have been developed over the years and that can be used to obtain practical least-squares methods. A simple, yet effective method of eliminating high-order derivatives is to rewrite the equations as an equivalent first-order system⁴. For the Poisson problem, instead of working with the functional (4.1), we consider an alternative one given by

$$J(\phi, \mathbf{u}; f, g) = \|\nabla \cdot \mathbf{u} - f\|_0^2 + \|\nabla \phi - \mathbf{u}\|_0^2 + \|\phi - g\|_0^2. \quad (4.13)$$

This functional is based on the equivalent first-order system (2.21) with an inhomogeneous boundary condition. Minimization of this functional results in a least-squares variational problem of the form (4.12), but now with

$$Q(U; V) = \int_{\Omega} (\nabla \cdot \mathbf{u})(\nabla \cdot \mathbf{v}) \, d\Omega + \int_{\Omega} (\nabla \phi - \mathbf{u}) \cdot (\nabla \psi - \mathbf{v}) \, d\Omega + \int_{\Gamma} \phi \psi \, d\Gamma$$

and

$$F(V) = \int_{\Omega} f \nabla \cdot \mathbf{v} \, d\Omega + \int_{\Gamma} g \psi \, d\Gamma,$$

⁴This can be done in many ways, so in a sense using first-order formulations increases the level of “ambiguity”. However, as already mentioned, this is in fact a flexibility of the approach instead.

where $U = (\phi, \mathbf{u})$ and $V = (\psi, \mathbf{v})$. The idea of using equivalent first-order formulations of second-order problems is reminiscent of the mixed-Galerkin methods of §2.2. However, now we can choose *any pair* of finite element spaces for approximating ϕ and \mathbf{u} since, unlike the mixed-Galerkin case, we are not required to satisfy an inf-sup stability condition. The first-order system based least-squares formulation also results in algebraic systems having condition numbers much the same as that for Galerkin methods. Thus, if we compare the two least-squares methods for the Poisson equation, i.e., one based on the functional (4.1), the other on (4.13), it is clear that the second one is superior and more likely to be competitive with, e.g., the mixed-Galerkin method.

The next question is that of norm-equivalence, i.e., whether

$$J(\phi, \mathbf{u}; 0, 0) = \|\nabla \cdot \mathbf{u}\|_0^2 + \|\nabla\phi - \mathbf{u}\|_0^2 + \|\phi\|_{0,\Gamma}^2$$

defines a norm on a suitable Hilbert space. If (4.13) were norm-equivalent, the resulting least-squares method would fit nicely in the same framework as that for the Rayleigh-Ritz problem: existence and uniqueness of solutions along with quasi-optimality of the finite element approximations are guaranteed for any conforming discretization of the weak problem. Unfortunately, (4.13) does not have this property. A norm-equivalent functional for the first-order system (2.21) is

$$J(\phi, \mathbf{u}; f, g) = \|\nabla \cdot \mathbf{u} - f\|_0^2 + \|\nabla\phi - \mathbf{u}\|_0^2 + \|\phi - g\|_{1/2,\Gamma}^2, \quad (4.14)$$

where the boundary residual is measured in a *fractional order* trace norm. The new obstacle here is the conflict between norm-equivalence and practicality: in order to achieve norm-equivalence, we had to include the trace norm in the functional; unfortunately, this norm is difficult to compute. This problem cannot be avoided by changing the formulation since boundary terms necessarily require fractional norms regardless of the order of the differential operator. The easiest remedy is simply to drop the boundary residual and enforce the boundary condition on the trial space. Another remedy is to replace the fractional norm by a *mesh-dependent* weighted L^2 -norm:

$$J(\phi, \mathbf{u}; f, g) = \|\nabla \cdot \mathbf{u} - f\|_0^2 + \|\nabla\phi - \mathbf{u}\|_0^2 + h^{-1}\|\phi - g\|_{0,\Gamma}^2. \quad (4.15)$$

In contrast to the functional (4.14), this *weighted* functional is not norm-equivalent on the same Hilbert space, but it has properties that resemble norm-equivalence when restricted to a finite element space.

The conflict between norm-equivalence and practicality is not necessarily caused by boundary residual terms. For example, assuming that boundary conditions are satisfied exactly,

$$J(\phi, \mathbf{u}; f) = \|\nabla \cdot \mathbf{u} - f\|_{-1}^2 + \|\nabla\phi - \mathbf{u}\|_0^2 \quad (4.16)$$

is another norm-equivalent functional for the first-order Poisson problem (2.21). This functional is no more practical than (4.14) because the *negative order* norm $\|\cdot\|_{-1}$ is again not easily computable. To get a practical functional, this norm must be replaced by some computable equivalent. One approach is to use a scaling argument and replace (4.16) by the weighted functional

$$J(\phi, \mathbf{u}; f) = h^2 \|\nabla \cdot \mathbf{u} - f\|_0^2 + \|\nabla\phi - \mathbf{u}\|_0^2. \quad (4.17)$$

Another approach is to consider a more sophisticated replacement for (4.16) which uses a *discrete negative norm* defined by means of preconditioners for the Poisson equation.

4.4.1 Some questions and answers

The basic components of a least-squares method can be summarized as follows:

- a (quadratic, convex) least-squares functional that measures the size of the equation residuals in appropriate norms;
- a minimization principle for the least-squares functional;
- a discretization step in which one minimizes the functional over a finite element trial space.

Obviously, this methodology can be applied to any given PDE. Therefore, the first question is:

- When is the least-squares approach justified?

We also saw that there are many freedoms in the way this methodology can be applied to a given PDE. Therefore, another question is:

- How to quantify the best possible least-squares setting for a given PDE?

The answer to the first question is quite obvious: attractiveness of least-squares depends on the type of quasi-projection that can be associated with the Galerkin method. In particular, the appeal of a least-squares method increases with the deviation of the naturally occurring variational setting from the Rayleigh-Ritz principle.

The answer to the second question is not hard too: since we wish to simulate a Rayleigh-Ritz setting the variational equation must correspond to a true inner product projection. This is the same as to say that the least-squares functional must be norm equivalent.

Having found answers to these two questions we see that another one immediately arises:

- Will the “best” least-squares principle, as dictated by analyses, be also the one that is most convenient to use in practice?

Our examples show that often the answer to this question is negative – high-order derivatives, fractional norms, negative norms, all conspire to make the best functional less and less practical. Thus, we have reached the crux of the matter in least-squares development:

- How does one reconcile the “best” and the “most convenient” principles?

This question has generated a tremendous amount of research activity, among practitioners and analysts of least-squares methods. The use of equivalent first-order reformulations (often dubbed FOSLS approach) proposed in the late 70’s has become a powerful and by now a standard tool in least-squares methodologies; see [65, 67, 68, 66, 69, 70], [75, 78, 79, 80, 81, 82, 83], [88, 92, 89, 90, 91] and [98, 99, 100], among others.

This idea is often combined with other tools such as weighted norms, [46, 56, 57] and more recently, discrete negative norms [62, 63, 64] and [49, 50, 53]. The purpose of these tools is to provide the desired reconciliation between the “most-convenient” and the “best” least-squares principles. Formalization of this concept is the subject of the next chapter.

Chapter 5

Continuous and discrete least-squares principles

This chapter discusses some universal principles that are encountered in the development of least-squares methods. In particular we will introduce the notions of *continuous* and *discrete* least-squares principles. In what follows we adopt the stance that the single most important characteristic of least-squares methods is the true projection property which creates a Rayleigh-Ritz-like environment whenever one is not available naturally.

Given a PDE problem our first task will be to identify all norm-equivalent functionals that can be associated with the differential equations. In section 5.1, we show that such functionals are induced by a priori estimates for the partial differential equation problem: the data spaces suggested by the estimate provide the appropriate norms for measuring the residual “energy” while the corresponding solution spaces provide the candidate minimizers. The class of all such *Continuous Least-Squares* (CLS) principles is generated by considering all equivalent forms of the partial differential equation together with their valid a priori estimates. Therefore, a CLS principle describes

an “ideal” setting in which the balance between the “artificial” residual energy and the solution norm is mathematically correct.

As we have already seen, mathematically ideal least-squares principles are not necessarily the most practical to implement. Therefore, the next item on our agenda will be to reconcile the theoretical demands with the practicality constraints. We will refer to the outcome of this process as a *Discrete Least-Squares* (DLS) principle. A DLS principle represents

a compromise between mathematically desirable setting and practically feasible algorithm.

It is a fact of life that “practicality” is a rigid constraint so the remedy must be sought by either enlarging the class of CLS principles until it contains a satisfactory one and/or by transforming a CLS principle into a DLS one via a process that may involve sacrificing some of the Rayleigh-Ritz-like properties.

Enlarging of the CLS class is accomplished by using equivalent reformulated problems. Typically, reformulation involves reduction to first-order systems, but another approaches like the \mathcal{LL}^* method (see [70]) are also possible. As a result, one often gains additional tangible benefits such as being able to obtain direct approximations of physically relevant variables.

Transformation of CLSP to a practical DLSP is usually much more trickier, especially if a good method is desired. This process calls for lots of ingenuity and often must be carried on a case by case basis. If such transformation is necessary it is almost always accompanied by some loss of desirable mathematical structure. Fundamental properties of resulting least-squares finite element methods depend upon the degree to which the mathematical structure imposed by the CLS principle has been compromised during its transformation to DLS principle.

In the ideal case, the CLSP class contains a principle which meets all practicality constraints without any further modifications so that the DLS principle is obtained by simple restriction to finite element spaces. Clearly, this situation describes a *conforming* finite element method, where

the discrete “energy” balance of the DLS principle represents restriction to finite element spaces of a mathematically correct relation between data and solution.

If this is not possible, then the next best thing is a CLS principle with a mathematical structure that can be recreated on finite element spaces in a manner that captures the essential “energy” balance of the continuous principle and reproduces it independently of any grid-size parameters. Transformation of this CLS principle involves the construction of sophisticated discrete norms which ensure that

the discrete “energy” balance of DLS principle represents a mathematically correct relation between data and solution on finite element spaces despite not being a restriction of a CLS principle.

We call resulting DLS principle and method *norm-equivalent*. While achieving norm-equivalence may not be trivial, these principles are capable of recovering all essential advantages of a Rayleigh-Ritz scheme.

A third pattern in the transformation occurs when norm-equivalence is not an option due to, e.g., the complexity of the required norms. An alternative then would be a simpler DLS principle for which

the discrete “energy” balance represents a mathematically correct relation between data and solution only asymptotically and involves explicit dependence on grid-size parameters.

We call such DLS principles and methods *quasi-norm-equivalent*. Dependence of the energy balance on the grid-size is the price that must be paid to satisfy the practicality constraint and it may or may not have some negative effects on the resulting method.

A fourth pattern in the transformation occurs when the mathematical structure of the CLS principle is completely disregarded resulting in a *non-equivalent* DLS principle for which

the discrete “energy” balance does not represent a mathematically correct relation between data and solution.

These principles create an “energy” imbalance relation in which data norms are bounded from below and above by different solution norms.

Except for the conforming DLS principle, all other transformations commit various variational crimes against the ideal CLS principle. However, departure from the ideal, mathematically correct setting does not automatically lead to the same disastrous results as say, violation of the inf-sup condition in the mixed method. In fact, even non-equivalent methods rarely fail in an obvious manner and their solutions are quite often good. This truly remarkable feature of least-squares principles is owed to their roots in inner product projections. This makes any least-squares method, including methods with substantial deviations from the mathematically correct setting, extremely robust and considerably less susceptible to variational crimes compared to other schemes.

5.1 The abstract problem

Throughout this chapter, \mathcal{L} denotes a linear differential operator that acts on functions defined on some bounded, open region $\Omega \subset \mathcal{R}^n$ and \mathcal{R} denotes a linear operator which is applied to functions defined on the boundary Γ

of Ω . Both \mathcal{L} and \mathcal{R} may depend on the spatial variable \mathbf{x} . We consider an abstract boundary value problem

$$\mathcal{L}\mathbf{u} = \mathbf{f} \quad \text{in } \Omega \tag{5.1}$$

$$\mathcal{R}\mathbf{u} = \mathbf{g} \quad \text{on } \Gamma, \tag{5.2}$$

where \mathbf{f} and \mathbf{g} denote data functions. Concerning (5.1)–(5.2), we make the following assumptions.

A.1. There exist Hilbert spaces $X = X(\Omega)$, $Y = Y(\Omega)$, and $Z = Z(\Gamma)$ ¹ such that the mapping $\mathbf{u} \mapsto (\mathcal{L}\mathbf{u}, \mathcal{R}\mathbf{u})$ is a homeomorphism $X \mapsto Y \times Z$.

A.2. The operator $(\mathcal{L}\mathbf{u}, \mathcal{R}\mathbf{u})$ is of Fredholm type, i.e., it has a closed range and both the kernel and the co-range are finite dimensional.

These assumptions are sufficiently general to include a wide range of partial differential equation problems. For example, **A.1**–**A.2** are valid for differential operators that are elliptic in the sense of Agmon, Douglis, and Nirenberg; see [11]. We will consider least-squares methods for such PDE’s in the next chapter.

The second hypothesis allows us to disregard the case of (5.1)–(5.2) possessing multiple solutions. Indeed, if $(\mathcal{L}, \mathcal{R})$ has a nontrivial kernel, then according to **A.2**, it must be finite dimensional. Consequently, $(\mathcal{L}, \mathcal{R})$ can be augmented by a finite number of constraints² in such a way that (5.1)–(5.2) always has a unique solution.

An important consequence of **A.1**–**A.2** is the existence of two positive constants C_1 and C_2 whose values are independent of \mathbf{u} and such that

$$C_2\|\mathbf{u}\|_X \leq \|\mathcal{L}\mathbf{u}\|_Y + \|\mathcal{R}\mathbf{u}\|_Z \leq C_1\|\mathbf{u}\|_X. \tag{5.3}$$

The inequalities in (5.3) describe a relation between the solution and data of a boundary value problem that is fundamental to least-squares principles. It defines the proper balance between the solution “energy” and the residual “energy.” Note that for any given partial differential equation problem, there may exist many combinations of data and solution spaces for which the problem is well posed and, in particular, for which (5.3) holds.

¹The symbols $\|\cdot\|_X$ and $(\cdot, \cdot)_X$ will denote the norm and inner product, respectively, on the space X ; analogous notations will be used for the spaces Y and Z .

²One example is given by the zero mean constraint $\int_{\Omega} p \, d\Omega = 0$ which is added to the Stokes equations to ensure the uniqueness of the pressure. A similar constraint can be added to a pure Neumann problem for the Poisson equation which also has a one dimensional null-space consisting of all constant functions.

One example is given by differential operators which have complete sets of homeomorphisms; see [13] and [14]. For such operators, energy bounds such as (5.3) hold on scales of Hilbert spaces, i.e., collections of spaces $\mathbf{X}_q(\Omega)$, $\mathbf{Y}_q(\Omega)$, and $\mathbf{Z}_q(\Gamma)$ parametrized by an integer parameter q ; see [8] or [12]. Then, every value of q gives rise to a valid a priori estimate for the partial differential equation problem.

5.2 Continuous least-squares principles

The *continuous least-squares principle* for (5.1)–(5.2) stems directly from the solution-data balance defined by (5.3). The data spaces Y and Z provide the norms for measuring the “energy” of the residuals while the solution space X serves as a trial space for candidate minimizers of the “energy” functional. Specifically, we define the artificial, quadratic, convex least-squares “energy” functional

$$J(\mathbf{u}; \mathbf{f}, \mathbf{g}) = \frac{1}{2} \left(\|\mathcal{L}\mathbf{u} - \mathbf{f}\|_Y^2 + \|\mathcal{R}\mathbf{u} - \mathbf{g}\|_Z^2 \right), \quad (5.4)$$

and the continuous least-squares principle for the problem (5.4):

$$\text{seek } \mathbf{u} \in X \quad \text{such that} \quad J(\mathbf{u}; \mathbf{f}, \mathbf{g}) \leq J(\mathbf{v}; \mathbf{f}, \mathbf{g}) \quad \forall \mathbf{v} \in X. \quad (5.5)$$

Whenever the data is identically zero, we will simply write $J(\mathbf{u})$ instead of $J(\mathbf{u}; 0, 0)$.

Let us now show that the continuous least-squares principle (5.5) is well posed and that the unique minimizer of (5.4) coincides with the unique solution $\mathbf{u} \in X$ of (5.1)–(5.2).

Theorem 1 *Assume that A.1 and A.2 hold. Then,*

1. *the functional (5.4) is norm-equivalent in the sense that*

$$\frac{1}{4} C_2^2 \|\mathbf{u}\|_X^2 \leq J(\mathbf{u}) \leq \frac{1}{2} C_1^2 \|\mathbf{u}\|_X^2 \quad \forall \mathbf{u} \in X; \quad (5.6)$$

2. *there exists a unique minimizer $\mathbf{u} \in X$ of (5.4); moreover, the unique minimizer \mathbf{u} depends continuously on the data, i.e., \mathbf{u} satisfies*

$$\|\mathbf{u}\|_X \leq C (\|\mathbf{f}\|_Y + \|\mathbf{g}\|_Z), \quad (5.7)$$

where C is a constant whose value is independent of \mathbf{f} , \mathbf{g} , and \mathbf{u} ;

3. *\mathbf{u} is the unique minimizer of (5.4) if and only if \mathbf{u} is the unique solution of (5.1)–(5.2).*

Proof. To show 1, it suffices to note that

$$J(\mathbf{u}) = \frac{1}{2} \left(\|\mathcal{L}\mathbf{u}\|_X^2 + \|\mathcal{R}\mathbf{u}\|_Z^2 \right)$$

so that the norm-equivalence (5.6) follows from (5.3).

To prove 2, standard techniques from the calculus of variations can be used to show that all minimizers \mathbf{u} of (5.4) necessarily satisfy the Euler-Lagrange equation

$$\delta J(\mathbf{u}) = \lim_{\varepsilon \rightarrow 0} \frac{dJ(\mathbf{u} + \varepsilon \mathbf{v})}{d\varepsilon} = 0 \quad \forall \mathbf{v} \in X.$$

A simple calculation shows that this equation is identical with the variational problem

$$\text{seek } \mathbf{u} \in X \quad \text{such that} \quad Q(\mathbf{u}; \mathbf{v}) = F(\mathbf{v}) \quad \forall \mathbf{v} \in X, \quad (5.8)$$

where the bilinear form $Q(\cdot; \cdot)$ and the linear functional $F(\cdot)$ are given by

$$Q(\mathbf{u}; \mathbf{v}) = (\mathcal{L}\mathbf{u}, \mathcal{L}\mathbf{v})_Y + (\mathcal{R}\mathbf{u}, \mathcal{R}\mathbf{v})_Z \quad (5.9)$$

and

$$F(\mathbf{v}) = (\mathbf{f}, \mathcal{L}\mathbf{v})_Y + (\mathbf{g}, \mathcal{R}\mathbf{v})_Z, \quad (5.10)$$

respectively. From the lower bound in (5.3), we obtain

$$Q(\mathbf{u}; \mathbf{u}) = \|\mathcal{L}\mathbf{u}\|_X^2 + \|\mathcal{R}\mathbf{u}\|_Z^2 \geq \frac{1}{2} C_2^2 \|\mathbf{u}\|_X^2$$

while the Cauchy inequality and the upper bound in (5.3) yield

$$Q(\mathbf{u}; \mathbf{v}) \leq \left(\|\mathcal{L}\mathbf{u}\|_Y + \|\mathcal{R}\mathbf{u}\|_Z \right) \left(\|\mathcal{L}\mathbf{v}\|_Y + \|\mathcal{R}\mathbf{v}\|_Z \right) \leq C_1^2 \|\mathbf{u}\|_X \|\mathbf{v}\|_X,$$

i.e., $Q(\cdot; \cdot)$ is a continuous and coercive bilinear form on $X \times X$. Again, using Cauchy's inequality and the upper bound in (5.3), it is easy to see that

$$\begin{aligned} F(\mathbf{v}) &\leq \left(\|\mathcal{L}\mathbf{v}\|_Y + \|\mathcal{R}\mathbf{v}\|_Z \right) \left(\|\mathbf{f}\|_Y + \|\mathbf{g}\|_Z \right) \\ &\leq C_1 \|\mathbf{v}\|_X \left(\|\mathbf{f}\|_Y + \|\mathbf{g}\|_Z \right), \end{aligned}$$

i.e., $F(\mathbf{v})$ is a bounded linear functional on X and

$$\|\mathcal{F}\| \leq C_1 \left(\|\mathbf{f}\|_Y + \|\mathbf{g}\|_Z \right).$$

As a result, the existence and uniqueness of a minimizer \mathbf{u} that solves (5.8) follows from the Riesz Representation Theorem. Finally, the coercivity of the bilinear form $Q(\cdot; \cdot)$ along with the continuity of $F(\cdot)$ implies

$$\frac{1}{2}C_2^2\|\mathbf{u}\|_X^2 \leq Q(\mathbf{u}; \mathbf{u}) = F(\mathbf{u}) \leq C_1\|\mathbf{u}\|_X(\|\mathbf{f}\|_Y + \|\mathbf{g}\|_Z)$$

so that

$$\|\mathbf{u}\|_X \leq \frac{2C_1}{C_2^2}(\|\mathbf{f}\|_Y + \|\mathbf{g}\|_Z)$$

which proves (5.7).

To show the last assertion, let \mathbf{u}_1 and \mathbf{u}_2 denote the minimizer of (5.8) and a solution of (5.1)–(5.2), respectively. Since \mathbf{u}_1 is the minimizer and since \mathbf{u}_2 causes the residuals of (5.1)–(5.2) to vanish,

$$J(\mathbf{u}_1; \mathbf{f}, \mathbf{g}) \leq J(\mathbf{u}_2; \mathbf{f}, \mathbf{g}) = 0.$$

As a result,

$$\begin{aligned} C_2\|\mathbf{u}_1 - \mathbf{u}_2\|_X &\leq \|\mathcal{L}(\mathbf{u}_1 - \mathbf{u}_2)\|_Y + \|\mathcal{R}(\mathbf{u}_1 - \mathbf{u}_2)\|_Z \\ &= \|\mathcal{L}(\mathbf{u}_1) - \mathbf{f}\|_Y + \|\mathcal{R}(\mathbf{u}_1) - \mathbf{g}\|_Z = 2J(\mathbf{u}_1; \mathbf{f}, \mathbf{g})^{1/2} = 0, \end{aligned}$$

i.e., $\mathbf{u}_1 = \mathbf{u}_2$. \square

Theorem 1 describes a Rayleigh-Ritz principle, albeit for an externally defined, artificial, “energy” functional. The “energy” inner product for this principle is $Q(\cdot; \cdot)$, while $\|\mathbf{u}\|^2 = Q(\mathbf{u}; \mathbf{u}) = 2J(\mathbf{u})$ is the “energy” norm. Thus, it is clear that we have succeeded in emulating the Rayleigh-Ritz principle for any given partial differential equation problem and under very general assumptions. In other words,

we have established a mathematical framework which allows us to take an arbitrary well-posed partial differential equation problem and replace it by an equivalent, well-posed, unconstrained minimization problem. This framework is completely determined by the pair $\{X, J(\cdot; \cdot, \cdot)\}$. The set of all such pairs forms the class of continuous least-squares (CLS) principles.

The least-squares problem (5.5) is equivalent to the original equations (5.1)–(5.2) in the sense that their solutions belonging to the space X coincide – each minimizer of (5.4) solves the differential equations and vice versa. However, it is important to remember that, as a rule, the variational problem

(5.8) is not a standard, e.g., Galerkin, *weak form* of (5.1)–(5.2). For example, if \mathcal{L} is such that the Green’s formula

$$(\mathbf{u}, \mathcal{L}\mathbf{v})_{\mathbf{Y}} - \langle \mathcal{L}^*\mathbf{u}, \mathbf{v} \rangle_{\Omega} = \langle \mathcal{R}^*\mathbf{u}, \mathbf{v} \rangle_{\Gamma} \quad (5.11)$$

holds, where $\langle \cdot, \cdot \rangle$ denotes an appropriate duality pairing, then smooth minimizers of (5.4) are not directly solutions of (5.1)–(5.2), instead they solve the *strong* problem (compare with (4.4)–(4.5) in §4.1)

$$\mathcal{L}^*\mathcal{L}\mathbf{v} = \mathcal{L}^*\mathbf{f} \quad \text{in } \Omega \quad (5.12)$$

augmented with the essential condition (5.2) and the natural condition

$$\mathcal{R}^*\mathcal{L}\mathbf{u} = \mathcal{R}^*\mathbf{f}. \quad (5.13)$$

Equations (5.12), (5.2) and (5.13) form the boundary value problem for which the least-squares functional (5.4) is the naturally occurring convex, quadratic, energy functional providing the Rayleigh-Ritz setting. In other words, the *strong* problem (5.12), (5.2) and (5.13) is the differential equation whose weak Galerkin form coincides with the *least-squares* variational problem (5.8). Thus, it is conceivable to develop a least-squares principle for (5.1)–(5.2) by immersion of these equations into the appropriate strong least-squares problem followed by a standard Galerkin procedure. Of course, this is hardly the most efficient or lucid method.

Finally, we draw attention to the fact that \mathcal{L}^* coincides with the usual dual only if $Y \equiv L^2(\Omega)$. In general, the problem (5.12) can be determined from (5.1)–(5.2) through differentiation and linear combinations that account for the norm structure of Y .

5.3 Discrete least-squares principles

Given a pair $\{X, J(\cdot)\}$ consider another pair $\{X^h, J_h(\cdot)\}$ consisting of

1. a discrete, e.g., finite element, space X^h parametrized by a “mesh-size” parameter h and which approximates X in a sense to be specified later;
2. a quadratic functional $J_h(\cdot; \cdot, \cdot) : X^h \times Y \times Z \mapsto \mathcal{R}$.

The pair $\{X^h, J_h(\cdot)\}$ gives rise to a *discrete* least-squares principle:

$$\text{seek } \mathbf{u}^h \text{ in } X^h \text{ such that } J_h(\mathbf{u}^h; \mathbf{f}, \mathbf{g}) \leq J(\mathbf{v}^h; \mathbf{f}, \mathbf{g}) \quad \forall \mathbf{v}^h \in X^h. \quad (5.14)$$

Since the objective is to use (5.14) to determine approximate solutions of (5.1)–(5.2), it is necessary to make additional assumptions concerning the pair $\{X^h, J_h(\cdot)\}$ that will connect the two problems.

D.1 The least-squares functional is *consistent* in the sense that for all smooth data \mathbf{f} and \mathbf{g} and all smooth solutions \mathbf{u} of (5.1)–(5.2), $J_h(\mathbf{u}; \mathbf{f}, \mathbf{g}) = 0$.

D.2 The least-squares functional is *positive*, i.e.,

$$J_h(\mathbf{v}^h; 0, 0) > 0 \quad \forall 0 \neq \mathbf{v}^h \in X^h.$$

The positivity assumption **D.2** implies that

$$\|\cdot\|_h \equiv J(\cdot; 0, 0)^{1/2} : X^h \mapsto \mathcal{R} \quad (5.15)$$

defines a norm on X^h which we refer to as the *discrete energy norm*. Since X^h is finite dimensional, we can also infer the existence of an inner product

$$((\cdot, \cdot))_h : X^h \times X^h \mapsto \mathcal{R}, \quad (5.16)$$

called *discrete energy inner product* such that

$$\|\mathbf{v}^h\|_h^2 = ((\mathbf{v}^h, \mathbf{v}^h))_h. \quad (5.17)$$

Let us show that **D.1** and **D.2** are by themselves sufficient to solve (5.14) and obtain “optimal” approximations.

Theorem 2 *Assume that **D.1** and **D.2** hold for the pair $\{X^h, J(\cdot)\}$ and let \mathbf{u} denote a smooth solution of (5.1)–(5.2). Then,*

1. *the problem (5.14) has a unique minimizer $\mathbf{u}^h \in X^h$;*
2. *\mathbf{u}^h is the orthogonal projection of \mathbf{u} with respect to the discrete energy inner product (5.16).*

Proof. From the consistency assumption **D.1** and the (5.17), it is not hard to see that the Euler-Lagrange equation for (5.14) is

seek \mathbf{u}^h in X^h such that

$$B^h(\mathbf{u}^h; \mathbf{v}^h) = F^h(\mathbf{v}^h) \quad \text{for all } \mathbf{v}^h \in X^h, \quad (5.18)$$

where

$$B^h(\cdot; \cdot) = ((\cdot, \cdot))_h \quad \text{and} \quad F^h(\cdot) = ((\mathbf{u}, \cdot))_h.$$

Let $\{\phi_i^h\}$ denote a basis for X^h so that $\mathbf{u}^h = \sum_{i=1}^N U_i \phi_i^h$. It is obvious that (5.18) is a linear system of algebraic equations for the unknown coefficient vector \vec{U} . The matrix and the right hand side of this system are

$$A_{ij} = ((\phi_j^h, \phi_i^h))_h \quad \text{and} \quad F_i = ((\mathbf{u}, \phi_i^h))_h.$$

Clearly A is symmetric and from the positivity assumption we conclude that A is also positive definite. As a result, the system $A\vec{U} = \vec{F}$ has a unique solution.

To prove the second part it suffices to notice that (5.18) can be recast as

$$((\mathbf{u}^h - \mathbf{u}, \mathbf{v}^h))_h = 0 \quad \text{for all } \mathbf{v}^h \in X^h.$$

from where it is immediately obvious that \mathbf{u}^h is orthogonal projection of \mathbf{u} relative to the energy inner product. \square

Corollary 1 *The least-squares solution \mathbf{u}^h minimizes the discrete energy norm error, that is*

$$|||\mathbf{u} - \mathbf{u}^h|||_h = \inf_{\mathbf{v}^h \in X^h} |||\mathbf{u} - \mathbf{v}^h|||_h. \quad (5.19)$$

Theorem 2 shows that a least-squares principle is capable of producing reasonable results under a very limited set of assumptions. This explains the remarkable robustness of least-squares methods – almost any sensible pair $\{X^h, J_h(\cdot)\}$ will satisfy both **D.1-D.2**, while pairs that violate one or both conditions are very rare. They are so rare that in fact, one would have to deliberately construct such a discrete least-squares principle! Another remarkable observation is that neither **D.1** nor **D.2** appeal in any way to a mathematically correct CLS principle. Does this mean that we can completely ignore CLS and not bother with finding proper function spaces or norms and just proceed directly to find a pair $\{X^h, J_h(\cdot)\}$ satisfying **D.1** and **D.2** which, in view of (5.19), appears to be enough to obtain good results? The answer, of course, is no and the reason is that so far we have avoided addressing a key issue, namely the asymptotic behavior of the method, including the convergence of \mathbf{u}^h to \mathbf{u} .

Let us now explain why violations of the correct energy balance are less transparent for moderate values of h and why they will amplify as h becomes smaller, e.g., when the grid is refined. Assume for a moment that the setting for the continuous least-squares principle is such that both the continuous energy norm $|||\cdot|||$ and the natural norm $\|\cdot\|_X$ are meaningful for $\mathbf{u}^h \in X^h$ so that their restrictions to X^h are well-defined norms on this space. The positivity assumption **D.2** means that $|||\cdot|||_h$ is another norm on this finite-dimensional space and as such, it must be equivalent to the restrictions of $|||\cdot|||$ and $\|\cdot\|_X$. As a result, for every fixed $h > 0$, there are constants $\gamma_1(h)$ and $\gamma_2(h)$ such that

$$\gamma_1(h)\|\mathbf{u}^h\|_X \leq |||\mathbf{u}^h|||_h \leq \gamma_2(h)\|\mathbf{u}^h\|_X \quad \forall \mathbf{u}^h \in X^h, \quad (5.20)$$

i.e., a version of the correct energy balance holds for any fixed h . Likewise, if \vec{U} denotes the coefficient vector corresponding to U^h it is not hard to see that

$$\delta_1(h)\vec{U}^T M\vec{U} \leq \vec{U}^T A\vec{U} \leq \delta_2(h)\vec{U}^T M\vec{U} \quad (5.21)$$

for some other constants $\delta_1(h)$ and $\delta_2(h)$, and where M denotes the Gram matrix of the finite element space bases *relative to the inner product* $\|\cdot\|_X$. However, the asymptotic behavior of these constants depends entirely on the relation between the two energy norms and neither **D.1** nor **D.2** can control the growth (or decay) of $\gamma_i(h)$ and $\delta_i(h)$. On the other hand, it is the asymptotic of $\gamma_i(h)$ that governs the convergence of least-squares approximations and it is the asymptotic of $\delta_i(h)$ that governs the matrix conditioning. These constants measure the deviation of the discrete least-squares principle from the ideal setting defined by a CLS principle. As the deviation from CLS principle grows, so does the dependence of these constants on h and the quality of least-squares approximations and algebraic systems deteriorates with h . Therefore, asymptotic behavior of discrete least-squares principles depends on the

equivalence relation between the discrete energy norm and the continuous energy norm.

At the beginning of this chapter we sketched four possible relations between the continuous and discrete least-squares principles. It should be clear by now that each one of these relations gives rise to a bound similar to (5.20) and that the asymptotic behavior of γ_1 and γ_2 will depend on how well the DLSP reproduces the correct energy balance (5.3).

Conforming DLSP are simply restrictions of a given CLSP. In this case the pair $\{X^h, J_h(\cdot)\}$ is identified with a subspace X^h of X and $J_h(\cdot) = J(\cdot)$. As a result, (5.20) is a restriction of the energy balance (5.3) to X^h which means that $\gamma_1(h)$ and $\gamma_2(h)$ are independent of h ; in fact they coincide with the constants C_1 and C_2 from (5.3).

Norm-equivalent DLSP are identified with pairs $\{X^h, J_h(\cdot)\}$ for which $X^h \subset X$ and (5.20) holds with γ_1 and γ_2 *independent* of h . In general, for such methods $J_h(\cdot) \neq J(\cdot)$, which means that (5.20) does not represent a restriction of (5.3). Nevertheless, *norm-equivalent* methods do recover all advantages of a Rayleigh-Ritz setting.

Quasi-norm-equivalent DLSP are identified with pairs $\{X^h, J_h(\cdot)\}$ for which $X^h \subset X$ but (5.20) holds with γ_1 and γ_2 which *depend* on the mesh parameter h . These methods are capable of producing optimal convergence rates, however, the dependence on h in the equivalence bound leads to higher

condition numbers and/or lack of spectral equivalence with the natural inner product on $X \times X$.

And lastly, *non-equivalent DLSP* are identified with pairs $\{X^h, J_h(\cdot)\}$ for which X^h is not necessarily a subspace of X and $J_h(\cdot) \neq J(\cdot)$. As a result, an equivalence relation like (5.20) exists, but it is stated in terms of different³ spaces for the lower and upper bounds. Thus, nothing much can be said about optimality of the convergence rates and the spectral equivalence of the algebraic problems.

In the next chapter we specialize this framework to the important class of differential operators that are elliptic in the sense of Agmon-Douglis and Nirenberg. Among the members of this class are the various forms of the Stokes operator, div-curl operators with the appropriate boundary conditions and many other examples of practically important PDE's.

³If X^h satisfies an inverse inequality these bounds may be converted to bounds in terms of the same function space. This necessarily will introduce dependence on h in the lower and/or upper equivalence constants.

Chapter 6

Least-squares methods for ADN systems

Theorem 2 shows that least-squares principles can lead to a sensible method under very limited set of assumptions. This, of course is one of the great appeals of least-squares methodology. However, if a least-squares method is based only on the expectation that hypothesis **D.1-D.2** hold, nothing much can be said beyond the fact that approximate solutions are projections of the exact solution with respect to the discrete inner product (5.16) and that the least-squares solution minimizes the error as measured by the discrete energy norm (5.15). In particular, no specific information can be obtained about asymptotic convergence rates. Furthermore, without knowing the asymptotic behavior of the “constants” in (5.21) it is hard to develop efficient preconditioners for the solution of the least-squares algebraic systems.

These issues become more tractable when the abstract development of least-squares methods is carried in the context of a particular class of differential equations. This is precisely the purpose of this chapter where we focus on first-order differential operators that are elliptic in the sense of Agmon, Douglis and Nirenberg; see [11]. The ADN theory is, perhaps, the most powerful and universal tool for the analysis of elliptic boundary value problems. It has been successfully used in the context of least-squares methods in [46], [48], [56], [58], and [75]. For elliptic problems in the plane a parallel theory exists; see [10], which also has been used in the development of least-squares methods. For examples the reader can consult Wendland’s book [10] and the papers [79], [80], among others.

The least-squares theory developed in this chapter stands apart from the approaches cited above in several aspects. First, it includes a very broad

class of problems, namely ADN elliptic systems. At the same time it is focused on first-order systems because these are the problems of practical interest in least-squares. Lastly, our treatment highlights the idea of the least-squares method as realization of a mathematically ideal Continuous Least Squares Principle through a practical Discrete Least Squares principle.

We begin the chapter with a brief summary of ADN elliptic theory. In the next section we use this theory to show that both **A.1** and **A.2** hold for ADN systems. (Because of the rather technical nature of ADN theory, specific details of its application are collected in Appendix A.) This fact immediately leads us to identification of all CLSP for a given ADN elliptic operator. Section 6.3 briefly discusses transformation of general ADN operator into an equivalent first-order system. Then, in §6.4 we show that first-order ADN systems give rise to two basic classes of CLS principles - one associated with homogeneous elliptic operators, and one associated with non-homogeneous elliptic operators.

The core of this chapter is section 6.5 where we formulate discrete least-squares principles for the two types of first-order ADN operators. In particular, we highlight the often forgotten fact that

a first-order system is not necessarily homogeneous elliptic!

The most important consequence of this fact is that a mathematically well-posed continuous least-squares principle for a first-order system may still be impractical, thus transformation to a DLSP may be required even for first-order systems.

6.1 ADN differential operators

We consider systems of partial differential equations of the form (5.1)-(5.2), that is

$$\begin{aligned}\mathcal{L}\mathbf{u} &= \mathbf{f} && \text{in } \Omega \\ \mathcal{R}\mathbf{u} &= \mathbf{g} && \text{on } \Gamma.\end{aligned}$$

Here $\mathbf{u} = (u_1, u_2, \dots, u_N)$ is a vector of dependent variables,

$$D = (\partial/\partial x_1, \dots, \partial/\partial x_n) = (\partial_1, \dots, \partial_n)$$

denotes a differentiation operator, and $\mathcal{L} = \mathcal{L}(\mathbf{x}, D) = \mathcal{L}_{ij}(\mathbf{x}, D)$, $i, j = 1, \dots, N$. Likewise, the boundary operator has the form $\mathcal{R} = \mathcal{R}(\mathbf{x}, D) = \mathcal{R}_{lj}(\mathbf{x}, D)$, $l = 1, \dots, m$, $j = 1, \dots, N$. Lastly, it is assumed that each $\mathcal{L}_{ij}(\mathbf{x}, \boldsymbol{\xi})$ and $\mathcal{R}_{lj}(\mathbf{x}, \boldsymbol{\xi})$ is a polynomial in $\boldsymbol{\xi}$. In what follows we restrict

our attention to differential operators that are elliptic in the sense of the following definition, due to Agmon, Douglis and Nirenberg; see [11]:

Definition 1 *The system (5.1) is ADN-elliptic if there exist integer weights $\{s_i\}$ and $\{t_j\}$, for the equations and the unknowns, respectively, such that*

1. $\deg \mathcal{L}_{ij}(\mathbf{x}, \boldsymbol{\xi}) \leq s_i + t_j$;
2. $\mathcal{L}_{ij} \equiv 0$ whenever $s_i + t_j < 0$;
3. $\det \mathcal{L}_{ij}^P(\mathbf{x}, \boldsymbol{\xi}) \neq 0$ for all real $\boldsymbol{\xi} \neq 0$;

where the principal part \mathcal{L}^P of \mathcal{L} is defined as all terms \mathcal{L}_{ij} for which $\deg \mathcal{L}_{ij}(\mathbf{x}, \boldsymbol{\xi}) = s_i + t_j$.

Although this definition may seem a little bit artificial, it can be shown that for nondegenerate systems one can always find s_i and t_j so that the principal part \mathcal{L}^P does not vanish identically; see [15]. For such systems the degree r of the determinant $L(\mathbf{x}, \boldsymbol{\xi}) = \det \mathcal{L}(\mathbf{x}, \boldsymbol{\xi})$ equals the maximum degree R of the terms forming $L(\mathbf{x}, \boldsymbol{\xi})$ (in general $r \leq R$). Furthermore, Volevich [15] has shown that Definition 1 is equivalent to ellipticity in the following sense: $r = R$ and $\boldsymbol{\xi} \equiv 0$ is the only real root of $L'(\mathbf{x}, \boldsymbol{\xi}) = 0$, where L' denotes the part of order r of L .

The orders of \mathcal{R}_{lj} will also depend on two sets of integer weights: the unknown's weights $\{t_j\}$ already defined for \mathcal{L} , and a new set $\{r_l\}$ where each r_l is attached to the l th condition in (5.2). As before, it will be required that

$$\deg \mathcal{R}_{lj}(\mathbf{x}, \boldsymbol{\xi}) \leq r_l + t_j,$$

with the understanding that $\mathcal{R}_{lj} \equiv 0$ when $r_l + t_j < 0$. Finally, the principal part \mathcal{R}^P of the boundary operator will be defined as all terms \mathcal{R}_{lj} such that $\deg \mathcal{R}_{lj}(\mathbf{x}, \boldsymbol{\xi}) = r_l + t_j$. The three sets of indices can always be normalized in such a way that $s_i \leq 0$, $r_l \leq 0$ and $t_j \geq 0$. However, the sets of indices may not be unique, even with that normalization, i.e., there are examples of operators which possess more than one principal parts and still satisfy Definition 1.

An important subset of ADN elliptic systems is the class of Petrovski systems; see [14].

Definition 2 *A system is elliptic in the sense of Petrovski if it is elliptic in the sense of ADN and $s_1 = \dots = s_N = 0$. If in addition $t_1 = \dots = t_N$, the system is called homogeneous elliptic.*

One additional condition, which is satisfied for all elliptic systems in three or more space dimensions but must be assumed in two-dimensions is the *supplementary condition* of [11].

Definition 3 (Supplementary Condition on \mathcal{L}) *det $\mathcal{L}^P(\mathbf{x}, \boldsymbol{\xi})$ is of even degree $2m$ with respect to $\boldsymbol{\xi}$. For every pair of linearly independent vectors $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$, the polynomial $\det \mathcal{L}^P(\mathbf{x}, \boldsymbol{\xi} + \tau \boldsymbol{\xi}')$ in the complex variable τ has exactly m roots with positive imaginary part.*

When an elliptic system satisfies the Supplementary condition it is also called *regularly elliptic*; see [14]. Our final assumption concerning the system (5.1) is that \mathcal{L} is uniformly elliptic in the sense that there exists a positive constant C such that

$$C^{-1}|\boldsymbol{\xi}|^{2m} \leq |\det \mathcal{L}^P(\mathbf{x}, \boldsymbol{\xi})| \leq C|\boldsymbol{\xi}|^{2m}. \quad (6.1)$$

When boundary conditions (5.2) are attached to the operator \mathcal{L} , resulting boundary value problem may or may not be well-posed. A well-posed problem will result only if \mathcal{R} “complements” \mathcal{L} in a proper way. A necessary and sufficient condition for this is given by an algebraic criterion involving the principal parts \mathcal{L}^P and \mathcal{R}^P . This criterion, known as the *complementing condition* is due to Agmon, Douglis and Nirenberg, [11].

To state this condition let $\tau_k^+(\mathbf{x}, \boldsymbol{\xi})$ denote the m roots of $\det \mathcal{L}^P(\mathbf{x}, \boldsymbol{\xi} + \tau \boldsymbol{\xi}')$ having positive imaginary part, \mathbf{n} denote the normal to Γ ;

$$M^+(\mathbf{x}, \boldsymbol{\xi}, \tau) = \prod_{k=1}^m (\tau - \tau_k^+(\boldsymbol{\xi})),$$

and, lastly, let \mathcal{L}' denote the adjoint matrix to \mathcal{L}^P . Then, we have the following definition [11].

Definition 4 (Complementing Condition) *For any point $\mathbf{x} \in \Gamma$ and any real, non-zero vector $\boldsymbol{\xi}$ tangent to Γ at \mathbf{x} , regard $M^+(\mathbf{x}, \boldsymbol{\xi}, \tau)$ and the elements of the matrix*

$$\sum_{j=1}^N \mathcal{R}_{lj}^P(\mathbf{x}, \boldsymbol{\xi} + \tau \mathbf{n}) \mathcal{L}'_{jk}(\mathbf{x}, \boldsymbol{\xi} + \tau \mathbf{n})$$

as polynomials in τ . The operators \mathcal{L} and \mathcal{R} satisfy the complementing condition if the rows of the latter matrix are linearly independent modulo $M^+(\boldsymbol{\xi}, \tau)$, i.e.,

$$\sum_{l=1}^m C_l \sum_{j=1}^N \mathcal{R}_{lj}^P(\mathbf{x}, \boldsymbol{\xi} + \tau \mathbf{n}) \mathcal{L}'_{jk}(\mathbf{x}, \boldsymbol{\xi} + \tau \mathbf{n}) \equiv 0 \pmod{M^+} \quad (6.2)$$

if and only if the constants C_l all vanish.

For simplicity, in what follows the boundary value problem (5.1)-(5.2) will be called *elliptic* if:

1. \mathcal{L} is elliptic in the sense of ADN;
2. \mathcal{L} is regularly elliptic;
3. \mathcal{L} is uniformly elliptic;
4. \mathcal{R} satisfies the complementing condition.

With the problem (5.1)-(5.2) we associate the function spaces

$$\mathbf{X}_q = \prod_{j=1}^N H^{q+t_j}(\Omega); \quad \mathbf{Y}_q = \prod_{i=1}^N H^{q-s_i}(\Omega); \quad \mathbf{Z}_q = \prod_{l=1}^m H^{q-r_l-1/2}(\Gamma). \quad (6.3)$$

We now proceed to show that ADN elliptic systems satisfy hypotheses **A.1**–**A.2** of §5.1. The first hypothesis follows from a general result due to Agmon, Douglis and Nirenberg [11]. In what follows we shall skip the reference to \mathbf{x} and D and simply write \mathcal{L} and \mathcal{R} .

Theorem 3 *Let $t' = \max t_j$, $q \geq d = \max(0, \max r_l + 1)$ and assume that Ω is a bounded domain of class $C^{q+t'}$. Furthermore, assume that the coefficients of \mathcal{L} are of class $C^{q-s_i}(\bar{\Omega})$ and that the coefficients of \mathcal{R} are of class $C^{q-r_l}(\Gamma)$. If (5.1)-(5.2) is elliptic and $\mathbf{f} \in \mathbf{Y}_q$, $\mathbf{g} \in \mathbf{Z}_q$ then*

1. Every solution $\mathbf{u} \in \mathbf{X}_d$ is in fact in \mathbf{X}_q .
2. There is a positive constant C , independent of \mathbf{u} , \mathbf{f} and \mathbf{g} , such that, for every solution $\mathbf{u} \in \mathbf{X}_q$

$$\sum_{j=1}^N \|\mathbf{u}_j\|_{q+t_j, \Omega} \leq C \left(\sum_{i=1}^N \|\mathbf{f}_i\|_{q-s_i, \Omega} + \sum_{l=1}^m \|\mathbf{g}_l\|_{q-r_l-1/2, \Gamma} + \sum_{j=1}^N \|\mathbf{u}_j\|_{0, \Omega} \right). \quad (6.4)$$

Moreover, if the problem (5.1)-(5.2) has a unique solution, then the L^2 -norm on the right-hand side of (6.4) can be omitted. \square

Clearly, **A.1** is implied by (6.4). Furthermore, it can be shown that ADN elliptic operators are of Fredholm type; see [13], [14], [10], i.e., their range is closed and both the kernel and the co-range are finite dimensional. Therefore, **A.2** is also satisfied and $(\mathcal{L}, \mathcal{R})$ can be augmented by a finite number of

constraints so that the problem (5.1)-(5.2) always has a unique solution. In addition to the uniqueness, it will be also assumed that (6.4) remains valid for $q < 0$. This assumption, which amounts to the existence of complete sets of homeomorphisms for (5.1)-(5.2), is known to hold for self-adjoint ADN operators, or Petrovski systems; see [13]-[14]. With these assumptions (6.4) can be restated as follows: for all smooth functions \mathbf{u} in Ω and all integers q

$$\|\mathbf{u}\|_{\mathbf{X}_q} \leq C (\|\mathcal{L}\mathbf{u}\|_{\mathbf{Y}_q} + \|\mathcal{R}\mathbf{u}\|_{\mathbf{Z}_q}). \quad (6.5)$$

6.2 Continuous least-squares principles for ADN operators

From §5.2 we know that the proper “energy” balance (5.3) for the PDE can be determined through a priori bounds, after which the proper least-squares functional can be easily identified according to (5.4). For ADN systems Theorem 3 provides us with precisely that tool in the form of the estimate (6.5). As a result, for any elliptic ADN system, the artificial energy functional that provides a mathematically correct measure of the residual “energy” is given by

$$J(\mathbf{u}; \mathbf{f}, \mathbf{g}) = \|\mathcal{L}\mathbf{u} - \mathbf{f}\|_{\mathbf{Y}_q}^2 + \|\mathcal{R}\mathbf{u} - \mathbf{g}\|_{\mathbf{Z}_q}^2. \quad (6.6)$$

The functional (6.6) is norm-equivalent in the sense that

$$C^{-1}\|\mathbf{u}\|_{\mathbf{X}_q}^2 \leq \|\mathcal{L}\mathbf{u}\|_{\mathbf{Y}_q}^2 + \|\mathcal{R}\mathbf{u}\|_{\mathbf{Z}_q}^2 = J(\mathbf{u}; \mathbf{0}, \mathbf{0}). \quad (6.7)$$

This functional also gives rise to a mathematically well-posed Continuous Least-Squares Principle:

$$\min_{\mathbf{u} \in \mathbf{X}_q} J(\mathbf{u}; \mathbf{f}, \mathbf{g}). \quad (6.8)$$

This principle provides the setting of (5.5) for ADN systems, while the analogue of (5.8) is

$$\text{seek } \mathbf{u} \in \mathbf{X}_q \text{ such that } Q(\mathbf{u}; \mathbf{v}) = F(\mathbf{v}) \quad \forall \mathbf{v} \in \mathbf{X}_q, \quad (6.9)$$

where now

$$Q(\mathbf{u}; \mathbf{v}) = (\mathcal{L}\mathbf{u}, \mathcal{L}\mathbf{v})_{\mathbf{Y}_q} + \langle \mathcal{R}\mathbf{u}, \mathcal{R}\mathbf{v} \rangle_{\mathbf{Z}_q}$$

and

$$F(\mathbf{v}) = (\mathbf{f}, \mathcal{L}\mathbf{v})_{\mathbf{Y}_q} + \langle \mathbf{g}, \mathcal{R}\mathbf{v} \rangle_{\mathbf{Z}_q}$$

Like in the abstract case, norm-equivalence of the artificial energy functional (6.6) implies that $Q(\cdot; \cdot)$ defines an equivalent “energy” inner product

on $\mathbf{X}_q \times \mathbf{X}_q$. Thus, as long as (6.9) is discretized using a finite element subspace \mathbf{X}^h of \mathbf{X}_q , all attractive features of Rayleigh-Ritz setting (non-restrictive choice of finite element space, symmetric and positive definite algebraic systems, quasioptimal error estimates) of energy minimization principles would be formally recovered by the least-squares method. In addition, equivalence of $Q(\cdot; \cdot)$ and the inner product on $\mathbf{X}_q \times \mathbf{X}_q$ may be used in the design of preconditioners for the least-squares algebraic system. More precisely, if $A^h = Q(\phi_i; \phi_j)$ and $K^h = (\phi_i, \phi_j)_{\mathbf{X}_q}$, then A^h and K^h are spectrally equivalent in the sense that

$$C^{-1}\boldsymbol{\xi}^T K^h \boldsymbol{\xi} \leq \boldsymbol{\xi}^T A^h \boldsymbol{\xi} \leq C\boldsymbol{\xi}^T K^h \boldsymbol{\xi}, \quad \forall \boldsymbol{\xi} \in \mathcal{R}^n$$

Thus, A^h can be preconditioned by any matrix that is spectrally equivalent with K^h .

However, the CLSP described above may fail to be practical in the sense discussed in §4.4.1. Let us recall that to deem a least-squares method practical we require that

- the discrete system can be obtained without difficulty, or at least, with no more difficulty than for a Galerkin method.
- this system should have a condition number comparable to the condition number of the system in the Galerkin method;
- discretization should be accomplished using standard, easy to use finite element spaces.

The first condition will be violated if the least-squares functional involves, fractional or negative order Sobolev space norms because such norms are not computable. The second and third conditions will be violated if for some s_i and t_j we have that $s_i + t_j \geq 2$. In this case the term $\|\mathcal{L}_{ij}u_j - f_i\|_{q-s_i}$ will effectively involve second, or higher order derivatives, i.e., it cannot be discretized using standard C^0 spaces. In what follows we focus on the development of practical least-squares methods when the “energy” functional does not involve fractional order trace norms. Such norms arise whenever the essential boundary conditions are enforced weakly through the least-squares minimization process. For examples of such least-squares methods we refer to [46], [10] and [105], while here, for simplicity, we restrict our attention to the case of homogeneous boundary conditions which are imposed on the space \mathbf{X}_q .

When the differential operator \mathcal{L} involves second or higher order derivatives, a standard approach in modern least-squares methods has been to

transform (5.1)-(5.2) into an equivalent first-order system. This step is motivated by the observation that $\|\mathcal{L}_{ij}u_j - f_i\|_0^2$ can be discretized by merely continuous finite element spaces, provided the order of \mathcal{L}_{ij} does not exceed one. We encountered some specific examples of this idea in §4.4. In the next section we discuss the transformation process in more general terms, specialize (6.5) to the case of first-order systems, and derive norm-equivalent functionals for these systems.

6.3 First-order ADN systems

Any ADN-elliptic system of order higher than one can be transformed into an equivalent first-order system which remains elliptic in the same sense (this is not true if the usual definition of ellipticity, involving only differentiated terms, is used; see the example below). This transformation can be effected through the following process; see [11]. First, all variables are divided into two sets according to their indices: a set $\{u_{k'}\}$ containing all variables for which $t_j > 1$ and a set $\{u_{k''}\}$ of all variables for which $t_j \leq 1$. Then, the new variables are introduced as

$$u_{k',j} = \partial_j u_{k'}$$

while at the same time the equations defining these variables are appended to the differential operator. The original operator \mathcal{L} itself also undergoes a transformation. All terms in which u'_k is not differentiated remain unchanged. A term in which u'_k is differentiated is substituted according to the rule

$$D^\alpha(\partial_j u_{k'}) \mapsto D^\alpha(u_{k',j}).$$

Although rewriting of \mathcal{L} is not unique it can be shown; see [11], that the new system is elliptic in the sense of ADN and that $\max t_j \leq 2$, $\min s_i \geq -1$. The original boundary conditions (5.2) are transformed in a similar fashion into equivalent boundary conditions for the first-order system. Again, this process is not unique and can be accomplished in several possible ways; the important fact is that if the Complementing Condition was satisfied by (5.2), then it will be also satisfied by the new boundary conditions; see [11]. As a result, we are guaranteed that the new operator augmented with the new boundary conditions (denoted again by \mathcal{L} and \mathcal{R} , respectively) is well-posed, so that (6.5) remains valid.

Example 1 (Laplace operator.) *In 2D, the second order problem (2.6) can be transformed into a first-order system with the help of the new dependent variables $u_1 = \phi_x$ and $u_2 = \phi_y$. The first-order equation is the familiar*

system (2.21) from §2.2:

$$\begin{aligned}\frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y} &= f \\ \frac{\partial \phi}{\partial x} - u_1 &= 0 \\ \frac{\partial \phi}{\partial y} - u_2 &= 0\end{aligned}$$

If the principal part of (2.21) is defined by taking only the highest order terms then

$$\det \mathcal{L}^P(\mathbf{x}, \boldsymbol{\xi}) = \begin{vmatrix} 0 & \xi_1 & \xi_2 \\ \xi_1 & 0 & 0 \\ \xi_2 & 0 & 0 \end{vmatrix} \equiv 0.$$

However, (2.21) is elliptic in the sense of Definition 1. Indeed, with the choice $t_1 = 2$, $t_2 = t_3 = 1$ and $s_1 = 0$, $s_2 = s_3 = -1$ the determinant of the principal part is

$$\det \mathcal{L}^P(\mathbf{x}, \boldsymbol{\xi}) = \begin{vmatrix} 0 & \xi_1 & \xi_2 \\ \xi_1 & -1 & 0 \\ \xi_2 & 0 & -1 \end{vmatrix} = |\boldsymbol{\xi}|^2.$$

Our next example concerns an important first-order form of the Stokes problem (2.17) which will be considered in detail in §7.1.1 of Chapter 7. This example shows the *non-uniqueness* of the ADN indices, i.e., the possibility that a differential operator may possess multiple principal parts! Subsequently we will see that this fact has a significant impact on the least-squares finite element method.

Example 2 (Velocity-Vorticity-Pressure Stokes system.) We consider the Stokes equations (2.17) in 2D. The second order term involves the velocity variable $\mathbf{u} = (u_1, u_2)$. In principle we could have introduced all four derivatives $u_{1,x}$, $u_{1,y}$, $u_{2,x}$ and $u_{2,y}$ as new dependent variables. Instead we choose to introduce their linear combination

$$\boldsymbol{\omega} = \frac{\partial u_2}{\partial x} - \frac{\partial u_1}{\partial y}$$

as the sole new variable. One reason is that $\boldsymbol{\omega}$ has a physical meaning - this is the vorticity “vector” of \mathbf{u} . Another reason is that we increase the size

of the system only by one equation and one variable. It can be shown that (2.17) transforms into the first-order system

$$\begin{aligned}
\frac{\partial \omega}{\partial y} + \frac{\partial p}{\partial x} &= f_1 \\
-\frac{\partial \omega}{\partial x} + \frac{\partial p}{\partial y} &= f_2 \\
\frac{\partial u_2}{\partial x} - \frac{\partial u_1}{\partial y} - \omega &= 0 \\
\frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y} &= 0
\end{aligned} \tag{6.10}$$

Let us assume that the unknowns are ordered as (ω, p, u_1, u_2) . If

$$t_1 = \dots = t_4 = 1 \quad \text{and} \quad s_1 = \dots = s_4 = 0$$

then the determinant of the principal part is

$$\det \mathcal{L}^p(\mathbf{x}, \boldsymbol{\xi}) = \begin{vmatrix} \xi_2 & \xi_1 & 0 & 0 \\ -\xi_1 & \xi_2 & 0 & 0 \\ 0 & 0 & -\xi_2 & \xi_1 \\ 0 & 0 & \xi_1 & \xi_2 \end{vmatrix} = -|\boldsymbol{\xi}|^4.$$

However, if $t_1 = t_2 = 1, t_3 = t_4 = 2$ and $s_1 = s_2 = 0, s_3 = s_4 = -1$,

$$\det \mathcal{L}^p(\mathbf{x}, \boldsymbol{\xi}) = \begin{vmatrix} \xi_2 & \xi_1 & 0 & 0 \\ -\xi_1 & \xi_2 & 0 & 0 \\ -1 & 0 & -\xi_2 & \xi_1 \\ 0 & 0 & \xi_1 & \xi_2 \end{vmatrix} = -|\boldsymbol{\xi}|^4,$$

that is system (6.10) satisfies Definition 1 with two different sets of weights.

6.4 Continuous least-squares principles for first-order systems

Let us first assume that the (first-order) problem (5.1)-(5.2) is elliptic in the sense of Petrovski. Then $s_i = 0$ for all $i = 1, \dots, N$ and therefore $t_j = 1$ for all $j = 1, \dots, N$. Consequently, assuming that \mathbf{X}_q is restricted by the homogeneous boundary condition $\mathcal{R}\mathbf{u} = 0$, estimate (6.5) specializes to

$$\|\mathbf{u}\|_{\mathbf{X}_q} = \sum_{j=1}^N \|u_j\|_{q+1} \leq C \sum_{i=1}^N \left\| \sum_{j=1}^N \mathcal{L}_{ij} u_j \right\|_q \tag{6.11}$$

If (5.1)-(5.2) is not Petrovski, then there will be at least one equation index $s_i = -1$. Since all \mathcal{L}_{ij} are at most of order one, there will be at least one unknowns index $t_j = 2$. Without loss of generality we can assume that for some k and l

$$s_1 = \dots = s_k = 0; \quad s_{k+1} = \dots = s_N = -1, \quad (6.12)$$

and

$$t_1 = \dots = t_l = 1; \quad t_{l+1} = \dots = t_N = 2, \quad (6.13)$$

respectively. As a result, for non-Petrovski first-order systems (6.5) specializes to

$$\begin{aligned} \|\mathbf{u}\|_{\mathbf{x}_q} &= \sum_{j=1}^l \|u_j\|_{q+1} + \sum_{j=l+1}^N \|u_j\|_{q+2} \\ &\leq C \left(\sum_{i=1}^k \left\| \sum_{j=1}^N \mathcal{L}_{ij} u_j \right\|_q + \sum_{i=k+1}^N \left\| \sum_{j=1}^N \mathcal{L}_{ij} u_j \right\|_{q+1} \right) \end{aligned} \quad (6.14)$$

To define the norm-equivalent functionals and the associated CLS principles we further restrict the range of q to -1 and 0. For Petrovski systems the choice $q = 0$ in (6.11) corresponds to the norm-equivalent functional

$$J_P(\mathbf{u}; \mathbf{f}) = \sum_{i=1}^N \left\| \sum_{j=1}^N \mathcal{L}_{ij} u_j - \mathbf{f}_i \right\|_0^2, \quad (6.15)$$

while for non-Petrovski systems the choices $q = -1$ or $q = 0$ in (6.14) yield the two norm-equivalent functionals

$$J_{-1}(\mathbf{u}; \mathbf{f}) = \sum_{i=1}^k \left\| \sum_{j=1}^N \mathcal{L}_{ij} u_j - \mathbf{f}_i \right\|_{-1}^2 + \sum_{i=k+1}^N \left\| \sum_{j=1}^N \mathcal{L}_{ij} u_j - \mathbf{f}_i \right\|_0^2 \quad (6.16)$$

and

$$J_0(\mathbf{u}; \mathbf{f}) = \sum_{i=1}^k \left\| \sum_{j=1}^N \mathcal{L}_{ij} u_j - \mathbf{f}_i \right\|_0^2 + \sum_{i=k+1}^N \left\| \sum_{j=1}^N \mathcal{L}_{ij} u_j - \mathbf{f}_i \right\|_1^2, \quad (6.17)$$

respectively. Each one of these three functionals gives rise to CLS principle in the manner outlined in §5.2.

6.5 Discrete least-squares principles for first-order systems

To discuss finite element methods based on the three functionals (6.15)-(6.17) let \mathcal{T}_h denote a regular triangulation of the domain Ω into finite elements. We consider spaces of continuous, piecewise polynomial functions defined with respect to \mathcal{T}_h and denoted by S_d^h . It is assumed that for every $u \in H^{d+1}(\Omega)$ there exists $u^h \in S_d^h$ with

$$\|u - u^h\|_0 + h\|u - u^h\|_1 \leq Ch^{d+1}\|u\|_{d+1}. \quad (6.18)$$

For example, the space P_1 (continuous, piecewise linear polynomials on triangles) satisfies (6.18) with $d = 1$, while the space P_2 (continuous, piecewise quadratic polynomials on triangles) satisfies (6.18) with $d = 2$. We also recall that for regular triangulations the Euclidean norm of the coefficient vector of u^h , denoted by $|u^h|$, and the L^2 norm of u^h are related by the inequality

$$C^{-1}h^M|u^h| \leq \|u^h\|_0 \leq Ch^M|u^h|, \quad (6.19)$$

where M denotes the dimension of S_d^h . We will also need the inverse inequality

$$\|u^h\|_1 \leq Ch^{-1}\|u^h\|_0 \quad (6.20)$$

which holds for most standard finite element spaces on regular triangulations; see [3].

All three mathematically correct functionals (6.15)-(6.17) are norm equivalent, however, only (6.15) is practical in the sense discussed in §4.4. Functional (6.16) contains negative order norms while (6.17) has terms with $t_j - s_i = 2$, i.e., their total order is two. The reason is that although \mathcal{L} involves only first-order terms, the problem (5.1)-(5.2) is not homogeneous elliptic, and therefore, the components of \mathbf{u} have different differentiability properties. As a result, transformation to first-order systems alone may not be sufficient to derive a practical least-squares method, unless the new system also happens to be of Petrovski type. Therefore, for non-Petrovski systems it is still necessary to effect a transition from the ideal CLSP to a practical DLSP. This means that we must define alternative, discrete least-squares functionals to replace (6.16) and (6.17).

6.5.1 Least-squares for Petrovski systems

Consider a first-order Petrovski system and the CLS principle

$$\min_{\mathbf{u} \in \mathbf{X}_0} J_P(\mathbf{u}; f). \quad (6.21)$$

associated with the least-squares functional (6.15). The minimization space in (6.21) is given by

$$\mathbf{X}_0 = \left\{ \mathbf{u} \mid \mathbf{u} \in \prod_{j=1}^N H^1(\Omega); \quad \mathcal{R}\mathbf{u} = 0 \quad \text{on } \Gamma \right\},$$

and the Euler-Lagrange equation for (6.21) is

$$\text{seek } \mathbf{u} \in \mathbf{X}_0 \text{ such that } Q(\mathbf{u}; \mathbf{v}) = F(\mathbf{v}) \quad \forall \mathbf{v} \in \mathbf{X}_0, \quad (6.22)$$

where now

$$Q(\mathbf{u}; \mathbf{v}) = \sum_{i=1}^N \left(\sum_{j=1}^N \mathcal{L}_{ij} u_j, \sum_{j=1}^N \mathcal{L}_{ij} v_j \right)_0 \quad \text{and} \quad F(\mathbf{v}) = (\mathbf{f}, \sum_{j=1}^N \mathcal{L}_{ij} v_j)_0.$$

The form in (6.22) involves only L^2 -inner products of first-order terms and the space \mathbf{X}_0 is a product of $H^1(\Omega)$ spaces. As a result, for first-order Petrovski systems a practical least-squares method can be derived directly from the CLS principle by choosing a finite element subspace \mathbf{X}^h of \mathbf{X}_0 and setting $J_h(\cdot) = J(\cdot)$. According to the terminology introduced in §5 we call the ensuing discrete least-squares principle $\{\mathbf{X}^h, J_h(\cdot)\}$ *conforming*. The next theorem shows that a least-squares method based on (6.21)-(6.22) does indeed meet all criteria for practicality - discretization is accomplished by standard C^0 finite element spaces, approximations are quasi-optimal, algebraic systems can be easily preconditioned and their condition numbers are similar to those of a standard Galerkin method.

Theorem 4 *Assume that (5.1)-(5.2) is a first-order Petrovski system and that \mathbf{X}_0 is the space defined above. Furthermore, let*

$$\mathbf{X}^h = \left\{ \mathbf{u}^h \mid \mathbf{u}^h \in \prod_{j=1}^N S_d^h, \quad \mathcal{R}\mathbf{u}^h = 0 \quad \text{on } \Gamma \right\}$$

for some integer $d \geq 1$ and assume that $\mathbf{u} \in \mathbf{X}_q$ for some $q \geq 0$. Then,

1. the least-squares variational problem (6.22) has a unique solution $\mathbf{u} \in \mathbf{X}_0$ for any $\mathbf{f} \in \mathbf{Y}_0$;
2. the discrete least-squares variational problem

$$\text{seek } \mathbf{u}^h \in \mathbf{X}^h \text{ such that } Q(\mathbf{u}^h; \mathbf{v}^h) = F(\mathbf{v}^h) \quad \forall \mathbf{v}^h \in \mathbf{X}^h, \quad (6.23)$$

has a unique solution \mathbf{u}^h such that

$$\|\mathbf{u} - \mathbf{u}^h\|_1 \leq C \inf_{\mathbf{v} \in \mathbf{X}^h} \|\mathbf{u} - \mathbf{v}^h\|_1. \quad (6.24)$$

and

$$\|\mathbf{u} - \mathbf{u}^h\|_1 \leq Ch^{\tilde{d}} \|\mathbf{u}\|_{\tilde{d}+1}, \quad \tilde{d} = \min\{d, q\}; \quad (6.25)$$

3. the least-squares discretization matrix A^h defined by $A_{ij}^h = Q(\boldsymbol{\xi}_i; \boldsymbol{\xi}_j)$ is spectrally equivalent to the block diagonal matrix $\text{diag}(D, \dots, D)$ with

$$D_{ij} = (\phi_i, \phi_j)_1.$$

Here $\{\boldsymbol{\xi}_i\}$ and $\{\phi_i\}$ denote standard nodal bases for \mathbf{X}^h and S_d^h , respectively. Furthermore, $\text{cond}(A) = O(h^{-2})$.

Proof. To prove 1. and 2. it suffices to show that $Q(\cdot; \cdot)$ is continuous and coercive on $\mathbf{X}_0 \times \mathbf{X}_0$. From the norm equivalence of (6.15)

$$C\|\mathbf{u}\|_1^2 \leq J_P(\mathbf{u}; \mathbf{0}) = Q(\mathbf{u}; \mathbf{u})$$

which establishes coercivity. Next, since each \mathcal{L}_{ij} is of order at most one,

$$\left(\mathcal{L}_{ij}u_j, \mathcal{L}_{kl}v_l \right)_0 \leq \|\mathcal{L}_{ij}u_j\|_0 \|\mathcal{L}_{kl}v_l\|_0 \leq C\|u_j\|_1 \|v_l\|_1,$$

which implies the continuity. As a result, existence and uniqueness of a solution to (6.22) follows from the Riezs representation theorem. To show that (6.23) also has a unique solution we note that $\mathbf{X}^h \subset \mathbf{X}_0$. Thus, $Q(\cdot; \cdot)$ remains continuous and coercive on $\mathbf{X}^h \times \mathbf{X}^h$ and (6.25) follows by a standard finite element argument.

For the proof of the last part we agree to use \mathbf{u}^h or u_i^h to denote both a finite element function and the coefficient vector of its nodal representation. From the identities

$$(\mathbf{u}^h)^T A^h \mathbf{u}^h = Q(\mathbf{u}^h; \mathbf{u}^h) \quad \text{and} \quad (u_i^h)^T D u_i^h = (u_i^h, u_i^h)_1$$

and the fact that $Q(\cdot; \cdot)$ is continuous and coercive it follows that

$$C^{-1} \sum_{i=1}^n u_i^h D u_i^h \leq (\mathbf{u}^h)^T A^h \mathbf{u}^h \leq C \sum_{i=1}^n (u_i^h)^T D u_i^h,$$

i.e., A^h and $\text{diag}(D, \dots, D)$ are spectrally equivalent.

To find a bound for the condition number of A^h , we assume that (6.19) is valid for S_d^h . Then

$$C^{-1}h^{2M}|\mathbf{u}^h|^2 \leq \|\mathbf{u}^h\|_0^2 \leq Q(\mathbf{u}^h; \mathbf{u}^h) \leq C\|\mathbf{u}^h\|_1^2 \leq Ch^{2M-2}|\mathbf{u}^h|^2$$

where the last inequality follows from (6.19) and (6.20). Thus, $\text{cond}(A^h) = O(h^{-2})$. \square

First-order Petrovski systems offer the most favorable setting for the development of least-squares methods in the sense that a practical method for such systems is derived directly from the ideal CLS principle. As a result, application of least-squares to Petrovski systems provides a variational setting that is essentially identical with that of a classical Rayleigh-Ritz method. Another advantage of such systems is the equivalence of $Q(\cdot; \cdot)$ and the standard inner product on $[H^1(\Omega)]^n$. As a result, least-squares algebraic problems for Petrovski systems can be preconditioned using any good preconditioner for the Poisson equation.

6.5.2 Least-squares for first-order ADN systems

In this section we develop least-squares methods for first-order systems that are not homogeneous elliptic. The CLS principle for such systems violates one or more of the practicality requirements. As a result, a least-squares method defined from this principle will lead to methods that are formally quasi-optimal but would not be useful in practice. To circumvent this problem we consider Discrete Least-Squares Principles derived from the CLSP, but based on practical discrete energy functionals. These functionals are not necessarily norm-equivalent on the same spaces as the primary ones and, as a result, their minimization may not be meaningful on \mathbf{X}_q .

Weighted least-squares principles

Weighted least-squares principles are based on the premise that in finite dimensional spaces all norms are equivalent. Thus, a norm which appears in a least-squares functional and is impractical can be replaced by an L^2 -norm scaled by the appropriate equivalence constant which usually depends on the mesh parameter h . For the norm-equivalent functionals (6.16) and (6.17) this leads to weighted L^2 functionals given by

$$J_h(\mathbf{u}; \mathbf{f}) = h^2 \sum_{i=1}^k \left\| \sum_{j=1}^N \mathcal{L}_{ij} u_j - \mathbf{f}_i \right\|_0^2 + \sum_{i=k+1}^N \left\| \sum_{j=1}^N \mathcal{L}_{ij} u_j - \mathbf{f}_i \right\|_0^2 \quad (6.26)$$

and

$$J_h(\mathbf{u}; \mathbf{f}) = \sum_{i=1}^k \left\| \sum_{j=1}^N \mathcal{L}_{ij} u_j - \mathbf{f}_i \right\|_0^2 + h^{-2} \sum_{i=k+1}^N \left\| \sum_{j=1}^N \mathcal{L}_{ij} u_j - \mathbf{f}_i \right\|_0^2, \quad (6.27)$$

respectively. We note that (6.26) and (6.27) differ only by the common (and unimportant for the minimization) factor h^2 , i.e., these two functionals are essentially the same. Thus, we consider only methods based on (6.27). A Discrete Least Squares principle associated with this functional is given by a pair $\{\mathbf{X}^h, J_h(\cdot)\}$, where \mathbf{X}^h is a finite dimensional space to be specified later and $J_h(\cdot)$ is the functional (6.27). The discrete minimization problem thus reads:

$$\min_{\mathbf{u}^h \in \mathbf{X}^h} J_h(\mathbf{u}^h; \mathbf{f}). \quad (6.28)$$

The corresponding discrete variational problem is

$$\text{seek } \mathbf{u}^h \in \mathbf{X}^h \text{ such that } B^h(\mathbf{u}^h; \mathbf{v}^h) = F^h(\mathbf{v}^h) \quad \forall \mathbf{v}^h \in \mathbf{X}^h, \quad (6.29)$$

where now

$$B^h(\mathbf{u}^h; \mathbf{v}^h) = \sum_{i=1}^k \left(\sum_{j=1}^N \mathcal{L}_{ij} u_j^h, \sum_{j=1}^N \mathcal{L}_{ij} v_j^h \right)_0 + h^{-2} \sum_{i=k+1}^N \left(\sum_{j=1}^N \mathcal{L}_{ij} u_j^h, \sum_{j=1}^N \mathcal{L}_{ij} v_j^h \right)_0$$

and

$$F^h(\mathbf{v}) = \sum_{i=1}^k (\mathbf{f}_i, \sum_{j=1}^N \mathcal{L}_{ij} v_j^h)_0 + h^{-2} \sum_{i=k+1}^N (\mathbf{f}_i, \sum_{j=1}^N \mathcal{L}_{ij} v_j^h)_0.$$

Approximations defined by (6.29) are studied in the next theorem.

Theorem 5 *Assume that the indices s_i, t_j are given by (6.12) and (6.13), respectively, and let*

$$\mathbf{X}^h = \{ \mathbf{u}^h \mid \mathbf{u}^h \in \prod_{j=1}^l S_d^h \times \prod_{j=l+1}^N S_{d+1}^h; \quad \mathcal{R} \mathbf{u}^h = 0 \quad \text{on } \Gamma \} \quad (6.30)$$

where S_d^h and S_{d+1}^h are finite element spaces satisfying (6.18) for some $d \geq 1$. Also, assume that there exists a positive integer $r \geq d$ such that the exact solution \mathbf{u} of (5.1)-(5.2) belongs to the space

$$\mathbf{X}_r = \{ \mathbf{u} \mid \mathbf{u} \in \prod_{j=1}^l H^{r+1}(\Omega) \times \prod_{j=l+1}^N H^{r+2}(\Omega); \quad \mathcal{R} \mathbf{u} = 0 \quad \text{on } \Gamma \}$$

Then,

1. the least-squares variational problem (6.29) has a unique solution \mathbf{u}^h and

$$\sum_{j=1}^l \|u_j - u_j^h\|_0 + \sum_{j=l+1}^N \|u_j - u_j^h\|_1 \leq h^{d+1} \left(\sum_{j=1}^l \|u_j\|_{d+1} + \sum_{j=l+1}^N \|u_j\|_{d+2} \right); \quad (6.31)$$

2. condition number of the least-squares discretization matrix for (6.29) is bounded by $O(h^{-4})$.

Proof. The first part of this theorem follows from a general result of Aziz et. al. [46]. To show the second part we proceed as in Theorem 4 to find that now

$$C^{-1}h^{2M}|\mathbf{u}^h|^2 \leq \|\mathbf{u}^h\|_0^2 \leq B^h(\mathbf{u}^h; \mathbf{u}^h) \leq Ch^{-2}\|\mathbf{u}^h\|_1^2 \leq Ch^{2M-4}|\mathbf{u}^h|^2,$$

i.e., $\text{cond}(A) = O(h^{-4})$. \square

Compared with the method from §6.5.1, the weighted method does not fit so nicely into a Rayleigh-Ritz-like framework. The principal reason is that by using a DLSP based on the weighted functional (6.27) we have deviated from the mathematically ideal framework prescribed by the CLSP for (6.16). In particular, the least-squares variational problem (6.29) does not represent a restriction to \mathbf{X}^h of a variational problem associated with the CLSP. In fact, the weighted method is nonconforming in the sense that while the CLSP functional (6.16) is minimized over the space $\mathbf{X}_0 = \prod_{j=1}^l H^1(\Omega) \times \prod_{j=l+1}^n H^2(\Omega)$, the discrete space \mathbf{X}^h is not contained in \mathbf{X}_0 but only in $\mathbf{X}_{-1} = \prod_{j=1}^l L^2(\Omega) \times \prod_{j=l+1}^n H^1(\Omega)$. Concerning the norm-equivalence of (6.27) one can show that

$$\sum_{j=1}^l \|u_j^h\|_0^2 + \sum_{j=l+1}^N \|u_j^h\|_1^2 \leq J_h(\mathbf{u}^h; \mathbf{0}) \leq h^{-2} \left(\sum_{j=1}^l \|u_j^h\|_0^2 + \sum_{j=l+1}^N \|u_j^h\|_1^2 \right), \quad (6.32)$$

provided (6.20) holds. Both the lower and the upper bounds in (6.32) are in the norm of \mathbf{X}_{-1} , but the upper bound is scaled by h^{-2} . This can be interpreted as an attempt to mimic the norm on \mathbf{X}_0 , however, this scaling causes the bound for the condition number to behave like $O(h^{-4})$. A similar “one-sided” norm-equivalence bounds can be established for the other weighted functional where now the scaling is h^2 and is applied to the lower

bound:

$$h^2 \left(\sum_{j=1}^l \|u_j^h\|_0^2 + \sum_{j=l+1}^N \|u_j^h\|_1^2 \right) \leq J_h(\mathbf{u}^h; \mathbf{0}) \leq \sum_{j=1}^l \|u_j^h\|_0^2 + \sum_{j=l+1}^N \|u_j^h\|_1^2. \quad (6.33)$$

According to the terminology of §5 we call such DLS principles *quasi norm-equivalent*. Both (6.32) and (6.33) provide an example of DLS principles for which the constants in the equivalence bound (5.21) are mesh-dependent. The fact that these constants depend on h means that there is no apparent spectral equivalence between the least-squares discretization matrix and the matrix associated with the standard inner product on \mathbf{X}_0 . This makes it harder to precondition efficiently the discrete equations.

Discrete negative norm least-squares principles

In this section we focus attention on the negative norm functional (6.16). Our goal is to find a discrete (mesh-dependent) replacement for the negative norm $\|\cdot\|_{-1}$ so that the resulting discrete functional retains the norm equivalence properties of the continuous functional, at least for discrete functions. We note that the weighted L^2 norm $h\|\cdot\|_0$ used in (6.26) can be viewed as one such replacement. However, this norm is not equivalent to $\|\cdot\|_{-1}$ which is reflected by the factor h^2 that appears in (6.33). To define a discrete negative norm with better equivalence properties we use an approach suggested by Bramble et. al. in [62]. As before, let $D_{ij} = (\phi_i, \phi_j)_1$ and let \mathbf{B}^h denote a symmetric and positive semidefinite operator that is spectrally equivalent to D^{-1} in the sense that

$$C^{-1}(D^{-1}v, v) \leq (\mathbf{B}^h v, v) \leq C(D^{-1}v, v), \quad \forall v \in L^2(\Omega). \quad (6.34)$$

We define the discrete negative norm as

$$\|v\|_{-h} = ((h^2\mathbf{I} + \mathbf{B}^h)v, v)^{1/2}, \quad \forall v \in L^2(\Omega). \quad (6.35)$$

Lemma 1 *There exists a positive constant C such that for any $u \in L^2(\Omega)$*

$$C^{-1}\|u\|_{-1} \leq \|u\|_{-h} \leq C(h\|u\|_0 + \|u\|_{-1}). \quad (6.36)$$

If the inverse inequality (6.20) holds for S_d^h then

$$C^{-1}\|u^h\|_{-1} \leq \|u^h\|_{-h} \leq C\|u^h\|_{-1}, \quad (6.37)$$

that is, $\|\cdot\|_{-h}$ is equivalent to $\|\cdot\|_{-1}$ on S_d^h . \square

For a proof of this lemma we refer to [50]. Note that without the term \mathbf{B}^h norm $\|\cdot\|_{-h}$ reduces to just a weighted L^2 norm, i.e., this term is critical for (6.36) and (6.37). To define the least-squares method we first replace the energy functional (6.16) by the discrete negative norm functional

$$J_{-h}(\mathbf{u}; \mathbf{f}) = \sum_{i=1}^k \left\| \sum_{j=1}^N \mathcal{L}_{ij} u_j - \mathbf{f}_i \right\|_{-h}^2 + \sum_{i=k+1}^N \left\| \sum_{j=1}^N \mathcal{L}_{ij} u_j - \mathbf{f}_i \right\|_0^2 \quad (6.38)$$

and then consider the Discrete Least Squares Principle

$$\min_{\mathbf{u}^h \in \mathbf{X}^h} J_{-h}(\mathbf{u}^h; \mathbf{f}) \quad (6.39)$$

where the space \mathbf{X}^h is defined as in (6.30). The discrete variational problem is then given by

$$\text{seek } \mathbf{u}^h \in \mathbf{X}^h \text{ such that } B^{-h}(\mathbf{u}^h; \mathbf{v}^h) = \mathcal{F}^h \mathbf{v}^h \quad \forall \mathbf{v}^h \in \mathbf{X}^h, \quad (6.40)$$

where

$$B^{-h}(\mathbf{u}^h; \mathbf{v}^h) = \sum_{i=1}^k \left(\sum_{j=1}^N \mathcal{L}_{ij} u_j^h, \sum_{j=1}^N \mathcal{L}_{ij} v_j^h \right)_{-h} + \sum_{i=k+1}^N \left(\sum_{j=1}^N \mathcal{L}_{ij} u_j^h, \sum_{j=1}^N \mathcal{L}_{ij} v_j^h \right)_0$$

and

$$\mathcal{F}^h \mathbf{v} = \sum_{i=1}^k \left(\mathbf{f}_i, \sum_{j=1}^N \mathcal{L}_{ij} v_j^h \right)_{-h} + \sum_{i=k+1}^N \left(\mathbf{f}_i, \sum_{j=1}^N \mathcal{L}_{ij} v_j^h \right)_0.$$

Theorem 6 *Assume that \mathbf{X}^h is defined by (6.30) for some integer $d \geq 1$ and that the exact solution \mathbf{u} of (5.1)-(5.2) belongs to the space \mathbf{X}_r , defined in Theorem 5, for some $r \geq 0$. Then,*

1. *the least-squares variational problem (6.40) has a unique solution \mathbf{u}^h and*

$$\sum_{j=1}^l \|u_j - u_j^h\|_0 + \sum_{j=l+1}^N \|u_j - u_j^h\|_1 \leq h^{\tilde{d}+1} \left(\sum_{j=1}^l \|u_j\|_{\tilde{d}+1} + \sum_{j=l+1}^N \|u_j\|_{\tilde{d}+2} \right) \quad (6.41)$$

where $\tilde{d} = \min\{r, d\}$;

2. the condition number of the least-squares discretization matrix for (6.40) is bounded by $O(h^{-2})$ and this matrix is spectrally equivalent to the block-diagonal matrix

$$M = \underbrace{(G, \dots, G)}_l \underbrace{(D, \dots, D)}_{N-l},$$

where $G = (\phi_i, \phi_j)_0$ is the Gramm matrix for the basis of S_d^h and $D = (\phi_i, \phi_j)_1$.

Proof. We first show that $B^{-h}(\cdot; \cdot)$ is continuous and coercive on $\mathbf{X}^h \times \mathbf{X}^h$, i.e.,

$$\begin{aligned} C^{-1} \left(\sum_{j=1}^l \|u_j^h\|_0^2 + \sum_{j=l+1}^N \|u_j^h\|_1^2 \right) &\leq B^{-h}(\mathbf{u}^h; \mathbf{u}^h) \\ &\leq C \left(\sum_{j=1}^l \|u_j^h\|_0^2 + \sum_{j=l+1}^N \|u_j^h\|_1^2 \right). \end{aligned} \quad (6.42)$$

Since $u_j^h \in S_d^h$ or S_{d+1}^h and the order of each \mathcal{L}_{ij} is at most one, it follows that $\mathcal{L}_{ij}u_j^h \in L^2(\Omega)$ for all $i, j = 1, \dots, n$. Then, using the lower bound in (6.36), the norm-equivalence of (6.16) and the fact that \mathbf{X}^h is a subspace of \mathbf{X}_{-1} yields

$$\begin{aligned} B^{-h}(\mathbf{u}^h; \mathbf{u}^h) &= \sum_{i=1}^k \left\| \sum_{j=1}^N \mathcal{L}_{ij}u_j^h \right\|_{-h}^2 + \sum_{i=k+1}^N \left\| \sum_{j=1}^N \mathcal{L}_{ij}u_j^h \right\|_0^2 \\ &\geq C \left(\sum_{i=1}^k \left\| \sum_{j=1}^N \mathcal{L}_{ij}u_j^h \right\|_{-1}^2 + \sum_{i=k+1}^N \left\| \sum_{j=1}^N \mathcal{L}_{ij}u_j^h \right\|_0^2 \right) \\ &= C J_{-1}(\mathbf{u}^h; \mathbf{0}) \\ &\geq C \left(\sum_{j=1}^l \|u_j^h\|_0^2 + \sum_{j=l+1}^N \|u_j^h\|_1^2 \right) = \|\mathbf{u}^h\|_{\mathbf{X}_{-1}}^2. \end{aligned}$$

To show continuity we note that all discrete negative norm terms in $B^{-h}(\cdot; \cdot)$ correspond to an equation index $s_i = 0$; $i = 1, \dots, k$, while all L^2 terms - to an equation index $s_i = -1$; $i = k+1, \dots, n$. Let us fix $1 \leq i \leq k$ so that $s_i = 0$. Then, using the Cauchy inequality, the fact that the order of each \mathcal{L}_{ij} is at most one, and the inverse inequality, the i th term in $B^{-h}(\cdot; \cdot)$ can

be bounded as follows:

$$\begin{aligned}
& \left(\sum_{j=1}^N \mathcal{L}_{ij} u_j^h, \sum_{j=1}^N \mathcal{L}_{ij} v_j^h \right)_{-h} \\
& \leq \left(\sum_{j=1}^N \|\mathcal{L}_{ij} u_j^h\|_{-h} \right) \left(\sum_{j=1}^N \|\mathcal{L}_{ij} v_j^h\|_{-h} \right) \\
& \leq \sum_{j=1}^N \left(h \|\mathcal{L}_{ij} u_j^h\|_0 + \|\mathcal{L}_{ij} u_j^h\|_{-1} \right) \sum_{j=1}^N \left(h \|\mathcal{L}_{ij} v_j^h\|_0 + \|\mathcal{L}_{ij} v_j^h\|_{-1} \right) \\
& \leq \sum_{j=1}^N \left(h \|u_j^h\|_1 + \|u_j^h\|_0 \right) \sum_{j=1}^N \left(h \|v_j^h\|_1 + \|v_j^h\|_0 \right) \\
& \leq \sum_{j=1}^N \|u_j^h\|_0 \sum_{j=1}^N \|v_j^h\|_0
\end{aligned}$$

Next consider a term with $k+1 \leq i \leq n$ so that $s_i = -1$. Since $\deg \mathcal{L}_{ij} \leq s_i + t_j$ and $t_j = 1$ for $j = 1, \dots, l$ it follows that the first l differential operators have order zero, while the last $N-l$ have orders bounded by 1, that is:

$$\deg \mathcal{L}_{ij} = 0; \quad j = 1, \dots, l; \quad \deg \mathcal{L}_{ij} \leq 1; \quad j = l+1, \dots, N.$$

Then,

$$\begin{aligned}
& \left(\sum_{j=1}^N \mathcal{L}_{ij} u_j^h, \sum_{j=1}^N \mathcal{L}_{ij} v_j^h \right)_0 \\
& \leq \sum_{j=1}^N \|\mathcal{L}_{ij} u_j^h\|_0 \sum_{j=1}^N \|\mathcal{L}_{ij} v_j^h\|_0 \\
& = \left(\sum_{j=1}^l \|\mathcal{L}_{ij} u_j^h\|_0 + \sum_{j=l+1}^N \|\mathcal{L}_{ij} u_j^h\|_0 \right) \left(\sum_{j=1}^l \|\mathcal{L}_{ij} v_j^h\|_0 + \sum_{j=l+1}^N \|\mathcal{L}_{ij} v_j^h\|_0 \right) \\
& \leq C \left(\sum_{j=1}^l \|u_j^h\|_0 + \sum_{j=l+1}^N \|u_j^h\|_1 \right) \left(\sum_{j=1}^l \|v_j^h\|_0 + \sum_{j=l+1}^N \|v_j^h\|_1 \right)
\end{aligned}$$

Combining both inequalities yields continuity in the norm of \mathbf{X}_{-1} :

$$B^{-h}(\mathbf{u}^h; \mathbf{v}^h) \leq \left(\sum_{j=1}^l \|u_j^h\|_0 + \sum_{j=l+1}^N \|u_j^h\|_1 \right) \left(\sum_{j=1}^l \|v_j^h\|_0 + \sum_{j=l+1}^N \|v_j^h\|_1 \right)$$

$$= \|\mathbf{u}^h\|_{\mathbf{X}_{-1}} \|\mathbf{v}^h\|_{\mathbf{X}_{-1}}. \quad (6.43)$$

This establishes existence and uniqueness of the least-squares solution \mathbf{u}^h . To prove the error estimate we note that (6.40) is a consistent scheme and thus, $B^{-h}(\mathbf{u} - \mathbf{u}^h; \mathbf{v}^h) = 0$ for all $\mathbf{v}^h \in \mathbf{X}^h$. However, the error estimate cannot be established using a standard finite element argument because $B^{-h}(\cdot; \cdot)$ is coercive and continuous only on $\mathbf{X}^h \times \mathbf{X}^h$. Thus, we proceed as follows. Let \mathbf{u}_I^h denote the interpolant of the exact solution \mathbf{u} so that from (6.18) it follows that

$$\|\mathbf{u} - \mathbf{u}_I^h\|_{\mathbf{X}_{-1}} \leq h^{\tilde{d}+1} \|\mathbf{u}\|_{\mathbf{X}_j}.$$

Since

$$\|\mathbf{u} - \mathbf{u}^h\|_{\mathbf{X}_{-1}} \leq \|\mathbf{u} - \mathbf{u}_I^h\|_{\mathbf{X}_{-1}} + \|\mathbf{u}^h - \mathbf{u}_I^h\|_{\mathbf{X}_{-1}}$$

we only need to bound the last term above, which belongs to \mathbf{X}^h :

$$\begin{aligned} \|\mathbf{u}^h - \mathbf{u}_I^h\|_{\mathbf{X}_{-1}}^2 &\leq CB^{-h}(\mathbf{u}^h - \mathbf{u}_I^h; \mathbf{u}^h - \mathbf{u}_I^h) \\ &= CB^{-h}(\mathbf{u}_I^h - \mathbf{u}; \mathbf{u}^h - \mathbf{u}_I^h) \\ &\leq CB^{-h}(\mathbf{u}_I^h - \mathbf{u}; \mathbf{u}_I^h - \mathbf{u})^{1/2} B^{-h}(\mathbf{u}^h - \mathbf{u}_I^h; \mathbf{u}^h - \mathbf{u}_I^h)^{1/2} \\ &\leq CB^{-h}(\mathbf{u}_I^h - \mathbf{u}; \mathbf{u}_I^h - \mathbf{u})^{1/2} \|\mathbf{u}^h - \mathbf{u}_I^h\|_{\mathbf{X}_{-1}}. \end{aligned}$$

Thus,

$$\|\mathbf{u}^h - \mathbf{u}_I^h\|_{\mathbf{X}_{-1}} \leq CB^{-h}(\mathbf{u}_I^h - \mathbf{u}; \mathbf{u}_I^h - \mathbf{u})^{1/2}.$$

To bound the energy norm of $\mathbf{u}_I^h - \mathbf{u} = E$ note that

$$B^{-h}(E; E)^{1/2} \leq C \left(\sum_{i=1}^k \left\| \sum_{j=1}^N \mathcal{L}_{ij} E_j \right\|_{-h} + \sum_{i=k+1}^N \left\| \sum_{j=1}^N \mathcal{L}_{ij} E_j \right\|_0 \right).$$

Using (6.36) for $1 \leq i \leq k$,

$$\begin{aligned} \left\| \sum_{j=1}^N \mathcal{L}_{ij} E_j \right\|_{-h} &\leq \sum_{j=1}^N (h \|\mathcal{L}_{ij} E_j\|_0 + \|\mathcal{L}_{ij} E_j\|_{-1}) \\ &\leq \sum_{j=1}^N (h \|E_j\|_1 + \|E_j\|_0) \\ &\leq h^{\tilde{d}+1} \sum_{j=1}^l \|u_j\|_{\tilde{d}+1} + h^{\tilde{d}+2} \sum_{j=l+1}^N \|u_j\|_{\tilde{d}+2}. \end{aligned}$$

For $k + 1 \leq i \leq N$, we separate terms of orders zero and one:

$$\begin{aligned}
\left\| \sum_{j=1}^N \mathcal{L}_{ij} E_j \right\|_0 &\leq \sum_{j=1}^l \|\mathcal{L}_{ij} E_j\|_0 + \sum_{j=l+1}^N \|\mathcal{L}_{ij} E_j\|_0 \\
&\leq C \left(\sum_{j=1}^l \|E_j\|_0 + \sum_{j=l+1}^N \|E_j\|_1 \right) \\
&\leq Ch^{\bar{d}+1} \left(\sum_{j=1}^l \|u_j\|_{\bar{d}+1} + \sum_{j=l+1}^N \|u_j\|_{\bar{d}+2} \right)
\end{aligned}$$

This establishes (6.41). Lastly, the spectral equivalence between the least-squares discretization matrix A^h and the matrix M follows from the identities

$$\begin{aligned}
(\mathbf{u}^h)^T \mathbf{A} \mathbf{u}^h &= B^{-h}(\mathbf{u}^h; \mathbf{u}^h), \quad (u_j^h)^T D u_j^h = (u_j^h, u_j^h)_1, \\
(u_j^h)^T G u_j^h &= (u_j^h, u_j^h)_0,
\end{aligned}$$

and (6.42). This also implies that $\text{cond}(A) = O(h^{-2})$. \square

Like the weighted method, the negative norm method is based on a DLS principle which does not represent a restriction to \mathbf{X}^h of a CLS principle. As a result, both methods are not conforming in the sense of §5. However, the negative norm functional (6.38) retains the norm equivalence properties of (6.16) for all discrete functions, while the weighted functional does not. As a result, the negative norm method leads to algebraic problems which have better condition numbers and are easier to precondition. Indeed, Theorem 6 shows that (5.21) holds for (6.38) with $\delta_1(h)$ and $\delta_2(h)$ independent of h , i.e., A^h and M are spectrally equivalent. This means that any matrix that is spectrally equivalent to M can be used to precondition A^h . It is easy to see that the Gramm matrix G is spectrally equivalent to $h^2 \mathbf{I}$ so that if T is any good preconditioner for the Poisson equation, the matrix

$$L = \text{diag}(\underbrace{h^2 \mathbf{I}, \dots, h^2 \mathbf{I}}_l, \underbrace{T, \dots, T}_{N-l}) \tag{6.44}$$

can be used for this purpose. On the other hand, the negative norm method is more complicated algorithmically and must be implemented in an assembly-free way because of the density of the matrix A^h . Thus, the possibility to devise efficient preconditioners for A^h is essential for the utility of this method.

6.6 Concluding remarks

In this chapter we demonstrated the use of ADN elliptic theory in the analysis and development of least-squares finite element methods. The principal role of ADN theory was to identify the proper balance (5.3) between solution and residual energies. For ADN elliptic systems this balance is given by (6.4) in Theorem 3 or, equivalently, by (6.5).

Our main focus was on first-order systems because first-order operators are most convenient from practical point of view. In §6.4 we specialized the results of Theorem 3 to such systems and identified two distinctive classes of first-order operators:

- the class of *homogeneous elliptic* first-order operators for which (6.5) is given by (6.11), and
- the class of *non-homogeneous elliptic* first-order operators, for which (6.5) is given by (6.14).

These classes resulted in two substantially different settings for the least-squares method. The *homogeneous elliptic* class leads to well-posed CLS principles $\{\mathbf{X}_q, J(\cdot)\}$ where $J(\cdot)$ is given by (6.15). For $q = 0$ the space \mathbf{X}_0 is a product of $H^1(\Omega)$ spaces and $J(\cdot)$ involves only L^2 -norms of first-order terms. As a result, the DLS principle $\{\mathbf{X}^h, J_h(\cdot)\}$ is merely a *restriction* of the CLS principle to the finite element subspace \mathbf{X}^h .

The *non-homogeneous elliptic* class leads to well-posed CLS principles $\{\mathbf{X}_q, J(\cdot)\}$ where $J(\cdot)$ is now given by (6.16) or (6.17). The second functional contains H^1 -norms; the first - H^{-1} -norms. In both cases these functionals are not practical. We considered two possibilities for transforming $\{\mathbf{X}_q, J(\cdot)\}$ into a practical DLS principle $\{\mathbf{X}^h, J_h(\cdot)\}$.

The first one was to replace (6.17) by a weighted L^2 -norm functional. This leads to *quasi norm-equivalent* DLS principles in which the upper and/or lower bounds in the equivalence relations (5.20) and (5.21) depend on h .

The second possibility was to replace (6.16) by the discrete negative norm functional (6.38). This leads to *norm-equivalent* DLS principles in which the equivalence relations (5.20) and (5.21) hold independently of h .

To summarize, among the main conclusions from this chapter is the observation that

a first-order reformulation will lead to a mathematically well-posed least-squares principle, which at the same time is practical,

if and only if the first-order differential operator is homogeneous elliptic.

Chapter 7

Least-squares for the Stokes and the Navier-Stokes equations

The Stokes equations (2.17), first encountered in §2.2, belong to the class of problems whose solutions can be characterized by constrained optimization of convex, quadratic functional. We also recall that the use of Lagrange multipliers led us to the mixed formulation (2.16). A finite element method based on these weak equations is subject to the restrictive inf-sup condition (2.25).

The nonlinear Navier-Stokes equations (2.37) are, on the other hand, an example of a system which is not associated with optimization problem. Here the weak formulation (2.38)-(2.39) was obtained by formal Galerkin procedure. Nevertheless, finite element methods based on this weak problem are also subject to the inf-sup condition.

As a result, application of least-squares principles for the design of finite element methods for (2.17) and (2.37) is justified. The next step is to apply the methodology developed in Chapter 6 in a manner which will allow one to fully utilize the potential of least-squares in the algorithmic design. From §4.2 in Chapter 4 we know that direct use of the second order system will not, in general, lead to a practical method. Therefore, our first task will be to enlarge the set of potentially practical CLS principles for the Stokes equations by developing a sufficient supply of equivalent first-order formulations. For each one of these formulations we use the ADN theory to identify the settings which verify hypotheses **A.1-A.2** of §5.1. For simplicity we consider homogeneous boundary conditions and assume that solutions spaces

are constrained by the boundary conditions. Therefore, our discussion focuses on proper choices of data spaces $Y(\Omega)$ and solution spaces $X(\Omega)$ which verify **A.1-A.2**. The choice of $Z(\Gamma)$ is only briefly discussed in section 7.2.

Each one of the first-order Stokes formulations can be extended to the nonlinear case in an obvious manner by including an appropriate form of advective term. This extension is not accompanied by introduction of new dependent variables and so we will use the same terms to denote both the linear and the nonlinear first-order systems.

7.1 First-order equations

Transformation of a high-order PDE to a first-order system can be accomplished in many different ways. The original procedure described in [11] (see §6.3) introduces as new dependent variables *all high order derivatives*. While this approach is universal in the sense that it can be applied to any ADN system and will result in an ADN system, it is not necessarily the best one. One reason is that the total number of variables in the new system can increase dramatically. Example 2 shows that transformation can also be effected using *linear combinations* of derivatives. This has the additional advantage of allowing direct approximation of physically meaningful variables represented by such linear combinations, and without a significant increase in the number of dependent variables.

For the Stokes equations (2.17) there exist three general categories of transformations to first-order systems. The first one, which is essentially the approach described in [11], is to use all partial derivatives of the vector valued field \mathbf{u} as new variables, i.e., to set $\underline{\mathbf{U}} = \nabla \mathbf{u}$. Another choice is to use as a new variable the axial vector of the skew-symmetric gradient tensor $\underline{\mathbf{U}} = (\nabla \mathbf{u} - \nabla \mathbf{u}^T)/2$. This variable was introduced in Example 2 and it leads to a vorticity based first-order system. A third choice is to use the symmetric gradient tensor $\underline{\mathbf{U}} = (\nabla \mathbf{u} + \nabla \mathbf{u}^T)/2$. This variable gives rise to stress-based Stokes system.

7.1.1 The velocity-vorticity-pressure equations

The velocity-vorticity-pressure first-order Stokes system and the companion Navier-Stokes formulation are by a wide margin the most popular in the context of least-squares methods for incompressible flows. It was introduced by Jiang and Chang in [98] and then explored by a number of researchers in [99], [100, 101, 102, 103], [104], [54], [55] and [50]. Theoretical analysis was carried by Bochev and Gunzburger [48], [47], and [56, 57].

To state this formulation recall the curl operator in three dimensions and its two-dimensional counterparts

$$\nabla \times \phi = \begin{pmatrix} \phi_y \\ -\phi_x \end{pmatrix} \quad \text{and} \quad \nabla \times \mathbf{u} = u_{2x} - u_{1y}.$$

The context should make clear which operator is relevant.

We also recall that the axial vector of the skew-symmetric part of the velocity gradient is given by $\boldsymbol{\omega} = \nabla \times \mathbf{u}$ and is called *vorticity* vector. Using this vector as a new dependent variable, the vector identity

$$\nabla \times \nabla \times \mathbf{u} = -\Delta \mathbf{u} + \nabla \nabla \cdot \mathbf{u},$$

and in view of the incompressibility constraint $\nabla \cdot \mathbf{u} = 0$ the Stokes equations (2.17) can be cast into the first-order system

$$\nu \nabla \times \boldsymbol{\omega} + \nabla p = \mathbf{f} \quad \text{in } \Omega \quad (7.1)$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega \quad (7.2)$$

$$\nabla \times \mathbf{u} - \boldsymbol{\omega} = \mathbf{0} \quad \text{in } \Omega \quad (7.3)$$

along with the velocity boundary condition

$$\mathbf{u} = \mathbf{0} \quad \text{on } \Gamma \quad (7.4)$$

and the zero mean pressure constraint (2.18).

In two dimensions, the system (7.1)-(7.3) contains four equations and four unknowns and is uniformly elliptic of total order four. In three dimensions, the number of equations and unknowns is seven, and the resulting system is not elliptic in the sense of ADN. By adding the redundant equation

$$\nabla \cdot \boldsymbol{\omega} = 0 \quad \text{in } \Omega \quad (7.5)$$

and the gradient of a “slack” variable ϕ to (7.3):

$$\nabla \times \mathbf{u} - \boldsymbol{\omega} + \nabla \phi = \mathbf{0} \quad \text{in } \Omega, \quad (7.6)$$

uniform ellipticity can be restored; see [75]. The augmented system (7.1), (7.2), (7.5), and (7.6) has total order eight, in contrast to the total order of the Stokes problem in primitive variables which is six in three dimensions. It should be noted that one also imposes homogeneous boundary conditions for the slack variable ϕ and that one can then show that $\phi \equiv 0$ so that, a posteriori, (7.6) is identical to (7.3). In fact, the addition of ϕ is needed

only for the purpose of analyses; it is not needed in the development or implementation of least-squares based algorithms for which one can safely use the system (7.1)-(7.5). However, the addition of (7.5) is crucial to the stability and accuracy of least-squares finite element methods for the Stokes problem in three dimensions.

To extend the velocity-vorticity-pressure formulation to the Navier-Stokes equations, one has to choose a particular form for the nonlinear term in (2.37). One possibility is to keep the nonlinear term in a form involving only the velocity field, i.e., to replace (7.1) by

$$\nu \nabla \times \boldsymbol{\omega} + \mathbf{u} \cdot \nabla \mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega. \quad (7.7)$$

Another possibility is to use the vector identity

$$\mathbf{u} \cdot \nabla \mathbf{u} = \frac{1}{2} \nabla |\mathbf{u}|^2 - \mathbf{u} \times \nabla \times \mathbf{u} = \frac{1}{2} \nabla |\mathbf{u}|^2 - \mathbf{u} \times \boldsymbol{\omega}$$

to replace (7.1) by

$$\nu \nabla \times \boldsymbol{\omega} + \boldsymbol{\omega} \times \mathbf{u} + \nabla P = \mathbf{f} \quad \text{in } \Omega, \quad (7.8)$$

where $P = p + 1/2|\mathbf{u}|^2$ denotes the total pressure.

Our ultimate goal is to use equations (7.1)-(7.3) to set up a least-squares principle for the Stokes equations. Thus, we turn attention to the abstract framework of §5.1, and especially the verification of the two hypothesis **A.1.** and **A.2.**. For simplicity, we restrict attention to the case of two space dimensions; most of the relevant results can be easily extended to the augmented system, i.e., including (7.5), in three dimensions.

Let us recall that the relevance of **A.1** stems from the fact that this hypothesis implies the correct “energy balance” (5.3) for the least-squares principle. In other words, **A.1** allows us to determine both the correct artificial least-squares energy functional, and the appropriate minimization space for this functional. To determine functional settings in which elliptic boundary value problems are well-posed, we will rely on the elliptic regularity theory of Agmon, Douglis and Nirenberg [11]. From Chapter 6 we know that well-posed problems are characterized as having uniformly elliptic principal parts and boundary conditions which satisfy the celebrated *complementing condition*. One advantage of this approach for finding the appropriate function setting in **A.1** is that ADN theory allows one to treat in a systematic way different choices of boundary conditions. As we shall see, the choice of boundary conditions has great importance to the validity of a priori estimates. At the same time, direct methods do not allow for

a unified treatment of several boundary conditions. Because of the rather complex nature of ADN theory here we present only a summary of results from the analysis. The technical details that accompany verification of **A.1** are summarized for the convenience of the reader in Appendix A. More details can be found in [56], [47] and [48].

From Example 2 we know that the velocity-vorticity-pressure Stokes problem admits two different principal parts given by

$$\mathcal{L}_1^p = \begin{pmatrix} \nu \nabla \times \boldsymbol{\omega} & + & \nabla p \\ & & \nabla \times \mathbf{u} \\ & & \nabla \cdot \mathbf{u} \end{pmatrix} \quad (7.9)$$

and

$$\mathcal{L}_2^p = \begin{pmatrix} \nu \nabla \times \boldsymbol{\omega} & + & \nabla p \\ -\boldsymbol{\omega} & + & \nabla \times \mathbf{u} \\ & & \nabla \cdot \mathbf{u} \end{pmatrix}, \quad (7.10)$$

respectively. In view of the boundary condition (7.4) and the zero mean condition (2.18), the function spaces in (6.3) corresponding to these principal parts specialize to

$$\mathbf{X}_q = H^{q+1}(\Omega) \times H^{q+1}(\Omega) \cap L_0^2(\Omega) \times \mathbf{H}^{q+1}(\Omega) \cap \mathbf{H}_0^1(\Omega) \quad (7.11)$$

and

$$\mathbf{Y}_q = H^q(\Omega) \times H^q(\Omega) \times \mathbf{H}^q(\Omega) \quad (7.12)$$

for (7.9) and

$$\mathbf{X}_q = H^{q+1}(\Omega) \times H^{q+1}(\Omega) \cap L_0^2(\Omega) \times \mathbf{H}^{q+2}(\Omega) \cap \mathbf{H}_0^1(\Omega) \quad (7.13)$$

and

$$\mathbf{Y}_q = H^q(\Omega) \times H^q(\Omega) \times \mathbf{H}^{q+1}(\Omega) \quad (7.14)$$

for (7.10). We recall that \mathbf{X}_q denotes the function space for the unknowns $(\boldsymbol{\omega}, p, \mathbf{u})$ and \mathbf{Y}_q denotes the function space for the data or equation residuals. Furthermore, since the pressure zero mean constraint (2.18) is imposed on the pressure space component in (7.11) and (7.13), the uniqueness of the solutions is guaranteed. As a result, the two a priori bounds (6.11) and (6.14) specialize to

$$\begin{aligned} & \|\boldsymbol{\omega}\|_{q+1} + \|p\|_{q+1} + \|\mathbf{u}\|_{q+1} \\ & \leq C (\|\nu \nabla \times \boldsymbol{\omega} + \nabla p\|_q + \|\nabla \times \mathbf{u} - \boldsymbol{\omega}\|_q + \|\nabla \cdot \mathbf{u}\|_q) \end{aligned} \quad (7.15)$$

and

$$\begin{aligned} & \|\boldsymbol{\omega}\|_{q+1} + \|p\|_{q+1} + \|\mathbf{u}\|_{q+2} \\ & \leq C (\|\nu \nabla \times \boldsymbol{\omega} + \nabla p\|_q + \|\nabla \times \mathbf{u} - \boldsymbol{\omega}\|_{q+1} + \|\nabla \cdot \mathbf{u}\|_{q+1}), \end{aligned} \quad (7.16)$$

respectively. In the context of §5.2, the a priori bounds (7.15) and (7.16) represent the least-squares energy balance (5.3). Likewise, the solution and data space pairs (7.11)-(7.12) and (7.13)-(7.14) provide the proper energy balance between residual energy and solution energy. Note also that the setting provided by (7.11)-(7.12) and (7.15) corresponds to a homogeneous elliptic problem. In contrast, the setting of (7.13)-(7.14) and (7.16) describes a non-homogeneous elliptic system.

Although both principal parts (7.9) and (7.10) are uniformly elliptic operators of total order four, not all boundary conditions for the system (7.1)-(7.3) will satisfy the complementing condition for both principal parts. For example, the boundary condition (7.4) on the velocity vector satisfies the complementing condition¹ only with the principal part (7.10). As a result, the a priori estimate for the system (7.1)-(7.3), (7.4), and (2.18) relevant to the least-squares methods, is given by (7.16). In fact, one can show that the estimate (7.15) cannot hold with the velocity boundary condition; see Example 3 in Appendix A.

An example of a boundary condition for which (7.15) is valid is provided by the *pressure-normal velocity* boundary condition

$$p = 0 \quad \text{and} \quad \mathbf{u} \cdot \mathbf{n} = 0 \quad \text{on } \Gamma. \quad (7.17)$$

The fact that (7.16) is not valid for velocity boundary conditions indicates that the corresponding boundary value problem is not well-posed in the spaces (7.11)-(7.12). This can also be seen by considering the principal part (7.9) along with the velocity boundary condition. The corresponding boundary value problem then uncouples into two ill-posed problems given by

$$\left\{ \begin{array}{l} \nu \nabla \times \boldsymbol{\omega} + \nabla p = \mathbf{f} \end{array} \right\} \quad \text{and} \quad \left\{ \begin{array}{l} \nabla \times \mathbf{u} = 0 \\ \nabla \cdot \mathbf{u} = 0 \\ \mathbf{u}|_{\Gamma} = \mathbf{0} \end{array} \right\};$$

the first is underdetermined and the second is overdetermined. In contrast, the same principal part with (7.17) uncouples into two well-posed problems:

$$\left\{ \begin{array}{l} \nu \nabla \times \boldsymbol{\omega} + \nabla p = \mathbf{f} \\ p|_{\Gamma} = 0 \end{array} \right\} \quad \text{and} \quad \left\{ \begin{array}{l} \nabla \times \mathbf{u} = 0 \\ \nabla \cdot \mathbf{u} = 0 \\ \mathbf{u} \cdot \mathbf{n}|_{\Gamma} = 0 \end{array} \right\}.$$

¹This is shown in detail in Appendix A.

We can conclude that

- the velocity-vorticity-pressure system satisfies **A.1** with two distinctively different functional settings;
- These settings are described by the solution and data space combinations given by the pairs (7.11)-(7.12) and (7.13)-(7.14), respectively;
- Validity of a specific setting depends critically on the particular set of boundary conditions;
- the use of the homogeneous elliptic setting (7.11)-(7.12) or the inhomogeneous elliptic setting (7.13)-(7.14) to define a CLSP will depend on the boundary condition.

What is even more striking, there are examples of boundary conditions for which the setting, that is validity of either (7.15) or (7.16) changes with the space dimension. Table 7.1.1, taken from [48], gives a list of boundary conditions classified according to the ellipticity setting for the velocity-vorticity-pressure equations. In Table 7.1.1, *Type 1* refers to boundary conditions for which (7.15) is valid; *Type 2* denotes boundary conditions for which (7.16) holds. Consider for instance the tangential velocity-pressure boundary condition

$$\mathbf{n} \times \mathbf{u} \times \mathbf{n} = \mathbf{0} \quad \text{and} \quad p = 0 \quad \text{on } \Gamma.$$

In two dimensions, this boundary operator satisfies the complementing condition with either of the principal parts (7.9) or (7.10), whereas in three dimensions it satisfies the same condition only with the principal part (7.10). As a result, the estimate (7.15) is valid only in two dimensions.

7.1.2 The velocity-pressure-stress equations

A first-order system with substantially different properties is obtained when the stress tensor scaled by $\sqrt{\nu/2}$

$$\underline{\mathbf{T}} = \sqrt{2\nu}\epsilon(\mathbf{u}), \quad \text{where} \quad \epsilon(\mathbf{u}) \equiv \frac{1}{2}(\nabla\mathbf{u} + (\nabla\mathbf{u})^T),$$

is used in the transformation of (2.17) into a first-order system. Here, the relevant vector identity is given by

$$\nabla \cdot \underline{\mathbf{T}} = \sqrt{2\nu}(\Delta\mathbf{u} + \nabla\nabla \cdot \mathbf{u}),$$

where $\nabla \cdot \underline{\mathbf{T}}$ denotes the vector whose components are the divergences of the corresponding rows of $\underline{\mathbf{T}}$. Then, in view of incompressibility constraint,

Table 7.1: Classification of boundary conditions for the Stokes and Navier-Stokes equations: velocity-vorticity-pressure formulation.

Boundary conditions		\mathcal{R}^3	\mathcal{R}^2	Type
BC1	Velocity	\mathbf{u}	\mathbf{u}	2
	Slack variable	ϕ	-	
BC1A	Velocity	\mathbf{u}	\mathbf{u}	2
	Normal vorticity	$\boldsymbol{\omega} \cdot \mathbf{n}$	-	
BC2	Normal velocity	$\mathbf{u} \cdot \mathbf{n}$	$\mathbf{u} \cdot \mathbf{n}$	1
	Normal vorticity	$\boldsymbol{\omega} \cdot \mathbf{n}$	-	
	Pressure	r	r	
	Slack variable	ϕ	-	
BC2A	Normal velocity	$\mathbf{u} \cdot \mathbf{n}$	$\mathbf{u} \cdot \mathbf{n}$	1
	Tangential vorticity	$\mathbf{n} \times \boldsymbol{\omega} \times \mathbf{n}$	$\boldsymbol{\omega}$	
	Slack variable	ϕ	-	
BC2B	Normal velocity	$\mathbf{u} \cdot \mathbf{n}$	$\mathbf{u} \cdot \mathbf{n}$	not well-posed (r is redundant in \mathcal{R}^2)
	Tangential vorticity	$\mathbf{n} \times \boldsymbol{\omega} \times \mathbf{n}$	$\boldsymbol{\omega}$	
	Pressure	r	r	
BC2C	Normal velocity	$\mathbf{u} \cdot \mathbf{n}$	$\mathbf{u} \cdot \mathbf{n}$	not well-posed in \mathcal{R}^3 1 in \mathcal{R}^2
	Vorticity	$\boldsymbol{\omega}$	$\boldsymbol{\omega}$	
BC3	Tangential velocity	$\mathbf{n} \times \mathbf{u} \times \mathbf{n}$	$\mathbf{u} \cdot \mathbf{t}$	2 in \mathcal{R}^3 1 in \mathcal{R}^2
	Pressure	r	r	
	Slack variable	ϕ	-	
BC3A	Tangential velocity	$\mathbf{n} \times \mathbf{u} \times \mathbf{n}$	$\mathbf{u} \cdot \mathbf{t}$	1
	Normal vorticity	$\boldsymbol{\omega} \cdot \mathbf{n}$	-	
	Pressure	r	r	
BC3B	Tangential velocity	$\mathbf{n} \times \mathbf{u} \times \mathbf{n}$	$\mathbf{u} \cdot \mathbf{t}$	not well-posed
	Normal vorticity	$\boldsymbol{\omega} \cdot \mathbf{n}$	-	
	Slack variable	ϕ	-	
BC3C	Tangential velocity	$\mathbf{n} \times \mathbf{u} \times \mathbf{n}$	$\mathbf{u} \cdot \mathbf{t}$	1
	Tangential vorticity	$\mathbf{n} \times \boldsymbol{\omega} \times \mathbf{n}$	$\boldsymbol{\omega}$	
BC4	Vorticity	$\boldsymbol{\omega}$	$\boldsymbol{\omega}$	not well-posed
	Pressure	r	r	
BC4A	Vorticity	$\boldsymbol{\omega}$	$\boldsymbol{\omega}$	not well-posed
	Slack variable	ϕ	-	
BC5	Tangential vorticity	$\mathbf{n} \times \boldsymbol{\omega} \times \mathbf{n}$	$\boldsymbol{\omega}$	not well-posed
	Pressure	r	r	
	Slack variable	ϕ	-	

the system (2.17) and (7.4) can be replaced by the *velocity-pressure-stress* system

$$\begin{aligned}
\sqrt{2\nu} \nabla \cdot \underline{\mathbf{T}} - \nabla p &= \mathbf{f} && \text{in } \Omega \\
\nabla \cdot \mathbf{u} &= 0 && \text{in } \Omega \\
\underline{\mathbf{T}} - \sqrt{2\nu} \boldsymbol{\epsilon}(\mathbf{u}) &= \underline{\mathbf{0}} && \text{in } \Omega \\
\mathbf{u} &= \mathbf{0} && \text{on } \Gamma.
\end{aligned} \tag{7.18}$$

The inclusion of the nonlinear term $\mathbf{u} \cdot \nabla \mathbf{u}$ into the first equation of (7.18) provides an extension of the velocity-pressure-stress system to the Navier-Stokes equations. As before, uniqueness of solutions to (7.18) can be guaranteed by imposing the zero mean constraint (2.18) on the pressure space.

Again, our main goal is to find settings in which **A.1-A.2** hold, so as to establish the proper theoretical setting for least-squares principles based on (7.18). As we saw in the last section this task can be effectively accomplished by the ADN elliptic theory. Here we use again the same approach. For the technical details the reader is referred to §A.2 in Appendix A or [58].

In two dimensions, the velocity-pressure-stress system has six equations and unknowns. In three dimensions, the number of unknowns and equations increases to ten. It can be shown that in 2D the principal part of (7.18) is given by the differential operator

$$\mathcal{L}^P U = \begin{pmatrix} T_1 - \sqrt{2\nu} \frac{\partial u_1}{\partial x} \\ 2T_2 - \sqrt{2\nu} \left(\frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right) \\ T_3 - \sqrt{2\nu} \frac{\partial u_2}{\partial y} \\ \frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y} \\ \sqrt{2\nu} \left(\frac{\partial T_1}{\partial x} + \frac{\partial T_2}{\partial y} \right) - \frac{\partial p}{\partial x} \\ \sqrt{2\nu} \left(\frac{\partial T_2}{\partial x} + \frac{\partial T_3}{\partial y} \right) - \frac{\partial p}{\partial y} \end{pmatrix}, \quad (7.19)$$

that is

$$\mathcal{L}^P = \mathcal{L}.$$

In contrast to the velocity-vorticity-pressure equations, the principal part (7.19) is unambiguously defined, and the total order of (7.18) coincides with the total order of the Stokes problem in primitive variables in both two and three dimensions; see §A.2.

The functional setting that provides verification of hypothesis **A.1** in two dimensions for the problem (7.18) is given by

$$\mathbf{X}_q = [H^{q+1}(\Omega)]^3 \times H^{q+1}(\Omega) \cap L_0^2(\Omega) \times [H^{q+2}(\Omega) \cap H_0^1(\Omega)]^2$$

for the unknowns $(\mathbf{T}, p, \mathbf{u})$ and

$$\mathbf{Y}_q = [H^q(\Omega)]^2 \times H^{q+1}(\Omega) \times [H^{q+1}(\Omega)]^3$$

for the data or equation residuals. As a result, the a priori estimate (6.14) now specializes to

$$\begin{aligned} & \|\mathbf{T}\|_{q+1} + \|p\|_{q+1} + \|\mathbf{u}\|_{q+2} \\ & \leq C \left(\|\sqrt{2\nu} \nabla \cdot \mathbf{T} - \nabla p\|_q + \|\nabla \cdot \mathbf{u}\|_{q+1} + \|\mathbf{T} - \sqrt{2\nu} \epsilon(\mathbf{u})\|_{q+1} \right). \end{aligned} \quad (7.20)$$

Note that the estimate (7.20) implies that regardless of the choice of boundary operators, the system (7.18) is not of Petrovsky type, i.e., it cannot be homogeneous elliptic.

7.1.3 Velocity gradient-based transformations

From the specializations of the energy balance (5.3) that we have encountered so far only (7.15) does not require combinations of different order spaces for the solution and the data. This was due to the fact that only the setting (7.11)-(7.12) corresponds to a homogeneous elliptic system. From Theorem 4 in §6.5.1 we also know that among all first-order ADN systems, homogeneous elliptic systems are the most appealing from a least-squares point of view. To recall, this was due to the fact that

solution energy can be measured in the L^2 -norm;

and in combination with the fact that only first-order derivatives appear in the equations it means that

only standard, C^0 finite element spaces are required.

Because lower and upper bounds in the energy balance (6.11) for homogeneous elliptic systems are given in terms of the H^1 -norm, in the least-squares literature such formulations are often called H^1 -coercive. A further advantage of least-squares methods based on H^1 -coercive systems is that the algebraic equations can be solved by efficient multilevel techniques; see [66].

However, neither velocity-vorticity-pressure system nor the velocity - pressure - stress equations possess this property, at least for the practically important velocity boundary condition. It turns out that in order to define a first-order form of the Stokes equations which at the same time is homogeneous elliptic, one has to introduce $\nabla \mathbf{u}$ as a new dependent variable and then augment the differential equations by a number of *compatibility conditions*. We call such systems *velocity gradient-based*. Essentially, the transformations we are about to discuss follow the original recipe of [11] in which all higher order derivatives are used as new variables. The point at which they depart from this recipe is the use of redundant relations to augment the equations until the new system becomes homogeneous elliptic.

Velocity gradient-velocity-pressure equations

To define the first *velocity gradient-velocity-pressure formulation*, one introduces *all* first derivatives of the velocity components as new dependent

variables, i.e., we set $\underline{\mathbf{U}} = (\nabla \mathbf{u})^t$ so that $V_{ij} = (\partial u_i / \partial x_j)$. In terms of $\underline{\mathbf{U}}$, the Stokes problem (2.17) is given by

$$-\nu \nabla \cdot \underline{\mathbf{U}} + \nabla p = \mathbf{f} \quad \text{in } \Omega \quad (7.21)$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega \quad (7.22)$$

$$\underline{\mathbf{U}} - (\nabla \mathbf{u})^t = \underline{\mathbf{0}} \quad \text{in } \Omega, \quad (7.23)$$

and (7.4). In (7.21) $\nabla \cdot \underline{\mathbf{U}}$ denotes the vector whose components are the divergences of the corresponding rows of $\underline{\mathbf{U}}$. The system (7.21)-(7.23) and (7.4) is not fully H^1 -coercive. It can be shown; see [66], [67] that the energy balance (5.3) for this system is

$$\begin{aligned} & \|\underline{\mathbf{U}}\|_{q+1} + \|\mathbf{u}\|_{q+2} + \|p\|_{q+1} \\ & \leq C (\|-\nu \nabla \cdot \underline{\mathbf{U}} + \nabla p\|_q + \|\underline{\mathbf{U}} - (\nabla \mathbf{u})^t\|_{q+1} + \|\nabla \cdot \mathbf{u}\|_{q+1}). \end{aligned} \quad (7.24)$$

In [67], the new variables $\underline{\mathbf{U}}$ are called ‘‘velocity fluxes;’’ since that terminology is usually reserved for a different physical quantity, we prefer the term ‘‘velocity gradient.’’

The main idea of [67] is that full H^1 -coercivity can be obtained by augmenting (7.21)-(7.23) with additional constraints. In particular, in view of the identity $tr \underline{\mathbf{U}} = \nabla \cdot \mathbf{u}$, the definition of $\underline{\mathbf{U}}$, and the boundary condition (7.4), one can add to (7.21)-(7.23) the equations

$$\nabla(tr \underline{\mathbf{U}}) = \mathbf{0} \quad \text{in } \Omega \quad (7.25)$$

and

$$\nabla \times \underline{\mathbf{U}} = \underline{\mathbf{0}} \quad \text{in } \Omega \quad (7.26)$$

and the boundary condition

$$\underline{\mathbf{U}} \times \mathbf{n} = \underline{\mathbf{0}} \quad \text{on } \Gamma, \quad (7.27)$$

where $\nabla \times \underline{\mathbf{U}}$ denotes the vector whose components are the curls of the corresponding rows of $\underline{\mathbf{U}}$ and $\underline{\mathbf{U}} \times \mathbf{n}$ denotes the vector whose components are the vector product of the rows of $\underline{\mathbf{U}}$ with the unit outer normal vector \mathbf{n} .

The resulting system (7.21)-(7.23) and (7.25)-(7.27) is overdetermined, but consistent. In two-dimensions, the number of unknowns equals seven, and the number of equations equals eleven. In three-dimensions, we have thirteen unknowns and twenty five equations². The system (7.21)-(7.23)

²Strictly speaking this means that the ADN theory cannot be applied directly to the overdetermined system. To apply this theory it is first necessary to augment the equations by one or more slack variables in a manner similar to the one described in §7.1.1. The slack variables are only needed for the analyses and can be safely omitted from computations.

augmented with (7.25)-(7.27) now admits a functional setting in which hypothesis **A.1.** is satisfied for

$$\mathbf{X}_q = \tilde{\mathbf{H}}^{q+1}(\Omega) \times \mathbf{H}^{q+1}(\Omega) \cap \mathbf{H}_0^1(\Omega) \times H^{q+1}(\Omega) \cap L_0^2(\Omega)$$

for the unknowns $(\underline{\mathbf{U}}, \mathbf{u}, p)$, where $\tilde{\mathbf{H}}^{q+1}(\Omega) = [H^{q+1}(\Omega)]^{n^2}$ constrained by (7.27), and

$$\mathbf{Y}_q = \mathbf{H}^q(\Omega) \times H^q(\Omega) \times [H^q(\Omega)]^{n^2} \times \mathbf{H}^q(\Omega) \times \mathbf{H}^q(\Omega)$$

for the equation residuals. The energy balance (6.5) for the augmented system specializes to

$$\begin{aligned} \|\underline{\mathbf{U}}\|_{q+1} + \|\mathbf{u}\|_{q+1} + \|p\|_{q+1} \leq C \left(\|\nu \nabla \cdot \underline{\mathbf{U}} + \nabla p\|_q + \|\nabla \cdot \mathbf{u}\|_q \right. \\ \left. + \|\underline{\mathbf{U}} - (\nabla \mathbf{u})^t\|_q + \|\nabla(\text{tr} \underline{\mathbf{U}})\|_q + \|\nabla \times \underline{\mathbf{U}}\|_q \right). \end{aligned} \quad (7.28)$$

The two velocity gradient-velocity-pressure formulations can be easily extended to the Navier-Stokes equations. In terms of the new variable $\underline{\mathbf{U}}$, the nonlinear term in (2.37) can be expressed as $\underline{\mathbf{U}} \cdot \mathbf{u}$ so that, for the Navier-Stokes problem, (7.21) is replaced by

$$-\nu \nabla \cdot \underline{\mathbf{U}} + \underline{\mathbf{U}} \cdot \mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega.$$

The constrained velocity gradient-pressure equations

This approach was suggested in [78] and here we present it in the case of two space dimensions. The new variables introduced to effect the transformation to a first-order system are the entries of the velocity gradient constrained by the incompressibility constraint $\nabla \cdot \mathbf{u} = 0$, i.e., they are given by

$$\underline{\mathbf{G}} = \begin{pmatrix} v_1 & v_2 \\ v_3 & -v_1 \end{pmatrix} \quad (7.29)$$

where

$$v_1 = \frac{\partial u_1}{\partial x_1} = -\frac{\partial u_2}{\partial x_2}, \quad v_2 = \frac{\partial u_1}{\partial x_2}, \quad \text{and} \quad v_3 = \frac{\partial u_2}{\partial x_1},$$

and where u_1 and u_2 denote the components of the velocity \mathbf{u} . Using the new variables and the equality of second mixed derivatives, the Stokes problem (2.17) in two dimensions can be written in the form (see [78])

$$\begin{aligned} -\nu \nabla \cdot \underline{\mathbf{G}} + \nabla p &= \mathbf{f} \quad \text{in } \Omega \\ \nabla \times \underline{\mathbf{G}} &= \mathbf{0} \quad \text{in } \Omega \\ \underline{\mathbf{G}} \times \mathbf{n} &= \mathbf{0} \quad \text{on } \Gamma. \end{aligned} \quad (7.30)$$

In [78], the new variables (7.29) are called “accelerations” and the system (7.30) the “acceleration-velocity” formulation of the Stokes equations. However, the new variables are not components of the acceleration vector so that, instead, we call the system (7.30) the *constrained velocity gradient-pressure* formulation of the Stokes problem.

The planar system (7.30) has four equations and four unknowns, and one can show that it is elliptic in the sense of Petrovsky so that, with $\tilde{\mathbf{H}}^{q+1}(\Omega) = [H^{q+1}(\Omega)]^3$ constrained by the boundary condition in (7.30), hypothesis **A.1** holds with

$$\mathbf{X}_q = \tilde{\mathbf{H}}^{q+1}(\Omega) \times H^{q+1}(\Omega) \cap L_0^2(\Omega) \quad \text{and} \quad \mathbf{Y}_q = [H^q(\Omega)]^2 \times [H^q(\Omega)]^2$$

for the unknowns (\mathbf{G}, p) and the equation residuals, respectively. The energy balance (6.5) for this functional setting specializes to

$$\|\underline{\mathbf{G}}\|_{q+1} + \|p\|_{q+1} \leq C (\|-\nu \nabla \cdot \underline{\mathbf{G}} + \nabla p\|_q + \|\nabla \times \underline{\mathbf{G}}\|_q). \quad (7.31)$$

The velocity has been eliminated from (7.30); it is recovered by solving the additional div-curl system

$$\begin{aligned} \nabla \times \mathbf{u} &= v_3 - v_2 \quad \text{in } \Omega \\ \nabla \cdot \mathbf{u} &= 0 \quad \text{in } \Omega \\ \mathbf{u} \cdot \mathbf{n} &= 0 \quad \text{on } \Gamma \end{aligned} \quad (7.32)$$

Although it is not obvious that the solution of (7.32) satisfies the boundary condition (7.4), it can be shown that this is indeed the case.

Although the system (7.30) is H^1 -coercive, owing to the elimination of the velocity field, this system cannot be extended to the Navier-Stokes equations. Elimination of the velocity field in (7.30) can be considered as an artifact since one can simply consider (7.30) together with (7.32). Such a first-order system is studied in [83], where the new variables are called “stresses” and the corresponding first-order system is called the “stress-velocity-pressure” Stokes system despite the fact that the new variables are not the components of the stress tensor. This is not to be confused with the formulation of §7.1.2 for which the true stresses are used.

7.1.4 First-order formulations: concluding remarks

In sections 7.1.1–7.1.3 we presented five different first-order systems that can be derived from the Stokes equations by introducing new dependent variables. In all five cases the new variables involve derivatives of the velocity

field. When new variables represent linear combinations of these derivatives, such as the vorticity or stresses, resulting systems are not always H^1 -coercive. This is due to the fact that interdependencies between the new variables and the velocity field remain coupled, i.e., formulations “remember” that some of the variables are actually velocity derivatives. To uncouple the variables and obtain homogeneous elliptic systems, the velocity-gradient and the constrained velocity-gradient approaches use the components of the velocity gradient as new dependent variables, and add new constraints until the dependencies between the variables become subdominant. This may lead to an overdetermined, but consistent, problem.

7.2 Inhomogeneous boundary conditions

Here we briefly discuss the proper choice of the boundary data space $Z(\Gamma)$ for the first-order Stokes systems presented above. This space is required if one wishes to set up a CLS principle in which boundary conditions are enforced weakly instead of being imposed on the solution space $X(\Omega)$.

Recall that the velocity-vorticity-pressure Stokes problem has an ambiguously defined principal part and that, as a result, there are two possible functional settings that verify the hypotheses of §5.1. These two settings are given by (7.11)-(7.12) and (7.13)-(7.14), respectively. When this problem is augmented with inhomogeneous boundary conditions, the data spaces are given by $\mathbf{Y}_q \times \mathbf{Z}_q$, where \mathbf{Z}_q is a trace space defined on Γ . The specific form of \mathbf{Z}_q then can be determined from (6.4) in Theorem 3. Of course, given a particular boundary operator, the form of \mathbf{Z}_q will depend on the principal part which verifies the complementing condition for this boundary operator. For example, the velocity-vorticity-pressure formulation of the Stokes equations with the pressure-normal velocity boundary condition (7.17) is homogeneous elliptic and this space is given by

$$\mathbf{Z}_q = H^{q+1/2}(\Gamma) \times H^{q+1/2}(\Gamma)$$

for $(\mathbf{u} \cdot \mathbf{n}, p)$, whereas for the velocity boundary condition (7.4) we have that

$$\mathbf{Z}_q = [H^{q+3/2}(\Gamma)]^n, \quad n = 2 \text{ or } 3,$$

for \mathbf{u} . As a result, the relevant a priori estimates corresponding to the two principal parts (7.9) and (7.10) are now given by

$$\|\omega\|_{q+1} + \|p\|_{q+1} + \|\mathbf{u}\|_{q+1} \tag{7.33}$$

$$\begin{aligned} \leq & C \left(\|\nu \nabla \times \omega + \nabla p\|_q + \|\nabla \times \mathbf{u} - \omega\|_q + \|\nabla \cdot \mathbf{u}\|_q \right. \\ & \left. + \|\mathbf{u} \cdot \mathbf{n}\|_{q+1/2, \Gamma} + \|p\|_{q+1/2, \Gamma} \right) \end{aligned}$$

and

$$\begin{aligned} & \|\omega\|_{q+1} + \|p\|_{q+1} + \|\mathbf{u}\|_{q+2} \tag{7.34} \\ \leq & C \left(\|\nu \nabla \times \omega + \nabla p\|_q + \|\nabla \times \mathbf{u} - \omega\|_{q+1} + \|\nabla \cdot \mathbf{u}\|_{q+1} \right. \\ & \left. + \|\mathbf{u}\|_{q+3/2, \Gamma} \right), \end{aligned}$$

respectively. A priori estimates for other first-order Stokes problems with inhomogeneous boundary conditions can be derived in a similar manner. For example, when the first-order Stokes problem is H^1 -coercive, e.g., the velocity gradient-velocity-pressure formulation, the space \mathbf{Z}_q for the inhomogeneous velocity boundary condition is given by $[H^{q+1/2}(\Gamma)]^n$, $n = 2$ or 3 . If the system is not H^1 -coercive, e.g., the velocity-stress-pressure formulation, then \mathbf{Z}_q is given by $[H^{q+3/2}(\Gamma)]^n$, $n = 2$ or 3 .

Therefore, (7.33) and (7.34) provide the energy balance (5.3) for least-squares principles in which essential boundary conditions are enforced variationally. In particular, these estimates indicate the appropriate norms that should be used to measure the energy of the boundary data.

In conclusion, we note that the Agmon-Douglis-Nirenberg theory also allows one to determine the form of the boundary data space \mathbf{Z}_q when the boundary condition involves differential operators. Such boundary conditions for the Stokes problem are, however, outside the scope of these notes.

7.3 Least-squares methods

Each one of the first-order systems considered in §§7.1.1–7.1.3 leads to a continuous least-squares principle (CLSP) by virtue of functional settings that verify hypotheses **A.1-A.2** in §5.2. All systems, that is the velocity-vorticity-pressure (7.1)-(7.3), the velocity-pressure stress (7.18) and the velocity - gradient equations (7.21)-(7.23), or (7.30) are first-order ADN systems³. As a result, least-squares methods for the Stokes equations based on these systems can be developed according to the approach described in

³With the possible addition of slack variables whenever the original first-order system is overdetermined.

Chapter 6. In particular, (5.4) specializes to the following least-squares functionals for the Stokes equations with the velocity boundary condition (7.4):

Velocity-vorticity-pressure functional:

$$J(\boldsymbol{\omega}, \mathbf{u}, p) = \frac{1}{2} \left(\|\nu \nabla \times \boldsymbol{\omega} + \nabla p - \mathbf{f}\|_q^2 + \|\nabla \times \mathbf{u} - \boldsymbol{\omega}\|_{q+1}^2 + \|\nabla \cdot \mathbf{u}\|_{q+1}^2 \right) \quad (7.35)$$

Velocity-pressure-stress functional:

$$J(\underline{\mathbf{T}}, \mathbf{u}, p) = \frac{1}{2} \left(\|\underline{\mathbf{T}} - \sqrt{2\nu} \epsilon(\mathbf{u})\|_{q+1}^2 + \|\nabla \cdot \mathbf{u}\|_{q+1}^2 + \|\sqrt{2\nu} \nabla \cdot \underline{\mathbf{T}} - \nabla p - \mathbf{f}\|_q^2 \right) \quad (7.36)$$

Constrained velocity gradient-pressure functional:

$$J(\underline{\mathbf{G}}, p) = \frac{1}{2} \left(\|\nu \nabla \cdot \underline{\mathbf{G}} + \nabla p - \mathbf{f}\|_q^2 + \|\nabla \times \underline{\mathbf{G}}\|_q^2 \right) \quad (7.37)$$

Velocity gradient-velocity-pressure functional I:

$$J(\underline{\mathbf{U}}, \mathbf{u}, p) = \frac{1}{2} \left(\|\nu \nabla \cdot \underline{\mathbf{U}} + \nabla p - \mathbf{f}\|_q^2 + \|\nabla \cdot \mathbf{u}\|_{q+1}^2 + \|\underline{\mathbf{U}} - (\nabla \mathbf{u})^t\|_{q+1}^2 \right) \quad (7.38)$$

Velocity gradient-velocity-pressure functional II:

$$J(\underline{\mathbf{U}}, \mathbf{u}, p) = \frac{1}{2} \left(\|\nu \nabla \cdot \underline{\mathbf{U}} + \nabla p - \mathbf{f}\|_q^2 + \|\nabla \cdot \mathbf{u}\|_q^2 + \|\underline{\mathbf{U}} - \nabla \mathbf{u}^t\|_q^2 + \|\nabla(\text{tr} \underline{\mathbf{U}})\|_q^2 + \|\nabla \times \underline{\mathbf{U}}\|_q^2 \right). \quad (7.39)$$

Only (7.37) and (7.39) are based on homogeneous elliptic first-order systems, i.e., only these functionals are H^1 -norm equivalent. Therefore, least-squares methods based on these two functionals can be developed according to §6.5.1. In particular, the CLS principles for these functionals are *practical* and no transformation to a DLSP is required⁴.

For all other functionals practical least-squares methods will necessarily involve a transformation of CLSP to a practical Discrete Least Squares Principle (DLSP). Here we will employ the techniques of §6.5.2, namely the *weighted norm* approach and the *negative norm* approach. Both the weighted

⁴In the sense that DLSP is simply a restriction of $\{\mathbf{X}_0, J(\cdot)\}$ to the finite element subspace \mathbf{X}^h of \mathbf{X}_0 .

and the negative norm approach will lead to least-squares methods that are optimally accurate.

However, a reasonable DLS principle and a sensible method can also be defined based only on the assumptions stated in §5.3. According to the terminology adopted in Chapter 5, we call methods based on such principles *non-equivalent* because they are not based on mathematically established energy balance for the PDE. In this case, the only requirements that must be met by the abstract DLSP represented by (5.14) were stated in **D.1-D.2**. While resulting least-squares methods may not be optimal, Theorem 2 indicates that they are still capable of producing approximate solutions to our problems. Moreover, methods based only on **D.1-D.2** are usually very straightforward to implement, especially when compared with negative norm methods. For this reason, we devote the next section to a brief discussion of such methods for the Stokes equations.

7.3.1 Non-equivalent least-squares

Historically, the first examples of least-squares methods for the Stokes and the Navier-Stokes equations were based on non-equivalent least-squares functionals; see, e.g., [98, 99, 101] and [103], among others. The reason for this was the fact that combination of first-order systems with L^2 -norms to measure the residual energy leads to a very simple and easy to implement scheme. However, as we saw in §7.1, not all first-order Stokes systems are homogeneous elliptic (or, which is the same, H^1 -coercive). This fact was first pointed out in [56] and [58]. As a result, the use of L^2 -norms for the residual energy does not necessarily lead to a mathematically correct energy balance. However, thanks to the generality of hypotheses **D.1-D.2** one can satisfy these two conditions almost automatically by any sensible definition of a least-squares functional. This fact has contributed significantly to the success of early least-squares methods based on first-order reformulations. To summarize, with the *velocity boundary condition* we have the following non-equivalent functionals:

Velocity-vorticity-pressure functional:

$$J(\boldsymbol{\omega}, \mathbf{u}, p) = \frac{1}{2} \left(\|\nu \nabla \times \boldsymbol{\omega} + \nabla p - \mathbf{f}\|_0^2 + \|\nabla \times \mathbf{u} - \boldsymbol{\omega}\|_0^2 + \|\nabla \cdot \mathbf{u}\|_0^2 \right) \quad (7.40)$$

Velocity-pressure-stress functional:

$$J(\underline{\mathbf{T}}, \mathbf{u}, p) = \frac{1}{2} \left(\|\underline{\mathbf{T}} - \sqrt{2\nu} \boldsymbol{\epsilon}(\mathbf{u})\|_0^2 + \|\nabla \cdot \mathbf{u}\|_0^2 + \|\sqrt{2\nu} \nabla \cdot \underline{\mathbf{T}} - \nabla p - \mathbf{f}\|_0^2 \right) \quad (7.41)$$

Velocity gradient-velocity-pressure functional I:

$$J(\underline{\mathbf{U}}, \mathbf{u}, p) = \frac{1}{2} \left(\| -\nu \nabla \cdot \underline{\mathbf{U}} + \nabla p - \mathbf{f} \|_0^2 + \|\nabla \cdot \mathbf{u}\|_0^2 + \|\underline{\mathbf{U}} - (\nabla \mathbf{u})^t\|_0^2 \right) \quad (7.42)$$

We remind the reader that if the boundary condition is changed the non-equivalence of these functionals may also change. Consider, for example, the functional (7.40). When the first-order system (7.1)-(7.3) is augmented by the normal velocity-pressure boundary condition (7.17), the corresponding boundary value problem is homogeneous elliptic. As a result, the system (7.1)-(7.3) is fully H^1 -coercive, and the relevant a priori estimate is given by (7.15). This means that for the boundary condition (7.17) the functional (7.40) represents the correct energy balance. Therefore, Theorem 4 is applicable and the error estimate (6.24) specializes to

$$\begin{aligned} & \|\mathbf{u} - \mathbf{u}^h\|_r + \|\boldsymbol{\omega} - \boldsymbol{\omega}^h\|_r + \|p - p^h\|_r \\ & \leq Ch^{k+1-r} \left(\|\mathbf{u}\|_{k+1} + \|\boldsymbol{\omega}\|_{k+1} + \|p\|_{k+1} \right), \quad r = 0, 1. \end{aligned} \quad (7.43)$$

This estimate is valid, e.g., if the standard finite element spaces P_k or Q_k are used for all variables.

Let us now suppose that (7.1)-(7.3) is instead augmented by the velocity boundary condition (7.4). Then, the corresponding boundary value problem is not homogeneous elliptic. Thus, the system (7.1)-(7.3) is not fully H^1 -coercive, and the relevant a priori estimate is now given by (7.16). This fact by itself does not immediately imply that the method is not optimal; it only indicates that standard finite element analyses cannot be used to show that the optimally accurate error estimates given by (7.43) are valid with the velocity boundary condition. A more careful analysis of this method does however reveal that it indeed is suboptimal; suboptimal convergence rates can be observed computationally as well. An example will be presented in the next section.

Consider next the functional (7.41). From §7.1.2, we know that the associated boundary value problem is not fully H^1 -coercive, regardless of the choice of boundary conditions. Similarly, the first-order system (7.21)-(7.23) is not fully H^1 -coercive and estimate (7.24) implies that the functional (7.42) is not norm equivalent. Thus, in both cases, the optimality of the resulting methods cannot be established using standard elliptic arguments. In fact, in both cases, one can devise counterexamples that will reveal sub-optimal convergence rates.

7.3.2 Weighted least-squares methods

In this section we discuss transformation to DLSP based on the use of weighted L^2 -norms. Resulting methods fall into the category analyzed in §6.5.2. According to the terminology in §5 we call such methods *quasi norm-equivalent* because their energy balance depends on the mesh parameter h .

We consider the first-order system (7.1)-(7.3) along with the boundary condition (7.4). In this case, the correct energy balance is given by (7.16) and the correct CLSP is based on the least-squares functional (7.35). Setting $q = 0$ in (7.35) implies that for the velocity boundary condition the correct least-squares functional is

$$J(\boldsymbol{\omega}, \mathbf{u}, p) = \frac{1}{2} \left(\|\nu \nabla \times \boldsymbol{\omega} + \nabla p - \mathbf{f}\|_0^2 + \|\nabla \times \mathbf{u} - \boldsymbol{\omega}\|_1^2 + \|\nabla \cdot \mathbf{u}\|_1^2 \right) \quad (7.44)$$

instead of (7.40). However, the use of H^1 -norms in (7.44) calls for discretization of *second-order* terms such as $(\nabla \nabla \cdot \mathbf{u})$ and $(\nabla \nabla \times \mathbf{u})$. Conforming discretizations of such terms can be handled using subspaces of $H^2(\Omega)$. In the finite element setting, this essentially requires the use of finite element spaces that are continuously differentiable across the element faces. Unfortunately, in two and three dimensions, such elements are impractical, which offsets the potential advantages of a least-squares formulation based on (7.44).

We have encountered the same situation in the abstract setting of §6.5.2 and the functional (6.16). Following the approach outlined in this section we replace (7.44) by the mesh-dependent functional

$$J_h(\boldsymbol{\omega}, \mathbf{u}, p) = \frac{1}{2} \left(\|\nu \nabla \times \boldsymbol{\omega} + \nabla p - \mathbf{f}\|_0^2 + h^{-2} \|\nabla \times \mathbf{u} - \boldsymbol{\omega}\|_0^2 + h^{-2} \|\nabla \cdot \mathbf{u}\|_0^2 \right). \quad (7.45)$$

This functional represent specialization of (6.27) to the velocity-vorticity-pressure Stokes equations. Therefore, Theorem 5 is applicable, and the error estimate (6.31) specializes to

$$\|\boldsymbol{\omega} - \boldsymbol{\omega}^h\|_0 + \|p - p^h\|_0 + \|\mathbf{u} - \mathbf{u}^h\|_1 \leq C h^k \left(\|\boldsymbol{\omega}\|_k + \|p\|_k + \|\mathbf{u}\|_{k+1} \right) \quad (7.46)$$

This error estimate is valid for $k \geq 2$ if one uses, e.g., the finite element spaces P_k or Q_k for the velocity and P_{k-1} and Q_{k-1} for the pressure and vorticity. Note that the error in the approximation is measured in norms corresponding to (7.16) with $q = -1$. As a result, for the approximation of the pressure and the vorticity one can use finite element spaces with interpolation order of one degree less than that used for the velocity approximation. This also means that (7.45) is not optimal if equal order interpolation is used for all dependent variables.

Let us now give an example which shows that without the weights (7.45) yields suboptimal convergence rates. Since (7.45) without the weights is simply the non-equivalent functional (7.40) this will also establish the fact that non-equivalence can affect convergence rates. Using the exact solution from Example 3 in Appendix A with $n = 1$, the functional (7.40), and discretization by quadratic elements on triangles, we have computationally obtained the (approximate) convergence rates as given in Table 7.2; one can conclude that the rates for the velocity boundary condition case are sub-optimal. In Table 7.2 the columns BC1 contain results for (7.40) with the velocity boundary condition, while the BC2 columns present the rates for (7.40) with the normal velocity-pressure boundary condition (7.17). We draw attention to the fact that all rates in the BC2 columns are optimal. This is due to the fact that the velocity-vorticity-pressure equations with (7.17) are homogeneous elliptic system and (7.40) is in actuality a norm-equivalent functional!

rates	L^2 error			H^1 error		
variable	BC1W	BC1	BC2	BC1W	BC1	BC2
u	3.64	2.71	3.11	2.15	2.03	2.04
v	3.31	2.37	3.10	2.10	2.06	2.02
ω	3.57	2.20	3.00	2.35	1.64	1.93
p	3.11	2.34	2.98	2.37	1.64	1.97

Table 7.2: Rates of convergence with and without the weights. Velocity-vorticity-pressure formulation with (7.4) and (7.17).

Another candidate for a similar treatment is the velocity-pressure-stress system (7.18). Recall that this system is not fully H^1 -coercive, i.e., the L^2 functional (7.41) is not norm equivalent. As a result, one can find smooth solutions such that the L^2 formulation (7.41) for (7.18) yields suboptimal convergence rates. At the same time, using (7.20) with $q = 0$ to define a norm-equivalent least-squares functional will lead to impractical methods. Following again the ideas of §6.5.2 we are led to the weighted functional for (7.18):

$$\begin{aligned}
 J_h(\underline{\mathbf{T}}, \mathbf{u}, p) &= \frac{1}{2} \left(h^{-2} \|\underline{\mathbf{T}} - \sqrt{2\nu} \epsilon(\mathbf{u})\|_0^2 + h^{-2} \|\nabla \cdot \mathbf{u}\|_0^2 \right. \\
 &\quad \left. + \|\sqrt{2\nu} \nabla \cdot \underline{\mathbf{T}} - \nabla p - \mathbf{f}\|_0^2 \right); \tag{7.47}
 \end{aligned}$$

see [58]. Again, this functional specializes (6.27) to the velocity-pressure-stress first-order system. The resulting finite element method shares many

common properties with the one for the velocity-vorticity-pressure system, including optimal error estimates in which the error in the approximations of $\underline{\mathbf{T}}$, \mathbf{u} , and p is measured in norms corresponding to (7.20) with $q = -1$, i.e.,

$$\|\underline{\mathbf{T}} - \underline{\mathbf{T}}^h\|_0 + \|p - p^h\|_0 + \|\mathbf{u} - \mathbf{u}^h\|_1 \leq C h^k \left(\|\underline{\mathbf{T}}\|_k + \|p\|_k + \|\mathbf{u}\|_{k+1} \right) \quad (7.48)$$

that is valid for $k \geq 2$ if one uses, e.g., the finite element spaces P_k or Q_k for the velocity and P_{k-1} and Q_{k-1} for the pressure and stress. As with (7.46), the estimate (7.48) is not optimal if equal order interpolation is used for all dependent variables.

One can also show that the weights in (7.47) are necessary for the optimal convergence rates in (7.48). For example, consider the following exact solution; [58] (compare with Example 3!)

$$\begin{aligned} u_1 = u_2 &= \sin(\pi x) \sin(\pi y) \\ T_1 = T_2 = T_3 &= \sin(\pi x) \exp(\pi y) \\ p &= \cos(\pi x) \exp(\pi y). \end{aligned}$$

and a method based on (7.47) implemented using P_1 elements for $\underline{\mathbf{T}}$ and p , and P_2 elements for the velocity. Numerical estimates of convergence rates with and without the weights are summarized in Table 2.

rates	L^2 error			H^1 error		
	WLS	LS	BA	WLS	LS	BA
u	3.59	1.11	3.00	2.85	1.00	2.00
v	3.13	1.28	3.00	2.77	1.17	2.00
T_{11}	2.42	1.25	2.00	0.99	0.94	1.00
T_{12}	2.48	1.14	2.00	1.01	0.99	1.00
T_{22}	2.34	1.26	2.00	1.05	0.76	1.00
p	2.40	0.94	2.00	1.10	0.92	1.00

Table 7.3: Convergence rates with and without the weights. Velocity-pressure-stress formulation.

Since without the weights (7.47) gives the non-equivalent functional (7.41) this table shows once again that non-equivalent discrete least-squares principles can lead to loss of convergence rate.

7.3.3 H^{-1} least-squares methods

In this section we consider another transformation to a DLSP, this time based on the use of discrete negative norms. This approach was developed for general first-order ADN systems in §6.5.2. It can be applied whenever the first-order system fails to be homogeneous elliptic, as in the case of the velocity-vorticity-pressure Stokes equations with the velocity boundary condition. The use of negative norms has certain advantages when compared with the weighted L^2 approach of the last section. Most notably, resulting algebraic equations have condition numbers comparable with the condition number of the systems resulting from Galerkin discretizations. In contrast, analysis of §6.5.2 shows that weighted functionals lead to algebraic equations with condition numbers of order $O(h^{-4})$. Of course, these advantages come at a certain price, and in the case of negative norm methods it is in the more complicated implementation along with the fact that the linear systems are dense and must necessarily be solved by assembly free methods.

Let us consider again the velocity-vorticity-pressure system (7.1)-(7.3) with the boundary condition (7.4). The fact that this system is not homogeneous elliptic implies that one cannot use the same norm to measure all residuals of the first-order equations. Recall that setting $q = 0$ in the a priori estimate (7.16) leads to the mathematically correct, but impractical⁵ functional (7.35) which has been used to motivate the weighted functional (7.45). If, on the other hand, one chooses $q = -1$ in the a priori estimate (7.16), this leads to CLSP based on the minimization of

$$J_{-1}(\boldsymbol{\omega}, \mathbf{u}, p) = \frac{1}{2} \left(\|\nu \nabla \times \boldsymbol{\omega} + \nabla p - \mathbf{f}\|_{-1}^2 + \|\nabla \times \mathbf{u} - \boldsymbol{\omega}\|_0^2 + \|\nabla \cdot \mathbf{u}\|_0^2 \right). \quad (7.49)$$

Obviously, this functional represents a specialization of (6.16) to the velocity-vorticity-pressure Stokes system. A CLS principle which requires minimization of this functional is hardly any more practical than a CLS principle based on (7.16) with $q = 0$. This is because negative norms are not easy to compute. Thus, transformation to a DLSP is still required and here we propose to use the discrete negative norm (6.35), introduced in §6.5.2. This leads to a DLS principle based on the minimization of

$$J_{-h}(\boldsymbol{\omega}, \mathbf{u}, p) = \frac{1}{2} \left(\|\nu \nabla \times \boldsymbol{\omega} + \nabla p - \mathbf{f}\|_{-h}^2 + \|\nabla \times \mathbf{u} - \boldsymbol{\omega}\|_0^2 + \|\nabla \cdot \mathbf{u}\|_0^2 \right). \quad (7.50)$$

This functional is clearly a specialization of (6.38). Note that with the trivial

⁵This functional was impractical because of the fact that it involved second order derivatives.

choice $B^h \equiv 0$, (7.50) reduces to the weighted functional (7.45). Least-squares methods can now be defined according to the recipe of §6.5.2. In particular, the error estimate from Theorem 6 specializes to

$$\|\omega - \omega^h\|_0 + \|p - p^h\|_0 + \|\mathbf{u} - \mathbf{u}^h\|_1 \leq C h^k \left(\|\omega\|_k + \|p\|_k + \|\mathbf{u}\|_{k+1} \right) \quad (7.51)$$

In contrast to the error estimate (7.46), the bound (7.51) holds for $k \geq 1$. This means that asymptotic convergence rate of the negative norm method can be established under less stringent regularity assumptions than those for the weighted method.

It should be noted that the use of discrete negative norms in (7.50) leads to algebraic problems with dense matrices. As a result, a practical implementation of corresponding finite element methods is necessarily restricted to the use of iterative solvers that do not require matrix assembly. On the positive side, the algebraic system for (7.50) has $O(h^{-2})$ condition number and can be preconditioned in a much more efficient manner using, e.g., a block preconditioner defined according to (6.44).

7.4 Least-squares methods for the Navier-Stokes equations

All classes of least-squares methods developed for the Stokes equations, i.e., non-equivalent, weighted and negative norm can be easily extended, at least in principle, to the nonlinear Navier-Stokes equations. Indeed, given a CLS principle for a first-order Stokes problem, the corresponding CLS principle for the Navier-Stokes equations is readily available by simply including an appropriate form of the nonlinear term into the residual of the momentum equation. From a practical point of view, the resulting methods differ from their Stokes counterparts in two aspects. First, the associated discrete problem now constitutes a nonlinear system of algebraic equations that must be solved in an iterative manner using, e.g., a Newton linearization. Second, solving the discrete system may not be straightforward for high values of the Reynolds number since it is well-known that the attraction ball for, e.g., Newton's method, decreases as the Reynolds number increases.

Most existing least-squares methods for the Navier-Stokes equations are based on the velocity-vorticity-pressure form of this problem, see e.g. [48, 49, 50], [54, 55, 57]. Exceptions include [51, 52] and [53] which consider velocity gradient methods, and [103] where a stress-based method is discussed. The differences among various least-squares methods involve the

choice of the discretization spaces, the treatment of the nonlinear term, and the method used for solution of the nonlinear discrete equations. For example, the methods of [100], [101], and [99] are based on *non-equivalent* DLS principles, discretization by piecewise linear finite elements, and the $\mathbf{u} \cdot \nabla \mathbf{u}$ form of the nonlinear term.

Other authors use instead the $\boldsymbol{\omega} \times \mathbf{u}$ form of the nonlinear term. Solution of the nonlinear discrete equations is by Newton linearization and solution of the linearized equations is by the conjugate gradient method with Jacobi preconditioning. The method of [104] is very similar; however, solution of the linearized problem now involves the conjugate gradient method preconditioned by incomplete Choleski factorization. The p -version of the finite element method has been used in [102]. The methods of [47] and [57] use weighted least-squares functionals similar to (7.45), where in addition to the mesh dependent weights h^{-2} , the residual of the momentum equation is weighted by the Reynolds number. To handle large values of the Reynolds number, these methods use Newton linearization combined with continuation with respect to the Reynolds number. Large scale computations and parallelization issues have been considered in [106], [107], [108], [109], [110] and [112]. Numerical comparison between velocity-vorticity-pressure, velocity-stress-pressure and velocity gradient formulations is given in [111]. Discussion of relative advantages and disadvantages of different forms of the nonlinear term can be found in [6].

The nonlinearity also considerably complicates the mathematical analysis of corresponding least-squares methods. At present, analyses available are limited to methods based on the velocity-vorticity-pressure (see [48], [47], and [50]) and velocity gradient (see [52] and [53]) forms of the Navier-Stokes equations. In both cases, analyses are based on the abstract approximation theory of Brezzi-Rappaz-Raviart [4] or its modifications. Since discussion of these results would require substantial amount of theoretical and technical background, it is beyond the scope of these lectures. Thus, in what follows we only outline the main idea of the error analysis.

It can be shown that the Euler-Lagrange equation associated with a least-squares functional for the Navier-Stokes equations can be cast into an abstract canonical form given by

$$F(\lambda, U) \equiv U + T \cdot G(\lambda, U) = 0, \quad (7.52)$$

where $\lambda = Re$, T corresponds to a least-squares solution operator for the associated Stokes problem, and G is a nonlinear operator. Similarly, the corresponding discrete nonlinear problem can be identified with an abstract

equation of the form

$$F^h(\lambda, U^h) \equiv U^h + T^h \cdot G(\lambda, U^h) = 0, \quad (7.53)$$

where T^h is a discrete counterpart of T . The importance of this abstract form is signified by the fact that discretization in (7.53) is introduced solely by means of an approximation to the *linear* operator T in (7.52). As a result, under some assumptions, one can show that the error in the nonlinear approximation defined by (7.53) is of the same order as the error in the least-squares solution of the linear Stokes problem.

Chapter 8

Least squares for $-\Delta u = f$

In this chapter we specialize the methods developed in Chapter 6 to the Poisson equation in 2D

$$-\Delta\phi = f \quad \text{in } \Omega \tag{8.1}$$

$$\phi = 0 \quad \text{on } \Gamma. \tag{8.2}$$

However, we also consider methods that do not fit completely into the abstract framework of Chapter 6. In these methods the energy balance (5.3) is not derived from the ADN theory which restricts the possible range of spaces in the a priori estimates to the standard Sobolev spaces $W_q^p(\Omega)$. Instead, the relevant a priori bounds are derived in a direct manner which makes it possible to obtain energy balance for the least-squares method in terms of spaces such as $H(\Omega, \text{div})$ or $H(\Omega, \text{div}) \cap H(\Omega, \mathbf{curl})$. For more details about such direct techniques we refer to [65], [71], [72], [73], [74], and [93]. Nevertheless, it should be pointed out that regardless of the method used to establish (5.3), formulation of a mathematically well-posed least-squares principle follows essentially the path outlined in Chapter 5.

Let us recall that the Poisson equation (8.1)-(8.2) was considered in section 2.1 as an example of a problem for which a natural unconstrained minimization principle exists. Also, in section 2.3 we saw that a standard Galerkin procedure applied to the second order Poisson problem will necessarily recover this optimization setting. As a result, the Galerkin method for (8.1)-(8.2) operates in the favorable Rayleigh-Ritz setting and application of a least-squares principle to the second order problem is not justified.

However, if the goal is to obtain an approximation to $\nabla\phi$ rather than to ϕ we may prefer to compute $\mathbf{v} = \nabla\phi$ directly rather than to differentiate the approximation of ϕ . Then, application of the standard Galerkin

procedure will inevitably lead us to the saddle-point weak problem (2.20). Now, the least-squares approach becomes an attractive alternative to the mixed method and its application is completely justified. Thus, we begin this chapter with a brief summary of the available first-order formulations for (8.1)-(8.2) and their properties.

8.1 First-order systems

A first-order form of (8.1)-(8.2) is given by the *div-grad* system

$$\nabla \cdot \mathbf{v} = f \quad \text{in } \Omega \quad (8.3)$$

$$\nabla \phi + \mathbf{v} = 0 \quad \text{in } \Omega \quad (8.4)$$

along with (8.2). In view of (8.4) and (8.2) this system can be augmented by an additional equation (curl constraint)

$$\nabla \times \mathbf{v} = 0 \quad \text{in } \Omega \quad (8.5)$$

and a boundary condition

$$\mathbf{n} \times \mathbf{v} = 0 \quad \text{on } \Gamma. \quad (8.6)$$

We shall refer to the augmented system (8.3)-(8.5), (8.2), and (8.6) as the *div-grad-curl* system. The system (8.3)-(8.4) is elliptic in the sense of ADN; see [11]. The appropriate indices (see Example 1 in §6.3) for the equations and the unknowns are given by $s_1 = 0$, $s_2 = s_3 = -1$ and $t_1 = t_2 = 1$, $t_3 = 2$, respectively (it is assumed that the equations are ordered as in (8.3)-(8.4) and that the unknowns are ordered as $(\mathbf{v}_1, \mathbf{v}_2, \phi)$). Thus, the div-grad system is not uniformly elliptic and the a priori estimates relevant to the least-squares method are

$$\|\mathbf{v}\|_0 + \|\phi\|_1 \leq C \left(\|\nabla \cdot \mathbf{v}\|_{-1} + \|\nabla \phi + \mathbf{v}\|_0 \right), \quad (8.7)$$

for all $(\mathbf{v}, \phi) \in \mathbf{L}^2(\Omega) \times H^1(\Omega)$ and

$$\|\mathbf{v}\|_1 + \|\phi\|_2 \leq C \left(\|\nabla \cdot \mathbf{v}\|_0 + \|\nabla \phi + \mathbf{v}\|_1 \right), \quad (8.8)$$

for all $(\mathbf{v}, \phi) \in \mathbf{H}^1(\Omega) \times H^2(\Omega)$. Formally, the div-grad-curl system is not ADN elliptic because it has 4 equations and 3 unknowns. By adding the gradient of a *slack*¹ variable ψ to equation (8.4) and a boundary condition

¹We encountered the same situation with several first-order formulations of the Stokes equations.

$\psi = 0$ this system becomes homogeneous elliptic; see [80]. It can be shown that the slack variable is identically zero and can be completely ignored so that the relevant a priori estimate is

$$\|\mathbf{v}\|_1 + \|\phi\|_1 \leq C \left(\|\nabla \cdot \mathbf{v}\|_0 + \|\nabla \phi + \mathbf{v}\|_0 + \|\nabla \times \mathbf{v}\|_0 \right). \quad (8.9)$$

for all $(\mathbf{v}, \phi) \in \mathbf{H}^1(\Omega) \times H^1(\Omega)$. In addition to (8.7)-(8.9) one can also show that

$$\|\mathbf{v}\|_{div} + \|\phi\|_1 \leq C \left(\|\nabla \cdot \mathbf{v}\|_0 + \|\nabla \phi + \mathbf{v}\|_0 \right) \quad (8.10)$$

for all $(\mathbf{v}, \phi) \in H(\Omega, \text{div}) \times H^1(\Omega)$. This a priori estimate does not follow from the ADN theory and must be established directly; see [65] or [71].

8.1.1 Inhomogeneous boundary conditions

To determine the appropriate norms for a given boundary operator one may rely again on the elliptic regularity theory of [11], or on various trace theorems relating boundary and interior norms of functions. For example, a result of [8] states that for every $g \in H^{1/2}(\Gamma)$ there is a unique $\phi \in H^1(\Omega)$ such that $\Delta \phi = 0$ in Ω , $\phi = g$ on Γ , and $\|\phi\|_1 \leq C \|g\|_{1/2, \Gamma}$. As a result, for the div-grad system with inhomogeneous Dirichlet boundary condition given by

$$-\nabla \cdot \mathbf{v} = f \quad \text{and} \quad \mathbf{v} = \nabla \phi \quad \text{in } \Omega \quad \text{and} \quad \phi = g \quad \text{on } \Gamma,$$

the relevant a priori estimate is given by

$$\|\phi\|_1 + \|\mathbf{v}\|_{H(\Omega, \text{div})} \leq C \left(\|\mathbf{v} - \nabla \phi\|_0 + \|\nabla \cdot \mathbf{v}\|_0 + \|\phi\|_{1/2, \Gamma} \right). \quad (8.11)$$

8.2 Continuous Least Squares Principles

The norm-equivalent functionals corresponding to (8.7)-(8.10) are given by

$$J_{-1}(\mathbf{v}, \phi; f) = \|\nabla \cdot \mathbf{v} - f\|_{-1}^2 + \|\nabla \phi + \mathbf{v}\|_0^2, \quad (8.12)$$

$$J_0(\mathbf{v}, \phi; f) = \|\nabla \cdot \mathbf{v} - f\|_0^2 + \|\nabla \phi + \mathbf{v}\|_1^2, \quad (8.13)$$

$$J_P(\mathbf{v}, \phi; f) = \|\nabla \cdot \mathbf{v} - f\|_0^2 + \|\nabla \phi + \mathbf{v}\|_0^2 + \|\nabla \times \mathbf{v}\|_0^2, \quad (8.14)$$

and

$$J(\mathbf{v}, \phi; f) = \|\nabla \cdot \mathbf{v} - f\|_0^2 + \|\nabla \phi + \mathbf{v}\|_0^2, \quad (8.15)$$

respectively. The last two functionals involve only L^2 norms of first-order terms. As a result, the CLS principles associated with (8.14)-(8.15) lead to practical least-squares methods, i.e., the DLS principle $\{\mathbf{X}^h, J_h(\cdot)\}$ is simply a restriction of $\{\mathbf{X}, J(\cdot)\}$ to the finite element subspace \mathbf{X}^h . The first two functionals lead to CLS principles that are not practical. Therefore, in these two cases a DLSP must be used to define the finite element methods. To find such DLSP we proceed to replace (8.12) and (8.13) by suitable mesh-dependent functionals. As a substitute for (8.12) we consider the functional

$$J_{-h}(\mathbf{v}, \phi; f) = \|\nabla \cdot \mathbf{v} - f\|_{-h}^2 + \|\nabla \phi + \mathbf{v}\|_0^2. \quad (8.16)$$

while for (8.13) we consider the weighted functional

$$J_h(\mathbf{v}, \phi; f) = \|\nabla \cdot \mathbf{v} - f\|_0^2 + h^{-2} \|\nabla \phi + \mathbf{v}\|_0^2. \quad (8.17)$$

We could have also considered

$$J_h(\mathbf{v}, \phi; f) = h^2 \|\nabla \cdot \mathbf{v} - f\|_0^2 + \|\nabla \phi + \mathbf{v}\|_0^2,$$

as a substitute for (8.12), but this functional is merely a scaled version of (8.17). Clearly, (8.16) specializes the negative norm functional (6.38) to the first-order Poisson system. Likewise, (8.17) is specialization of the weighted functional (6.27).

8.2.1 Error estimates

With each one of the functionals (8.14)-(8.17) we associate a DLS principle, that is a pair $\{\mathbf{X}^h, J_h(\cdot)\}$, where \mathbf{X}^h is a suitable finite element space and $J_h(\cdot)$ is one of the least-squares functionals. Then, a least-squares method is defined in the usual manner by computing the minimizer of each functional over \mathbf{X}^h , i.e., by solving the problem

$$\min_{\mathbf{X}^h} J_h(\mathbf{v}, \phi; f).$$

For functionals (8.14)-(8.15) the appropriate space \mathbf{X}^h can be defined using equal order interpolation for all variables

$$\mathbf{X}^h = \{(\mathbf{v}^h, \phi^h) \mid (\mathbf{v}^h, \phi^h) \in \prod_{j=1}^3 S_d^h, \quad \phi^h = 0 \quad \text{on } \Gamma\}.$$

Resulting finite element methods are conforming in the sense that \mathbf{X}^h is contained both in $\mathbf{H}^1(\Omega) \times H^1(\Omega)$ and $H(\Omega, \text{div}) \times H^1(\Omega)$, which are exactly

the appropriate minimization spaces for the energy functionals (8.14) and (8.15), respectively. For the div-grad-curl system, which is homogeneous elliptic, the error bound (6.25) from Theorem 4 specializes to

$$\|\phi - \phi^h\|_1 + \|\mathbf{v} - \mathbf{v}^h\|_1 \leq Ch^d(\|\phi\|_{d+1} + \|\mathbf{v}\|_{d+1})$$

provided the exact solution is in $[H^{d+1}(\Omega)]^3$. The error estimate for the approximations defined by (8.15) is

$$\|\phi - \phi^h\|_1 + \|\mathbf{v} - \mathbf{v}^h\|_{H(\Omega, \text{div})} \leq Ch^d(\|\phi\|_{d+1} + \|\mathbf{v}\|_{d+1})$$

see [71]). Under additional regularity assumptions on the dual problems one can also establish error estimates in L^2 .

Consider now methods based on (8.17) and (8.16). According to Theorems 5 and 6, ϕ must be approximated by finite elements of one order higher than those used for \mathbf{v} . As a result, the proper choice of the space \mathbf{X}^h for the DLS principle $\{\mathbf{X}^h, J_h(\cdot)\}$ is now given by:

$$\mathbf{X}^h = \{(\mathbf{v}^h, \phi^h) \mid (\mathbf{v}^h, \phi^h) \in [S_d^h]^2 \times S_{d+1}^h, \quad \phi^h = 0 \quad \text{on } \Gamma\}.$$

For example, if \mathbf{v} is approximated using piecewise linear elements, then ϕ must be approximated by piecewise quadratic elements. Of course, one may as well use quadratics for all unknowns, but then error bounds (6.31) and (6.41) will not be optimal with respect to the spaces used. Assuming that $d \geq 1$ and $(\mathbf{v}, \phi) \in [H^{d+1}(\Omega)]^2 \times H^{d+2}(\Omega)$, both (6.31) and (6.41) specialize to

$$\|\phi - \phi^h\|_1 + \|\mathbf{v} - \mathbf{v}^h\|_0 \leq Ch^{d+1}(\|\mathbf{v}\|_{d+1} + \|\phi\|_{d+2}).$$

However, for the negative norm method convergence can be still established if $(\mathbf{v}, \phi) \in [H^1(\Omega)]^2 \times H^2(\Omega)$, while for the weighted method Theorem 5 requires that (\mathbf{v}, ϕ) is at least in $[H^2(\Omega)]^2 \times H^3(\Omega)$.

8.2.2 Conditioning and preconditioning of discrete systems

Condition numbers for all matrices, except the one associated with (8.17), are of order $O(h^{-2})$. For this matrix the weight h^{-2} causes an increase of the condition number order to $O(h^{-4})$.

Next, consider design of preconditioners for each one of the four systems. The norm equivalence of (8.15) implies that the associated form $Q(\cdot; \cdot)$ is equivalent to an $H^1(\Omega) \times H(\Omega, \text{div})$ inner product. As a result, the matrix A^h in the algebraic problem can be preconditioned by a block diagonal matrix consisting of a Poisson preconditioner and a discrete divergence block.

Similarly, the matrix A^h associated with (8.14) is equivalent to a block diagonal matrix of discrete Laplace operators. As a result, this system can be preconditioned by

$$\text{diag}(T^h, T^h, T^h)$$

where T^h is a preconditioner for the Poisson equation. The matrix for (8.16) is norm equivalent with respect to the norm on $\mathbf{L}^2(\Omega) \times H^1(\Omega)$ and as a result, the corresponding algebraic problem can be preconditioned by

$$\text{diag}(h^2\mathbf{I}, h^2\mathbf{I}, T^h).$$

Preconditioning of the algebraic system arising from the weighted functional (8.17) is more difficult due to the lack of norm equivalence. Combined with the higher condition number of this system this makes its numerical solution more complicated and time consuming than that of the other three systems.

Chapter 9

Least-squares methods that stand apart

In this chapter we examine three examples of least-squares methods that do not fit directly into the framework developed in Chapters 5–6. The first is represented by collocation least-squares methods. Here we consider examples of point and subdomain collocation methods. The second includes a method that combines least-squares ideas with the technique of Lagrange multipliers in order to enhance mass conservation. The third unconventional least-squares method casts the original boundary value problem into the framework of an optimal control or optimization problem with a least-squares functional serving the role of the cost or objective functional.

9.1 Least-squares collocation methods

In this section, we briefly review a class of least-squares methods in which the discretization step is taken prior to the least-squares step. Such methods are commonly known as *least-squares collocation*, *point least-squares*, *point matching*, or *overdetermined collocation* methods; see [119], [120]. The main idea is as follows. Consider again the linear boundary value problem (5.1)–(5.2). We assume that an approximate solution is sought in the form

$$U(x) \approx U_N(\mathbf{a}, x),$$

where $\mathbf{a} = (a_1, a_2, \dots, a_N)$ is a vector of unknown coefficients. Let $R_{\mathcal{L}}^j(\mathbf{a}, x)$, $j = 1, \dots, K$, and $R_{\mathcal{R}}^j(\mathbf{a}, x)$, $j = 1, \dots, L$ denote residuals of the equations in (5.1) and (5.2), respectively. To define a least-squares collocation method, one chooses a finite set of points $\{x_i\}_{i=1}^{M_1}$ in Ω , and another set of points

$\{x_i\}_{i=M_1+1}^M$ on Γ . Then, a least-squares functional is defined by summing the weighted squares of the residuals evaluated at the points x_i :

$$J(\mathbf{a}) = \sum_{j=1}^K \sum_{i=1}^{M_1} \alpha_{ji} \left(R_{\mathcal{L}}^j(\mathbf{a}, x_i) \right)^2 + \sum_{j=1}^L \sum_{i=M_1+1}^M \beta_{ji} \left(R_{\mathcal{R}}^j(\mathbf{a}, x_i) \right)^2. \quad (9.1)$$

The weights α_{ji} and β_{ji} may depend on both the particular equation and collocation point. Minimization of (9.1) with respect to the parameters in \mathbf{a} leads to (a usually overdetermined) algebraic system of the form $A\mathbf{a} = \mathbf{b}$, where A is an M by N matrix. Then, a discrete solution is determined by solving the normal equations $A^T A\mathbf{a} = A^T \mathbf{b}$. Methods formulated along these lines have been used for the numerical solution of the Navier-Stokes equations (see [120]) and hyperbolic problems, including the shallow water equations (see [121], [122], [123], and [124].) For numerous other applications of collocation least-squares, see [119].

Evidently, when the number of collocation points M equals the number of degrees of freedom N in $U_N(\mathbf{a}, x)$, the above methods reduce to a standard collocation procedure. Similarly, if $U_N(\mathbf{a}, x)$ is defined using a finite element space and the collocation points and weights correspond to a quadrature rule, then collocation is equivalent to a finite element least-squares method in which integration has been replaced by quadrature. Collocation least-squares methods offer some specific advantages. For example, since only a finite set of points x_i in the domain Ω need be specified, collocation least-squares are attractive for problems posed on irregularly shaped domains; see [123]. On the other hand, since the normal equations tend to become ill-conditioned, such methods require additional techniques, like scaling, or orthonormalization, in order to obtain a reliable solution; see [119].

Standard collocation, as well as collocation least-squares methods, use point-by-point matching criteria to define the discrete problem. Instead of a set of points one can also consider collocation over a set of subdomains of Ω . In such a case, the discrete problems are obtained by averaging differential equations over each subdomain. Here, for an illustration of this approach, we consider the subdomain Galerkin least-squares method of [79]. Let (5.1)-(5.2) correspond to a first-order homogeneous elliptic boundary value problem with $C = 0$, i.e., $\mathcal{L}\mathbf{u} = A\mathbf{u}_x + B\mathbf{u}_y$, $\mathcal{R}\mathbf{u} = R\mathbf{u}$ where R is a full-rank n by $2n$ matrix. To define the subdomain Galerkin/least-squares method for (5.1)-(5.2), we consider a finite element space \mathbf{X}^h consisting of continuous piecewise linear functions defined on a regular triangulation \mathcal{T}_h of the domain Ω into triangles Ω_k . These triangles will also serve as collocation subdomains. We let K and N denote the number of triangles and vertices,

respectively, in \mathcal{T}_h . For simplicity, we shall assume that the finite element functions in \mathbf{X}^h satisfy the essential boundary conditions (5.2). Then, a set of discrete equations is formed by averaging separately the components of the differential system (5.1)-(5.2) over each of the triangles $\Omega_k \in \mathcal{T}_h$:

$$\int_{\Omega_k} (\mathcal{L}\mathbf{u}^h)_j d\Omega = \int_{\Omega_k} (f)_j d\Omega \quad \text{for } k = 1, \dots, K \quad \text{and } j = 1, \dots, 2n. \quad (9.2)$$

Once a basis for \mathbf{X}^h is chosen, it is not difficult to see that (9.2) is equivalent to a rectangular linear algebraic system of the form $CU = F$ which consists of $2nK$ equations in approximately $2nN$ unknowns, i.e., there are about twice as many equations as unknowns. The subdomain-Galerkin/least-squares method of [79] consists per se of forming the matrix C and subsequently solving the above linear system by a discrete least-squares technique. If the data F is sufficiently smooth, one can show (see [79]) that the resulting method is optimal in the sense that

$$\|\mathbf{u} - \mathbf{u}^h\|_1 \leq C_1 h \|F\|_1 \quad \text{and} \quad \|\mathbf{u} - \mathbf{u}^h\|_0 \leq C_0 h^2 \|F\|_1.$$

We note that the discretization step in (9.2) can also be interpreted as an application of a nonstandard Galerkin method to the system (5.1)-(5.2) in which the test space consists of piecewise constant test functions with respect to \mathcal{T}_h . Similar subdomain collocation least-squares methods have also been developed for the numerical solution of Maxwell's equations; see [76].

9.2 Restricted least-squares methods

In general, when a least-squares method is used for the numerical solution of incompressible flow problems, computed velocity fields do not exactly satisfy the continuity equation. As a result, least-squares methods conserve mass only in an approximate manner and usually one can show that $\|\nabla \cdot \mathbf{u}^h\|_0 = O(h^r)$, where $r > 0$ depends on the particular finite element space employed. One way to enhance mass conservation involves the use of local mesh dependent weights along with special weights for the continuity equation. For example, the weighted functional (7.45) can be modified as follows (see [87]):

$$\begin{aligned} J_K(\boldsymbol{\omega}, p, \mathbf{u}) &= \frac{1}{2} \left(\|\nu \nabla \times \boldsymbol{\omega} + \nabla p - \mathbf{f}\|_0^2 \right. \\ &\quad \left. + \sum_j^J h_j^2 (W \|\nabla \cdot \mathbf{u}\|_{0, \Omega_j}^2 + \|\nabla \times \mathbf{u} - \boldsymbol{\omega}\|_{0, \Omega_j}^2) \right), \end{aligned} \quad (9.3)$$

where Ω_j , $j = 1, \dots, J$, denotes the j -th finite element, h_j denotes the diameter of Ω_j , and W is a weight for the continuity equation. Computational results with the corresponding finite element method reported in [87] indicate very good mass conservation properties with a moderate continuity equation weight ($W = 10$). Note that finite element methods based on the functional (9.3) do fit into the framework of Chapters 5 and 6 in the sense that these methods can be viewed as being based on DLSP derived from a CLSP for the correct least-squares functional (7.44).

Another approach, suggested in [85], which does not fit into the framework of Chapters 5–6, combines least-squares and Lagrange multiplier techniques into a method called *restricted least-squares*. The main idea of this method is to consider the continuity equation as a constraint that is enforced on each finite element via Lagrange multipliers. To state the method of [85], let \mathcal{T}_h denote a triangulation of Ω with n finite elements, \mathcal{L} denote a first-order Stokes differential operator, and \mathbf{X}^h denote a suitable finite element space defined over \mathcal{T}_h . The variational problem associated with the restricted least-squares method for the Stokes equations is then given by

seek $U^h \in \mathbf{X}^h$, and $\lambda_j \in \mathcal{R}$, $j = 1 \dots, J$, such that

$$\int_{\Omega} \mathcal{L}U^h \cdot \mathcal{L}V^h d\Omega + \sum_j^J \left(\lambda_j \int_{\Omega_j} \nabla \cdot \mathbf{v}^h d\Omega + \mu_j \int_{\Omega_j} \nabla \cdot \mathbf{u}^h d\Omega \right) = \int_{\Omega} \mathcal{L}V^h \cdot F$$

$\forall V^h \in \mathbf{X}^h$, $\mu_j \in \mathcal{R}$, $j = 1 \dots, J$. Although computational results obtained with the restricted method are very satisfactory, it also has some shortcomings. The use of Lagrange multipliers leads to a linear algebraic system with a symmetric but indefinite matrix that has a structure very similar to the matrices arising in mixed methods. Likewise, the size of the discrete problem increases by the number of additional constraints. Thus, at present it remains unclear whether the advantages of the restricted method outweigh the problems associated with imposing constraints on the velocity approximation. In particular, the loss of positive definiteness negates the main advantage of the least-squares formalism.

9.3 Least-squares optimization methods

The main idea of least-squares/optimization methods is to transform the original boundary value problem into an optimal control or optimization problem for which a cost functional is given by a least-squares type functional. To describe the method consider the following nonlinear Dirichlet

problem:

$$-\Delta\phi - G(\phi) = 0 \quad \text{in } \Omega \quad (9.4)$$

along with the boundary condition $\phi = 0$ on Γ . Then, an H^{-1} least-squares functional for (9.4) is given by

$$J(\phi) = \|\Delta\phi + G(\phi)\|_{-1}^2, \quad (9.5)$$

where $\|\cdot\|_{-1}$ denotes the negative norm. Minimization of (9.5) over $H_0^1(\Omega)$ would lead to a least-squares principle that is similar to the principles of Chapter 6.

The least-squares-optimization approach, however, considers minimization of

$$\mathcal{K}(\phi, \xi) = \|\Delta(\phi - \xi)\|_{-1}^2, \quad (9.6)$$

where $\xi \in H_0^1(\Omega)$ is a solution of

$$-\Delta\xi = G(\phi) \quad \text{in } \Omega \quad \text{and} \quad \xi = 0 \quad \text{on } \Gamma. \quad (9.7)$$

In the context of optimal control problems, one can identify ϕ with the control vector, ξ with the state variable, (9.7) with the state equation, and (9.6) with the cost functional. Furthermore, using the identity

$$\|\Delta\phi\|_{-1} = \|\nabla\phi\|_0 \quad \forall\phi \in H_0^1(\Omega),$$

one can replace (9.6) with the more easily computable (and therefore practical) cost functional

$$\mathcal{K}(\phi, \xi) = \|\nabla(\phi - \xi)\|_0^2. \quad (9.8)$$

To summarize, the least-squares/optimization method for (9.4) can be stated as follows:

minimize $\mathcal{K}(\phi, \xi)$ given by (9.8) over $\phi \in H_0^1(\Omega)$, subject to the state equation (9.7).

To solve the above optimization problem one can use an abstract version of the conjugate gradient method; see [125]. At each iteration, this method would require solution of two Dirichlet problems (9.7) for the computation of the descent direction.

This class of methods has been developed for nonlinear flow problems, including compressible flows (see [125, 126], and [127]) and the Navier-Stokes equations (see [125] and [128].) For example, to derive the least-squares/optimization method for the Navier-Stokes equations (2.37), let

$$\mathbf{Z} = \{\mathbf{u} \in \mathbf{H}_0^1(\Omega) \mid \nabla \cdot \mathbf{u} = 0 \text{ in } \Omega\}$$

and

$$\mathcal{K}(\mathbf{u}, \boldsymbol{\xi}) = \frac{\nu}{2} \|\Delta(\boldsymbol{\xi} - \mathbf{u})\|_{-1}^2 = \frac{\nu}{2} \int_{\Omega} |\nabla(\boldsymbol{\xi} - \mathbf{u})|^2 dx \quad (9.9)$$

and consider the Stokes problem

$$\begin{aligned} -\nu \Delta \boldsymbol{\xi} + \nabla q &= -\mathbf{u} \cdot \nabla \mathbf{u} && \text{in } \Omega \\ \nabla \cdot \boldsymbol{\xi} &= 0 && \text{in } \Omega \\ \boldsymbol{\xi} &= \mathbf{0} && \text{on } \Gamma. \end{aligned} \quad (9.10)$$

Then, the least-squares/optimization method for (2.37) is given by:

minimize $\mathcal{K}(\mathbf{u}, \boldsymbol{\xi})$ *given by* (9.9) *over* $\mathbf{u} \in \mathbf{Z}$, *subject to the state equation* (9.10).

To solve the above optimal control problem, one can again use an abstract conjugate gradients process. Now, computation of the descent direction at each iteration involves the solution of several Stokes problems; see [125] and [128].

Appendix A

The Complementing Condition

This appendix demonstrates verification of the celebrated Complementing Condition (see Definition 4 in Chapter 6) for the velocity-vorticity-pressure and the velocity-pressure-stress Stokes operators. Here the reader will find most of the technical details that accompany this task and which were omitted from the main text.

Before we proceed any further, let us point out that the Complementing Condition can also be described in the following non-algebraic way; see [9]. Let us assume that in a neighborhood of P the boundary Γ is flattened so that it lies on the plane $z = 0$. Then on $z \geq 0$ we consider a homogeneous, constant coefficient (frozen at P) system of partial differential equations corresponding to the principal part of the original system (5.1) with homogeneous (also constant coefficient) boundary conditions corresponding to the principal part of the boundary operator (5.2):

$$\mathcal{L}^P(P)\mathbf{u} = 0 \quad \text{in } z \geq 0 \quad (\text{A.1})$$

$$\mathcal{R}^P(P)\mathbf{u} = 0 \quad \text{on } z = 0 \quad (\text{A.2})$$

Let now $\mathbf{x} = (x, y, 0)$ and $\boldsymbol{\xi}$ be any real vector in the plane $z = 0$. The Complementing Condition requires that all solutions to (A.1) - (A.2) of the form $\mathbf{u} = e^{i \mathbf{x} \cdot \boldsymbol{\xi}} \mathbf{v}(z)$ must be identically zero, i.e. $\mathbf{v} \equiv 0$. Note that the ansatz $\mathbf{u} = e^{i \mathbf{x} \cdot \boldsymbol{\xi}} \mathbf{v}(z)$ reduces the homogeneous problem to a system of ODE's for \mathbf{v} . In addition to direct verification of Definition 4, this characterization provides an alternative way for establishing the Complementing Condition.

A.1 Velocity-Vorticity-Pressure Equations

In this section we continue the discussion started in Example 2, Chapter 6 and proceed to verify the Complementing Condition for the velocity-vorticity-pressure Stokes equations (7.1)-(7.3) with the velocity boundary condition (7.4), in two dimensions. The symbol of the operator \mathcal{L} in (7.1)-(7.3) is given by

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\xi}) = \begin{pmatrix} \xi_2 & \xi_1 & 0 & 0 \\ -\xi_1 & \xi_2 & 0 & 0 \\ -1 & 0 & -\xi_2 & \xi_1 \\ 0 & 0 & \xi_1 & \xi_2 \end{pmatrix} \quad (\text{A.3})$$

and the symbol of the boundary operator \mathcal{R} is

$$\mathcal{R}(\mathbf{x}, \boldsymbol{\xi}) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \quad (\text{A.4})$$

Let us now show that the first-order operator (7.1)-(7.3), augmented with the velocity boundary condition (7.4) cannot be homogeneous elliptic. To do this we assign the same weight to all equations and the same weight to all unknowns. In particular, we can choose $s_1 = s_2 = s_3 = s_4 = 0$ for the equations and $t_1 = t_2 = t_3 = t_4 = 1$ for the unknowns:

0	$\boldsymbol{\omega}_y$	p_x	0	0
0	$-\boldsymbol{\omega}_x$	p_y	0	0
0	$-\boldsymbol{\omega}$	0	$-u_{1y}$	u_{2x}
0	0	0	u_{1x}	u_{2y}
s/t	1	1	1	1

The symbol of the principal part according to these weights will be

$$\mathcal{L}^p(\mathbf{x}, \boldsymbol{\xi}) = \begin{pmatrix} \xi_2 & \xi_1 & 0 & 0 \\ -\xi_1 & \xi_2 & 0 & 0 \\ 0 & 0 & -\xi_2 & \xi_1 \\ 0 & 0 & \xi_1 & \xi_2 \end{pmatrix}. \quad (\text{A.5})$$

The weights s_i and t_j must be such that \mathcal{L}^p is uniformly elliptic. A simple calculation shows that

$$\det \mathcal{L}^p(\mathbf{x}, \boldsymbol{\xi}) = -(\xi_1^2 + \xi_2^2)^2 = -|\boldsymbol{\xi}|^4$$

and hence the uniform ellipticity condition

$$A^{-1}|\boldsymbol{\xi}|^{2m} \leq |\det \mathcal{L}^p(\mathbf{x}, \boldsymbol{\xi})| \leq A|\boldsymbol{\xi}|^{2m}$$

holds for $m = 2$ with $A = 1$. Before we proceed with the Complementing Condition we recall that in the two dimensions we must also check the Supplementary Condition (see Definition 3, Chapter 6).

Proposition 1 \mathcal{L}^p satisfies the Supplementary Condition.

Proof. We must show that for every pair of linearly independent real vectors $\boldsymbol{\xi}$, $\boldsymbol{\xi}'$ the polynomial $\det \mathcal{L}(\mathbf{x}, \boldsymbol{\xi} + \tau \boldsymbol{\xi}')$ in the complex variable τ has exactly m roots with positive imaginary part. Consider the equation

$$\begin{aligned} \det \mathcal{L}^p(\mathbf{x}, \boldsymbol{\xi} + \tau \boldsymbol{\xi}') &= -|\boldsymbol{\xi} + \tau \boldsymbol{\xi}'|^4 = \\ &= -(|\boldsymbol{\xi}|^2 + 2\tau(\boldsymbol{\xi}, \boldsymbol{\xi}') + \tau^2|\boldsymbol{\xi}'|^2)^2 = 0 \end{aligned}$$

The roots of the quadratic equation inside are

$$\tau_{1,2} = \frac{-(\boldsymbol{\xi}, \boldsymbol{\xi}') \pm \sqrt{(\boldsymbol{\xi}, \boldsymbol{\xi}')^2 - |\boldsymbol{\xi}|^2|\boldsymbol{\xi}'|^2}}{|\boldsymbol{\xi}'|^2}$$

We note that

$$(\boldsymbol{\xi}, \boldsymbol{\xi}')^2 < |\boldsymbol{\xi}|^2|\boldsymbol{\xi}'|^2$$

whenever $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ are linearly independent. Hence there will be exactly two roots with positive imaginary parts as required by the Supplementary Condition. \square

We are now prepared to show that the Complementing Condition does not hold for the velocity boundary condition (7.4) and the principal part (A.5). This principal part corresponds to the assumption that (7.1)-(7.3) with the velocity boundary condition is *homogeneous elliptic*.

Recall that for an invertible matrix A the adjoint A' is defined by $A' = \det A \cdot A^{-1}$. A tedious calculation shows that the adjoint of $\mathcal{L}^p(\mathbf{x}, \boldsymbol{\xi} + \tau \mathbf{n})$ is given by

$$\begin{aligned} \mathcal{L}'(\mathbf{x}, \boldsymbol{\xi} + \tau \mathbf{n}) &= \tag{A.6} \\ -|\boldsymbol{\xi}_1 + \tau \mathbf{n}|^2 &\begin{pmatrix} (\xi_2 + \tau n_2) & -(\xi_1 + \tau n_1) & 0 & 0 \\ (\xi_1 + \tau n_1) & (\xi_2 + \tau n_2) & 0 & 0 \\ 0 & 0 & -(\xi_2 + \tau n_2) & (\xi_1 + \tau n_1) \\ 0 & 0 & (\xi_1 + \tau n_1) & (\xi_2 + \tau n_2) \end{pmatrix} \end{aligned}$$

For simplicity let $|\boldsymbol{\xi}| = 1$, $|\mathbf{n}| = 1$, then since $(\boldsymbol{\xi}, \mathbf{n}) = 0$

$$\det \mathcal{L}^p(\mathbf{x}, \boldsymbol{\xi} + \tau \mathbf{n}) = -|\boldsymbol{\xi} + \tau \mathbf{n}|^4 = -(1 + \tau^2)^2$$

Therefore $\tau_1^+ = \tau_2^+ = i$ and $M^+(\boldsymbol{\xi}, \tau) = (\tau - i)^2$. The velocity boundary conditions do not involve differentiation and $\mathcal{R}^p(\mathbf{x}, \boldsymbol{\xi} + \tau \mathbf{n})$ is identical to (A.4):

$$\mathcal{R}^p(\mathbf{x}, \boldsymbol{\xi} + \tau \mathbf{n}) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

A simple calculation shows that $\mathcal{R}^p(\mathbf{x}, \boldsymbol{\xi} + \tau \mathbf{n}) \cdot \mathcal{L}'(\mathbf{x}, \boldsymbol{\xi} + \tau \mathbf{n})$ is the following matrix

$$(1 + \tau^2) \begin{pmatrix} -(\xi_2 + \tau n_2) & (\xi_1 + \tau n_1) & 0 & 0 \\ (\xi_1 + \tau n_1) & (\xi_2 + \tau n_2) & 0 & 0 \end{pmatrix}$$

The Complementing Condition will hold if the rows of this matrix are linearly independent modulo M^+ , i.e., one must verify that

$$(1 + \tau^2)(-C_1(\xi_2 + \tau n_2) + C_2(\xi_1 + \tau n_1)) = (\tau - i)^2 p_1(\tau) \quad (\text{A.7})$$

$$(1 + \tau^2)(C_1(\xi_1 + \tau n_1) + C_2(\xi_2 + \tau n_2)) = (\tau - i)^2 p_2(\tau) \quad (\text{A.8})$$

$$0 = (\tau - i)^2 p_3(\tau) \quad (\text{A.9})$$

$$0 = (\tau - i)^2 p_4(\tau) \quad (\text{A.10})$$

is possible only when $C_1 = C_2 = 0$ where $p_i(\tau)$; $i = 1, 4$ are some polynomials. By choosing $p_3(\tau) = p_4(\tau) \equiv 0$ (A.9) and (A.10) are trivially satisfied for all possible C_1 and C_2 so we may disregard them. A further simplification occurs when $(\tau - i)$ is factored from (A.7), (A.8). Then the left hand sides in (A.7) and (A.8) become a second degree polynomials in τ and we may set $p_1(\tau) = A_1(\tau + i)$; $p_2(\tau) = A_2(\tau + i)$ and factor it immediately. All this simplifies (A.7)-(A.10) to

$$(-C_1(\xi_2 + \tau n_2) + C_2(\xi_1 + \tau n_1)) = (\tau - i)A_1 \quad (\text{A.11})$$

$$(C_1(\xi_1 + \tau n_1) + C_2(\xi_2 + \tau n_2)) = (\tau - i)A_2 \quad (\text{A.12})$$

Without loss of generality we may assume that the coordinate axes are aligned with the directions of $\boldsymbol{\xi}$ and \mathbf{n} so that $\boldsymbol{\xi} = (1, 0)$ and $\mathbf{n} = (0, -1)$. Then (A.11) and (A.12) will hold for $C_1 = i$, $C_2 = 1$ and $A_1 = i$, $A_2 = -1$ and therefore the Complementing Condition is not satisfied.

Let us now show that if we assume different orders of differentiability for the unknown functions, i.e., that (7.1)-(7.3) is not homogeneous elliptic, then the Complementing Condition will hold for the velocity boundary condition. This requires us to choose different weights for the equations and different weights for the unknowns. In particular, we choose $s_1 = s_2 = -1$, $s_3 = s_4 = 0$ for the equations and $t_1 = t_2 = 2$, $t_3 = t_4 = 1$ for the unknowns. Now

from

0	ω_y	p_x	0	0
0	$-\omega_x$	p_y	0	0
-1	$-\omega$	0	$-u_{1y}$	u_{2x}
-1	0	0	u_{1x}	u_{2y}
s/t	1	1	2	2

it is easy to see that the symbol of the new principal part is given by

$$\mathcal{L}^p(\mathbf{x}, \boldsymbol{\xi}) = \begin{pmatrix} \xi_2 & \xi_1 & 0 & 0 \\ -\xi_1 & \xi_2 & 0 & 0 \\ -1 & 0 & -\xi_2 & \xi_1 \\ 0 & 0 & \xi_1 & \xi_2 \end{pmatrix}. \quad (\text{A.13})$$

Again

$$\det \mathcal{L}^p(\mathbf{x}, \boldsymbol{\xi}) = -(\xi_1^2 + \xi_2^2)^2 = -|\boldsymbol{\xi}|^4$$

and the uniform ellipticity and the Supplementary Condition clearly hold.

Let $\boldsymbol{\eta} = \boldsymbol{\xi} + \tau \mathbf{n}$, i.e.,

$$\eta_1 = \xi_1 + \tau n_1; \quad \eta_2 = \xi_2 + \tau n_2$$

Then for the adjoint of $\mathcal{L}^p(\mathbf{x}, \boldsymbol{\xi} + \tau \mathbf{n})$ we find

$$\mathcal{L}'(\mathbf{x}, \boldsymbol{\xi} + \tau \mathbf{n}) = - \begin{pmatrix} \eta_2 |\boldsymbol{\eta}|^2 & -\eta_1 |\boldsymbol{\eta}|^2 & 0 & 0 \\ \eta_1 |\boldsymbol{\eta}|^2 & \eta_2 |\boldsymbol{\eta}|^2 & 0 & 0 \\ -\eta_2^2 & \eta_1 \eta_2 & -\eta_2 |\boldsymbol{\eta}|^2 & \eta_1 |\boldsymbol{\eta}|^2 \\ \eta_1 \eta_2 & \eta_1^2 & \eta_1 |\boldsymbol{\eta}|^2 & \eta_2 |\boldsymbol{\eta}|^2 \end{pmatrix}. \quad (\text{A.14})$$

Let us choose again $|\boldsymbol{\xi}| = 1$, $|\mathbf{n}| = 1$, then $|\boldsymbol{\eta}|^2 = (1 + \tau^2)$ and $\mathcal{R}^p(\mathbf{x}, \boldsymbol{\xi} + \tau \mathbf{n}) \cdot \mathcal{L}'_A(\mathbf{x}, \boldsymbol{\xi} + \tau \mathbf{n})$ will be the following matrix:

$$\begin{pmatrix} -\eta_2^2 & \eta_1 \eta_2 & -\eta_2(1 + \tau^2) & \eta_1(1 + \tau^2) \\ \eta_1 \eta_2 & -\eta_1^2 & \eta_1(1 + \tau^2) & \eta_2(1 + \tau^2) \end{pmatrix}.$$

The rows of the latter matrix will be linearly independent modulo $M^+ = (\tau - i)^2$ if the identities

$$-C_1(\xi_2 + \tau n_2)^2 + C_2(\xi_1 + \tau n_1)(\xi_2 + \tau n_2) = (\tau - i)^2 p_1(\tau) \quad (\text{A.15})$$

$$C_1(\xi_1 + \tau n_1)(\xi_2 + \tau n_2) - C_2(\xi_1 + \tau n_1)^2 = (\tau - i)^2 p_2(\tau) \quad (\text{A.16})$$

$$(1 + \tau^2)(-C_1(\xi_2 + \tau n_2) + C_2(\xi_1 + \tau n_1)) = (\tau - i)^2 p_3(\tau) \quad (\text{A.17})$$

$$(1 + \tau^2)(C_1(\xi_1 + \tau n_1) + C_2(\xi_2 + \tau n_2)) = (\tau - i)^2 p_4(\tau) \quad (\text{A.18})$$

can only hold with $C_1 = C_2 = 0$.

We will show that (A.15) and (A.16) cannot be verified unless $C_1 = C_2 = 0$. Indeed the left hand sides in (A.15) and (A.16) are second degree polynomials, hence $p_3(\tau)$ and $p_4(\tau)$ must be constant polynomials. Again, without loss of generality we may assume that the coordinate axes are aligned with the directions of $\boldsymbol{\xi}$ and \mathbf{n} so that $\boldsymbol{\xi} = (1, 0)$ and $\mathbf{n} = (0, -1)$. With this assumption (A.15) and (A.16) become

$$-C_1\tau^2 - C_2\tau = A_3(\tau - i)^2 \quad (\text{A.19})$$

$$-C_1\tau - C_2 = A_4(\tau - i)^2 \quad (\text{A.20})$$

The right hand side of (A.20) is a second degree polynomial and an equality is possible if and only if $A_4 = C_1 = C_2 \equiv 0$. Hence the Complementing Condition holds.

Example 3 *Let us show that the result concerning validity of the Complementing Condition under equal differentiability assumption is sharp. More precisely; see [56], consider Ω given by the unit square and let $\nu = 1$, $q = 0$, $\boldsymbol{\omega}_n = -\cos(nx) \exp(ny)$, $p_n = \sin(nx) \exp(ny)$, and $\mathbf{u}_n \equiv \mathbf{0}$. Then, (7.15) would imply that*

$$\begin{aligned} O(\exp(n)) &\sim \|\mathbf{curl} \boldsymbol{\omega}_n + \mathbf{grad} p_n\|_0 + \|\mathbf{curl} \mathbf{u}_n - \boldsymbol{\omega}_n\|_0 + \|\mathbf{div} \mathbf{u}_n\|_0 \\ &\geq C(\|\mathbf{u}_n\|_1 + \|\boldsymbol{\omega}_n\|_1 + \|p_n\|_1) \sim O(n \exp(n)) \end{aligned}$$

which is a contradiction. This counterexample can also be extended to three dimensions; see [48].

Remark 1 *Along similar lines one can verify that the boundary operator (7.17) satisfies the complementing condition with both principal parts.*

A.2 Velocity-Pressure-Stress Equations

In this section we present some of the details concerning application of ADN theory to the velocity-pressure-stress equations (7.18). For the sake of brevity we shall limit our discussion to the case of two-dimensions. We assume that the unknowns are ordered as:

$$U = (T_1, T_2, T_3, p, u_1, u_2),$$

where $T_1 = T_{11}$, $T_2 = T_{12}$ and $T_3 = T_{22}$, and that the six differential equations in (7.18) are ordered as

$$\mathcal{L}U = \begin{pmatrix} T_1 - \sqrt{2\nu} \frac{\partial u_1}{\partial x} \\ 2T_2 - \sqrt{2\nu} \left(\frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right) \\ T_3 - \sqrt{2\nu} \frac{\partial u_2}{\partial y} \\ \frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y} \\ \sqrt{2\nu} \left(\frac{\partial T_1}{\partial x} + \frac{\partial T_2}{\partial y} \right) - \frac{\partial p}{\partial x} \\ \sqrt{2\nu} \left(\frac{\partial T_2}{\partial x} + \frac{\partial T_3}{\partial y} \right) - \frac{\partial p}{\partial y} \end{pmatrix}. \quad (\text{A.21})$$

According to these ordering agreements we choose the following indices

$$t_1 = t_2 = t_3 = t_4 = 1, t_5 = t_6 = 2$$

$$s_1 = s_2 = s_3 = s_4 = -1, s_5 = s_6 = 0$$

for the unknowns and the differential equations, respectively. For this choice of indices we have that

$$\mathcal{L}^P = \mathcal{L},$$

where \mathcal{L} is defined in (A.21) and that

$$\det \mathcal{L}^p(\mathbf{x}, \boldsymbol{\xi}) = \det \mathcal{L}(\mathbf{x}, \boldsymbol{\xi}) = -\nu(\xi_1^2 + \xi_2^2)^2 = -\nu|\boldsymbol{\xi}|^4.$$

As a result, the uniform ellipticity condition

$$C_e^{-1}|\boldsymbol{\xi}|^{2m} \leq |\det \mathcal{L}^p(\mathbf{x}, \boldsymbol{\xi})| \leq C_e|\boldsymbol{\xi}|^{2m}$$

holds with $m = 2$ and $C_e = \nu$. In other words, the velocity-pressure-stress system in two-dimensions is uniformly elliptic of total order four and one must specify two conditions on the boundary Γ . This total order is the same as for the Stokes problem (2.17) in the primitive variables and therefore, one can use the same boundary operator (7.4). The boundary operator (7.4) does not involve differentiation and the choice of $t_5 = t_6 = 2$ implies that one has to take $r_1 = r_2 = -2$. Finally, it is also easy to see that \mathcal{L}^p satisfies the supplementary condition.

Note that the choice of t_j s above implies different orders of differentiability for the pressure and the stress components and the velocity field. If we assume equal orders of differentiability, *i.e.*, if we choose $t_1 = \dots = t_6 = 1$

then we must take $s_1 = \dots = s_6 = 0$ and the principal part becomes

$$\mathcal{L}^p U = \begin{pmatrix} -\sqrt{2\nu} \frac{\partial u_1}{\partial x} \\ -\sqrt{2\nu} \left(\frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right) \\ -\sqrt{2\nu} \frac{\partial u_2}{\partial y} \\ \frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y} \\ \sqrt{2\nu} \left(\frac{\partial T_1}{\partial x} + \frac{\partial T_2}{\partial y} \right) - \frac{\partial p}{\partial x} \\ \sqrt{2\nu} \left(\frac{\partial T_2}{\partial x} + \frac{\partial T_3}{\partial y} \right) - \frac{\partial p}{\partial y} \end{pmatrix}.$$

This principal part corresponds to a hypothesis that the velocity-pressure-stress Stokes operator may also be homogeneous elliptic. A simple calculation however, shows that $\det \mathcal{L}^p(\mathbf{x}, \boldsymbol{\xi}) = 0$ for all $\boldsymbol{\xi}$, i.e., the problem (7.18) is not elliptic in the sense of [11] *under the assumption of an equal differentiability*. The interpretation of this fact is that the velocity-pressure-stress system is not well posed if one assumes that all unknowns belong to $H^1(\Omega)$. As a result,

for this system the possibility that the differential operator may be homogeneous elliptic is already ruled out by the fact that the principal part corresponding to such an assumption is not elliptic!

A well-posed system will result if one assumes that $T_{ij}, p \in H^1(\Omega)$ and that $\mathbf{u} \in H^2(\Omega)^2$. This situation is quite different compared with the velocity-vorticity-pressure form of the Stokes equations considered in the previous section. For this first-order system there exist *two distinct* sets of indices and two distinct elliptic principal parts, one of which corresponds to a homogeneous elliptic operator!

Next we verify the complementing condition. Let \mathbf{n} be the outer unit normal vector to Γ at some point P and let $\boldsymbol{\xi}$ be a unit tangent vector to Γ at the same point. Then

$$\det \mathcal{L}^p(\mathbf{x}, \boldsymbol{\xi} + \tau \mathbf{n}) = \nu(1 + \tau^2)^2$$

and $M^+(\boldsymbol{\xi}, \tau) = (\tau - i)^2$. Without loss of generality we may assume that the coordinate axes are aligned with the directions of $\boldsymbol{\xi}$ and \mathbf{n} so that $\boldsymbol{\xi} = (1, 0)$ and $\mathbf{n} = (0, -1)$. Then, (6.2) reduces to

$$\begin{aligned} c_1 \tau^2 - c_2 \tau &= (\tau - i)^2 p_1(\tau) \\ c_1(\tau^3 - \tau) + c_2(\tau^2 - \tau) &= (\tau - i)^2 p_2(\tau) \\ c_1 \nu(\tau^2 + 1) - c_2 \nu \tau(\tau^2 + 1) &= (\tau - i)^2 p_3(\tau) \\ c_1 \tau - c_2 &= (\tau - i)^2 p_4(\tau) \end{aligned}$$

where c_i are constants and $p_i(\tau)$ are polynomials. Note that on the last line the right-hand side is at least a second degree polynomial, whereas the left-hand side is at most a first degree polynomial. Hence identity is possible if and only if $c_1 = c_2 = 0$, *i.e.* the complementing condition holds.

Bibliography

- [1] R. Adams, *Sobolev Spaces*, Academic, New York, 1975.
- [2] A. Aziz (editor), *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, Academic Press, 1972.
- [3] P. Ciarlet, *Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
- [4] V. Girault and P. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer, Berlin, 1986.
- [5] M. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows*, Academic, Boston, 1989.
- [6] B. N. Jiang, *The least-squares finite element method. Theory and applications in computational fluid dynamics and electromagnetics*. Springer, 1998.
- [7] Ladyzhenskaya, O. *The Mathematical Theory of Viscous Incompressible Flows*, Gordon and Breach, New York, 1969.
- [8] J. Lions and E. Magenes, *Nonhomogeneous Elliptic Boundary Value Problems and Applications*, Vol. I, Springer-Verlag, Berlin 1972.
- [9] M. Renardy and R. Rogers, *Introduction to Partial Differential Equations*, TAM 13, Springer-Verlag, Berlin, 1993.
- [10] W. Wedland, *Elliptic Systems in the Plane*, Pitman, London, 1979.
- [11] S. Agmon, A. Douglis, and L. Nirenberg; Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions II, *Comm. Pure Appl. Math.*, 17 (1964) pp. 35-92.

- [12] S. G. Krein, Yu. I. Petunin; Scales of Banach spaces, *Russian Math. Surveys*, 21/2 (1966) pp. 85–160.
- [13] Ya. A. Roitberg and Z. Seftel; A theorem on homeomorphisms for elliptic systems and its applications, *Math. USSR Sbornik*, 7 (1969) pp. 439–465.
- [14] Ya. A. Roitberg; A theorem about the complete set of isomorphisms for systems elliptic in the sense of Douglis and Nirenberg, *Ukrain. Mat. Zh.* (1975) pp. 447–450.
- [15] L. R. Volevich; A problem of linear programming arising in differential equations, *Uspekhi Mat. Nauk*, Vol. 18, No.3, (1963) pp. 155–162.
- [16] Babuška, I., Error bounds for finite element method, *Numer. Math.* 16 (1971) pp. 322–333.
- [17] I. Babuška, The finite element method with Lagrange multipliers, *Numer. Math.*, 20 (1973) pp. 179–192.
- [18] M. Bercovier, Perturbation of mixed variational problems. Application to mixed finite element methods, *RAIRO Anal. Numer.*, 12 (1978) pp. 211–236.
- [19] F. Brezzi, On existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers, *RAIRO Model. Math. Anal. Numer.*, 21 (1974) pp. 129–151.
- [20] R. Falk, An analysis of the penalty method and extrapolation for the stationary Stokes equations, in *Advances in Computer Methods for Partial Differential Equations*, ed. by R. Vichnevetsky, AICA, (1975) pp. 66–69.
- [21] M. Fortin and R. Glowinski, Augmented Lagrangian methods: applications to numerical solution of boundary value problems, *Studies in Mathematics and its applications*, Vol. 15, Eds. J. L. Lions, G. Papanicolaou, R. T. Rockafellar and H. Fujita, North-Holland 1983.
- [22] 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.

- [23] J.T. Oden, N. Kikuchi, Finite element methods for constrained problems in elasticity, *Int. J. Numer. Meth. Engrg.* 18 (1982) pp. 701–725.
- [24] R. C. Almeida and R. S. Silva, A stable Petrov-Galerkin method for convection-dominated problems, *Comput. Meth. Appl. Mech. Engrg.*, 140 (1997) pp. 291–304.
- [25] R. Becker, M. Braack, A modification of the least-squares stabilization for the Stokes equations, Report 03/00 (2000) University of Heidelberg,
- [26] 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 (1993) pp. 31–48.
- [27] F. Brezzi, M.-O. Bristeau, L. Franca, M. Mallet and G. Roge. A relationship between stabilized finite element methods and the Galerkin method with bubble functions, *Comput. Meth. Appl. Mech. Engrg.*, 96 (1992) pp. 117–129.
- [28] F. Brezzi and J. Douglas, Stabilized mixed methods for the Stokes problem, *Numer. Math.*, 53 (1988) pp. 225–235.
- [29] F. Brezzi and J. Pitkaranta, On the stabilization of finite element approximations of the Stokes problem, in *Efficient Solutions of elliptic systems*, ed. by W. Hackbush, Vieweg, Braunschweig (1984) pp. 11–19.
- [30] E. G. Do Carmo and A. Galeao, A consistent upwind Petrov-Galerkin method for convection-dominated problems, *Comput. Meth. Appl. Mech. Engrg.* 68 (1988) pp. 83–95.
- [31] E. G. Do Carmo and A. Galeao, Feedback Petrov-Galerkin methods for convection-dominated problems, *Comput. Meth. Appl. Mech. Engrg.* 88 (1991) pp. 1–16.
- [32] J. Douglas and J. Wang, An absolutely stabilized finite element method for the Stokes problem. *Math. Comp.*, 52 (1989) pp. 495–508.
- [33] L. P. Franca and R. Stenberg; Error analysis of some Galerkin least-squares methods for the elasticity equations, *SIAM J. Numer. Anal.*, 28/6 (1991) pp. 1680–1697.
- [34] L. P. Franca, S. Frey, and T.J.R. Hughes, Stabilized finite element methods: I. Application to the advective-diffusive model, *Comput. Meth. Appl. Mech. Engrg.* 95 (1992) pp. 253–276.

- [35] L. Franca and A. Russo, Recovering SUPG using Petrov-Galerkin formulations enriched with adjoint residual free bubbles, *Comput. Meth. Appl. Mech. Engrg.*, 182 (2000) pp. 333–339.
- [36] L.P. Franca and A. Russo, Deriving Upwinding, Mass Lumping and Selective Reduced Integration by Residual-Free Bubbles, *Appl. Math. Lett.* 9/5 (1996) pp. 83–88.
- [37] T.J.R. Hughes and D.S. Malkus, Mixed finite element methods - reduced and selective integration techniques - a unification of concepts, *Comp. Meth. Appl. Mech. Engrg.*, 15 (1978)
- [38] T. Hughes, W. Liu and A. Brooks, Finite element analysis of incompressible viscous flows by the penalty function formulation, *J. Comput. Phys.*, 30 (1979) pp. 1–60.
- [39] 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 (1982) pp. 47-65.
- [40] 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 (1986) pp. 341–355.
- [41] 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 (1986) pp. 85–99.
- [42] 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 (1987) pp. 85–96.
- [43] K. Jansen, S. Collis, C. Whiting and F. Shakib, *A better consistency for low-order stabilized finite element methods*, submitted.
- [44] C. Johnson, U. Navert and J. Pitkaranta, Finite element methods for linear hyperbolic problems, *Comput. Meth. Appl. Mech. Engrg.* 45 (1984) pp. 285–312.

- [45] C. Johnson and J. Pitkaranta, Analysis of some mixed finite element methods related to reduced integration, *Research Report*, Dept. of Comp. Science, Chalmers University of Technology, Fall 1980.
- [46] A. Aziz, R. Kellogg, and A. Stephens; Least-squares methods for elliptic systems, *Math. of Comp.*, 44/169 (1985) pp. 53-70.
- [47] P. Bochev, *Least-squares finite element methods for the Stokes and Navier-Stokes equations*, Ph.D. Thesis, Virginia Tech, Blacksburg, 1994.
- [48] P. Bochev, Analysis of least-squares finite element methods for the Navier-Stokes equations, *SIAM J. Num. Anal.*, 34/5 (1997) pp. 1817–1844.
- [49] P. Bochev, Experiences with negative norm least-squares methods for the Navier-Stokes equations, *ETNA*, 6 (1997) pp. 44-62.
- [50] P. Bochev, Negative norm least-squares methods for the velocity-vorticity-pressure Navier-Stokes equations, *Numerical Methods in PDE's*, 15 (1999) pp. 237–256.
- [51] P. Bochev, Z. Cai, T. Manteuffel, and S. McCormick, First order least squares for the Navier-Stokes equations, *Proc. Seventh Copper Mountain Multigrid Conference*, NASA Conference Publication 3339, Part 1 (1996) pp. 41–55.
- [52] P. Bochev, Z. Cai, T. Manteuffel, and S. McCormick, Analysis of velocity-flux least squares methods for the Navier-Stokes equations, Part-I *SIAM. J. Num. Anal.* 35/3 (1998) pp. 990-1009.
- [53] P. Bochev, T. Manteuffel, and S. McCormick Analysis of velocity-flux least squares methods for the Navier-Stokes equations, Part-II *SIAM. J. Num. Anal.*, 36/4 (1999) pp. 1125–1144,
- [54] P. Bochev and M. Gunzburger, A least-squares finite element method for the Navier-Stokes equations, *Appl. Math. Lett.*, 6 (1993) pp. 27–30.
- [55] P. Bochev and M. Gunzburger, Accuracy of least-squares methods for the Navier-Stokes equations, *Comput. & Fluids*, 22 (1993) pp. 549–563
- [56] P. Bochev and M. Gunzburger, Analysis of least-squares finite element methods for the Stokes equations, *Math. Comp.*, 63 (1994) pp. 479–506.

- [57] P. Bochev and M. Gunzburger, Analysis of weighted least-squares finite element method for the Navier-Stokes equations, *Proceedings of the 14th IMACS World Congress*, Georgia Tech, Atlanta, (1994) pp. 584–587.
- [58] P. Bochev and M. Gunzburger, Least-squares for the velocity-pressure-stress formulation of the Stokes equations, *Comput. Meth. Appl. Mech. Engrg.*, 126 (1995) pp. 267–287.
- [59] P. Bochev and M. Gunzburger, Least-squares finite element methods for elliptic equations, *SIAM Review*, 40/4 (1998) pp. 789–837.
- [60] J. Bramble and J. Nitsche, A generalized Ritz-least-squares method for Dirichlet problems, *SIAM J. Numer. Anal.*, 10 (1973) pp. 81–93.
- [61] J. Bramble and A. Schatz, Least-squares methods for $2m$ th order elliptic boundary value problems, *Math. Comp.*, 25 (1971) pp. 1–32.
- [62] J. Bramble, R. Lazarov, and J. Pasciak, A least squares approach based on a discrete minus one inner product for first order systems, *Technical Report 94-32*, Mathematical Science Institute, Cornell University, 1994.
- [63] J. Bramble and J. Pasciak, Least-squares methods for Stokes equations based on a discrete minus one inner product, *J. Comp. App. Math.*, 74 (1996) pp. 155–173.
- [64] J. Bramble, R. Lazarov, and J. Pasciak, Least-squares for second order elliptic problems, *Computer Methods in Appl. Mech. Engrg.*, 152 (1998) pp. 195–210.
- [65] 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 (1994) pp. 1785–1799.
- [66] Z. Cai, T. Manteuffel, and S. McCormick, First-order system least-squares for second order partial differential equations: Part II, *SIAM J. Numer. Anal.*, 34 (1997) pp. 425–454.
- [67] Z. Cai, T. Manteuffel, and S. McCormick, First-order system least-squares for the Stokes equations, with application to linear elasticity, *SIAM J. Num. Anal.*, 34 (1997) pp. 1727–1741.
- [68] Z. Cai, T. Manteuffel, and S. McCormick, First-order system least-squares for velocity-vorticity-pressure form of the Stokes equations, with application to linear elasticity, *ETNA*, 3 (1995) pp. 150–159.

- [69] Z. Cai, T. Manteuffel, S. McCormick, and S. Parter, First-order system least-squares for planar elasticity - pure traction, *SIAM J. Numer. Anal.*, 35 (1998) pp. 320–335.
- [70] Z. Cai, T. Manteuffel, S. McCormick, and J. Ruge, First-order system \mathcal{LL}^* (FOSLL*): scalar elliptic partial differential equations, submitted.
- [71] G. Carey, A. Pehlivanov, and R. Lazarov, Least-squares mixed finite element methods for second order elliptic problems, *SIAM J. Numer. Anal.*, 31 (1994) pp. 1368–1377.
- [72] G. Carey and A. Pehlivanov Error estimates for least-squares mixed finite elements, *Math. Mod. Numer. Anal.*, 28 (1994) pp. 499–516
- [73] G. Carey, A. Pehlivanov, and P. Vassilevski, Least-squares mixed finite element methods for non-self-adjoint problems: I. Error estimates, *Numer. Math.* (1996) pp. 501–522.
- [74] G. Carey, A. Pehlivanov, and P. Vassilevski, Least-squares mixed finite element methods for non-self-adjoint problems: II. Performance of block-ILU factorization methods, *SIAM J. Sci. Comput.*, 16 (1995) pp. 1126–1136.
- [75] C. Chang and M. Gunzburger, A finite element method for first order elliptic systems in three dimensions, *Appl. Math. Comp.*, 23 (1987) pp. 171–184.
- [76] C. Chang, Finite element method for the solution of Maxwell’s equations in multiple media, *Appl. Math. Comp.*, 25 (1988) pp. 89–99.
- [77] C. Chang, A least squares finite element method for the Helmholtz equation, *Comp. Meth. Appl. Mech. Engrg.*, 83 (1990) pp. 1–7.
- [78] C. Chang, A mixed finite element method for the Stokes problem: an acceleration pressure formulation, *Appl. Math. Comp.*, 36 (1990) pp. 135–146.
- [79] C. Chang and M. Gunzburger, A subdomain-Galerkin/least squares method for first order elliptic systems in the plane, *SIAM J. Numer. Anal.*, 27 (1990) pp. 1197–1211.
- [80] C. Chang, Finite element approximation for grad-div type systems in the plane, *SIAM J. Numer. Anal.*, 29 (1992) pp. 452–461.

- [81] C. Chang, An error estimate of the least squares finite element method for the Stokes problem in three dimensions, *Math. Comp.*, 63 (1994) pp. 41–50.
- [82] C. Chang, An error analysis of least squares finite element method of velocity-pressure-vorticity formulation for Stokes problem: correction, *Mathematics Research Report # 95-53*, Department of Mathematics, Cleveland State University, Cleveland, 1995.
- [83] C. Chang, A least squares finite element method for incompressible flow in stress-velocity-pressure version, *Comp. Meth. Appl. Mech. Engrg.*, to appear.
- [84] C. Chang and B. Jiang, An error analysis of least squares finite element method of velocity-pressure-vorticity formulation for Stokes problem, *Comp. Meth. Appl. Mech. Engrg.*, 84 (1990) pp. 247–255.
- [85] C. Chang J. Nelson, Least squares finite element method for the Stokes problem with zero residual of mass conservation, *SIAM J. Numer. Anal.*, 34 (1997) pp. 480–489.
- [86] C. Chang, Least squares finite element for second order boundary value problems, *Appl. Math. Comp.*, to appear.
- [87] J. Deang and M.D. Gunzburger, Issues related to least-squares finite element methods for the Stokes equations, *SIAM J. Sci. Comp.* 20 (1998) pp. 878–906.
- [88] G. J. Fix, E. Stephan, On the finite element least squares approximation to higher order elliptic systems, *Arch. Rat. Mech. Anal.*, 91/2 (1986) pp. 137–151.
- [89] G. Fix and M. Gunzburger, On least squares approximations to indefinite problems of the mixed type, *Inter. J. Numer. Meth. Engng.*, 12 (1978) pp. 453–469.
- [90] G. Fix, M. Gunzburger and R. Nicolaides, On finite element methods of the least-squares type, *Comput. Math. Appl.*, 5 (1979) pp. 87–98.
- [91] G. Fix, M. Gunzburger and R. Nicolaides, On mixed finite element methods for first-order elliptic systems, *Numer. Math.*, 37 (1981) pp. 29–48.

- [92] D. Jespersen; A least-squares decomposition method for solving elliptic equations, *Math. Comp.*, 31 (1977) pp. 873–880.
- [93] B. N. Jiang and L. Povinelli, Optimal least-squares finite element methods for elliptic problems, *Comp. Meth. Appl. Mech. Engrg.*, 102 (1993) pp. 199–212.
- [94] B. N. Jiang and G. F. Carey, A stable least-squares finite element method for nonlinear hyperbolic problems, *Int. J. Num. Meth. Fluids*, 8 (1988) pp. 933–942.
- [95] B. N. Jiang and G. F. Carey, Least-squares finite elements for first-order hyperbolic systems, *Int. J. Num. Meth. Engrg.*, 26 (1988) pp. 81–93.
- [96] B. N. Jiang and G. F. Carey, Least-squares finite element methods for compressible Euler equations, *Int. J. Num. Meth. Fluids*, 10 (1990) pp. 557–568.
- [97] B. N. Jiang, Non-oscillatory and non-diffusive solution of convection problems by the iteratively reweighted least-squares finite element method, *J. Comp. Phys.*, 105/1 (1993) pp. 108–121.
- [98] B. Jiang and C. Chang, Least-squares finite elements for the Stokes problem, *Comput. Meth. Appl. Mech. Engrg.*, 78 (1990) pp. 297–311.
- [99] B.-N. Jiang and L. Povinelli; Least-squares finite element method for fluid dynamics, *Comput. Meth. Appl. Mech. Engrg.*, 81 (1990) pp. 13–37.
- [100] B.-N. Jiang; A least-squares finite element method for incompressible Navier-Stokes problems, *Inter. J. Numer. Meth. Fluids*, 14 (1992) pp. 843–859.
- [101] B.-N. Jiang, T. Lin, and L. Povinelli; A least-squares finite element method for 3D incompressible Navier-Stokes equations, *AIAA Report 93-0338*, (1993).
- [102] B.-N. Jiang and V. Sonnad; Least-squares solution of incompressible Navier-Stokes equations with the p-version of finite elements, *NASA TM 105203 (ICOMP Report 91-14)*, NASA, Cleveland, (1991).
- [103] B. Jiang, T. Lin, and L. Povinelli, Large-scale computation of incompressible viscous flow by least-squares finite element method, *Comput. Meth. Appl. Mech. Engrg.*, 114 (1994) pp. 213–231.

- [104] D. Lefebvre, J. Peraire, and K. Morgan; Least-squares finite element solution of compressible and incompressible flows, *Int. J. Num. Meth. Heat Fluid Flow*, 2 (1992) pp. 99–113.
- [105] G. Starke, Multilevel boundary functionals for least-squares mixed finite element methods, *SIAM, J. Num. Anal.*, 36 (1999) pp. 1065–1077.
- [106] L. Tang, T. Cheng and T. Tsang, Transient solutions for three-dimensional lid-driven cavity flows by a least-squares finite element method, *Int. J. Numer. Methods Fluids* 21 (1995) pp. 413–432.
- [107] L. Tang and T. Tsang, Transient solutions by a least-squares finite element method and Jacobi conjugate gradient technique, *Numer. Heat Trans. Part B*, 28 (1995) pp. 183–198.
- [108] L. Tang and T. Tsang, A least-squares finite element method for for time dependent incompressible flows with thermal convection, *Int. J. Numer. Methods Fluids*, 17 (1993) pp. 271–289.
- [109] L. Tang and T. Tsang, Temporal, spatial and thermal features of 3-D Rayleigh-Benard convection by least-squares finite element method, *Comp. Meth. Appl. Mech. Engrg.*, 140 (1997) pp. 201–219.
- [110] T. Tsang, X. Ding, Large eddy simulation of turbulent flows by a least-squares finite element method, submitted.
- [111] T. Tsang, X. Ding, On first-order formulations of the least-squares finite element method for incompressible flows, submitted.
- [112] T. Tsang, X. Ding, Parallelization of the least-squares finite element method, submitted.
- [113] P. Bochev and J. Choi, A Comparative Numerical Study of Least-Squares, SUPG and Galerkin Methods for Convection Problems, *Int. Jour. Comp. Fluid Dynamics*, to appear.
- [114] P. Bochev and J. Choi, Improved least-squares error estimates for scalar hyperbolic problems, *Comp. Meth. in Appl. Math.* 1/2 (2001) pp. 115–124.
- [115] R.D. Lazarov, L. Tobiska, and P. S. Vassilevski, Streamline-diffusion least-squares mixed finite element methods for convection-diffusion problems, *East-West J. Numer. Math*, 5/4 (1997) pp. 249–264.

- [116] R.D. Lazarov and P.S. Vassilevski, Least-Squares Streamline Diffusion Finite Element Approximation to Singularly Perturbed Convection-Diffusion Problems, *to appear*.
- [117] K. Miller and M. Baines, Least-Squares Moving Finite Elements, *Report No.98/06*, Oxford University Computing Laboratory, Numerical Analysis Group, 1998.
- [118] F. Taghaddosi, W. Habashi, G. Guevremont and D. Ait-ali-yahia, An Adaptive Least-Squares Method for the Compressible Euler Equations, *Int. Jour. Numer. Meth. Fluids*, 31 (1999) pp. 1121-1139.
- [119] E. Eason, A review of least-squares methods for solving partial differential equations, *Int. J. Numer. Meth. Engrg.*, 10 (1976) pp. 1021-1046
- [120] J. Milthorpe and G. Steven, On a least squares approach to the integration of the Navier-Stokes equations, in: *Finite Elements in Fluids*, Volume 3, Ed. R. H. Gallagher, Wiley (1978) pp. 89-103.
- [121] D. Zeitoun, J. Liable, and G. Pinder, A weighted least-squares method for first order hyperbolic systems, *to appear*.
- [122] D. Zeitoun and G. Pinder, A least squares approach for solving remediation problems of contaminated aquifers, *Numerical methods in water resources*, 4 (1989) pp. 329-335.
- [123] J. Liable and G. Pinder, Least-squares collocation solution of differential equations on irregularly shaped domains using orthogonal meshes, *Numer. Meth. PDE's*, 5 (1989) pp. 347-361.
- [124] J. Liable and G. Pinder, Solution of shallow water equations by least squares collocation, *Adv. Water Res.*, *to appear*.
- [125] M. Bristeau, O. Pironneau, R. Glowinski, J. Periaux, P. Perrier, and G. Poirier, Application of optimal control and finite element methods to the calculation of transonic flows, in: *Numerical Methods in Applied Fluid Dynamics*, Ed. B. Hunt, Academic Press (1980), pp. 203-312.
- [126] M. Bristeau, O. Pironneau, R. Glowinski, J. Periaux, P. Perrier, and G. Poirier, Finite element methods for transonic flow calculations, in: *Recent Advances in Numerical Methods in Fluids*; Volume 4, Ed. W. G. Habashi, Pineridge Press (1985) pp. 703-731.

- [127] R. Glowinski, J. Periaux and O. Pironneau, Transonic flow simulation by the finite element method via optimal control, in *Finite Elements in Fluids*, Volume 3, Ed. R. H. Gallagher, Wiley (1978) pp. 205–217.
- [128] R. Glowinski, B. Mantel, J. Periaux, P. Perrier and O. Pironneau, On an efficient new preconditioned conjugate gradient method. Application to the in-core solution of the Navier-Stokes equations via non-linear least-squares and finite element methods, in *Finite Elements in Fluids*, Volume 4, Edited by R. H. Gallagher, Wiley (1982) pp. 365–401.