Stable Variational Formulations

SS-2017

Wolfgang Dahmen, RWTH Aachen

Contents

1 Introduction

1.1 Preliminary Remarks

This course addresses the numerical solution of different types of partial differential equations (PDEs) under suitable side conditions such as boundary value or initial value conditions. Specifically, when the side conditions are homogeneous the PDE can be written as *operator equation*

$$\mathcal{F}(u) = f, \tag{1.1.1}$$

where the data f are given. The standard procedure is then to choose a *discretization* of the operator $\mathcal{F}$, denoted e.g. as $\mathcal{F}_h$ and replace (1.1.1) by a *finite dimensional* (linear or nonlinear) system

$$\mathcal{F}_h(\mathbf{u}) = \mathbf{f}, \tag{1.1.2}$$

where the vector of unknown coefficients $\mathbf{u}$ representa a *finite dimensional* approximation to the unknown u in (1.1.1). Assume for the moment that $\mathcal{F}$ is linear the operator $\mathcal{F}_h$ has a matrix representation (once one has fixed the a numbering of the unknowns)

$$\mathbf{A}\mathbf{u} = \mathbf{f}, \tag{1.1.3}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$, say. Recall that, according to Hadamard, a problem is *well-posed* provided that

i) There exists a solution;

ii) the solution is unique;

iii) it depends continously on the data.

Is well-posedness a simple issue for finite-dimensional problems?

Remark 1.1.1 (1.1.3) is well-posed if and only if $\det \mathbf{A} \neq 0$. In fact, then $\mathbf{A}^{-1}$ exists ($\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^n$ is a bijection) and as a linear operator is bounded with respect to any norm $\|\cdot\|$ on $\mathbb{R}^n$ (all norms on $\mathbb{R}^n$ are equivalent). □

In fact, for any perturbed data $\tilde{\mathbf{f}}$ and corresponding solution $\tilde{\mathbf{u}} = \mathbf{A}^{-1}\tilde{\mathbf{f}}$, one has

$$\|\mathbf{u} - \tilde{\mathbf{u}}\| = \|\mathbf{A}^{-1}\mathbf{f} - \mathbf{A}^{-1}\tilde{\mathbf{f}}\| = \|\mathbf{A}^{-1}(\mathbf{f} - \tilde{\mathbf{f}})\| \leq \|\mathbf{A}^{-1}\|\|\mathbf{f} - \tilde{\mathbf{f}}\|, \quad (1.1.4)$$

and $\tilde{\mathbf{u}}$ tends to $\mathbf{u}$ when $\tilde{\mathbf{f}}$ tends to $\mathbf{f}$. Here for any matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$

$$\|\mathbf{B}\| := \sup_{\mathbf{u} \in \mathbb{R}^n} \frac{\|\mathbf{B}\mathbf{u}\|}{\|\mathbf{u}\|}, \quad (1.1.5)$$

denotes the operator norm of $\mathbf{B}$ with respect to the norm $\|\cdot\|$.

Thus, it seems that, once a problem has been discretized, the issue of well-posedness is simple and reduces to invertibility of the discrete operator.

Stability Unfortunately, a closer look reveals that things are not that simple. For instance, the matrix $\mathbf{A} = \mathbf{A}_n$ depends on the dimension n of the discretization and it could well be that $\|\mathbf{A}_n^{-1}\|$ tends to infnity when n grows, impeding the continuous dependence. The discretization is therefore called *stable* if

$$\|\mathbf{A}_n^{-1}\| = \mathcal{O}(1), \quad \textbf{uniformly in } n \textbf{ as } n \to \infty. \quad (1.1.6)$$

Unfortunately, this is still not the end of the story yet. In fact, for realistic problems n is typically very large so that one cannot use direct solvers for (1.1.3). Instead, one applies an *iterative* method.

Condition numbers When $\mathbf{A}$ is symmetric positive definite the conjugate gradient scheme is an option. It is well-known that the convergence speed then depends, however, on the condition number of $\mathbf{A}_n$

$$\kappa_{\|\cdot\|}(\mathbf{A}_n) := \|\mathbf{A}_n\|\|\mathbf{A}_n^{-1}\|. \tag{1.1.7}$$

For instance, when $\mathcal{F}$ is, a second order differential operator, such as the Laplacian, as in *Poisson's equation* with homogeneous Dirichlet boundary conditions

$$-\Delta u = f \quad \text{in } \Omega, \quad u|_{\partial\Omega} = 0, \tag{1.1.8}$$

one may manage to keep the $\|\mathbf{A}_n^{-1}\|$ uniformly bounded, but one cannot expect that simultaneously the $\|\mathbf{A}_n\|$ remain uniformly bounded as well. In fact, when n grows the $\mathbf{A}_n$ better and better approximate a differential operator which is not bounded as a mapping from a given space into itself. Hence, the condition numbers $\kappa_{\|\cdot\|}(\mathbf{A}_n)$ increase with n which degrades the performance of an iterative scheme. Therefore, a lot of effort has been devoted to the development of *preconditioners* which transform the linear system into an equivalent one but with smaller (sometimes even uniformly bounded in n) condition number.

Note: *The preconditioner has to use in one way or the other the properties of the underlying* infinite-dimensional *continuous problem.*

Upshot Abandoning the underlying continuous problem in favor of a seemingly simple finite-dimesnsional discrete problem may entail severe numerical issues which are difficult to solve without using the underlying continuous model. Instead the particular structure of the discretization should exploit knowledge about the original continuous problem to a best possibel extent.

Main problems:

- "Continuous dependence" is a notion that depends on the particular norm that measures accuracy. In the infinite-dimensional case norms are no longer equivalent.

- When viewing the continuous problem (PDE) as an operator equation, the choice of norms is equivalent to saying which normed linear space $\mathbb{U}$ is mapped by the operator into which normed space $\mathbb{W}$.
- Unique solvabiliy of the operator equation (1.1.1) for *all* data in $\mathbb{W}$ means that $\mathcal{F} : \mathbb{U} \to \mathbb{W}$ should be a *bijection.* When $\mathcal{F} = \mathcal{A}$ is linear, once $\mathbb{U}$, $\mathbb{W}$ have been chosen, one can define a condition number for $\mathcal{A}$, see (1.1.10) below.

A central theme in this course is to tightly interrelate discrete and continuous models to arrive at efficient numerical techniques.

Note: *One cannot expect a numerical scheme to work well if the underlying continuous problem is not well-posed in a suitable sense.*

To explain what we mean by a well-posed continuous problem assume that $\mathcal{F} = \mathcal{A}$ is a *linear* operator (as the Laplacian in (1.1.8)). The basic conceptual ingredients can be summarized as follows:

- Interpret a given PDE as an *operator equation* (1.1.1). Specifically, when $\mathcal{F} = \mathcal{A}$ is a *linear operator* this means to *identify* a suitable pair of (normed linear) spaces $\mathbb{U}, \mathbb{W}$ for which $\mathcal{A} : \mathbb{U} \to \mathbb{W}$. Then, unique solvability (i), (ii) means that $\mathcal{A}$ is actually a bijection. Continuous dependence means that $\mathcal{A} \in \mathcal{L}(\mathbb{U}, \mathbb{W})$ (the space of bounded linear operators from $\mathbb{U}$ to $\mathbb{W}$) meaning that

$$\|\mathcal{A}\|_{\mathcal{L}(\mathbb{U},\mathbb{W})} := \sup_{u\in\mathbb{U}} \frac{\|\mathcal{A}u\|_{\mathbb{W}}}{\|u\|_{\mathbb{U}}} < \infty, \tag{1.1.9}$$

  and $\mathcal{A}^{-1} \in \mathcal{L}(\mathbb{W}, \mathbb{U})$. In this case $\mathcal{A}$ has a bounded *condition number*

$$\kappa_{\mathbb{U},\mathbb{W}}(\mathcal{A}) := \|\mathcal{A}\|_{\mathcal{L}(\mathbb{U},\mathbb{W})}\|\mathcal{A}^{-1}\|_{\mathcal{L}(\mathbb{W},\mathbb{U})}, \tag{1.1.10}$$

  and a continuous dependence of the solution in $\mathbb{U}$ is obtained in analogy to (1.1.4) with a stability constant $\|\mathcal{A}^{-1}\|_{\mathcal{L}(\mathbb{W},\mathbb{U})}$.
- The identification of the domain $\mathbb{U}$ and range $\mathbb{W}$ of the operator $\mathcal{A}$ is part of the problem, since these spaces generally depend on

the structure of the operator $\mathcal{A}$. The most powerful framework for finding suitable pairs of spaces is the concept of *weak formulations* or *variational formulation* of the operator equation. This is a central topic in this course.

- Nonlinear problems can be treated (as in Newton's method) by *linearization.* The Frechét derivative $D\mathcal{F}(v)$ at a point $v \in \mathbb{U}$ is (as is easily shown in the context of variational formulations) a mapping into $\mathbb{W}$. So well-posedness can often be reduced to well-posedness of the linearized problem for suitable neigborhoods of linrearization points v in a neighborhood of the solution.

1.2 Some Principal Consequences

In the above terminology $\kappa_{\mathbb{U},\mathbb{W}}(\mathcal{A}) < \infty$ means that $\mathcal{A}$ has a finite relative condition and hence is well-posed. The following observation indicates that it is also important to keep $\kappa_{\mathbb{U},\mathbb{W}}(\mathcal{A}) < \infty$ at moderate size.

A Posteriori Error Bounds Suppose $u_h \in \mathbb{U}$ is an approximation to the solution $u \in \mathbb{U}$ of

$$\mathcal{A}u = f. \tag{1.2.1}$$

Then

$$\begin{aligned}\|u - u_h\|_{\mathbb{U}} &= \|\mathcal{A}^{-1}(\mathcal{A}(u - u_h))\|_{\mathbb{U}} \leq \|\mathcal{A}^{-1}\|_{\mathcal{L}(\mathbb{W},\mathbb{U})}\|\mathcal{A}(u - u_h)\|_{\mathbb{W}}\\ &= \|\mathcal{A}^{-1}\|_{\mathcal{L}(\mathbb{W},\mathbb{U})}\|f - \mathcal{A}u_h\|_{\mathbb{W}}\\ &\leq \|\mathcal{A}^{-1}\|_{\mathcal{L}(\mathbb{W},\mathbb{U})}\|\mathcal{A}\|_{\mathcal{L}(\mathbb{U},\mathbb{W})}\|u - u_h\|_{\mathbb{U}}.\end{aligned} \tag{1.2.2}$$

It will be convenient to reexpress this as follows. If one has bounds

$$\|\mathcal{A}^{-1}\|_{\mathcal{L}(\mathbb{W},\mathbb{U})} \leq c_{\mathcal{A}}^{-1}, \quad \|\mathcal{A}\|_{\mathcal{L}(\mathbb{U},\mathbb{W})} \leq C_{\mathcal{A}} \tag{1.2.3}$$

for some finite constants $c_{\mathcal{A}}, C_{\mathcal{A}}$, then the error in $\mathbb{U}$ is sandwiched by the residual in $\mathbb{W}$

$$C_{\mathcal{A}}^{-1}\|f - \mathcal{A}u_h\|_{\mathbb{W}} \leq \|u - u_h\|_{\mathbb{U}} \leq c_{\mathcal{A}}^{-1}\|f - \mathcal{A}u_h\|_{\mathbb{W}}. \tag{1.2.4}$$

The significance of this relation is that

- the residual involves only known quantities and hence can, in principle, be evaluated. (It wil be seen that in practice the computation of the norm for $\mathbb{W}$ may be tricky because it is typically a dual norm).
- These computable bounds are lower and upper bounds and hence are "equivalent" to the error.
- However, the estimate of $\|u - u_h\|_{\mathbb{U}}$ is the sharper the smaller the condition number $\kappa_{\mathbb{U},\mathbb{W}}(\mathcal{A})$.

Remark 1.2.1 Therefore, from a numerical point of view well-posedness is not quite sufficient because $\kappa_{\mathbb{U},\mathbb{W}}(\mathcal{A})$ could be very large (and we'll encounter such examples). Instead, we are interested in arranging things so that $\kappa_{\mathbb{U},\mathbb{W}}(\mathcal{A})$ has a moderate size. In this case we speak of a *well-conditioned* problem. We'll discuss ways of *preconditioning* the original problem on the infinite-dimensional level. □

1.3 A Guiding Example

There are (at least) three major discretization concepts, *Finite Difference methods, Finite Volume schemes* and *Galerkin-type* methods based on weak formulations of the underlying PDE.

In this course, we confine the discussion entirely to the latter class for two major reasons:

a) The PDE often represents a physical model only under an additional regularity assumption of the solution. This is often not satisfied and a weak formulation tries to recover all physically relevant cases by relaxing the notion of solution.

b) As indicated in the previous chapter we wish to tightly relate discretizations to the mapping properties of the operator in the given operator equation. This is supported in a anatural way by schemes based on weak formulations of a PDE.

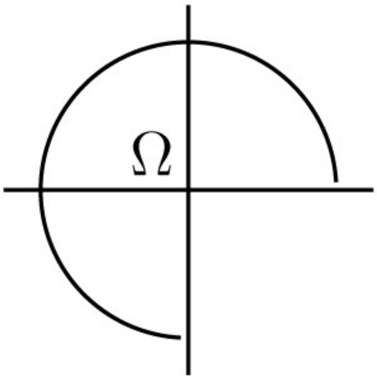
1.3.1 A Regularity Issue and Some Prerequisits

We begin with an example that illustrates a) for instance, when $\mathcal{A} = -\Delta$ is the Laplacian. In this case one could think of letting it act on $\mathbb{U} = C^2(\Omega)$, the space of twice continuously differential functions in Ω. Then Δ maps this space into the space of continuous functions $\mathbb{W} = C(\Omega)$. However, as shown below, for any data in $C(\Omega)$ one finds a solution only under additional assumptions on Ω. Instead, it will be seen that *Sobolev spaces* are better suited.

Example 1.3.1 Let

$$\Omega = \left\{(x, y) \in \mathbb{R}^2 : \; x^2 + y^2 < 1, \; x < 0 \text{ or } y > 0\right\}.$$

We can identify the complex plane $\mathbb{C}$ and $\mathbb{R}^2$ through $\mathbb{C} \ni z = x + iy \leftrightarrow (x, y) \in \mathbb{R}^2$.



Then $w(z) := z^{2/3}$ is analytic in Ω and

$$u\,(z) := \operatorname{Im}\, w\,(z)$$

is a harmonic function, i.e., $\Delta u = 0$. Hence with $g = u|_{\partial\Omega}$ this function solves

$$-\Delta u = 0, \text{ in } \Omega, \; u = g \text{ on } \partial\Omega.$$

But, since

$$w'(z) = \frac{2}{3} z^{-1/3}$$

not even the first derivatives of u remain bounded when $z \to 0$. Hence the previous consistency bounds requiring, for instance, boundedness of $\sup_{x\in\Omega} \max_{k_1+k_2\leq 4} |\partial_{x_1}^{k_1} \partial_{x_2}^{k_2} u(x)|$ are not applicable. □

This shows that different strategies are needed that work with different *regularity* notions. In fact, recall that the derivation of Poisson's equation from a diffusion model worked under the *assumption* that certain integrands are continuous. To cover also physically meaningful scenarios for which this assumption is violated the key is to *weaken* the notion of "solution".

The basic idea is easy to explain for the Poisson equation with homogeneous Dirichlet boundary conditions

$$-\Delta u(x) = f(x), \ x \in \Omega, \ u(x) = 0, \ x \in \partial\Omega. \tag{1.3.1}$$

$u \in C^2(\Omega) \cap C(\overline{\Omega})$ is called "classical" or "strong" solution, i.e., the equation is required to hold at *each point* $x \in \Omega$.

To prepare for later developments this can be re expressed as follows: recall that the operator

$$\delta_x : f \to f(x) =: \delta_x f =: \langle f, \delta_x \rangle \tag{1.3.2}$$

is a linear and bounded (hence continuous) operator from $C(\Omega)$ to $\mathbb{R}$. In fact

$$|\delta_x f| = |f(x)| \leq \sup_{x\in\Omega} |f(x)| = \|f\|_{L_\infty(\Omega)},$$

i.e.,

$$\|\delta_x\|_{L_\infty(\Omega)\to\mathbb{R}} = 1. \tag{1.3.3}$$

One also says that the Dirac functional belongs to the *dual* of $C(\Omega)$, the space of bounded linear functionals on $C(\Omega)$.

One can now say that (1.3.1) means that the equation

$$-\Delta u = f$$

holds under any "test" with Dirac functionals

$$\langle -\Delta u, \delta_x \rangle = \langle f, \delta_x \rangle, \ x \in \Omega. \tag{1.3.4}$$

As the initial example shows thus way of testing requires too much regularity of the solution and is therefore in general inappropriate.

Such "strong" tests can be weakened as follows. Recall that $\Omega \subset \mathbb{R}^d$ is open. Let $C_0^\infty(\Omega)$ denote the space of infinitely differentiable functions with compact support in Ω. Of course, whenever u satisfies (1.3.1) one certainly has

$$\int\limits_\Omega -\Delta u\left(x\right) v\left(x\right) \mathrm{d}x = \int\limits_\Omega f\left(x\right) v\left(x\right) \mathrm{d}x, \ \forall v \in C_0^\infty\left(\Omega\right). \tag{1.3.5}$$

If one interprets derivatives - in this case Δ - in the distributional sense and requires the validity of $-\Delta u = f$ only in the above sense under testing with "smooth" functions (which could be understood as "local averaging") one may indeed relax the requirements on u. We discuss next in which sense this is indeed the case.

Domains: In what follows Ω will always denote an open bounded connected domain in $\mathbb{R}^d$ whose boundary $\partial\Omega$ is a piecewise smooth and Lipshitz-Graph. This means that there exists a finite covering of $\partial\Omega$ by open sets such that on each of them the portion of $\partial\Omega$ is the graph of a Lipshitz function.

A frequently used tool is

Green's Formulae: Denoting by $n(x)$ for each point $x \in \partial\Omega$ possessing a unique tangent plane the outward unit normal vector, one has for any $v, w \in C^1(\Omega)$

$$\int_\Omega \partial_{x_i} w v \mathrm{d}x = -\int_\Omega w \partial_{x_i} v \mathrm{d}x + \int_{\partial\Omega} v w n_i \mathrm{d}s, \tag{1.3.6}$$

where n_i is the ith component of n. As a consequence one obtains for vector fields $\underline{w}$

$$\int_\Omega (\operatorname{div} \underline{w}) v \mathrm{d}x = -\int_\Omega \underline{w}^T \nabla v \mathrm{d}x + \int_{\partial\Omega} n^T \underline{w} v \mathrm{d}s. \tag{1.3.7}$$

In particular, this yields for $v \equiv 1$ Gauß' Theorem

$$\int_\Omega (\operatorname{div} \underline{w}) \mathrm{d}x = \int_{\partial\Omega} n^T \underline{w} \mathrm{d}s. \tag{1.3.8}$$

To see that the tests in (1.3.5) are indeed weaker than point-wise tests in (1.3.4) we apply (1.3.7) to obtain

$$\begin{aligned} -\int\limits_\Omega \Delta u(x)\, v(x)\, \mathrm{d}x &= \int\limits_\Omega \operatorname{div}(\nabla u(x))\, v(x)\, \mathrm{d}x \\ &= \int\limits_\Omega (\nabla u(x))^T \nabla v(x)\, \mathrm{d}x - \int\limits_{\partial\Omega} n^T \nabla u(x) \underbrace{v(x)}_{=0} \mathrm{d}x \\ &= \int\limits_\Omega (\nabla u(x))^T \nabla v(x)\, \mathrm{d}x, \end{aligned} \tag{1.3.9}$$

since $v(x) = 0,\ x \in \partial\Omega$, for $v \in C_0^\infty(\Omega)$. Thus (1.3.5) reads: find u with $u|_{\partial\Omega} = 0$, such that

$$\int\limits_\Omega (\nabla u(x))^T \nabla v(x)\, \mathrm{d}x = \int\limits_\Omega f(x)\, v(x)\, \mathrm{d}x, \quad \forall v \in C_0^\infty(\Omega). \tag{1.3.10}$$

This is called the *weak formulation* of (1.3.1).

Note that, for (1.3.10) to make sense, it is no longer required to have $u \in C^2(\Omega)$. In fact, using Cauchy-Schwarz

$$\left| \int\limits_\Omega (\nabla u(x))^T \nabla v(x)\, \mathrm{d}x \right| \le \left(\int\limits_\Omega |\nabla u(x)|^2\, \mathrm{d}x \right)^{\frac{1}{2}} \left(\int\limits_\Omega |\nabla v(x)|^2\, \mathrm{d}x \right)^{\frac{1}{2}}, \tag{1.3.11}$$

(where $|\nabla u(x)|^2 = \sum_{i=1}^{d} |\frac{\partial}{\partial x_i} u(x)|^2$ stands for the Euclidean norm) one sees that the gradient of u just needs to be *square integrable.*

Moreover, the test functions v were chosen from $C_0^\infty(\Omega)$ just for convenience. One can do integration by parts and saying that $v|_{\partial\Omega} = 0$ makes (even point-wise) sense. However, defining (with Lebesque integration)

$$\begin{cases} \|v\|_{0,\Omega} = \|v\|_{L_2(\Omega)} := \left(\int\limits_\Omega |v(x)|^2 \, \mathrm{d}x \right)^{1/2} \\ \|v\|_{1,\Omega}^2 = \|v\|_{H^1(\Omega)} := \|v\|_{0,\Omega}^2 + \|\nabla v\|_{0,\Omega}^2 , \end{cases} \tag{1.3.12}$$

it is clear that $\|\cdot\|_{0,\Omega}$ and $\|\cdot\|_{1,\Omega}$ are both norms (defined in particular, on $C_0^\infty(\Omega)$). Now, let v^* be any limit of a sequence $v_j \in C_0^\infty(\Omega)$ with respect to $\|\cdot\|_{1,\Omega}$, i.e.

$$\|v^* - v_j\|_{1,\Omega} \to 0, \; j \to \infty, \tag{1.3.13}$$

then we still have

$$\int\limits_\Omega (\nabla u(x))^T \nabla v^*(x) \, \mathrm{d}x = \int\limits_\Omega f(x) \, v^*(x) \, \mathrm{d}x. \tag{1.3.14}$$

Exercise 1.3.1 *Prove* (1.3.14) *formally.*

Hence, once u satisfies (1.3.10), it still holds under tests by all *limits* of elements from $C_0^\infty(\Omega)$ in the norm $\|\cdot\|_{1,\Omega}$.

1.3.2 Weak Formulation of the Poisson Equation

This suggests defining

$$H_0^1(\Omega) := H_0^1(\Omega; \mathbb{R}) := \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{1,\Omega}} \tag{1.3.15}$$

as the *closure* of $C_0^\infty(\Omega)$ under $\|\cdot\|_{1,\Omega}$. Here $C_0^\infty(\Omega)$ denotes the space of infinitel often differentiable functions with compact support in Ω. Note: such functions must vanish on $\partial\Omega$ since the support is closed by definition.

$H_0^1(\Omega)$ is an example of a *Sobolev space.* The subscript "0" indicates that (in a sense to be made precise later) its elements vanish on $\partial\Omega$. Therefore it makes sense to look for the solution u also in $H_0^1(\Omega)$. Obviously, $H_0^1(\Omega)$ is a closed subspace of the correspondingunconstrained *Sobolev space*

$$H^1(\Omega) := H^1(\Omega;\mathbb{R}) := \overline{C^\infty(\Omega)}^{\|\cdot\|_{1,\Omega}}. \tag{1.3.16}$$

As will be discussed later in more detail $H_0^1(\Omega)$, $H^1(\Omega)$ are examples of *Hilbert spaces* which, in particular, are *complete normed linear spaces.* By the above limiting argument the validity of (1.3.6), (1.3.7), (1.3.8) extends to elements in $H^1(\Omega)$ which will be used frequently.

The above considerations lead to an abstract formulation of (1.3.10) that will guide subsequent developments. To describe this we need a few elementary functional analytic notions. Recall that a linear space $\mathbb{X}$ (or more precisely the pair $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$) endowed with a norm $\|\cdot\|_{\mathbb{X}}$ is called a *Banach space* if it is complete, i.e., Cauchy sequences in $\mathbb{X}$ have a limit in $\mathbb{X}$.

A Banach space is a *Hilbert space* if the norm is induced by an inner product $\|v\|_{\mathbb{X}} = (v,v)_{\mathbb{X}}^{1/2}$. Moreover, given a Banach space $\mathbb{X}$, we denote by $\mathbb{X}'$ its *normed dual*, which is the space of all bounded linear functionals $w : \mathbb{X} \to \mathbb{R}$, endowed with the norm

$$\|w\|_{\mathbb{X}'} := \sup_{v\in\mathbb{X}\setminus\{0\}} \frac{w(v)}{\|v\|_{\mathbb{X}}}. \tag{1.3.17}$$

We sometimes write $\langle w, v\rangle = w(v)$. Now note that

$$f(v) := (f,v)_{0,\Omega} := \int\limits_{\Omega} f(x)\, v(x)\, \mathrm{d}x \tag{1.3.18}$$

is a *linear functional* on

$$L_2(\Omega) := \left\{ w : \Omega \to \mathbb{R},\ \text{measurable}:\ \|w\|_{0,\Omega} < \infty \right\}.$$

Just as the Dirac functional is bounded on $C(\Omega)$ the functional represented by f is bounded on $L_2(\Omega)$, (i.e. by Cauchy-Schwarz)

$$|f(v)| = \left|(f,v)_{0,\Omega}\right| \le \|f\|_{0,\Omega}\, \|v\|_{0,\Omega}$$

as long as f itself, as a function, belongs to $L_2(\Omega)$. Therefore, f can also be identified with a bounded linear functional on $H_0^1(\Omega) \subset L_2(\Omega)$, or in the above terms

$$f \in (H_0^1(\Omega))' =: H^{-1}(\Omega). \tag{1.3.19}$$

It will later be seen that the dual $H^{-1}(\Omega)$ of $H_0^1(\Omega)$ is actually strictly larger than $L_2(\Omega)$, i.e., contains elements that cannot be identified with a function in $L_2(\Omega)$.

Next note that

$$a\left(u,v\right) := \int\limits_{\Omega} \left(\nabla u\left(x\right)\right)^T \nabla v\left(x\right) \mathrm{d}x \tag{1.3.20}$$

is a *symmetric bilinear form* on $H_0^1(\Omega)$. In fact, by (1.3.11), we have

$$|a(v,w)| \leq \|\nabla v\|_{0,\Omega}\|\nabla w\|_{0,\Omega} \leq \|v\|_{1,\Omega}\|w\|_{1,\Omega}, \quad \forall\ v,w \in H_0^1(\Omega). \tag{1.3.21}$$

We say that the bilinear form $a(\cdot,\cdot)$ is *continuous* on $H_0^1(\Omega) \times H_0^1(\Omega)$. In fact, $a(\cdot,\cdot)$ is also continuous on the larger space $H^1(\Omega)$.

Weak Formulation: Therefore, (1.3.10) can be restated as:

Given $f \in (H_0^1(\Omega))'$ find $u \in H_0^1(\Omega)$, such that

$$a\left(u,v\right) = f\left(v\right),\ \forall v \in H_0^1\left(\Omega\right). \tag{1.3.22}$$

This is a first, and perhaps simplest version of a weak formulation of a PDE. It is a special instance of the general

Variational Problem: Let $\mathbb{U}$ be a Hilbert space and $a(\cdot,\cdot)$ be a continuous bilinear form on $\mathbb{U} \times \mathbb{U}$. Given $f \in \mathbb{U}'$ find $u \in \mathbb{U}$ such that

$$a(u,v) = f(v), \quad \forall\ v \in \mathbb{U}. \tag{1.3.23}$$

Remark 1.3.1 We shall refer to $\mathbb{U}$ as the (infinite-dimensional) *trial space* in this weak formulation. Note that we use here $\mathbb{U}$ also as the (infinite-dimensional) *test space.* In that sense the formulation is *symmetric*, i.e., trial and test space are the same. Such formulations are also called *Galerkin formulation.* □

Remark 1.3.2 The above scenario can be extended as follows: let $\mathbb{U}, \mathbb{V}$ be Hilbert spaces and suppose that $B(\cdot, \cdot) : \mathbb{U} \times \mathbb{V} \to \mathbb{R}$ is bilinear. For each $u \in \mathbb{U}$ define the functional

$$(\mathcal{B}_u)(v) := B(u, v), \quad v \in \mathbb{V}. \tag{1.3.24}$$

i) Bilinearity of $B(\cdot, \cdot)$ means that $\mathcal{B}_u$ is (for each fixed $u \in \mathbb{U}$) a linear functional.

ii) $\mathcal{B}_u$ is bounded, i.e., $\mathcal{B}_u \in \mathbb{V}'$, if and only if there exists a constant $C(u)$ (depending on u) such that

$$|\mathcal{B}_u(v)| = |B(u, v)| \leq C(u)\|v\|_{\mathbb{V}}. \tag{1.3.25}$$

iii) Again, since $B(\cdot, \cdot)$ is bilinear, the assignement

$$u \mapsto \mathcal{B}_u$$

is a *linear mapping* that takes any $u \in \mathbb{U}$ to a functional $\mathcal{B}_u \in \mathbb{V}'$. Therefore, we can write

$$\mathcal{B}_u = \mathcal{B}u, \quad \mathcal{B} : \mathbb{U} \to \mathbb{V}'. \tag{1.3.26}$$

iv) This allows us to reinterpret the variational problem: for $f \in \mathbb{V}'$ find $u \in \mathbb{U}$ such that

$$B(u, v) = f(v), \quad v \in \mathbb{V}',$$

as an operator equation

$$\mathcal{B}u = f, \tag{1.3.27}$$

i.e., we have the situation discussed before where in (1.2.1) we now have (for $\mathcal{A} = \mathcal{B}$)

$$\mathbb{W} = \mathbb{V}'. \tag{1.3.28}$$

In such a setting the range $\mathbb{W}$ of an operator arises naturally as a *dual space* of the test space.

v) As an example, one can interpret (1.3.9) as follows: $-\Delta u$ acts as a linear functional on $H_0^1(\Omega)$ which, by (1.3.21), is bounded. Therefore, we can view $-\Delta$ as an *operator* that maps $H_0^1(\Omega)$ into $H^{-1}(\Omega) := (H_0^1(\Omega))'$, i.e., $\mathbb{U} = \mathbb{V}$, $\mathbb{W} = \mathbb{U}' = \mathbb{V}'$.

vi) In the case of Poisson's equation we have taken $\mathbb{U} = \mathbb{V}$, trial- and test-space are the same, which is the case of a *symmetric* variational formulation. We will see later that in some important cases one must take $\mathbb{V} \neq \mathbb{U}$ in order to arrive at a well-conditioned variational formulation or even at formulation with a continuous bilinear form. Corresponding weak formulations are also called *Petrov-Galerkin* formulations

vii) For the operator $\mathcal{B}$ to have a bounded condition it must be bounded. This means $C(u) \leq C_{\mathcal{B}}\|u\|_{\mathbb{U}}$ must hold for some constant $C_{\mathcal{B}} < \infty$. Therefore, for a reasonable variational formulation, we must have that the bilinear form is continuous

$$|B(u,v)| \leq C_{\mathcal{B}}\|u\|_{\mathbb{U}}\|v\|_{\mathbb{V}}, \quad u \in \mathbb{U},\, v \in \mathbb{V}. \tag{1.3.29}$$

Thus, in subsequent examples continuity of the bilinear form is the first property to be checked. By the above reasoning (1.3.29) is equivalent to saying

$$\mathcal{B} \in \mathcal{L}(\mathbb{U}, \mathbb{V}'), \quad \|\mathcal{B}\|_{\mathcal{L}(\mathbb{U},\mathbb{V}')} \leq C_{\mathcal{B}}. \tag{1.3.30}$$

viii) To show that $\mathcal{B}^{-1}$ exists and is also bounded is in general more difficult to check and will be discussed later in detail. □

2 Sobolev Spaces

The previous discussion indicates the relevance of "function spaces" of the type (1.3.15) and related objects. An in-depth treatment would go beyond the scope of this lecture and is typically part of a course on partial differential equations. One can consult for instance [Alt85, Ada75, AU10] for more details. To provide some underpinning for those who haven't taken such a course, some relevant facts and ideas are collected in this chapter, but some proofs will have to be skipped.

The notion of *Hilbert space* as a generalization of Euclidean spaces plays a pivotal role.

2.1 Hilbert spaces

Recall that a vector space $\mathbb{H}$ over $\mathbb{R}$, equipped with a scalar product $(\cdot,\cdot)_{\mathbb{H}}$ is called a Hilbert space, if $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ with $\|v\|_{\mathbb{H}} := (v,v)_H^{1/2}$ is a *complete* normed linear space. As metioned earlier, "complete" means that every Cauchy sequence in $\mathbb{H}$ has a limit in $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$. As a simplest example $(\mathbb{R}^d, |\cdot|)$ is a Hilbert space over $\mathbb{R}$, where $|x| = (\sum_{i=1}^d x_i^2)^{1/2}$. It is sometimes necessary to consider Hilbert spaces over the complex field $\mathbb{C}$. In this case the inner product is a sesquilinear form with the property

$$(\cdot,\cdot)_{\mathbb{H}} : \mathbb{H} \times \mathbb{H} \to \mathbb{C}, \quad (v,w)_{\mathbb{H}} = \overline{(w,v)_{\mathbb{H}}}, \quad v,w \in \mathbb{H}, \tag{2.1.1}$$

where $\overline{a}$ is the complex conjugate of $a \in \mathbb{C}$. For instance, for *any countable* index set $\mathcal{I}$ let $\mathbb{C}^{\mathcal{I}}$ denote the set of all sequences $(c_\lambda)_{\lambda\in\mathcal{I}}$ and define for $|a|^2 := a\overline{a}$, $a \in \mathbb{C}$,

$$(\mathbf{v},\mathbf{w})_{\ell_2(\mathcal{I})} := \sum_{\lambda\in\mathcal{I}} v_\lambda \overline{w}_\lambda, \quad \mathbf{v},\, \mathbf{w} \in \mathbb{C}^{\mathcal{I}}, \quad \|\mathbf{v}\|_{\ell_2(\mathcal{I})} := \Big(\sum_{\lambda\in\mathcal{I}} |v_\lambda|^2\Big)^{1/2}, \tag{2.1.2}$$

the space

$$\ell_2(\mathcal{I}) = \left\{ \mathbf{v} \in \mathbb{C}^{\mathcal{I}} : \ \|\mathbf{v}\|_{\ell_2(\mathcal{I})} < \infty \right\}.$$

Then $(\ell_2(\mathcal{I}), \|\cdot\|_{\ell_2(\mathcal{I})})$ is a Hilbert space over $\mathbb{C}$.

Note that $\mathcal{I}$ is allowed to have *infinite* cardinality $\#(\mathcal{I}) = \infty$. When $\#(\mathcal{I}) = d$, $\ell_2(\mathcal{I})$ can, of course, be identified with $\mathbb{C}^d$ endowed with the *Euclidean norm* $|\cdot| = \|\cdot\|_{\ell_2(\mathcal{I})}$. This particular *sequence space* (admitting infinite sets $\mathcal{I}$) is important for the following reason. One can show that every *separable* Hilbert space $\mathbb{H}$, which means that $\mathbb{H}$ contains a countable dense subset, possesses an *orthonormal basis* $\Psi = \{\psi_\lambda : \lambda \in \mathcal{I}\}$ for some countable index set $\mathcal{I}$, such that by orthonormality

$$\|v\|_{\mathbb{H}} = \Big(\sum_{\lambda \in \mathcal{I}} \big|(v, \psi_\lambda)_{\mathbb{H}}\big|^2 \Big)^{1/2} = \|\mathbf{v}\|_{\ell_2(\mathcal{I})}, \quad \mathbf{v} = \big((v, \psi_\lambda)_{\mathbb{H}}\big)_{\lambda \in \mathcal{I}}.$$

i.e., the sequence norm induces the Hilbert space topology. In this sense, Hilbert spaces are a natural generalization of Euclidean spaces.

This is, of course, important in Harmonic Analysis but also in modern wavelet-based solvers for PDEs.

We will primarily be concerned with *function spaces* that are are Hilbert spaces.

2.2 Lebesgue spaces

Defining

$$\|f\|_{L_\infty(\Omega)} := \begin{cases} \Big(\int\limits_{\Omega} |f(x)|^p \, \mathrm{d}x \Big)^{1/p}, & 1 \leq p < \infty, \\ \operatorname*{ess\,sup}_{x \in \Omega} |f(x)|, & p = \infty, \end{cases}$$

where integration is always understood in the Lebesgue sense, the *Lebesgue-spaces*

$$L_p(\Omega) = \Big\{ f : \Omega \to \mathbb{R} \, (\text{or } \mathbb{C}) \quad \text{measurable} : \ \|f\|_{L_p(\Omega)} < \infty \Big\},$$

are for $1 \leq p \leq \infty$ complete normed linear spaces - *Banach spaces.* Lebesgue integration, as opposed to Riemann integration, is needed to guarantee *completeness.*

Strictly speaking, the elements of $L_p(\Omega)$ are *equivalence classes* containing all functions that differ only on sets of measure zero. Nevertheless, we call $f \in L_p(\Omega)$ a "function" instead of an equivalence class where f is just a representer.

Here we are mainly interested in $p = 2$. In this case, the norm $\|\cdot\|_{L_2(\Omega)}$ is already defined by (1.3.12) and is often denoted as

$$\|\cdot\|_{0,\Omega} = \|\cdot\|_{L_2(\Omega)}$$

for conventional reasons. The point that distinguishes $L_2(\Omega)$ from $p \neq 2$ is that

$$\|v\|_{0,\Omega}^2 = (v, v)_{0,\Omega} \tag{2.2.1}$$

is induced by an inner product, namely

$$(v, w)_{0,\Omega} := \int\limits_{\Omega} v(x)\,\overline{w(x)}\mathrm{d}x, \tag{2.2.2}$$

which is a scalar product on $L_2(\Omega)$. Thus, since $L_2(\Omega)$ is complete

$$\left(L_2(\Omega), \|\cdot\|_{0,\Omega}\right) \text{ is a Hilbert space.}$$

2.3 Weak derivatives and Sobolev spaces

Classical (strong) derivatives are defined in a point-wise sense, weak derivatives in an *average* sense. To motivate the definition suppose that $v \in C^k(\Omega)$ and $\alpha \in \mathbb{Z}_+^d$, $|\alpha| = \alpha_1 + \ldots + \alpha_d = k$. Then, using repeated integration by parts, one has for any $\phi \in C_0^\infty(\Omega)$

$$\int\limits_{\Omega} \left(\frac{\partial^\alpha}{\partial x^\alpha} v(x)\right) \phi(x)\,\mathrm{d}x = (-1)^{|\alpha|} \int\limits_{\Omega} v(x) \frac{\partial^\alpha}{\partial x^\alpha} \phi(x)\,\mathrm{d}x.$$

The right hand side still makes sense when v does not belong to $C^k(\Omega)$. This leads to the following:

Definition 2.3.1 Let $v \in L_p(\Omega)$, $1 \leq p < \infty$. $D^\alpha v \in L_p(\Omega)$ is called α-th weak derivative of v in $L_p(\Omega)$ if

$$\int\limits_\Omega D^\alpha v\,(x)\,\phi\,(x)\,\mathrm{d}x = (-1)^{|\alpha|} \int\limits_\Omega v\,(x)\,\frac{\partial^\alpha}{\partial x^\alpha}\phi\,(x)\,\mathrm{d}x. \tag{2.3.1}$$

□

Exercise 2.3.1 $v(x) := |x| \in L_p((-1,1))$ *has a weak derivative*

$$D^1 v\,(x) = \begin{cases} -1; & x \in (-1,0)\,, \\ 1; & x \in (0,1)\,. \end{cases}$$

which belongs to $L_p\,((-1,1))$ *for any* $1 \leq p < \infty$.

Show that v *has no second derivative in* $L_p\,(\Omega)$.

The k-th order Sobolev space in $L_p(\Omega)$ is often denoted by

$$W^{k,p}\,(\Omega) := \{v \in L_p\,(\Omega) : D^\alpha v \in L_p\,(\Omega)\,,\, |\alpha| = k\}\,.$$

It is a Banach space when endowed with the norm

$$\|v\|_{W^{k,p}(\Omega)} := \left(\sum_{j=0}^{k} \sum_{|\alpha|=j} \|D^\alpha v\|^p_{L_p(\Omega)} \right)^{1/p}. \tag{2.3.2}$$

In the special case $p = 2$, one often uses the notation

$$W^{k,2}\,(\Omega) := H^k\,(\Omega)\,,\ H^0\,(\Omega) = L_2\,(\Omega)\,.$$

$H^k(\Omega)$ is also a Hilbert space with scalar product

$$(v,w)_{k,\Omega} := \sum_{j=0}^{k} \sum_{|\alpha|=k} (D^\alpha v, D^\alpha w)_{0,\Omega}\,. \tag{2.3.3}$$

The expression

$$|v|^2_{k,\Omega} := \sum_{|\alpha|=k} \|D^\alpha v\|^2_{0,\Omega} \tag{2.3.4}$$

is only a *semi-norm* when $k > 0$.

Thus, we have

$$\|v\|_{k,\Omega} := \|v\|_{H^k(\Omega)} = \left(\sum_{j=0}^{k} |v|_{j,\Omega}^2 \right)^{1/2} . \qquad (2.3.5)$$

Remark 2.3.1 One can show that

$$\left(\|v\|_{0,\Omega}^2 + |v|_{k,\Omega}^2 \right)^{1/2}$$

is an equivalent norm on $H^k(\Omega)$, i.e. one can omit the intermediate seminorms $|v|_{j,\Omega}$, $0 < j < k$. □

Remark 2.3.2 Under the above assumptions on Ω one can show that

$$H^k\left(\Omega\right) = \overline{C^\infty\left(\Omega\right)}^{\|\cdot\|_{k,\Omega}}, \qquad (2.3.6)$$

i.e., smooth functions are dense in $H^k(\Omega)$. This was the definition used in Section 1.3.2, see (1.3.16). □

It is not clear yet how to deal with *boundary conditions* in the context of weak derivatives. For instance, it makes no sense to say "$v \in H_0^1(\Omega)$ when v vanishes on $\partial\Omega$ and possesses first order weak derivatives". We will see later that the existence of weak derivatives will allow us to say something about values on the boundary of a domain. So for the moment, we are content with taking up (1.3.15) and define in agreement with (2.3.6)

$$H_0^k\left(\Omega\right) = \overline{C_0^\infty\left(\Omega\right)}^{\|\cdot\|_{k,\Omega}} \subseteq H^k\left(\Omega\right), \qquad (2.3.7)$$

as a closed subspace of $H^k(\Omega)$, it is again a Hilbert space, see (3.1.6) in the context of the *biharmonic equaltion.*

We will have to compare different spaces with each other regarding the "strength" of the respective norms. A space $(\mathbb{Y}, \|\cdot\|_{\mathbb{Y}})$ is said to be *continuously embedded* in $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$ if

$$\mathbb{Y} \subseteq \mathbb{X} \quad \text{and} \quad \|v\|_{\mathbb{X}} \leq C \, \|v\|_{\mathbb{Y}} \quad \forall \, v \in \mathbb{Y}$$

holds for some constant $C < \infty$, i.e., the "smaller" space has a "stronger" norm.

One has the following continuous embeddings

$$\begin{array}{ccccccccccc} L_2\left(\Omega\right) & = & H^0\left(\Omega\right) & \supsetneqq & H^1\left(\Omega\right) & \supsetneqq & \ldots & \supsetneqq & H^k\left(\Omega\right) & \supsetneqq & \ldots \\ & & \Vert & & \rotatebox{-90}{$\supsetneqq$} & & & & \rotatebox{-90}{$\supsetneqq$} & & \\ & & H_0^0\left(\Omega\right) & \supsetneqq & H_0^1\left(\Omega\right) & \supsetneqq & \ldots & \supsetneqq & H_0^k\left(\Omega\right) & \supsetneqq & \ldots \end{array} \tag{2.3.8}$$

Non-integer orders:

It is possible (in fact necessary) to define Sobolev spaces of *non-integer* order $H^s(\Omega)$, $s \geq 0$, which in some sense "interpolate" the $H^k(\Omega)$ $k \in \mathbb{N}$. (Corresponding spaces of "negative order" $s < 0$ are then defined as the duals to the ones with positive order.) The s between two integers offer a refined measure of smoothness for the L_2-norm like the Hölder spaces for the L_∞-norm

$$C^s\left(\Omega\right) = \left\{v \in C\left(\Omega\right) : \ \left|v\left(x\right) - v\left(y\right)\right| \leq C\left|x-y\right|^s, \ x, y \in \Omega\right\}, 0 < s < 1.$$

Such a non-integer smoothness order is easy to define when $\Omega = \mathbb{R}^d$, using the *Fourier transform*

$$\hat{f}\left(\xi\right) = \frac{1}{\left(2\pi\right)^{d/2}} \int\limits_{\mathbb{R}^d} f\left(x\right) e^{-i\xi^T x} \mathrm{d}x.$$

In fact, recalling that

$$\left(\frac{\partial^\alpha}{\partial x^\alpha} f\right)\hat{}(\xi) = \left(i\xi\right)^\alpha \hat{f}\left(\xi\right), \quad \alpha \in \mathbb{Z}_+^d,$$

one can show that

$$\left\|v\right\|_{k,\mathbb{R}^d}^2 \sim \int\limits_{\mathbb{R}^d} \left(1 + \left|\xi\right|^k\right)^2 \left|\hat{v}\left(\xi\right)\right|^2 \mathrm{d}\xi.$$

It is then natural to define

$$\left\|v\right\|_{s,\mathbb{R}^d}^2 := \int\limits_{\mathbb{R}^d} \left(1 + \left|\xi\right|^s\right)^2 \left|\hat{v}\left(\xi\right)\right|^2 \mathrm{d}\xi, \tag{2.3.9}$$

which describes "smoothness" or "regularity" by a certain *decay* of the Fourier transform, because high-frequency components ($\hat{f}(\xi)$, $|\xi|$ large) are required to decay rapidly when s is large for v to be in

$$H^s(\mathbb{R}^d) := \{v \in L_2(\mathbb{R}^d) : \left\|v\right\|_{s,\mathbb{R}^d} < \infty\}. \tag{2.3.10}$$

Unfortunately, on bounded domains this is less straight-forward. There are several possible strategies which can be shown to be all equivalent under mild conditions on Ω.

The first is to use *Whitney's Extension Theorem*, that says that $v \in H^k(\Omega)$ can be extended to $H^k(\mathbb{R}^d)$ to some $\tilde{v} \in H^k(\mathbb{R}^d)$ such that $\|\tilde{v}\|_{k,\mathbb{R}^d} \leq C\|v\|_{k,\Omega}$ for some C depending only on Ω and k. Then one can define for $s \leq k$

$$|||v|||_{s,\Omega} = |||\tilde{v}|||^2_{s,\mathbb{R}^d}$$

where the right hand expression is given by (2.3.9). This is an important theoretical fact but of little help in practical computations.

The second possible strategy is to use the concept of "interpolation of Banach spaces" which is a systematic way of filling "gaps" in a scale of Banach spaces, see [BL76].

A third possibility is based on an explicit expression for the norm: Let $k < s < k+1$:

$$\|v\|^2_{s,\Omega} := \|v\|^2_{k,\Omega} + \sum_{|\alpha|=k} \int_\Omega \int_\Omega \frac{|D^\alpha v(x) - D^\alpha v(y)|^2}{|x-y|^{d+2(s-k)}} \mathrm{d}x\mathrm{d}y \qquad (2.3.11)$$

which is reminiscent of the Hölder norms.

Such non-integer regularity measures are not introduced for curiosity. They are needed to properly deal with the restrictions of Sobolev functions to lower dimensional manifolds like domain boundaries.

Obviously, $\|\cdot\|_{s+\varepsilon,\Omega}$ is a stronger norm than $\|\cdot\|_{s,\Omega}$ for any $\varepsilon > 0$. Hence $H^{s+\varepsilon}(\Omega)$ is continuously embedded in $H^s(\Omega)$ for any $\varepsilon > 0$. This can be made more precise. To this end, we need the notion of *compact embedding*.

A Banach space $(\mathbb{Y}, \|\cdot\|_\mathbb{Y})$ is compactly embedded in another Banach space $(\mathbb{X}, \|\cdot\|_\mathbb{X})$, writing $\mathbb{Y} \hookrightarrow \mathbb{X}$, if $\mathbb{Y} \subset \mathbb{X}$ and the unit ball

$$U\mathbb{Y} := \{y \in \mathbb{Y} : \ \|y\|_\mathbb{Y} \leq 1\}$$

in $\mathbb{Y}$ is a *compact subset* of $\mathbb{X}$, i.e., the canonical injection is a compact mapping. Compact embeddings play a crucial role when establishing later *convergence rates* for numerical schemes.

In the finite dimensional case a continuous embedding is also compact (by Heine-Borel). In the infinite dimensional case this is not true, i.e., a closed bounded set is not necessarily compact. Then compactness typically results from *higher regularity.* The following (special case of) Rellich's Theorem is a typical example. (There are more general variants for $W^{k,p}(\Omega)$.)

Theorem 2.3.1 (Rellich) *Whenever $s < t$ one has the compact embedding*

$$H^t(\Omega) \hookrightarrow H^s(\Omega).$$

□

2.4 Classical versus Weak Regularity

Weak differentiability is - as the term indicates - a weaker regularity notion than classical point-wise regularity. The interesting point is that the discrepancy between these notions depends on the spatial dimension d.

Remark 2.4.1

$$H^1((a,b)) \subset C((a,b)).$$

□

Proof Exercise. Hint: use that $C^\infty((a,b))$ is a dense subset of $H^1((a,b))$ and employ the Arzela-Ascoli Theorem stating that a bounded equicontinuous sequence has a convergent subsequence. ■

However, already for $d = 2$ an element of $H^1(\Omega)$, $\Omega \subset \mathbb{R}^2$, need *not* be continuous.

Example 2.4.1 Let $\Omega = \{x \in \mathbb{R}^2 : \ |x| < 1\}$ be the open unit ball. Then

$$u(x) := \log\left(\log\frac{2}{|x|}\right) \in H^1(\Omega)$$

but u is not bounded in Ω. □

Proof Exercise. Hint: $u \in H^1(\Omega)$ follows from

$$\int\limits_0^{1/2} \frac{1}{r\log^2 r}\mathrm{d}r < \infty.$$

■

Example 2.4.2 For $d \geq 3$ one has

$$u(x) := |x|^{-t}, \ 0 < t < \frac{d-2}{2}$$

belongs to $H^1(\Omega)$. Obviously, the singularity at $x = 0$ becomes the stronger the larger d. □

The general situation is described by the *Sobolev Embedding Theorem*, a very special case of which reads as follows.

Theorem 2.4.1 *Whenever* $s > \frac{d}{2}$ *one has* $H^s(\Omega) \subset C(\Omega)$, $(\Omega \subset \mathbb{R}^d)$ *and the embedding is compact. More generally*

$$W^{s,q}(\Omega) \hookrightarrow W^{s',p}(\Omega) \quad \textit{if} \quad d\Big(\frac{1}{q} - \frac{1}{p}\Big) < s - s'. \tag{2.4.1}$$

Hence, the larger d the larger is the discrepancy between pointwise (classical) and weak differentiability.

2.5 Some Important Inequalities

The next result says that the seminorm $|\cdot|_{1,\Omega}$ (see (2.3.5)) is even a norm on the subspace $H_0^1(\Omega)$ of $H^1(\Omega)$.

Proposition 2.5.1 (Poincaré-Friedrichs-Inequality) *There exists a constant c_Ω depending only on Ω, such that*

$$\|v\|_{0,\Omega} \leq c_\Omega \|\nabla v\|_{0,\Omega} = c_\Omega |v|_{1,\Omega}, \quad v \in H_0^1(\Omega). \tag{2.5.1}$$

□

Remark 2.5.1 (2.5.1) does not hold for $v \in H^1(\Omega)$, because the constant functions belong to $H^1(\Omega)$, and $\|c\|_{0,\Omega} > 0$ while $|c|_{1,\Omega} = 0$. This does not contradict (2.5.1) though since the only constant in $H_0^1(\Omega)$ is zero. □

Proof (of Proposition 2.5.1) Since $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$ it suffices to confirm (2.5.1) for $v \in C_0^\infty(\Omega)$ and then pass to the limit. Moreover, without loss of generality we can assume that $\Omega \subseteq (0,s)^d$. Moreover, any $v \in C_0^\infty(\Omega)$ can be extended to $C_0^\infty((0,s)^d)$ by setting $v(x) = 0,\ x \in (0,s)^d \backslash \Omega$. Then

$$v(x_1, x_2, \ldots, x_d) = \underbrace{v(0, x_2, \ldots, x_d)}_{=0} + \int_0^{x_1} \partial_{x_1} v(y, x_2, \ldots, x_d)\,\mathrm{d}y.$$

Thus

$$|v(x)|^2 \leq \left(\int_0^s 1\mathrm{d}y\right)\left(\int_0^s |\partial_{x_1} v(y, x_2, \ldots, x_d)|^2 \,\mathrm{d}y\right)$$

and

$$\int_0^s |v(x)|^2 \,\mathrm{d}x_1 \leq s^2 \left(\int_0^s |\partial_{x_1} v(x_1, x_2, \ldots, x_d)|^2 \,\mathrm{d}x_1\right)$$

which gives

$$\begin{aligned}\int_{(0,s)^d} |v(x)|^2 \,\mathrm{d}x &\leq s^2 \int_{(0,s)^d} |\partial_{x_1} v(x)|^2 \,\mathrm{d}x \\ &\leq s^2 \|\nabla v\|_{0,\Omega}^2 .\end{aligned} \tag{2.5.2}$$

■

Corollary 2.5.1 *For Ω as above, $\|\cdot\|_{m,\Omega}$ and $|\cdot|_{m,\Omega}$ are equivalent norms on $H_0^m(\Omega)$, $m \in \mathbb{N}$. In particular, when $\Omega \subseteq (0,s)^d$ one has*

$$|v|_{m,\Omega} \leq \|v\|_{m,\Omega} \leq \left(1+s^2\right)^{m/2} |v|_{m,\Omega} \,. \tag{2.5.3}$$

□

Proof The first inequality is trivial. We verify the second one by induction. By (2.5.1) we have

$$\|v\|^2_{0,(0,s)^d} + |v|^2_{1,(0,s)^d} \leq \left(1+s^2\right) |v|^2_{1,(0,s)^d} \,,$$

which is (2.5.3) for $m=1$. Suppose it holds for some $m \geq 1$. Then by (2.5.1), one obtains for any $\alpha \in \mathbb{Z}^d_+$, $|\alpha| = m$,

$$\|\partial^\alpha v\|^2_{0,\Omega} \leq s^2 \left\|\partial_{x_j} \partial^\alpha v\right\|^2_{0,\Omega} , \quad j = 1, \ldots, d. \tag{2.5.4}$$

Therefore

$$\begin{aligned} \|v\|^2_{m+1,\Omega} &= \|v\|^2_{m,\Omega} + |v|^2_{m+1,\Omega} \\ &\leq \left(1+s^2\right)^m |v|^2_{m,\Omega} + |v|^2_{m+1,\Omega} \\ &\overset{(2.5.4)}{\leq} \left(1+s^2\right)^m s^2 |v|^2_{m+1,\Omega} + |v|^2_{m+1,\Omega} \\ &= \left(1+s^2\right)^{m+1} |v|^2_{m+1,\Omega} \,. \end{aligned}$$

■

The fact that a semi-norm becomes a norm on a subspace of $H^1(\Omega)$ can be extended in several ways. One such extension concerns homogeneous boundary conditions on *part* of $\partial\Omega$. Let

$$\Gamma_D \subset \partial\Omega, \quad \mathrm{vol}_{d-1}\left(\Gamma_D\right) > 0,$$

and consider

$$H^1_{0,\Gamma_D}\left(\Omega\right) := \overline{\left\{\phi \in C^\infty\left(\Omega\right) : \operatorname{supp}\phi \cap \Gamma_D = \emptyset\right\}}^{\|\cdot\|_{1,\Omega}}. \tag{2.5.5}$$

The following extension of Proposition 2.5.1 will later be needed.

Proposition 2.5.2 *Under the above assumptions on Γ_D one still has*

$$\|v\|_{0,\Omega} \leq c_\Omega \, |v|_{1,\Omega} \, , \;\; v \in H^1_{0,\Gamma_D}(\Omega) \tag{2.5.6}$$

for some constant c_Ω. □

Proof We sketch a proof that uses "soft" functional analytic arguments. Suppose that (2.5.6) does not hold for any constant c_Ω. Hence, there must exist a sequence $(v_j)_{j\in\mathbb{N}} \subset H^1_{0,\Gamma_D}(\Omega)$ such that

$$1 = \|v_n\|_{0,\Omega} \geq n|v_n|_{1,\Omega}, \quad n \in \mathbb{N}, \tag{2.5.7}$$

which implies that

$$|v_n|_{1,\Omega} \leq 1/n, \quad n \in \mathbb{N}.$$

Hence

$$\|v_n\|^2_{1,\Omega} \leq 1 + n^{-2} \leq 2, \quad n \in \mathbb{N},$$

i.e., the v_n are uniformly bounded in $H^1(\Omega)$ and hence, by Rellich's Theorem 2.3.1, belong to a compact set in $L_2(\Omega)$. Therefore, there exists a convergent subsequence (again denoted by $(v_j)_{j\in\mathbb{N}}$) with limit v^*, i.e.,

$$\|v_n - v^*\|_{0,\Omega} \to 0, \quad n \to \infty.$$

Now we use an additional result from Functional Analysis, namely that bounded sets are *weakly compact*. Hence, a subsequence of $(v_j)_{j\in\mathbb{N}}$ (again denoted by $(v_j)_{j\in\mathbb{N}}$) must have a weak limit in $H^1_{0,\Gamma_D}(\Omega)$ which must agree with the strong limit v^* in $L_2(\Omega)$. This means, in particular, that for each (fixed) $\phi \in \left(C_0^\infty(\Omega)\right)^d$ one has

$$\begin{aligned}
\underbrace{|v_n|_{1,\Omega}\|\phi\|_{0,\Omega}}_{\to 0,\, n\to\infty} &\geq \left|(\nabla v_n, \phi)_{0,\Omega}\right| = \left|\int_\Omega v_n \mathrm{div}\, \phi dx\right| \\
&= \left|\int_\Omega (v_n - v^*)\mathrm{div}\, \phi dx + \int_\Omega v^* \mathrm{div}\, \phi dx\right| \\
&\geq \left|\int_\Omega v^* \mathrm{div}\, \phi dx\right| - \underbrace{\|v_n - v^*\|_{0,\Omega}\|\mathrm{div}\, \phi\|_{0,\Omega}}_{\to 0,\, n\to\infty} .
\end{aligned}$$

Hence, for each fixed $\phi \in \left(C_0^\infty(\Omega)\right)^d$ one has

$$\int_\Omega v^* \mathrm{div}\, \phi dx = 0.$$

which, in turn, means that the weak gradient of v^* vanishes. We now invoke another theorem that says that if the weak gradient of a function vanishes, this function must be a constant. Since the only constant in $H^1_{0,\Gamma_D}(\Omega)$ is the zero function we arrive at a contradiction to (2.5.7), proving the assertion. ■

The next result is often attributed to *Deny-Lions* and follows from a more general Theorem by *Whitney*. It will later be crucial for establishing error bounds for Finite Element schemes. Moreover, a special case yields another Poincaré-type inequality.

Theorem 2.5.1 *Assume that $\Omega \subset \mathbb{R}^d$ is a bounded Lipshitz-domain. For any $1 \leq p \leq \infty$, $l \leq k \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$ there exists a constant $C = C(\Omega, d, l, k, d)$ depending only on the listed parameters such that (with $|\cdot|_{W^0(L_p(\Omega))} = \|\cdot\|_{L_p(\Omega)}$)*

$$\inf_{P \in \mathbb{P}_k} \|v - P\|_{W^{l,p}(\Omega)} \leq C|v|_{W^{k,p}(\Omega)}, \quad \forall\, v \in W^{k,p}(\Omega)), \tag{2.5.8}$$

where $\mathbb{P}_k$ denotes the space of all polynomials of total order k *which means* total degree *(sum of exponents) $k-1$.* □

Proof Since the kth order derivatives of a polynomial of degree $k-1$ vanish there is nothing to prove for $l = k$. So assume that $l < k$. Again we sketch a proof by contraposition. Suppose (2.5.8) were not true. Then there exists a sequence $(v_j)_{j\in\mathbb{N}} \subset W^{k,p}(\Omega)$ such that (upon possibly modifying each v_j by subtracting a polynomial in $\mathbb{P}_k$)

$$\inf_{P \in \mathbb{P}_k} \|v_j - P\|_{W^{l,p}(\Omega)} = \|v_j\|_{W^{l,p}(\Omega)} = 1 \geq \frac{1}{j}|v_j|_{W^{k,p}(\Omega)}, \quad j \in \mathbb{N}. \tag{2.5.9}$$

Hence the v_j belong to a bounded set in $W^{k,p}(\Omega)$ which is compactly embedded in $W^{l,p}(\Omega)$ (Theorem 2.4.1). Hence, there exists a convergent subsequence (again denoted by $(v_j)_{j\in\mathbb{N}}$) in $W^l(L_p(\Omega))$ with limit $v^* \in W^{l,p}(\Omega)$, i.e.,

$$\|v^* - v_j\|_{W^{l,p}(\Omega)} \to 0,\, j \to 0. \tag{2.5.10}$$

We wish to show that

$$\int_\Omega v^* D^\alpha \phi dx = 0, \quad \forall\, \phi \in C_0^\infty(\Omega),\, |\alpha| = k. \tag{2.5.11}$$

To see this note that for each $\alpha \in \mathbb{N}_0^d$, $|\alpha| = k$, $\beta \succ \alpha, |\beta| = l$,

$$\begin{aligned}\underbrace{\|D^\alpha v_n\|_{L_p(\Omega)}\|\phi\|_{L_p(\Omega)}}_{\to 0,\, n\to\infty} &\geq \big|(D^\alpha v_n, \phi)_{0,\Omega} = \Big|(-1)^{k-l}\int_\Omega D^\beta(v_n - v^*)D^{\alpha-\beta}\phi dx \\ &\qquad +(-1)^{k-l}\int_\Omega D^\beta v^* D^{\alpha-\beta}\phi dx\Big| \\ &\geq \Big|\int_\Omega v^* D^\alpha \phi dx\Big| - \underbrace{|v^* - v_j|_{W^{l,p}(\Omega)}\|D^{\alpha-\beta}\phi\|_{L_p(\Omega)}}_{\to 0,\, n\to\infty,\text{ by (2.5.10)}}.\end{aligned}$$

Again we use that v^* is the weak limit in $W^{k,p}(\Omega)$ of a subsequence which shows (2.5.11). We now use the fact whenever all kth order weak derivatives of a function vanish then this function must be a polynomial of degree at most $k-1$ and hence belongs to $\mathbb{P}_k$. This is a contradiction to the left part of (2.5.9). ■

Remark 2.5.2 One application of Theorem 2.5.1 uses the fact that when Ω is an affine shrinkage of a domain Ω^* of unit size the constant $C(\Omega, d, l, k, d)$ in (2.5.8) scales like

$$C(\Omega, d, l, k, d) \sim (\operatorname{diam}\Omega)^{k-l} C(d, l, k, d). \tag{2.5.12}$$

This follows from (2.5.8) by a corresponding affine change of variables. □

An important consequence of Theorem 2.5.1 reads as follows.

Corollary 2.5.2 (Poincaré-Inequality) *There exists a constant $C_\Omega < \infty$ such that*

$$\Big\|v - |\Omega|^{-1}\int_\Omega v dx\Big\|_{0,\Omega} \leq C_\Omega |v|_{1,\Omega}, \quad \forall\, v \in H^1(\Omega). \tag{2.5.13}$$

In particular, defining

$$\tilde{H}(\Omega) := \{v \in H^1(\Omega) : \int_\Omega v dx = 0\}, \tag{2.5.14}$$

which is a closed subspace of $H^1(\Omega)$ and hence again a Hilbert space, one has

$$\|v\|_{0,\Omega} \leq C_\Omega |v|_{1,\Omega}, \quad v \in \tilde{H}^1(\Omega). \tag{2.5.15}$$

The constant C_Ω depends on Ω as stated in Remark 2.5.2. □

PROOF Exercise ■

All three inequalities (2.5.1), (2.5.6), and (2.5.15) state conditions under which the semi-norm $|\cdot|_{1,\Omega}$ is a norm on some subspace of $H^1(\Omega)$.

Remark 2.5.3 It is sometimes important to know how the constants in (2.5.1), (2.5.6), (2.5.8), (2.5.13), depend on the domain Ω. The proof of (2.5.1) shows that it scales like diam(Ω), in agreement with (2.5.12), if one has some additional knowledge about Ω such as being an element of a family of affine images of a fixed reference domain, where the affine mappings have uniformly boun ded condition numbers. This will later be used in analyzing finite element methods. □

We conclude this section with a slightly specialized version of the above Poincaré-Inequality and an elementary proof which also confirms the scaling relation (2.5.12).

Lemma 2.5.1 *Assume that $\Omega \subset (0,\delta)^d$ is convex. Then one has for every $v \in H^1(\Omega)$*

$$\|v\|_{0,\Omega} \leq C \left(\frac{1}{|\Omega|^{1/2}} \left| \int\limits_{\Omega} v\,(x)\, dx \right| + \delta\, |v|_{1,\Omega} \right),\ v \in H^1(\Omega)\,, \qquad (2.5.16)$$

where C is independent of δ. □

PROOF By Cauchy Schwarz

$$\left| \int\limits_{\Omega} v\,(x)\, \mathrm{d}x \right| \leq \int\limits_{\Omega} |v\,(x)|\, \mathrm{d}x \leq |\Omega|^{1/2}\, \|v\|_{0,\Omega}$$

$$\leq |\Omega|^{1/2}\, \|v\|_1 \,.$$

Hence, the average $A_\Omega v := |\Omega|^{-1} \int_\Omega v dx$ is well defined. Since $C^1(\overline{\Omega})$ is dense in $H^1(\Omega)$ it suffices to prove (2.5.16) for $v \in C^1(\overline{\Omega})$ and then pass to the limit. To this end, let $x_0 \in \overline{\Omega}$ with $|v(x_0)| = \min_{x\in\overline{\Omega}} |v(x)|$. Then

$$v\,(x) = v\,(x_0) + \int\limits_0^1 (x - x_0)^T\, \nabla v\,(x_0 + t\,(x - x_0))\, \mathrm{d}t$$

so that

$$
\begin{aligned}
|v(x)| &\leq |v(x_0)| + \int\limits_0^1 |x - x_0| \, |\nabla v(x_0 + t(x - x_0))| \, \mathrm{d}t \\
&\leq \frac{1}{|\Omega|} \left| \int\limits_\Omega v(x) \, \mathrm{d}x \right| + \operatorname{diam}(\Omega) \int\limits_0^1 |\nabla v(x_0 + t(x - x_0))| \, \mathrm{d}t.
\end{aligned}
$$

Hence (using $(a+b)^2 \leq 2a^2 + 2b^2$)

$$
|v(x)|^2 \leq 2 \left(\frac{\left| \int\limits_\Omega v(x) \, \mathrm{d}x \right|}{|\Omega|} \right)^2 + 2\operatorname{diam}(\Omega)^2 \int\limits_0^1 |\nabla v((1-t)x_0 + tx)|^2 \, \mathrm{d}x\mathrm{d}t.
$$

Integrating over Ω yields

$$
\int\limits_\Omega |v(x)|^2 \, \mathrm{d}x \leq \frac{2}{|\Omega|} \left| \int\limits_\Omega v(x) \, \mathrm{d}x \right|^2 + 2\operatorname{diam}(\Omega)^2 \int\limits_0^1 \int\limits_\Omega |\nabla v((1-t)x_0 + tx)|^2 \, \mathrm{d}x\mathrm{d}t.
$$

Since Ω is convex $z = (1-t)x_0 + tx \in \Omega$ whenever $x \in \Omega$. Thus $\int_0^1 \int_\Omega |\nabla v((1-t)x_0 + tx)|^2 \mathrm{d}x\mathrm{d}t \leq |v|_{1,\Omega}^2$ which concludes the proof. ■

2.6 Traces

For point-wise defined continuous functions it is clear what their restriction to a lower dimensional set means. For elements of $L_2(\Omega)$ such a restriction has no meaning because such lower dimensional sets have measure zero. Since elements of $H^1(\Omega)$, $\Omega \subset \mathbb{R}^d$, $d > 1$, need not be continuous it is not clear in which sense the restriction of $u \in H^1(\Omega)$ to $\partial\Omega$ is meaningful. This, however, is needed when looking for solutions to weak formulations like (1.3.10) in the space $H^1(\Omega)$ which are required to have "zero boundary values" or later inhomogeneous boundary values. The answer to this question is provided by the so called *Trace Theorem* for Sobolev spaces. We are content here with a weaker version and a special case.

Theorem 2.6.1 *There exists a continuous linear operator* $T : H^1(\Omega) \to L_2(\partial\Omega)$ *such that* $Tu = u|_{\partial\Omega}$ *whenever* $u \in H^1(\Omega) \cap C(\overline{\Omega})$ *and*

$$\|Tu\|_{0,\partial\Omega} \leq C\|u\|_{1,\Omega}, \quad \forall\, u \in H^1(\Omega), \tag{2.6.1}$$

where C *is independent of* u. □

Proof Suppose that we have already shown that for $u \in C^\infty(\overline{\Omega})$ (or $C^1(\overline{\Omega})$)

$$\|u|_{\partial\Omega}\|_{L_2(\partial\Omega)} \leq C\,\|u\|_{1,\Omega} \tag{2.6.2}$$

holds for some constant C independent of u. Now, by Remark 2.3.2, we know that for any $v \in H^1(\Omega)$ there exists a sequence $\{v_j\}_{j\in\mathbb{N}}$, with $v_j \in C^\infty(\overline{\Omega})$ such that

$$\|v - v_j\|_{1,\Omega} \to 0,\ j \to \infty. \tag{2.6.3}$$

Hence $\{v_j\}$ is a Cauchy sequence in $H^1(\Omega)$ and, by (2.6.2), $\{v_j|_{\partial\Omega}\}_{j\in\mathbb{N}}$ is a Cauchy sequence in $L_2(\partial\Omega)$. Since $L_2(\partial\Omega)$ is complete, there exists a limit of $v_j|_{\partial\Omega}$ in $L_2(\partial\Omega)$ which we call Tv.

Exercise 2.6.1 *Show that* $v \to Tv$ *is well defined, i.e.,* Tv *is independent of the particular sequence. Moreover* T *is a linear operator from* $H^1(\Omega)$ *into* $L_2(\partial\Omega)$ *satisfying*

$$\|Tv\|_{0.\partial\Omega} \leq C\,\|v\|_{1,\Omega}\,. \tag{2.6.4}$$

Therefore, it remains to prove the validity of (2.6.2) for any $u \in C^\infty(\overline{\Omega})$.

This is typically done in two steps. To simplify technicalities we assume that $\partial\Omega$ is a C^1 boundary. This means there exists a neighborhood $U \subset \mathbb{R}^d$ such that $\partial\Omega \cap U$ is the graph of a C^1-function $g(y)$, $|y| < r$, with respect to some suitable local coordinate system.

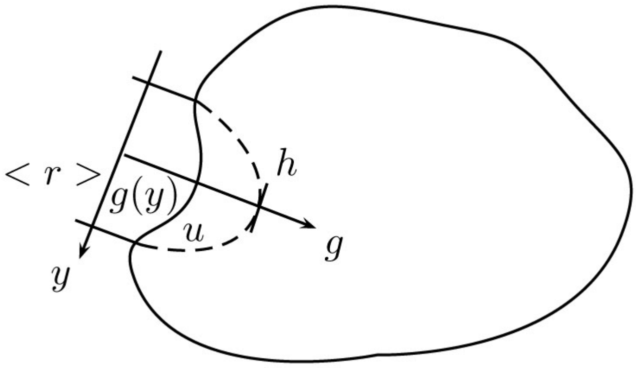
To prepare for the computation of an L_2-norm on $\partial\Omega$, suppose that $\Omega \subset \mathbb{R}^2$, i.e., $\partial\Omega$ is a curve, parametrized locally by the function $g(y)$. Then, a small line segment $d\Gamma$ on $\partial\Omega$ has by Pythagoras' Theorem length

$d\Gamma \approx h\sqrt{1+\frac{(g(y+h)-g(y))^2}{h^2}}$ so that $\int_{\partial\Omega} |u(x)|^2 d\Gamma = \int_{|y|<r} |u(y,g(y))|^2 \big(1+g'(y)^2\big)^{1/2} dy$. Below we use the corresponding multivariate analog for $y \in \mathbb{R}^{d-1}$.

<u>*Localization:*</u> suppose that $\operatorname{supp}(u) \subset U \cap \overline{\Omega}$ so that

$$
\begin{aligned}
\|u|_{\partial\Omega}\|_{0,\partial\Omega}^2 =& \int\limits_{|y|<r,\ y\in\mathbb{R}^{d-1}} |u\left(y,g\left(y\right)\right)|^2 \sqrt{1+|\nabla g\left(y\right)|^2}\mathrm{d}y \\
\leq& c_1^* \int\limits_{|y|<r} |u\left(y,g\left(y\right)\right)|^2 \mathrm{d}y, \quad c_1^* := \sup_{|y|<r} \sqrt{1+|\nabla g\left(y\right)|^2} \\
=& c_1^* \int\limits_{|y|<r} \int\limits_0^h -\frac{\partial}{\partial s}\Big[u\left(y,g\left(y\right)+s\right)^2\Big]\mathrm{d}s\mathrm{d}y \quad (u(y,g(y)+h)=0) \\
=& c_1^* \int\limits_{|y|<r} \int\limits_0^h -\left\{2u\left(y,g\left(y\right)+s\right)\frac{\partial}{\partial y_d}u\left(y,g\left(y\right)+s\right)\right\}\mathrm{d}s\mathrm{d}y \\
\overset{|2ab|\leq a^2+b^2}{\leq}& c_1^* \int\limits_{|y|<r} \int\limits_0^h \left\{u\left(y,g\left(y\right)+s\right)^2 + \left(\frac{\partial}{\partial y_d}u\left(y,g\left(y\right)+s\right)\right)^2\right\}\mathrm{d}s\mathrm{d}y \\
=& c_1^* \int\limits_{\Omega\cap U} u\left(x\right)^2 + |\nabla u\left(x\right)|^2 \mathrm{d}x \\
=& c_1^* \|u\|_{1,\Omega\cap U}^2
\end{aligned}
\tag{2.6.5}
$$

<u>*Partition of unity:*</u> Since $\partial\Omega$ is compact, one can find neighborhoods $U_k,\ k = 1,\dots,N$, such that $\partial\Omega \subset \bigcup\limits_{k=1}^N U_k$ and each $\partial\Omega \cap U_k$ is a C^1-graph (in a suitable coordinate system). Consider a partition of unity $\{\eta_k\}_{k=1}^N$,



$\eta_k \in C_0^\infty(\mathbb{R}^d)$ such that

$$\operatorname{supp} \eta_k \subset U_k, \quad 0 \le \eta_k \le 1$$
$$\sum_{k=1}^{N} \eta_k (x) = 1, \qquad x \in \overline{\Omega}.$$

Now for $u \in H^1(\Omega) \cap C^\infty(\overline{\Omega})$ we define $u_k := \eta_k u \in C^\infty(\overline{\Omega}) \cap H^1(\Omega)$ and conclude that supp $u_k \subset U_k$ while $\sum_{k=1}^N u_k = u$. Therefore

$$\begin{aligned} \|u|_{\partial\Omega}\|_{0,\partial\Omega} &= \left\| \sum_{k=1}^{N} u_k|_{\partial\Omega} \right\|_{0,\partial\Omega} \\ &\le \sum_{k=1}^{N} \|u_k|_{\partial\Omega}\|_{0,\partial\Omega} \,. \end{aligned}$$

Applying (2.6.5) to each u_k with $c_k = c_1^*$ we conclude that

$$\|u|_{\partial\Omega}\|_{0,\partial\Omega} \le \sum_{k=1}^{N} c_k \|u_k\|_{1,\Omega} \le C \|u\|_{1,\Omega} \,.$$

Since

$$\left\|\partial_{x_j} (\eta_k u)\right\|_{0,\Omega} \le \left\|\partial_{x_j} \eta_k\right\|_{L_\infty(\Omega)} \|u\|_{0,\Omega} + \left\|\partial_{x_j} u\right\|_{0,\Omega}$$

we conclude that

$$\|u_k\|_{1,\Omega} \leq \tilde{C}\,\|u\|_{1,\Omega}\,,$$

which completes the proof. Note that C depends on Ω through the derivatives of the η_k. ■

For later purposes we record the following "localized version" of the above Trace-inequality.

Remark 2.6.1 Suppose that $u \in C^\infty(\overline{\Omega})$ satisfies $\operatorname{supp} u \subset U \cap \Omega$ for some open subset U and $\operatorname{diam}(U \cap \Omega) = h$. Then there is a constant C depending only on $\partial\Omega$ such that

$$\big\|u|_{\partial\Omega}\big\|^2_{0,\partial\Omega} \leq C\big\{h^{-1}\|u\|^2_{0,\Omega\cap U} + h\|\nabla u\|^2_{0,\Omega\cap U}\big\}. \qquad (2.6.6)$$

Proof In the step from line 4 to line 5 in (2.6.5) we use the Young inequality $|2ab| \leq h^{-1}a^2 + hb^2$ to arrive at (2.6.6). ■

The following theorem explains in which sense we can say that $H^1_0(\Omega)$ consists exactly of thos functions in $H^1(\Omega)$ which have zero boundary values.

Theorem 2.6.2 *Asume that Ω is a Lipshitz domain. Then one has*

$$H^1_0(\Omega) = \{u \in H^1(\Omega) : Tu = 0\}.$$

The direction $\underline{u \in H^1_0(\Omega) \;\Rightarrow\; Tu = 0}$ is simple: in fact, since $H^1_0(\Omega) = \overline{C^\infty_0(\Omega)}^{\|\cdot\|_{1,\Omega}}$ there exists a sequence $(v_j)_{j\in\mathbb{N}} \subset C^\infty_0(\Omega)$ with $\|u - v_j\|_{1,\Omega} \to 0$, $j \to \infty$. Hence

$$\|Tu\|_{0,\partial\Omega} = \|Tu - Tv_j\|_{0,\partial\Omega} = \|T(u - v_j)\|_{0,\partial\Omega} \overset{(2.6.1)}{\leq} C\|u - v_j\|_{1,\Omega} \to 0,\ j \to \infty.$$

<u>The converse direction:</u> $u \in H^1(\Omega)$, $Tu = 0 \;\Rightarrow\; u \in H^1_0(\Omega)$ can be found e.g. in [Ada75, Eva98].

It is important to note that the trace operator $T : H^1(\Omega) \to L_2(\partial\Omega)$ is *not* surjective, i.e., the range of T is a *strict* subspace of $L_2(\partial\Omega)$. In fact, the traces of H^1-functions have some regularity. Consider for the range $T\left(H^1\left(\Omega\right)\right)$ of the trace operator on $H^1(\Omega)$

$$|||w|||_{\partial\Omega} := \inf_{v\in H^1(\Omega),\ Tv=w} \|v\|_{1,\Omega} \,. \tag{2.6.7}$$

It is easy to see that this defines a norm on the *trace space* $T(H^1(\Omega)) \subset L_2(\partial\Omega)$. To characterize this subspace by a "smoothness property" one needs Sobolev regularity of *non-integer* order. Without proof we record the following fundamental result because it is essential for dealing with boundary conditions, see [Eva98].

Theorem 2.6.3 *The norm* $|||\cdot|||_{\partial\Omega}$*, defined by* (2.6.7)*, is equivalent to* $\|\cdot\|_{1/2,\partial\Omega}$ *and* $H^{1/2}(\partial\Omega) = T(H^1(\Omega))$*, i.e., the two spaces agree as sets and both norms are equivalent.* □

The general Trace-Theorem says that $T\big(H^k(\Omega)\big) = H^{k-1/2}(\partial\Omega)$, $k \geq 1$. Even more generally, for $1 \leq p \leq \infty$ $T\big(W^{k,p}((\Omega)\big) = W^{k-1/p,p}(\partial\Omega)$, [Ada75].

2.7 Duality and Gelfand-Tripel

We have already seen that the right hand side f in (1.3.10) can be identified with a bounded linear functional on $H_0^1(\Omega)$ when $f \in L_2(\Omega)$. Generally speaking, in weak formulations the right hand side data act as *functionals* on the test space. The notion of "dual" space provides an appropriate framework. We recall from (1.3.17) that for any Banach space $\mathbb{X}$ with norm $\|\cdot\|_{\mathbb{X}}$ the set of all bounded linear functionals on $\mathbb{X}$ is denoted by $\mathbb{X}'$. $\mathbb{X}'$ - the normed dual - is also a Banach space under the norm

$$\|w\|_{\mathbb{X}'} := \|w\|_{\mathcal{L}(\mathbb{X},\mathbb{R})} := \sup_{v\in\mathbb{X}} \frac{w\left(v\right)}{\|v\|_{\mathbb{X}}}.$$

As before we are mostly interested in the case that $\mathbb{X}$ is a Hilbert space $\mathbb{H}$ (like $H^1(\Omega)$, or $H_0^1(\Omega)$ or more generally $H^s(\Omega)$) with norm $\|\cdot\|_{\mathbb{H}} = (\cdot,\cdot)_{\mathbb{H}}^{1/2}$, induced by an inner product on. When $\mathbb{H} = \mathbb{R}^d$ endowed with the standard inner product, $\mathbb{H}^{'}$ can be identified with itself, i.e., every linear (automatically bounded, since $\mathbb{R}^d$ is finite dimensional) functional ℓ on $\mathbb{R}^d$ can be realized as a standard inner product with an element of $\mathbb{R}^d$, i.e., for $\ell \in (\mathbb{R}^d)'$ there exists $y_\ell \in \mathbb{R}^d$, such that

$$\ell\,(x) = y_\ell^T x, \quad \forall x \in \mathbb{R}^d.$$

This is a special case of the *Riesz-Representation Theorem.* One says: y_ℓ is a representer of ℓ with respect to the *dual pairing*

$$\langle \ell, x\rangle = \ell(x) = y_\ell^T x.$$

Things are slightly more involved in infinite dimensions. In all the previous examples we have

$$\mathbb{H} \subseteq L_2(\Omega)$$

with (at least) a continuous embedding. Since in this case for $f \in L_2(\Omega)$, $v \in \mathbb{H}$,

$$\begin{aligned} |f\,(v)| := \left| \int\limits_\Omega f\,(x)\, v\,(x)\, \mathrm{d}x \right| &= \left| (f, v)_{0,\Omega} \right| \leq \|f\|_{0,\Omega} \|v\|_{0,\Omega} \\ &\leq \|f\|_{0,\Omega} \|v\|_{\mathbb{H}} \end{aligned}$$

we see that $L_2(\Omega)$ is (can be identified with) a subspace of $\mathbb{H}^{'}$, i.e.

$$\mathbb{H} \subseteq L_2\,(\Omega) \subseteq \mathbb{H}^{'}, \tag{2.7.1}$$

with continuous (actually dense) embeddings. A triple $(\mathbb{H}, L_2(\Omega), \mathbb{H}^{'})$ with the embeddings (2.7.1) is called a *Gelfand-triple.*

A common notation: In the case (2.7.1) one often expresses the action of a functional $f \in \mathbb{H}^{'}$ on $\mathbb{H}$ as

$$f : v \to f\,(v) = \langle f, v\rangle.$$

We shall see later that this notation reflects a certain representation of the functional f depending on the *pivot space* $L_2(\Omega)$.

In fact, it means that whenever $f \in L_2(\Omega) \subset \mathbb{H}^{'}$ one has

$$\langle f, v\rangle = (f, v)_{0,\Omega} = \int\limits_\Omega f\,(x)\, v\,(x)\, \mathrm{d}x,$$

i.e., f is realized through the standard inner product in $L_2(\Omega)$.

Exercise 2.7.1 *Find other possible representations.*

In view of the weak formulation (1.3.10), we are particularly interested in $\mathbb{H} = H_0'(\Omega)$ and recall the notation

$$\left(H_0^1(\Omega)\right)' =: H^{-1}(\Omega).$$

We have already seen that $L_2(\Omega) \subset H^{-1}(\Omega)$. Here are some examples that show that $L_2(\Omega)$ is a *strict subset* of $H^{-1}(\Omega)$.

Example 2.7.1 Let $\Omega = (-1,1) \times (0,1) \subset \mathbb{R}^2$ and

$$s(x,y) = \begin{cases} 0, & x \in (-1,0) \times (0,1) \\ 1, & x \in (0,1) \times (0,1). \end{cases} \tag{2.7.2}$$

Clearly, $s \in L_2(\Omega)$, so that

$$\ell(v) := \langle \ell, v \rangle := -(s, \partial_x v)_{0,\Omega}$$

belongs to $H^{-1}(\Omega)$ because, by Cauchy-Schwarz,

$$\begin{aligned} |\langle \ell, v \rangle| &= \left|(s, \partial_x v)_{0,\Omega}\right| \le \|s\|_{0,\Omega} \|\partial_x v\|_{0,\Omega} \\ &\le \|s\|_{0,\Omega} \|v\|_{1,\Omega}, \end{aligned}$$

i.e., ℓ is bounded. Note that, if s was differentiable, one would have

$$\langle \ell, v \rangle = (\partial_x s, v)_{0,\Omega},$$

i.e., $\ell = \partial_x s$ is the *distributional derivative* of s which no longer belongs to $L_2(\Omega)$ but still to $H^{-1}(\Omega)$.

However, s can be *approximated* in $L_2(\Omega)$ by a sequence $s_j \in C^\infty(\overline{\Omega})$ (exercise). Let

$$\langle \ell_j, v \rangle := -(s_j, \partial_x v)_{0,\Omega} = (\partial_x s_j, v)_{0,\Omega} = \langle \partial_x s_j, v \rangle$$

be the corresponding (smoothed) functionals.

Then

$$\begin{aligned} \|\ell - \ell_j\|_{-1,\Omega} := \|\ell - \ell_j\|_{H^{-1}(\Omega)} &= \sup_{v \in H_0^1(\Omega)} \frac{\langle \ell - \ell_j, v \rangle}{\|v\|_{1,\Omega}} = \sup_{v \in H_0^1(\Omega)} \frac{(s_j - s, \partial_x v)_{0,\Omega}}{\|v\|_{1,\Omega}} \\ &\le \sup_{v \in H_0'(\Omega)} \frac{\|s_j - s\|_{0,\Omega} \|\partial_x v\|_{0,\Omega}}{\|v\|_{1,\Omega}} \\ &\le \|s_j - s\|_{0,\Omega} \to 0, \ j \to \infty, \end{aligned}$$

which shows that ℓ can be approximated in $H^{-1}(\Omega)$ by functionals represented by C^∞ functions. □

This example shows that $H^{-1}(\Omega)$ contains functionals that *cannot* be represented directly through the L_2-scalar product with an L_2-function ($\partial_x s \notin L_2(\Omega)$). However, the functional ℓ defined above, that as a distribution agrees with $\partial_x s$ can be approximated with respect to the norm $\|\cdot\|_{-1,\Omega}$ arbitrarily well by functionals that do have infinitely often differentiable representers in $L_2(\Omega)$. We shall see soon that in this sense $C^\infty(\overline{\Omega})$ is dense in $H^{-1}(\Omega) = (H_0^1(\Omega))'$.

Finally, we have the following representation of ℓ:

$$\begin{aligned} -\left(s, \partial_x v\right)_{0,\Omega} &= -\int_0^1 \int_0^1 \partial_x v\left(x,y\right) \mathrm{d}x\mathrm{d}y \\ &= -\int_0^1 v\left(1,y\right) - v\left(0,y\right) \mathrm{d}y \qquad (2.7.3) \\ &= \int_0^1 v\left(0,y\right) \mathrm{d}y, \end{aligned}$$

since v has compact support in Ω. Hence ℓ is the trace integral over the interface $\Gamma := \{(0,y) \ : \ y \in (0,1)\} \subset \Omega$. So we could also argue that

$$\begin{aligned} |\langle \ell, v\rangle| = \left| \int_0^1 v\left(0,y\right) \mathrm{d}y \right| &\overset{\text{C.S.}}{\leq} \left(\int_0^1 \left|v\left(0,y\right)\right|^2 \mathrm{d}y \right)^{1/2} \\ &= \left\|Tv\right\|_{0,\Gamma}. \end{aligned}$$

More generally, for any $r \in L_2(\Gamma)$ we could define (since $v \in H^1(\Omega)$ possesses a trace in $L_2(\Gamma)$)

$$\ell\left(v\right) := \int_\Gamma r\left(y\right) Tv\left(y\right) \mathrm{d}y, \qquad (2.7.4)$$

which is clearly a linear functional on $H_0^1(\Omega)$. By Theorem 2.6.1 we obtain as before

$$|\ell(v)| \leq \|r\|_{0,\Gamma} \|Tv\|_{0,\Gamma} \overset{(2.6.4)}{\leq} C \|r\|_{0,\Gamma} \|v\|_{1,\Omega}$$

i.e. $\|\ell\|_{-1,\Omega} \leq c\|r\|_{0,\Gamma}$ and $\ell \in H^{-1}(\Omega)$.

Example 2.7.2 The stronger trace Theorem 2.6.3 says that r in (2.7.4) need not even be in $L_2(\Gamma)$. In fact, let for Ω as before

$$r \in \left(H^{1/2}(\partial\Omega)\right)' =: H^{-1/2}(\partial\Omega) \tag{2.7.5}$$

be a bounded linear functional on the trace space $H^{1/2}(\partial\Omega)$, then

$$\ell(v) := r(Tv) \tag{2.7.6}$$

belongs to $H^{-1}(\Omega)$. In this sense one has

$$H^{-1/2}(\partial\Omega) \subset \left(H^1(\Omega)\right)' \tag{2.7.7}$$

with a continuous embedding (Proof Exercise). □

Remark 2.7.1 The space $H^{-1/2}(\partial\Omega) = \left(H^{1/2}(\partial\Omega)\right)'$ can be characterized as the space of *normal traces* of $H(\mathrm{div})$-functions. More procesely, let

$$H(\mathrm{div};\Omega) := \{w \in L_2(\Omega)^d : \mathrm{div}\, w \in L_2(\Omega)\}$$

(always understood in the sense of weak derivatives). Then, the space spanned by the *normal traces* of vector fields in $H(\mathrm{div};\Omega)$ can be identified with the dual of $H^{1/2}(\Omega)$, i.e.,

$$H^{-1/2}(\partial\Omega) := \left(H^{1/2}(\Omega)\right)' = \{g = T(n \cdot w) : w \in H(\mathrm{div};\Omega)\}, \tag{2.7.8}$$

see [BF91]. □

3 Variational Formulations and the Inf-Sup-Condition

3.1 Further Important Examples

Before addressing the solvability of problems of the form (1.3.23) and their numerical solution, we collect several further important examples of PDEs and present their weak formulation. We then proceed characterizing the solvability of such variational problems.

3.1.1 2nd Order Elliptic Boundary Value Problems

Let $A(x) \in L_\infty(\Omega; \mathbb{R}^{d\times d})$ be a uniformly positive definite symmetrix matrix on Ω, $b(x) \in L_\infty(\Omega; \mathbb{R}^d)$, $c(x) \in L_\infty(\Omega)$. By the same use of Green's formulas as above the weak formulation of

$$\begin{aligned} -\mathrm{div}(A(x)\nabla u)(x) + b(x)\cdot\nabla u(x) + c(x)u(x) = f(x), &\qquad x \in \Omega, \\ u(x) = 0, &\qquad x \in \partial\Omega, \end{aligned} \tag{3.1.1}$$

for some $f \in L_2(\Omega)$ takes the form: find $u \in \mathbb{U} := H_0^1(\Omega)$ such that

$$B(u,v) := \int_\Omega \nabla v \cdot A\nabla u + vb\cdot\nabla u + cuv\mathrm{d}x = f(v), \quad v \in \mathbb{U} := H_0^1(\Omega). \tag{3.1.2}$$

Note that by Cauchy-Schwarz

$$\begin{aligned} |B(u,v)| &\leq |A|\,|u|_{1,\Omega}|v|_{1,\Omega} + \|b\|_{L_\infty(\Omega)}\|v\|_{0,\Omega}|u|_{1,\Omega} + \|c\|_{L_\infty(\Omega)}\|u\|_{0,\Omega}\|v\|_{0,\Omega} \\ &\leq 3\max\left\{|A|, \|b\|_{L_\infty(\Omega)}, \|c\|_{L_\infty(\Omega)}\right\}\|u\|_{1,\Omega}\|v\|_{1,\Omega} \\ &=: C_B\|u\|_{1,\Omega}\|v\|_{1,\Omega}, \quad u,v \in H_0^1(\Omega), \end{aligned} \tag{3.1.3}$$

i.e., the bilinear form is continuous on $H_0^1(\Omega) \times H_0^1(\Omega)$.

We shall later discuss other boundary conditions as well which effects the choice of $\mathbb{U}$ since in the above case the homogeneous boundary conditions are incorporated in $\mathbb{U}$. As in the case of the Poisson problem we have chosen here the trial space equal to the test space. We will see later that this turns out to be problematic when the convection b strongly dominates the diffusion A.

3.1.2 The Biharmonic Equation

The displacement of (the middle surface) of a *clamped* plate covering $\Omega \subset \mathbb{R}^2$ (thin enough to neglect shear effects) under a vertically acting force $f \in L_2(\Omega)$ can be modeled by the *biharmonic equation*

$$\begin{aligned} \Delta^2 u &= f \text{ in } \Omega, \\ u = \partial_n u &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{3.1.4}$$

The classical solution would have to belong to $C^4(\Omega)$. A weak formulation requires *higher order Sobolev spaces* $H^k(\Omega)$ to be discussed later in more detail. The corresponding norm for $H^k(\Omega)$ is

$$\|v\|_{k,\Omega}^2 := \|v\|_{0,\Omega}^2 + |v|_{k,\Omega}^2, \quad |v|_{k,\Omega}^2 := \sum_{|\alpha|=k} \|\partial^\alpha v\|_{0,\Omega}^2. \tag{3.1.5}$$

In analogy to the preceding discussion we can then consider the spaces

$$H^2(\Omega) := \overline{C^\infty(\Omega)}^{\|\cdot\|_{k,\Omega}}, \quad H_0^2(\Omega) := \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{2,\Omega}}. \tag{3.1.6}$$

Note that the elements v in $H_0^2(\Omega)$ now satisfy $v = \nabla_n v = 0$ on $\partial\Omega$, i.e., not only v but also the normal derivative $\nabla_n v$ vanishes on the boundary $\partial\Omega$ (again in a sense to be made precise later).

Exercise 3.1.1 *Use the Green's formulas to show that the weak formulation of* (3.1.4) *is: find $u \in \mathbb{U} := H_0^2(\Omega)$ such that*

$$B(u,v) := \int_\Omega \Delta u \Delta v dx = \int_\Omega v f dx =: f(v), \quad v \in H_0^2(\Omega), \tag{3.1.7}$$

and that

$$|B(u,v)| \leq d\|u\|_{2,\Omega}\|v\|_{2,\Omega}, \quad \forall\, u,v \in H^2(\Omega). \tag{3.1.8}$$

3.1.3 3D-Eddy Current Equations

The next example is a *system*, namely a special case of *Maxwell's equations* where the unknowns are 3D-vector fields representing (electric) currents. Defining the "curl" operator of a smooth vector field

$$\operatorname{curl} v := \nabla \wedge v := \Big(\partial_{x_2} v_3 - \partial_{x_3} v_2, \partial_{x_3} v_1 - \partial_{x_1} v_3, \partial_{x_1} v_2 - \partial_{x_2} v_1\Big)^T,$$

with the vector product

$$w \wedge v := (w_2 v_3 - w_3 v_2, w_3 v_1 - w_1 v_3, w_1 v_2 - w_2 v_1)^T,$$

for $f \in L_2(\Omega; \mathbb{R}^3)$ find

$$\begin{aligned} u \in \mathbb{U} &:= H_0(\mathrm{curl}; \Omega) := \overline{C_0^\infty(\Omega; \mathbb{R}^3)}^{\|\cdot\|_{H(\mathrm{curl};\Omega)}}, & (3.1.9) \\ \|v\|^2_{H(\mathrm{curl};\Omega)} &:= \|v\|^2_{0,\Omega} + \|\operatorname{curl} v\|^2_{0,\Omega}, & (3.1.10) \end{aligned}$$

such that

$$\begin{aligned} \operatorname{curl}(\mu \operatorname{curl} u) + \kappa u = f \quad &\text{in } \Omega \\ u \wedge n = 0 \quad &\text{on } \partial\Omega. \end{aligned} \qquad (3.1.11)$$

Note that $u \wedge n$ is tangent to $\partial\Omega$ so that we encounter here a different type of boundary condition. Note also that $H_0^1(\Omega; \mathbb{R}^3) \subset H(\mathrm{curl}; \Omega)$.

Exercise 3.1.2 *Defining* $H(\mathrm{curl}; \Omega)$ *as in* (3.1.9) *with* $C_0^\infty(\Omega; \mathbb{R}^3)$ *replaced by* $C^\infty(\Omega; \mathbb{R}^3)$*, show that*

$$\int_\Omega (\operatorname{curl} w) \cdot v dx = \int_\Omega w \cdot \operatorname{curl} v dx + \int_{\partial\Omega} w \cdot (v \wedge n) ds.$$

The weak formulation of (3.1.11) then reads: for $f \in \mathbb{U}'$, $\mathbb{U} = H_0(\mathrm{curl}; \Omega)$, find $u \in \mathbb{U}$ such that

$$B(u, v) := \int_\Omega \mu \operatorname{curl} u \cdot \operatorname{curl} v + \kappa u v dx = \int_\Omega f \cdot v dx =: f(v), \quad v \in \mathbb{U}. \qquad (3.1.12)$$

Again we obtain by Cauchy-Schwarz

$$|B(u, v)| \leq \max \big\{ \|\mu\|_{L_\infty(\Omega)}, \|\kappa\|_{L_\infty(\Omega)} \big\} \|u\|_{H(\mathrm{curl};\Omega)} \|v\|_{H(\mathrm{curl};\Omega)}, \qquad (3.1.13)$$

holds for all $u, v \in H(\mathrm{curl}; \Omega)$.

3.1.4 The Stokes System

The next example is the classical model for a very viscous stationary incompressible fluid flow where the unknowns is the *velocity vector field* u with values in $\mathbb{R}^d$ and the scalar pressure field p. Here *no-slip* (Dirichlet) boundary conditions are imposed on the velocity field:

$$\begin{array}{rcll} -\Delta u \quad +\nabla p & = & f & \text{in } \Omega, \\ \operatorname{div} u \qquad\quad & = & 0 & \text{in } \Omega, \\ u & = & 0 & \text{on } \partial\Omega, \end{array} \tag{3.1.14}$$

Note that the pressure is only defined up to a constant factor which has to be accounted for in a weak formulation. The first equation represents the momentum balance while the second equation implies preservation of mass for incompressible fluids.

Exercise 3.1.3 *To derive such a weak formulation, show that for two sufficiently smooth vector fields v, w vanishing on $\partial\Omega$ one has*

$$-\int_\Omega \Delta v \cdot w dx \;=\; \int_\Omega \nabla v : \nabla w dx = \sum_{i=1}^{d} \nabla v_i \cdot \nabla w_i dx. \tag{3.1.15}$$

Since the right hand side of (3.1.15) makes sense for $v, w \in H_0^1(\Omega; \mathbb{R}^d)$ we will seek the solution component u in $\mathbb{H} := H_0^1(\Omega; \mathbb{R}^d) = (H_0^1(\Omega))^d$. Moreover, when $v \in \mathbb{H} = H_0^1(\Omega; \mathbb{R}^d)$ we can use Green's Theorem to write

$$\int_\Omega \nabla p \cdot v dx \;=\; -\int_\Omega p \operatorname{div} v dx =: b(p, v). \tag{3.1.16}$$

Since

$$|b(p, v)| \le \|p\|_{0,\Omega} |v|_{1,\Omega}, \quad v \in H_0^1(\Omega; \mathbb{R}^d), \tag{3.1.17}$$

the pressure p need only belong to $L_2(\Omega)$. Since we have to factor out constance, let

$$\mathbb{Q} := \{q \in L_2(\Omega) : \int_\Omega q dx = 0\} \equiv L_2(\Omega)/\mathbb{R}. \tag{3.1.18}$$

Setting in this case

$$\mathbb{U} := \mathbb{H} \times \mathbb{Q}, \quad \|[v,q]\|_{\mathbb{U}} := \|v\|_{\mathbb{H}} + \|q\|_{\mathbb{Q}}, \tag{3.1.19}$$

the weak formulation of (3.1.14) reads: given

$$[f,0] \in \mathbb{H}' \times \mathbb{Q}' = (\mathbb{H} \times \mathbb{Q})' = \mathbb{U}',$$

find $[u,p] \in \mathbb{U}$ such that

$$\begin{array}{lclcll} a(u,v) & + & b(p,v) & = & f(v), & \forall\, v \in \mathbb{H}, \\ b(q,u) & & & = & 0, & \forall\, q \in \mathbb{Q}. \end{array} \tag{3.1.20}$$

It will later be convenient to write (3.1.20) in a more compact form, collecting the unknowns, test functions, and data in

$$U := [u,p], \quad V := [v,q] \in \mathbb{U} = \mathbb{H} \times \mathbb{Q}, \quad F := [f,0] \in \mathbb{U}' = \mathbb{H}' \times \mathbb{Q}',$$

and defining the bilinear form $B(U,V) := a(u,v) + b(p,v) + b(q,u)$ we see that (3.1.20) is equivalent to finding $U \in \mathbb{U}$ such that

$$B(U,V) \;=\; F(V), \quad \forall\, V \in \mathbb{U}', \tag{3.1.21}$$

which formally resembles to all earlier examples. In fact, defining now for the vector fields $v \in H_0^1(\Omega;\mathbb{R}^d)$ in analogy to the scalar case $|v|_{1,\Omega} := \|\nabla v\|_{0,\Omega}$ and recalling (3.1.17), we obtain

$$\begin{aligned} |B(U,V)| &\le |a(u,v)| + |b(p,v)| + |b(q,u)| \\ &\le |u\|_{1,\Omega}|v\|_{1,\Omega} + \|p\|_{0,\Omega}|v|_{1,\Omega} + \|q\|_{0,\Omega}|u|_{1,\Omega} \\ &= \big(|u|_{1,\Omega} + \|p\|_{0,\Omega}\big)\,|v|_{1,\Omega} + |u|_{1,\Omega}\|q\|_{0,\Omega} \\ &\le \big(|u|_{1,\Omega} + \|p\|_{0,\Omega}\big)\,|v|_{1,\Omega} + \big(|u|_{1,\Omega} + \|p\|_{0,\Omega}\big)\|q\|_{0,\Omega} \\ &\le \big(|u|_{1,\Omega} + \|p\|_{0,\Omega}\big)\big(|v|_{1,\Omega} + \|q\|_{0,\Omega}\big) \\ &\le \|U\|_{\mathbb{U}}\|V\|_{\mathbb{U}} \quad \text{(here } \mathbb{U} = \mathbb{V}\text{)}. \end{aligned} \tag{3.1.22}$$

Thus, also in this case the bilinear form $B(\cdot,\cdot)$ is continuous on $\mathbb{U} \times \mathbb{U} = \big(H^1(\Omega;\mathbb{R}^d) \times L_{2,0}(\Omega)\big) \times \big(H^1(\Omega;\mathbb{R}^d) \times L_{2,0}(\Omega)\big)$.

Nevertheless, there are essential structural differences between (3.1.21) and earlier examples of variational formulations that will be discussed later.

3.1.5 Time-Dependent Incompressible Navier-Stokes Equations

The Stokes System is a simplified special case of the full *incompressible* time-dependent *Navier-Stokes equations* which models also less viscous flows where transport effects can no longer be neglected. u, p denote again the velocity and pressure fields but are allowed to depend on time t as well. Allowing the viscosity ε to become small the momentum and continuity equations now read

$$\begin{array}{rcll} \partial_t u - \operatorname{div}(\varepsilon \nabla u) + u \cdot \nabla u + \nabla p &=& f & \text{in } \Omega, \\ \operatorname{div} u &=& 0 & \text{in } \Omega, \\ u &=& 0 & \text{on } \partial\Omega, \\ u(\cdot, 0) &=& u_0 & \text{in } \Omega. \end{array} \tag{3.1.23}$$

Aside from the fact that the unknowns $u(x,t), p(x,t)$ are now time dependent the main distinction from the preceding examples is that the system is now *nonlinear* (because of the third term in the first equation). Nevertheless, for large viscosity ε the first equation is essentially *parabolic* so that the initial-boundary conditions in the last two lines of (3.1.23) appear to be reasonable.

To relate the above system to the preceding examples one can first discretize in time and *linearize* which leads to the following *Oseen* formulation of the Navier-Stokes equations. To describe this, let us set

$$u^n(x) = u(x, t_n), \quad p^n(x) := p(x, t_n), \quad f^n(x) := f(x, t_n),$$

for a sequence of discrete time steps

$$0 = t_0 < t_1 < \cdots < t_N = T, \quad \tau := t_{n+1} - t_n,$$

with time increments τ which could vary but are here kept constant for convenience. Then, taking for convenience $\varepsilon \equiv$ constant, the time evolution in (3.1.23) can be approximated by

$$\begin{array}{rcll} u^{n+1} - \tau\varepsilon\Delta u + \tau u^n \cdot \nabla u^{n+1} + \tau \nabla p^{n+1} &=& u^n + \tau f^{n+1} & \text{in } \Omega, \\ \operatorname{div} u^{n+1} &=& 0 & \text{in } \Omega, \\ u^{n+1} &=& 0 & \text{on } \partial\Omega, \\ u^{n+1}(\cdot, t_n) &=& u^n & \text{in } \Omega. \end{array} \tag{3.1.24}$$

Since u^n in the third summand of the first equation is now a *known* velocity field from the preceding time step, the first equation in the unknwon u^{n+1} is now a *linear convection diffusion equation* with $c(x) = 1, A = \tau\varepsilon, b = \tau u^n$, and right hand side $u^n + \tau f^{n+1}$ in the convection-diffusion equation (3.1.1) (ignoring for the moment ∇p^{n+1}). Defining the bilinear form

$$\bar{a}(v, w) = \bar{a}_{\varepsilon,\tau}(v, w) = \int_\Omega \varepsilon\tau\nabla v : \nabla w + \tau(u^n \cdot \nabla v)w + vwdx, \qquad (3.1.25)$$

and $b(q, v)$ as in (3.1.16), the weak formulation of (3.1.24) becomes with the same choice of trial spaces $\mathbb{U}$

$$\begin{array}{lclcl} \bar{a}_{\varepsilon,\tau}(u^{n+1}, v) & + & b(p^{n+1}, v) & = & f^{n+1}, \quad \forall\, v \in \mathbb{H}, \\ b(u^{n+1}, q) & & & = & 0, \quad \forall\, q \in \mathbb{Q}, \end{array} \qquad (3.1.26)$$

and hence can again be cast in the form (3.1.21). Accordingly, the linearized problems have continuous bilinear forms over the same spaces as the Stokes problem.

The particular difficulty when trying to solve the Navier-Stokes equations numerically arises for small viscosity (large “Reynolds numbers” proporial to the inverse of the kinematic viscosity) because the underlying convection-diffusion equation becomes then strongly convection dominant. Again, the above choice of $\mathbb{U}$ then turns out to become problematic.

3.1.6 Pure Transport - an Unsymmetric Formulation

In the limit $\varepsilon \to 0$ the bilinear form $\bar{a}_{\varepsilon,\tau}(v, w)$ from (3.1.25) changes its character. It is therefore instructive to discuss as the last example the pure transport equation

$$b \cdot \nabla u + cu \;=\; f \quad \text{in } \Omega, \;\; u = 0 \quad \text{on } \Gamma_-, \qquad (3.1.27)$$

which is well-posed only when the boundary conditions are confined to the *inflow-boundary*

$$\Gamma_- := \{x \in \partial\Omega : n(x) \cdot b(x) < 0\}.$$

Note that the time-dependent analog $\partial_t u + b \cdot \nabla u + cu = f$ can be treated in exactly the same manner by replacing x by $\hat{x} := (x, t)$ and b by $(b^T, 1)^t$ with a corresponding inflow-boundary of the space-time cylinder $\Omega \times [0, T)$.

To arrive at a weak formulation of (3.1.27) one can again multiply the equation by test functions. But now we have two options, namely considering the bilinear form

$$B_1(u, v) := \int_\Omega (b \cdot \nabla u)v + cuvdx, \tag{3.1.28}$$

or, by applying Green's Theorem to obtain

$$\begin{aligned} B_2(u, v) &:= \int_\Omega u(-\mathrm{div}(bv)) + cuvdx + \int_{\partial\Omega} b \cdot nuvds \\ &= \int_\Omega -u(b \cdot \nabla v)u + (c - \mathrm{div}\, b)uvdx \\ &\qquad + \int_{\partial\Omega} b \cdot nuvds, \end{aligned} \tag{3.1.29}$$

When using $B_1(\cdot, \cdot)$ we are led to seek the solution in the space

$$\begin{aligned} \mathbb{U} &= H^b_-(\Omega) := \overline{\{v \in C^1(\Omega) : v|_{\Gamma_-} = 0\}}^{\|\cdot\|_{H^b(\Omega)}}, \\ \|v\|^2_{H^b(\Omega)} &:= \|v\|^2_{0,\Omega} + \|b \cdot \nabla u\|^2_{0,\Omega}. \end{aligned} \tag{3.1.30}$$

Roughly speaking the trial space is comprised of those functions in $L_2(\Omega)$ whose streamwise directional derivatives also belong to $L_2(\Omega)$ and which vanish on Γ_-. The latter fact is actually an issue because it is not clear beforehand whether the restriction of an $L_2(\Omega)$-function to Γ_-, which has measure zero, is meaningful. Note that this permits discontinuities along characteristic curves but not across.

For this choice of $\mathbb{U}$ the problem $B_1(u, v) = f(v) = \int_\Omega fvdx$ makes sense when the test functions should be taken as *arbitrary* elements in

$$\mathbb{V} = \mathbb{V}_1 := L_2(\Omega),$$

i.e., trial and test space are now *different*, because

$$|B_1(v, w)| \leq \|v\|_{H^b(\Omega)} \|w\|_{0,\Omega}, \quad v \in \mathbb{U}_1 := H^b(\Omega),\ w \in \mathbb{V}_1 := L_2(\Omega). \tag{3.1.31}$$

For the second choice $B_2(\cdot,\cdot)$ from (3.1.29) the unknown u has been "freed" from all derivatives. Defining in this case the test space

$$\mathbb{V}_2 := H^b_+(\Omega) = \overline{\{v \in C^1(\Omega) : v|_{\Gamma_+} = 0\}}^{\|\cdot\|_{H^b(\Omega)}}, \tag{3.1.32}$$

where $\Gamma_+ := \{x \in \partial\Omega : n(x) \cdot b(x) > 0\}$ is the *outflow-boundary* and since u is to vanish on Γ_-, the weak formulation according to (3.1.29) becomes

$$\int_\Omega -u(b \cdot \nabla v)u + (c - \operatorname{div} b)uvdx = \int_\Omega fvdx, \quad v \in \mathbb{V}_2, \tag{3.1.33}$$

and for $\mathbb{V}_2 := H^b_+(\Omega)$ one has

$$|B_2(v,w)| \leq \|v\|_{0,\Omega}\|w\|_{H^b(\Omega)}, \quad v \in \mathbb{U}_2 := L_2(\Omega),\, w \in \mathbb{V}_2. \tag{3.1.34}$$

In this case essentially no regularity properties are imposed on the solution. This is often referred to as *ultra-weak formulation* of (3.1.27).

In either case, it seems that a meaningful variational formulation, namely ensuring that the involved bilinear forms are continuous, requires in this case to take the trial space $\mathbb{U}$ different from the test space $\mathbb{V}$.

3.1.7 Discussion

The common abstract formulation of all the preceding examples is the following: given a bilinear form

$$B(\cdot,\cdot) : \mathbb{U} \times \mathbb{V} \to \mathbb{R}$$

where $\mathbb{U}, \mathbb{V}$ are Hilbert spaces endowed with inner products and norms

$$(\cdot,\cdot)_\mathbb{U}, \quad \|u\|^2_\mathbb{U} = (u,u)_\mathbb{U}, \quad (\cdot,\cdot)_\mathbb{V}, \quad \|v\|^2_\mathbb{V} = (v,v)_\mathbb{V},$$

and given $f \in \mathbb{V}'$, find $u \in \mathbb{U}$ such that

$$B(u,v) \;=\; f(v), \quad v \in \mathbb{V}, \tag{3.1.35}$$

noting that both sides of the equation are well-defined when the bilinear form $B(\cdot,\cdot)$ is <u>continuous</u> on $\mathbb{U} \times \mathbb{V}$, see (1.3.21).

When $\mathbb{U} = \mathbb{V}$ and $\mathbb{U}_n = \mathbb{V}_n$ in (4.9.7), this is called *Galerkin formulation.* When $\mathbb{U}_n \neq \mathbb{V}_n$ it is called *Petrov-Galerkin formulation.*

The subsequent discussions will be guided by the following issues:

- Although in most of the examples we had $\mathbb{U} = \mathbb{V}$, the last example indicates that it may be important to cover the case where test and trial spaces are different. We refer to this as an *unsymmetric* formulation. In a sense, the original classical Poisson problem can be viewed as an unsymmetric formulation where the test space is spanned by Dirac-functionals.
- In the pure transport case unsymmetric formulations cannot be avoided. In the convection dominant cases it will be seen later that unsymmetric formulations may be highly beneficial. After all, the underlying equations are unsymmetric.
- Starting with the strong classical formulation of a PDE, it may be an important part of the solution to *find* suitable pairs $\mathbb{U}, \mathbb{V}$ of trial and test spaces for which the weak formulation can be shown to be well-posed or even well-conditioned in a sense to be explained later.
- It is plausible that a stable numerical scheme as a discrete analog to the infinite dimensional problem has a better chance to be stable if the infinite dimensional formulation is well-conditioned or stable.

One may be deterred at the first glance by an abstract formulation like (4.9.7) because it involves a continuum of test conditions. However, such a formulation actually leads in a very natural way to finite dimensional analogs and *numerical schemes.* In fact, given (4.9.7), choose *finite dimensional* subspaces $\mathbb{U}_n \subset \mathbb{U}$, $\mathbb{V}_n \subset$ of equal dimension n, spanned by respective bases

$$\Phi_n = \{\phi_1, \dots, \phi_n\}, \quad \Psi_n = \{\psi_1, \dots, \psi_n\}, \tag{3.1.36}$$

and consider the finite dimensional analog: find $u_n \in \mathbb{U}_n$ such that

$$B(u_n, v) \ = \ f(v), \quad \forall \ v \in \mathbb{V}_n. \tag{3.1.37}$$

Although this looks exactly the same as (4.9.7), it is actually a *linear system of equations.* In fact, making the ansatz

$$u_n = \sum_{k=1}^{n} u_{n,k}\phi_k$$

with unknown coefficient vector $\mathbf{u}_n := (u_{n,k})_{k=1}^n \in \mathbb{R}^n$, plugging the ansatz for u_n into (3.1.37), using the fact that "testing with all $v \in \mathbb{V}$ is equivalent to testing by all *test-basis functions* ψ_i, (3.1.37) takes the form

$$\mathbf{A}_n\mathbf{u}_n = \mathbf{f}_n, \quad \text{where} \quad \mathbf{A}_n := \big(B(\phi_k, \psi_i)\big)_{i,k=1}^n, \ \mathbf{f}_n = \big(f(\psi_i)\big)_{i=1}^n. \quad (3.1.38)$$

Here are some of the issues arising in this context.

- In the above cases the trial spaces $\mathbb{U}_n$ are by construction *contained* in the (infinite dimensional) trial space $\mathbb{U}$ for the infinite dimensional problem. Such a method is called *conforming*. There are actually alternative options that will be discussed later.
- In contrast to finite difference methods the numerical solution is not only defined on a grid but is defined as a function "living" - in the conforming case - already in the function space which hosts also the exact solution. It is therefore natural to estimate errors in the corresponding norms.
- The estimation of these errors as well as the solvability and condition of the linear systems (3.1.37) will therefore, of course, depend strongly on the underlying problem and the governing norms.

3.2 The Inf-Sup-Condition

In this section we characterize the unique solvability of the abstract variational problem (4.9.7). These results apply to the infinite dimensional problem as well as to the finite dimensional counterpart (3.1.37). First, for any two Banach spaces $(\mathbb{X}, \|\cdot\|_{\mathbb{X}}), (\mathbb{Y}, \|\cdot\|_{\mathbb{Y}})$ we denote by $\mathcal{L}(\mathbb{X},\mathbb{Y})$ the space of all bounded linear operators from $\mathbb{X}$ to $\mathbb{Y}$ which is again a normed space under the norm

$$\|\mathcal{B}\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})} := \sup_{v\in\mathbb{X}} \frac{\|\mathcal{B}v\|_{\mathbb{Y}}}{\|v\|_{\mathbb{X}}}.$$

The next theorem clarifies under which circumstances for a bilinear form $B(\cdot,\cdot) : \mathbb{U}\times\mathbb{V} \to \mathbb{R}$ the variational problem

$$B(u, v) \ = \ f(v), \quad \forall\, v \in \mathbb{V}, \qquad (3.2.1)$$

has for any $f \in \mathbb{V}'$ a unique solution $u \in \mathbb{U}$.

Theorem 3.2.1 (Banach-Nečas) *Let $B(\cdot,\cdot) : \mathbb{U} \times \mathbb{V} \to \mathbb{R}$ be a continuous bilinear form, i.e.,*

$$\|B\| \;:=\; \sup_{u\in\mathbb{U}} \sup_{v\in\mathbb{V}} \frac{B(u,v)}{\|u\|_{\mathbb{U}}\,\|v\|_{\mathbb{V}}} < \infty. \tag{3.2.2}$$

Then there exists a unique linear operator $\mathcal{B} \in \mathcal{L}(\mathbb{U},\mathbb{V}')$ such that

$$\langle \mathcal{B}w, v\rangle := (\mathcal{B}w)(v) = B(w,v), \quad \forall\, w \in \mathbb{U},\; v \in \mathbb{V}, \tag{3.2.3}$$

with operator norm

$$\|\mathcal{B}\|_{\mathcal{L}(\mathbb{U},\mathbb{V}')} \;=\; \|B\|. \tag{3.2.4}$$

Moreover, the operator $\mathcal{B} : \mathbb{U} \to \mathbb{V}'$ is a (norm-) isomorphism (i.e., injective and surjective with bounded inverse) if and only if there exists a $c_{\mathcal{B}} > 0$ such that

$$\inf_{w\in\mathbb{U}} \sup_{v\in\mathbb{V}} \frac{B(w,v)}{\|w\|_{\mathbb{U}}\,\|v\|_{\mathbb{V}}} \;\geq\; c_{\mathcal{B}}, \tag{3.2.5}$$

and

$$\forall\, v \in \mathbb{V}, v \neq 0 \;\exists\; w \in \mathbb{U} \text{ such that } B(w,v) \neq 0. \tag{3.2.6}$$

Moreover, one has then

$$\|\mathcal{B}^{-1}\|_{\mathcal{L}(\mathbb{V}',\mathbb{U})} \;\leq\; \frac{1}{c_{\mathcal{B}}}, \tag{3.2.7}$$

i.e., the infs-up constant $c_{\mathcal{B}}$ determines the bound for $\|\mathcal{B}^{-1}\|_{\mathcal{L}(\mathbb{V}',\mathbb{U})}$. □

Proof The proof will consist of several steps.

Existence and Boundedness of $\mathcal{B}$: Fix $w \in \mathbb{U}$. Then

$$F_w : v \in \mathbb{V} \to F_w(v) := B(w,v) \in \mathbb{R}$$

defines a functional on $\mathbb{V}$ which, by bilinearity of $B(\cdot,\cdot)$, is linear. Since by (3.2.2)

$$|F_w(v)| \;=\; |B(w,v)| \leq \|B\|\|w\|_{\mathbb{U}}\|v\|_{\mathbb{V}}, \tag{3.2.8}$$

F_w is for each $w \in \mathbb{U}$ bounded, we have $F_w \in \mathbb{V}'$ with norm

$$\|F_w\|_{\mathcal{L}(\mathbb{V},\mathbb{R})} = \|F_w\|_{\mathbb{V}'} = \sup_{v\in\mathbb{V}} \frac{B(w,v)}{\|v\|_{\mathbb{V}}}.$$

Thus, the mapping $\mathcal{B} : w \in \mathbb{U} \to F_w \in \mathbb{V}'$ is well defined and, because of $F_{w_1+w_2}(v) = B(w_1+w_2, v) = F_{w_1}(v)+F_{w_2}(v)$, obviously linear. Moreover, by (3.2.8),

$$\|\mathcal{B}\|_{\mathcal{L}(\mathbb{U},\mathbb{V}')} = \sup_{w\in\mathbb{U}} \frac{\|\mathcal{B}w\|_{\mathbb{V}'}}{\|w\|_{\mathbb{U}}} = \sup_{w\in\mathbb{U}} \frac{\|F_w\|_{\mathbb{V}'}}{\|w\|_{\mathbb{U}}} = \sup_{w\in\mathbb{U}} \sup_{v\in\mathbb{V}} \frac{B(w,v)}{\|w\|_{\mathbb{U}}\|v\|_{\mathbb{V}}} = \|B\|$$

which shows (3.2.4).

Assume now the validity of (3.2.5) and (3.2.6). We wish to show that $\mathcal{B}$ is an isomorphism satisfying (3.2.7). We proceed in several steps:

Injectivity of $\mathcal{B}$: Condition (3.2.5) says that

$$c_{\mathcal{B}}\|w\|_{\mathbb{U}} \ \le\ \sup_{v\in\mathbb{V}} \frac{(\mathcal{B}w)(v)}{\|v\|_{\mathbb{V}}} = \|\mathcal{B}w\|_{\mathbb{V}'}, \quad \forall\, w \in \mathbb{U}, \tag{3.2.9}$$

which means that $\mathcal{B}$ is *injective.*

$\mathcal{B}$ has closed range: To prove the closedness of $\mathcal{B}(\mathbb{U})$ consider a sequence $(w_k)_{k\in\mathbb{N}} \subset \mathbb{U}$ for which $z_k := \mathcal{B}w_k$ converges in $\mathbb{V}'$ to some $z \in \mathbb{V}'$. We have to show that there exists a $w \in \mathbb{U}$ such that $z = \mathcal{B}w$. By (3.2.9), we have

$$c_{\mathcal{B}}\|w_k - w_j\|_{\mathbb{U}} \le \|\mathcal{B}(w_k - w_j)\|_{\mathbb{V}'} = \|z_k - z_j\|_{\mathbb{V}'} \to 0, \quad k, j \to \infty.$$

Thus, $(w_k)_{k\in\mathbb{N}}$ is a Cauchy sequence which, by completeness of $\mathbb{U}$ has a limit $w \in \mathbb{U}$. Boundedness and hence continuity of $\mathcal{B}$ yields $w = \mathcal{B}z$ which confirms closedness of the range $\mathcal{B}(\mathbb{U})$.

Surjectivity of $\mathcal{B}$: Suppose $\mathcal{B}$ were not surjective, i.e., $\mathcal{B}(\mathbb{U}) \ne \mathbb{V}'$. Since $\mathcal{B}(\mathbb{U})$ is closed we can decompose $\mathbb{V}'$ as $\mathbb{V}' = \mathcal{B}(\mathbb{U}) \oplus \mathcal{B}(\mathbb{U})^\perp$ where orthogonality refers to the scalar product $(\cdot,\cdot)_{\mathbb{V}'}$ in $\mathbb{V}'$, see e.g. [Ha06, Prop. 3.6.9] or [Alt85] or any textbook on Functional Analysis. By assumption $\mathcal{B}(\mathbb{U})^\perp \ne \{0\}$. Hence, there exists a $0 \ne z_0 \in \mathcal{B}(\mathbb{U})^\perp$ such that

$$(z, z_0)_{\mathbb{V}'} = 0, \quad \forall\, z \in \mathcal{B}(\mathbb{U}) \ \Leftrightarrow\ (\mathcal{B}w, z_0)_{\mathbb{V}'} = 0, \quad \forall\, w \in \mathbb{U}.$$

By the Riesz-Representation Theorem, there exists a $v_0 \in \mathbb{V}$ such that $(z, z_0)_{\mathbb{V}'} = z(v_0)$ for all $z \in \mathbb{V}'$. Thus

$$(\mathcal{B}w)(v_0) = B(w, v_0) = 0, \quad \forall\ w \in \mathbb{U},$$

which contradics (3.2.6). Thus $\mathcal{B} : \mathbb{U} \to \mathbb{V}'$ is an isomorphism.

The Bound (3.2.7): We can now rewrite (3.2.9) as

$$c_{\mathcal{B}} \|\mathcal{B}^{-1} z\|_{\mathbb{U}} \le \|z\|_{\mathbb{V}'}, \quad \forall\ z \in \mathbb{V}',$$

which is (3.2.7).

(3.2.7) *implies* (3.2.5)*:* By definition of the operator norm, we have

$$\begin{aligned}
\inf_{w\in\mathbb{U}} \sup_{v\in\mathbb{V}} \frac{B(w,v)}{\|w\|_{\mathbb{U}}\,\|v\|_{\mathbb{V}}} &= \inf_{w\in\mathbb{U}} \sup_{v\in\mathbb{V}} \frac{(\mathcal{B}w)(v)}{\|w\|_{\mathbb{U}}\,\|v\|_{\mathbb{V}}} \\
&= \inf_{w\in\mathbb{U}} \frac{\|\mathcal{B}w\|_{\mathbb{V}'}}{\|w\|_{\mathbb{U}}} \\
&= \inf_{z\in\mathbb{V}'} \frac{\|z\|_{\mathbb{V}'}}{\|\mathcal{B}^{-1}z\|_{\mathbb{U}}} \\
&= \left(\sup_{z\in\mathbb{V}'} \frac{\|\mathcal{B}^{-1}z\|_{\mathbb{U}}}{\|z\|_{\mathbb{V}'}} \right)^{-1} \\
&= \frac{1}{\|\mathcal{B}^{-1}\|_{\mathcal{L}(\mathbb{V}',\mathbb{U})}} \\
&\overset{(3.2.7)}{\ge} c_{\mathcal{B}},
\end{aligned}$$

which confirms the validity of (3.2.5).

Finally, and (3.2.6) is an immediate consequence of the bijectivity of $\mathcal{B}$.$\blacksquare$

Comments:

(i) The first part of Theorem 3.2.1 says that whenever the bilinear form $B(\cdot,\cdot)$ is continuous on $\mathbb{U}\times\mathbb{V}$ then the variational problem (3.2.1) is equivalent to the *operator equation*: for a given $f \in \mathbb{V}'$ find $u \in \mathbb{U}$ satisfying

$$\mathcal{B}u \ = \ f, \tag{3.2.10}$$

where $\mathcal{B} : \mathbb{U} \to \mathbb{V}'$ is *bounded.* Note that in all examples discussed in Section 3.1 the spaces $\mathbb{U}, \mathbb{V}$ (mostly $\mathbb{U} = \mathbb{V}$) were chosen in a way that the respective bilinear forms are indeed continuous. For instance, the discussion in Section 1.3.2 says that the weak formulation (1.3.22) of Poisson's equation can be interpreted as an operator equation for the Laplacian as a mapping from $H_0^1(\Omega)$ onto $H^{-1}(\Omega) = (H_0^1(\Omega))'$.

(ii) The second part of Theorem 3.2.1 characterizes when (3.2.10) has a unique solution for every right hand side. Clearly, unique solvability for *any* right hand side $f \in \mathbb{V}'$ is equivalent to bijectivity of $\mathcal{B}$. The *condition* of the problem can now be bounded by the *condition number*

$$\kappa_{\mathbb{U},\mathbb{V}'}(\mathcal{B}) := \|\mathcal{B}\|_{\mathcal{L}(\mathbb{U},\mathbb{V}')} \|\mathcal{B}^{-1}\|_{\mathcal{L}(\mathbb{V}',\mathbb{U})} \leq \frac{\|B\|}{c_{\mathcal{B}}}. \tag{3.2.11}$$

Hence, the larger $c_{\mathcal{B}}$ the better conditioned is (3.2.10). In particular, one has the *stability relation*

$$\|u\|_{\mathbb{U}} \leq c_{\mathcal{B}}^{-1} \|f\|_{\mathbb{V}'}. \tag{3.2.12}$$

Hence, Theorem 3.2.1 not only ensures unique solvability but also continuous dependence on the data and hence *well-posedness.*

In the finite dimensional case the inf-sup condition has a simple interpretation.

Exercise 3.2.1 *When $\mathbb{U} = \mathbb{V} = \mathbb{R}^d$, $\|\cdot\|_{\mathbb{U}} = \|\cdot\|_{\mathbb{V}} = |\cdot|$ (Euclidean norm), $\mathcal{B} \in \mathbb{R}^{d\times d}$, then*

$$\inf_{w\in\mathbb{R}^d} \sup_{v\in\mathbb{R}^d} \frac{v^T B w}{|w|\,|v|} = \sigma_d(\mathcal{B}), \tag{3.2.13}$$

where $\sigma_d(\mathcal{B})$ is the smallest singular value of $\mathcal{B}$.

The above results are often formulated in the following way.

Theorem 3.2.2 (Nečas, 1962) *Let $B(\cdot,\cdot) : \mathbb{U} \times \mathbb{V} \to \mathbb{R}$ be continuous, where $\mathbb{U}, \mathbb{V}$ are Hilbert spaces. The unique solvability of the variational problem:*

$$B(u, v) = f(v), \quad \forall\, v \in \mathbb{V},$$

for each $f \in \mathbb{V}'$ where u depends continuously on f, is equivalent to each one of the following conditions:

(i) There exists a $c_{\mathcal{B}} > 0$ *such that*

$$\inf_{w\in\mathbb{U}} \sup_{v\in\mathbb{V}} \tfrac{B(w,v)}{\|w\|_{\mathbb{U}}\,\|v\|_{\mathbb{V}}} > c_{\mathcal{B}},$$

for every $v \in \mathbb{V}, v \neq 0$ *there exists a* $w \in \mathbb{U}$ *such that*

$$B(w, v) \neq 0.$$

(ii) There exists an $\alpha > 0$ *such that*

$$\inf_{w\in\mathbb{U}} \sup_{v\in\mathbb{V}} \frac{B(w,v)}{\|w\|_{\mathbb{U}}\,\|v\|_{\mathbb{V}}} \; = \; \inf_{v\in\mathbb{V}} \sup_{w\in\mathbb{U}} \frac{B(w,v)}{\|w\|_{\mathbb{U}}\,\|v\|_{\mathbb{V}}} = \alpha. \quad (3.2.14)$$

□

In addition, for u *satisfying* (3.2.1), *one has*

$$\|u\|_{\mathbb{U}} \leq \alpha^{-1}\|f\|_{\mathbb{V}'}.$$

Proof One only has to verify the equivalence of (i) and (ii). To this end, recall that the adjoint $\mathcal{B}'$ of $\mathcal{B}$ maps $\mathbb{V}$ onto $\mathbb{U}'$ and is defined by

$$(\mathcal{B}w)(v) = \langle \mathcal{B}w, v\rangle = \langle w, \mathcal{B}'v\rangle = (\mathcal{B}'v)(w).$$

It is well-known (see any Functional Analysis textbook) that

$$\|\mathcal{B}\|_{\mathcal{L}(\mathbb{U},\mathbb{V}')} = \|\mathcal{B}'\|_{\mathcal{L}(\mathbb{V},\mathbb{U}')}. \quad (3.2.15)$$

Furthermore, recall from the above proof that

$$\inf_{w\in\mathbb{U}} \sup_{v\in\mathbb{V}} \frac{B(w,v)}{\|w\|_{\mathbb{U}}\,\|v\|_{\mathbb{V}}} = \|\mathcal{B}^{-1}\|^{-1}_{\mathcal{L}(\mathbb{V}',\mathbb{U})}.$$

The assertion follows then from $\|\mathcal{B}\|_{\mathcal{L}(\mathbb{U},\mathbb{V}')} = \|\mathcal{B}'\|_{\mathcal{L}(\mathbb{V},\mathbb{U}')}$. ■

The above theorems will be employed in the following sections to establish well-posedness of variational formulations for both, the infinite as well as finite dimensional problems.

4 Projection Methods

In this chapter we discuss the basic principles of discretization schemes based on the underlying weak formulation of the PDE.

Throughout this section we make the following

Assumption 4.0.3 *Suppose that for a given pair of Hilbert spaces $\mathbb{U}, \mathbb{V}$ and a bilinear form $B(\cdot,\cdot) : \mathbb{U} \times \mathbb{V}$ the problem: for $f \in \mathbb{V}'$ find $u \in \mathbb{U}$ such that*

$$B(u, v) = f(v), \quad v \in \mathbb{V} \tag{4.0.1}$$

is well-posed. That is, the conditions (3.2.2)*,* (3.2.5)*, and* (3.2.6) *hold with constants $C_{\mathcal{B}} = \|B\| < \infty$, $c_{\mathcal{B}} > 0$.*

Remark 4.0.1 Recall that under the above assumption one has

$$\kappa_{\mathbb{U},\mathbb{V}'}(\mathcal{B}) \leq \frac{C_{\mathcal{B}}}{c_{\mathcal{B}}}, \quad (\mathcal{B}w)(v) = B(w, v), \ w|in\mathbb{U}, \, v \in \mathbb{V}.$$

In particular, the *error-residual* relations (1.2.4) hold for $\mathbb{W} = \mathbb{V}'$, i.e.,

$$C_{\mathcal{B}}^{-1}\|f - \mathcal{B}w\|_{\mathbb{V}'} \leq \|u - w\|_{\mathbb{U}} \leq c_{\mathcal{B}}^{-1}\|f - \mathcal{B}w\|_{\mathbb{V}'}, \quad w \in \mathbb{U}, \tag{4.0.2}$$

where $\|f - \mathcal{B}w\|_{\mathbb{V}'} = \sup_{v\in\mathbb{V}} \frac{f(v) - B(w,v)}{\|v\|_{\mathbb{V}}}$. □

4.1 Petrov-Galerkin Scheme

We take up the brief discussion in Section 3.1.7. A natural way of deriving from (4.0.1) a finite-dimensional discrete problem yielding approximate solutions to (4.0.1) looks as follows:

i) Choose finite-dimensional subspaces $\mathbb{U}_n \subset \mathbb{U}$, $\mathbb{V}_h \subset \mathbb{V}$ such that

$$\dim \mathbb{U}_h = \dim \mathbb{V}_h. \tag{4.1.1}$$

ii) compute $u_h \in \mathbb{U}_h$ satisfying

$$B(u_h, v_h) = f(v_h), \quad v_h \in \mathbb{V}_h. \tag{4.1.2}$$

Condition (4.1.1) ensures that (4.1.2) is a quadratic linear system, i.e., we have as many unknows as linear equations. As indicated in Section 3.1.7, computationally (4.1.2) means the following.

Choose bases

$$\Phi_n = \{\phi_1, \dots, \phi_n\}, \quad \Psi_n = \{\psi_1, \dots, \psi_n\}, \quad n = n_h. \tag{4.1.3}$$

Then (4.1.2) is equivalent to

$$B(u_h, \psi_j) = f(\psi_j) =: f_j, \quad j = 1, \dots, n.$$

Substituting the ansatz

$$u_h = \sum_{k=1}^{n} u_{h,k} \phi_k$$

for u_h, bilinearity of $B(\cdot, \cdot)$ yields the linear system

$$\mathbf{A}_h \mathbf{u}_h = \mathbf{f}_n, \tag{4.1.4}$$

with an unknown coefficient vector $\mathbf{u}_h := (u_{h,k})_{k=1}^n \in \mathbb{R}^n$ where

$$\mathbf{A}_h := \big(B(\phi_k, \psi_i)\big)_{i,k=1}^n, \ \mathbf{f}_h = \big(f(\psi_i)\big)_{i=1}^n.$$

In analogy to the infinite-dimensional case

$$(\mathcal{B}_h(w_h)(v_h) = B(w_h, v_h), \quad w_h \in \mathbb{U}_h, \, v_h \in \mathbb{V}_h,$$

defines a linear operator, taking $\mathbb{U}_h \subset \mathbb{U}$ into a subspace of $\mathbb{V}'$. Defining

$$\kappa_{\mathbb{U},\mathbb{V}'}(\mathcal{B}_h) := \frac{\sup_{w_h \in \mathbb{U}_h, \|w_h\|_{\mathbb{U}}=1} \|\mathcal{B}_h w_h\|_{\mathbb{V}'}}{\inf_{w_h \in \mathbb{U}_h, \|w_h\|_{\mathbb{U}}=1} \|\mathcal{B}_h w_h\|_{\mathbb{V}'}}, \tag{4.1.5}$$

(4.1.2) is well-posed if and only if $\kappa_{\mathbb{U},\mathbb{V}'}(\mathcal{B}_h)$ is finite.

Main Questions:

(Q1) Does (4.1.4) (or (4.1.2)) have a unique solution and if so what can be said about $\kappa_{\mathbb{U},\mathbb{V}'}(\mathcal{B}_h)$?

(Q2) How to choose the test space $\mathbb{V}_h$, in particular, so that (Q!) has a positive answer?

(Q3) How to bound the error $\|u - u_h\|_{\mathbb{U}}$ in comparison with the *best approximation* $\inf_{\bar{u}\in\mathbb{U}_h} \|u - \bar{u}\|_{\mathbb{U}}$?

4.2 The Galerkin Case

When the variational formulation (4.0.1) is *symmetric*, i.e., $\mathbb{U} = \mathbb{V}$, and the problem is coercive, the above questions (Q1) - (Q3) have rather simple answers (given already in Numa IV). Taking then $\mathbb{U}_h = \mathbb{V}_h$, i.e., trial- and test-space are *the same*, the discrertization (4.1.2) reduces to the classical *Galerkin* scheme: find $u_u \in \mathbb{U}_h$ such that

$$B(u_h, v_h) = f(v_h), \quad v_h \in \mathbb{U}_h. \tag{4.2.1}$$

The main facts are collected in the following exercise:

Exercise 4.2.1 *Assume that $B(\cdot,\cdot)$ is a continuous and coercive bilinear form on $\mathbb{U}\times\mathbb{U}$ ($\mathbb{U}$ a Hilbert space), i.e.,*

$$|B(w,v)| \le C_{\mathcal{B}}\|w\|_{\mathbb{U}}\|v\|_{\mathbb{U}}, \quad B(w,w) \ge c_{\mathcal{B}}\|w\|_{\mathbb{U}}^2, \quad v,w \in \mathbb{U}. \tag{4.2.2}$$

i) Show that $\mathcal{B} : \mathbb{U} \to \mathbb{U}'$, defined by $(\mathcal{B}w)(v) = B(w,v)$, $w, v \in \mathbb{U}$, has a bounded condition

$$\kappa_{\mathbb{U},\mathbb{U}'}(\mathcal{B}) \le \frac{C_{\mathcal{B}}}{c_{\mathcal{B}}}. \tag{4.2.3}$$

ii) For any subspace $\mathbb{U}_h \subset \mathbb{U}$, the Galerkin scheme (4.2.1) *has a unique solution $u_h \in \mathbb{U}_h$ and the induced discrete operator*

$$(\mathcal{B}_h w_h)(v_h) = f(v_h), \quad w_h, v_h \in \mathbb{U}_h$$

has the same condition bound

$$\kappa_{\mathbb{U},\mathbb{U}'}(\mathcal{B}_h) \le \frac{C_{\mathcal{B}}}{c_{\mathcal{B}}}, \tag{4.2.4}$$

independent of $\mathbb{U}_h$.

iii) (Ceà-Lemma) One has the best approximation property (BAP)

$$\|u - u_h\|_{\mathbb{U}} \leq \frac{C_{\mathcal{B}}}{c_{\mathcal{B}}} \inf_{\bar{u}_h \in \mathbb{U}_h} \|u - \bar{u}_h\|_{\mathbb{U}}. \tag{4.2.5}$$

iv) Show that when $B(\cdot,\cdot)$ is in addition symmetric one even has

$$\|u - u_h\|_{\mathbb{U}} \leq \sqrt{\frac{C_{\mathcal{B}}}{c_{\mathcal{B}}}} \inf_{\bar{u}_h \in \mathbb{U}_h} \|u - \bar{u}_h\|_{\mathbb{U}}. \tag{4.2.6}$$

Remark 4.2.1 i) The condition number estimate in (4.2.4) refers to $\mathcal{B}_h$ as mapping from $\mathbb{U}_h \subset \mathbb{U}$ to a subspace of $\mathbb{U}'$. That does **not** mean that the matrix representation $\mathbf{A}_h$ from (4.1.4) is independent of the discretization. In fact, $\kappa_2(\mathbf{A}_h) = \kappa_{\ell_2(\mathbb{R}^n),\ell_2(\mathbb{R}^n)}(\mathbf{A}_h)$ will typically grow towards infinity when $\dim \mathbb{U}_h$ increases, because $\mathbf{A}_h$ is treated as an operator from $\ell_2(\mathbb{R}^n)$ into the same space $\ell_2(\mathbb{R}^n)$, ignoring the mapping properties of the underlying continuous operator. A preconditioner can only retrieve at best the condition of the infinite-dimensional variational problem!

ii) Under the abobe hypotheses the accuracy of the Galrkin methods is determined by the approximation properties of the trial space $\mathbb{U}_h$ alone, i.e., up to a fixed constant $\kappa_{\mathbb{U},\mathbb{U}'}(\mathcal{B})$ (the condition of the continuous problem) the error behaves like the best approximation error.

iii) The larger $\kappa_{\mathbb{U},\mathbb{U}'}(\mathcal{B})$ the worse gets the error bound. That is, a large condition number of the infinite-dimensional continuous problem directly impedes the accuracy of the Galerkin approximation. Also from this point of view it is important that the infinite-dimensional problem is not only well-posed but also well-conditioned. □

Questions (Q1) - (Q3) for the general Petrov-Galerkin case are discussed next.

4.3 Well-Posedness of Petrov-Galerkin Schemes

The Banach-Nečas-Theorem 3.2.1 applies, of course, in the same way to the pair $\mathbb{U}_h \times \mathbb{V}_h$ (simply restricting the bilinear form $B(\cdot,\cdot)$). However, there is one noteworthy distinction from the Galerkin case for coercive problems.

Remark 4.3.1 Coercivity of the bilinear form $B(\cdot,\cdot)$ is trivially inherited by any subspace $\mathbb{U}_h \subset \mathbb{U}$, i.e., the choice $\mathbb{V}_h = \mathbb{U}_h$ guaranties the same positive inf-sup constants for any subspace $\mathbb{U}_h$. For an indefinite problem with unsymmetric variational formulation ($\mathbb{U} \neq \mathbb{V}$) the situation is fundamnentally different. Given $\mathbb{U}_h \subset \mathbb{U}$ it is, in general, by no means clear which test-space $\mathbb{V}_h$ ensures a positive inf-sup constant. □

Exercise 4.3.1 *Assume that for* $B(\cdot,\cdot)$, $\mathbb{U}$, $\mathbb{V}$, *the conditions* (3.2.2), (3.2.5), *and* (3.2.6) *hold with continuity constant* $\|\mathcal{B}\| = C_{\mathcal{B}}$. *Show that for a given pair of finite-dimensional subspaces* $\mathbb{U}_h \subset \mathbb{U}$, $\mathbb{V}_h \subset \mathbb{V}$

$$\kappa_{\mathbb{U},\mathbb{V}'}(\mathcal{B}_h) \le \frac{C_{\mathcal{B}}}{c_{\mathcal{B}_h}}, \tag{4.3.1}$$

holds if and only if

$$\inf_{w_h \in \mathbb{U}_h} \sup_{v_h \in \mathbb{V}_h} \frac{B(w_h, v_h)}{\|w_h\|_{\mathbb{U}} \, \|v_h\|_{\mathbb{V}}} \ge c_{\mathcal{B}_h} \tag{4.3.2}$$

holds for some postive $c_{\mathcal{B}_h} > 0$. *Moreover, if for a given sequence of pairs* $\mathbb{U}_n \subset \mathbb{U}, \mathbb{V}_n \subset \mathbb{V}$, $n \in \mathbb{N}$, *of trial- and test-spaces the union of the trial spaces* $\mathbb{U}_n$ *is dense in* $\mathbb{U}$, *then one has*

$$\limsup_{n\in\mathbb{N}} c_{\mathcal{B}_n} \le c_{\mathcal{B}}. \tag{4.3.3}$$

Here, the induced finite-dimensional operators $\mathcal{B}_n$ *are again defined by*

$$(\mathcal{B}_n w_h)(v_h) = B(w_n, v_n), \quad w_n \in \mathbb{U}_n, \, v_n \in \mathbb{V}_n.$$

In brief: *A positive answer to (Q1) is equivalent to the validity of* (4.3.2) *for some positive* $c_{\mathcal{B}_h}$.

Remark 4.3.2 Thus, once we have a well-posed infinite-dimensional problem, to ensure well-posedness of a Petrov-Galerkin scheme one only has to re-check one inf-sup condition, rather than all three conditions in Theorem 3.2.1. But in contrast to the Galerkin case for coercive problems, this single inf-sup condition is not automatically satisfied and its validity depends on the choice of the test-space $\mathbb{V}_h$. A systematic choice of a good test-space (see (Q2)) will be discussed in a later section. □

Remark 4.3.3 When writing in what follows $\mathcal{B}_h \in \mathcal{L}(\mathbb{U}, \mathbb{V}')$ this is to express that the operator norm is defined with respect to the norms for $\mathbb{U}$ and $\mathbb{V}'$, i.e., its range $\mathcal{B}_h(\mathbb{U}_h)$ is viewed as a (closed) subspace of $\mathbb{V}'$ endowed with the norm $\|\cdot\|_{\mathbb{V}'}$. □

4.4 BAP of Petrov-Galerkin Solutions

We postpone the choice of $\mathbb{V}_h$ and assume throughout this section that for a given pair $\mathbb{U}_h \subset \mathbb{U}$, $\mathbb{V}_h \subset \mathbb{V}$ the inf-sup condition (4.3.2) is valid.

Exercise 4.4.1 *Assume that under the hypotheses of Exercise 4.3.1 the discrete inf-sup condition* (4.3.2) *holds for the pair* $\mathbb{U}_h \subset \mathbb{U}$, $\mathbb{V}_h \subset \mathbb{V}$*. Let* u, u_h *denote the solutions of* (4.0.1), (4.1.2)*, respectively. Then, the mapping*

$$\Pi_{\mathbb{U}_h,\mathbb{V}_h} : u \to u_h \tag{4.4.1}$$

is a well defined linear projector, i.e., in particular,

$$\Pi_{\mathbb{U}_h,\mathbb{V}_h}(\Pi_{\mathbb{U}_h,\mathbb{V}_h} u) = \Pi_{\mathbb{U}_h,\mathbb{V}_h} u.$$

Hint: Note first that

$$B(\Pi_{\mathbb{U}_h,\mathbb{V}_h} u, v_h) = B(u, v_h), \quad v_h \in \mathbb{V}_h. \tag{4.4.2}$$

Theorem 4.4.1 *When* (4.3.2) *holds the PG-scheme*

$$B(u_h, v_h) = f(v_h), \quad v_h \in \mathbb{V}_h,$$

is stable, i.e.,

$$\|u_h\|_{\mathbb{U}} = \|\Pi_{\mathbb{U}_h,\mathbb{V}_h} u\|_{\mathbb{U}} \leq c_{\mathcal{B}_h}^{-1} \|f\|_{\mathbb{V}'}. \tag{4.4.3}$$

Moreover, one has

$$\|u - \Pi_{\mathbb{U}_h,\mathbb{V}_h} u\|_{\mathbb{U}} \leq \frac{\|\mathcal{B}\|_{\mathcal{L}(\mathbb{U},\mathbb{V}')}}{c_{\mathcal{B}_h}} \inf_{\bar{u}_h \in \mathbb{U}_h} \|u - \bar{u}_h\|_{\mathbb{U}}, \tag{4.4.4}$$

i.e., the Petrov-Galerkin approximation behaves like the best approximation up to a constant factor given by the condition of the discrete operator $\mathcal{B}_h$. □

Proof Abbreviate for convenience $\Pi = \Pi_{\mathbb{U}_h,\mathbb{V}_h}$. By (4.3.2), we have

$$\begin{aligned} c_{\mathcal{B}_h}\|u_h\|_{\mathbb{U}} &\leq \sup_{v_h\in\mathbb{V}_h} \frac{B(u_h,v_h)}{\|v_h\|_{\mathbb{V}}} = \sup_{v_h\in\mathbb{V}_h} \frac{B(\Pi u,v_h)}{\|v_h\|_{\mathbb{V}}} \\ &= \sup_{v_h\in\mathbb{V}_h} \frac{B(u,v_h)}{\|v_h\|_{\mathbb{V}}} \leq \sup_{v\in\mathbb{V}} \frac{B(u,v)}{\|v\|_{\mathbb{V}}} \\ &= \sup_{v\in\mathbb{V}} \frac{f(v)}{\|v_h\|_{\mathbb{V}}} = \|f\|_{\mathbb{V}'}, \end{aligned} \tag{4.4.5}$$

which shows (4.4.3).

By Exercise 4.4.1, $\Pi \in \mathcal{L}(\mathbb{U},\mathbb{U}_h)$ is a projector. By Kato's Theorem (see [Szy06]) one has

$$\|\Pi\|_{\mathcal{L}(\mathbb{U},\mathbb{U})} = \|I-\Pi\|_{\mathcal{L}(\mathbb{U},\mathbb{U})}, \tag{4.4.6}$$

where I denotes the identity. Thus,

$$\begin{aligned} \|u-\Pi u\|_{\mathbb{U}} &= \|(I-\Pi)u\|_{\mathbb{U}} = \|(I-\Pi)(u-\bar{u}_h)\|_{\mathbb{U}} \\ &\leq \|I-\Pi\|_{\mathcal{L}(\mathbb{U},\mathbb{U}_h)}\|u-\bar{u}_h\|_{\mathbb{U}} = \|\Pi\|_{\mathcal{L}(\mathbb{U},\mathbb{U}_h)}\|u-\bar{u}_h\|_{\mathbb{U}} \end{aligned}$$

holds for any $\bar{u}_h \in \mathbb{U}_h$. Therefore, it remains to show that

$$\|\Pi\|_{\mathcal{L}(\mathbb{U},\mathbb{U})} \leq \frac{\|\mathcal{B}\|_{\mathcal{L}(\mathbb{U},\mathbb{V}')}}{c_{\mathcal{B}_h}}. \tag{4.4.7}$$

Noting that

$$\|f\|_{\mathbb{V}'} = \|\mathcal{B}u\|_{\mathbb{V}'} \leq \|\mathcal{B}\|_{\mathcal{L}(\mathbb{U},\mathbb{V}')}\|u\|_{\mathbb{U}},$$

(4.4.7) follows directly from (4.4.5), which completes the proof. ■

Formally, the error bound (4.4.4) looks the same as for the Galerkin case for coercive problems. Again there is an important difference.

Remark 4.4.1 The quantitative performance is again determined by the condition of the discrete problem. In contrast to the Galerkin case the condition of the discrete problem, however, is not determined by the underlying infinite-dimensional problem but depends crucially on the choice of the test-space $\mathbb{V}_h$. In fact, by (4.3.3) one always has for any family of trial-spaces whose union is dense in $\mathbb{U}$ that

$$\limsup_{h\geq 0} \kappa_{\mathbb{U},\mathbb{V}}(\mathcal{B}_h) \geq \kappa_{\mathbb{U},\mathbb{V}}(\mathcal{B}), \tag{4.4.8}$$

and the left hand side may be much larger than the right hand side, depending on the choice of $\mathbb{V}_h$. □

4.5 Optimal Test Spaces

Recall Assumption 4.0.3. Suppose further that $\mathbb{U}_h \subset \mathbb{U}$ is a finite dimensional subspace in which we wish to find an approximation to the exact solution u of (**??**). When $\mathbb{U} = \mathbb{V}$ and $B(\cdot, \cdot)$ is coercive, a Galerkin discretization with $\mathbb{V}_h = \mathbb{U}_h$ as the test-space is the natural way to go.

Central Issue: In general, when $\mathbb{V} \neq \mathbb{U}$, the main question is to choose a "good" test-space $\mathbb{V}_{\mathbb{U}_h} \subset \mathbb{V}$ such that the corresponding Petrov-Galerkin scheme is stable.

As a first step we identify in this section for a given trial space $\mathbb{U}_h \subset \mathbb{U}$ and *optimal test-space* $\mathbb{V}_h(\mathbb{U}_h)$ in the sense that the Petrov-Galerkin scheme

$$B(u_h, v_h) = f(v_h), \quad v_h \in \mathbb{V}_h(\mathbb{U}_h), \tag{4.5.1}$$

inherits the stability properties of the infinite-dimensional problem (**??**).

We will then see that these optimal spaces are not practical but will serve as a starting point for generating practically feasible *near-optimal* test-spaces.

To that end, let $(\cdot, \cdot)_{\mathbb{H}}$ denote the inner product on the Hilbert-space $\mathbb{H}$. An important tool in what follows is the notion of *Riesz-map*

$$\mathcal{R}_{\mathbb{H}} : \mathbb{H}' \to \mathbb{H},$$

defined as follows: for any linear functional $\ell \in \mathbb{H}'$, $z_\ell = \mathcal{R}_{\mathbb{H}}\ell \in \mathbb{H}$ is the unique element in $\mathbb{H}$ such that

$$(z_\ell, v)_{\mathbb{H}} = \ell(v), \quad v \in \mathbb{H}. \tag{4.5.2}$$

Remark 4.5.1 (4.5.2) has a unique solution. Moreover, $\mathcal{R}_{\mathbb{H}}$ is an *isometry*, i.e,,

$$\|\mathcal{R}_{\mathbb{H}}\ell\|_{\mathbb{H}} = \|\ell\|_{\mathbb{H}'}, \quad \ell \in \mathbb{H}', \quad \mathcal{R}_{\mathbb{H}}^{-1} = \mathcal{R}_{\mathbb{H}'}. \tag{4.5.3}$$

This is just the formalization of the Riesz-Representation Theorem, saying that any bounded linear functional on a Hilbert-space is represented by a unique element of the Hilbert space and realized through the scalar-product on $\mathbb{H}$. Computing the Riesz-representer means to solve the elliptic problem (4.5.2). □

Proof The bilinear form $(\cdot,\cdot)_{\mathbb{H}}$ trivially satisfies all conditions in the Banach-Nečas-Theorem 3.2.1 with continuity- and inf-sup constants both equal to one. This proves the claim. (Lax-Milgram already would have worked because a scalar product is triavially coercive and continuous with continuity and coercivity constants both equal to one). ■

Exercise 4.5.1 *Show that one has for any* $v \in \mathbb{H}, \ell, \ell' \in \mathbb{H}'$

$$(\ell, \ell')_{\mathbb{H}'} = \ell(\mathcal{R}_{\mathbb{H}}\ell') = \ell'(\mathcal{R}_{\mathbb{H}}\ell) = (\mathcal{R}_{\mathbb{H}}\ell, \mathcal{R}_{\mathbb{H}}\ell')_{\mathbb{H}}. \tag{4.5.4}$$

The next important notion is the so called *trial-to-test map* $\mathcal{T} : \mathbb{U} \to \mathbb{V}$ associated with the bilinear form $B(\cdot,\cdot)$:

$$(\mathcal{T}w, v)_{\mathbb{V}} = B(w, v), \quad v \in \mathbb{V}. \tag{4.5.5}$$

Remark 4.5.2 $\mathcal{T}$ defined by (4.5.5) is an isomorphism, i.e.,

$$\mathcal{T} \in \mathcal{L}(\mathbb{U}, \mathbb{V}), \quad \mathcal{T}^{-1} \in \mathcal{L}(\mathbb{V}, \mathbb{U}),$$

and

$$\mathcal{T} = \mathcal{R}_{\mathbb{H}} \circ \mathcal{B}. \tag{4.5.6}$$

Proof (4.5.6) follows from $B(w, v) = (\mathcal{B}w)(v) = (\mathcal{R}_{\mathbb{V}}\mathcal{B}w, v)_{\mathbb{V}}$ and Remark 4.5.1. Under the given assumptions on $B(\cdot,\cdot)$ $\mathcal{B} : \mathbb{U} \to \mathbb{V}'$ is an isomorphism. The compositions of isomorphisms is an isomorphism. ■

Remark 4.5.3 The mapping $\mathcal{T}$ is sometimes called "supremizer" because

$$\sup_{v\in\mathbb{V}} \frac{B(w, v)}{\|v\|_{\mathbb{V}}} = \|\mathcal{T}w\|_{\mathbb{V}}, \quad w \in \mathbb{U}. \tag{4.5.7}$$

Moreover, under Assumption 4.0.3, the inf-sup condition combined with (4.5.5) yields

$$\|\mathcal{T}w\|_{\mathbb{V}} \geq c_{\mathcal{B}}\|w\|_{\mathbb{U}}, \quad w \in \mathbb{U}, \tag{4.5.8}$$

while continuity of $\mathcal{B}$ yields

$$\|\mathcal{T}w\|_{\mathbb{V}} \leq C_{\mathcal{B}}\|w\|_{\mathbb{U}}. \tag{4.5.9}$$

Clearly, since $\mathcal{R}_{\mathbb{V}}$ is an isometry, $\mathcal{T}$ has the same condition number as $\mathcal{B}$.□

Proof In fact, one has

$$\sup_{v\in\mathbb{V}} \frac{B(w,v)}{\|v\|_{\mathbb{V}}} \overset{(4.5.5)}{=} \sup_{v\in\mathbb{V}} \frac{(\mathcal{T}w,v)_{\mathbb{V}}}{\|v\|_{\mathbb{V}}} = \|\mathcal{T}w\|_{\mathbb{V}},$$

which yields (4.5.7). (4.5.14) follows from (4.5.7) and Assumption 4.0.3.■

We can describe now the optimal test space associated with a given trial space $\mathbb{U}_h$.

Proposition 4.5.1 *Let Assumption 4.0.3 hold and define the space*

$$\mathbb{V}_h := \mathcal{T}(\mathbb{U}_h) = \{\mathcal{T}w_h : w_h \in \mathbb{U}_h\} \subset \mathbb{V}. \tag{4.5.10}$$

Then, the Petrov-Galerkin problem: find $u_h \in \mathbb{U}_h$ such that

$$B(u_h, v_h) = f(v_h), \quad v_h \in \mathcal{T}(\mathbb{U}_h), \tag{4.5.11}$$

is inf-sup stable, i.e.,

$$\inf_{w_h\in\mathbb{U}_h} \sup_{v_h\in\mathcal{T}(\mathbb{U}_h)} \frac{B(w_h,v_h)}{\|w_h\|_{\mathbb{U}}\|v_h\|_{\mathbb{V}}} \geq c_{\mathcal{B}}, \tag{4.5.12}$$

where $c_{\mathcal{B}}$ is the inf-sup constant of the infinite-dimensional problem (4.5.1). *In particular, it follows from* (4.4.3) *in Theorem 4.4.1 that*

$$\|u_h\|_{\mathbb{U}} \leq c_{\mathcal{B}}^{-1}\|f\|_{\mathbb{V}'}. \tag{4.5.13}$$

Proof Since $\mathcal{T}$ is an isomorphism we have $\dim \mathcal{T}(\mathbb{U}_h) = \dim \mathbb{U}_h$, so that $\mathcal{T}(\mathbb{U}_h)$ is viable. Moreover

$$\sup_{v_h \in \mathcal{T}(\mathbb{U}_h)} \frac{B(w_h, v_h)}{\|v_h\|_{\mathbb{V}}} = \sup_{v_h \in \mathcal{T}(\mathbb{U}_h)} \frac{(\mathcal{T} w_h, v_h)_{\mathbb{V}}}{\|v_h\|_{\mathbb{V}}} \geq \frac{(\mathcal{T} w_h, \mathcal{T} w_h)_{\mathbb{V}}}{\|\mathcal{T} w_h\|_{\mathbb{V}}}$$
$$= \|\mathcal{T} w_h\|_{\mathbb{V}} \overset{(4.5.7)}{=} \sup_{v \in \mathbb{V}} \frac{B(w_h, v)}{\|v\|_{\mathbb{V}}} \geq c_{\mathcal{B}} \|w_h\|_{\mathbb{U}}, \quad (4.5.14)$$

where we have used Assumption 4.0.3 in the last step. This shows (4.5.12). By Theorem 3.2.1, (4.5.11) has a unique solution $u_h \in \mathbb{U}_h$. The rest follows from (4.4.3) in Theorem 4.4.1. ■

Exercise 4.5.2 *Suppose that $\mathbb{S}_h \subset \mathbb{V}$ is a closed subspace and let $P_{\mathbb{S}_h} = P_{\mathbb{V},\mathbb{S}_h}$ denote the $\mathbb{V}$-orthogonal projector into $\mathbb{S}_h$ (i.e., $(v - P_{\mathbb{S}_h} v, v_h)_{\mathbb{V}} = 0$, for all $v_h \in \mathbb{S}_h$). Show that*

$$Q := \mathcal{R}_{\mathbb{V}'} P_{\mathbb{S}_h} \mathcal{R}_{\mathbb{V}}$$

is a $\mathbb{V}'$-orthogonal projector into $\mathcal{R}_{\mathbb{V}'}(\mathbb{S}_h) \subset \mathbb{V}'$.

Can you improve the estimate (4.5.13) somewhat, i.e., which portion of the data f is actually "seen" by the method?

A direct way of expressing (4.5.11) as a linear system is to choose a basis

$$\Phi_h = \{\phi_{h,1}, \ldots, \phi_{h,n}\} \subset \mathbb{U}_h, \quad n = n_h = \dim \mathbb{U}_h,$$

and compute

$$\Psi_h = \{\psi_{h,1}, \ldots, \psi_{h,n}\} \subset \mathcal{T}(\mathbb{U}_h), \quad \psi_{h,k} = \mathcal{T} \phi_{h,k}, \;\; 1 \leq k \leq n. \quad (4.5.15)$$

Thus each test-basis function is the solution of the (infinite-dimensional) variational (elliptic) problem

$$(\psi_{h,k}, v)_{\mathbb{V}} = B(\phi_{h,k}, v), \quad v \in \mathbb{V}. \quad (4.5.16)$$

Then (4.5.11) is equivalent to solving the linear system

$$\mathbf{B}_h \mathbf{u}_h = \mathbf{f}_h, \quad (4.5.17)$$

where $\mathbf{u}_h = (u_{h,1}, \ldots, u_{h,n})^T$ is the unknown coefficient vector of the Petrov-Galerkin solution $u_h = \sum_{k=1}^n u_{h,k} \phi_{h,k}$ and

$$\mathbf{B}_h = \big(B(\phi_{h,i}, \psi_{h,k})\big)_{i,k=1}^n, \quad \mathbf{f}_h = (f(\psi_{h,1}), \ldots, f(\psi_{h,n})^T. \quad (4.5.18)$$

Remark 4.5.4 The system matrix $\mathbf{B}_n$ is symmetric positive definite, regardless of whether the bilinear form $B(\cdot,\cdot)$ is symmetric positive definite or not, because

$$B(\phi_{h,i},\psi_{h,k}) \overset{(4.5.5)}{=} (\mathcal{T}\phi_{h,i},\psi_{h,k})_{\mathbb{V}} \overset{(4.5.15)}{=} (\mathcal{T}\phi_{h,i},\mathcal{T}\phi_{h,k})_{\mathbb{V}}. \qquad (4.5.19)$$

Thus, one can use (preconditioned) conjugate gradient methods for solving (4.5.17). □

Unfortunately, the scheme (4.5.11) is completely impractical for the following reasons:

(I) The computation of each test-basis function requires solving an *infinite-dimensional elliptic* problem (4.5.16).

(II) Even if one replaces (4.5.16) by a finite-dimensional discretized problem, then this problem is in general a global one and of comparable complexity as the original one, i.e., the overall complexity scales at least as $(\mathrm{dim}\mathbb{U}_h)^2$ which is not acceptable.

Nevertheless, the scheme serves as a starting point for the development of practically feasible versions that still exhibit nearly optimal stability.

4.6 Near-Optimal Test Spaces

We address first issue (I) and replace (4.5.16) by a finite-dimensional problem. The principal idea is quite simple, namely one picks

a sufficiently large but finite-dimensional subspace $\mathbb{S}_h \subset \mathbb{V}$, called *test-search space*

(which does **not** play the role of the test space, it will **contain** the test-space).

Since the scalar product $(\cdot,\cdot)_{\mathbb{V}}$ is trivially $\mathbb{V}$-elliptic and $B(w_h,\cdot)$ is a continuous linear functional on $\mathbb{V}$ there exists for every $w_h \in \mathbb{U}_h$ a unique $\mathcal{T}^h w_h \in \mathbb{S}_h$ such that

$$(\mathcal{T}^h w_h, v_h)_{\mathbb{V}} = B(w_h, v_h), \quad v_h \in \mathbb{S}_h. \qquad (4.6.1)$$

Thus, (4.6.1) defines a linear mapping $\mathcal{T}^h := \mathcal{T}_{\mathbb{S}_h} : \mathbb{U}_h \to \mathbb{S}_h$.

Remark 4.6.1 One has

$$\mathcal{T}^h = P_{\mathbb{S}_h} \circ \mathcal{T} = P_{\mathbb{S}_h} \circ \mathcal{R}_{\mathbb{V}} \circ \mathcal{B}. \tag{4.6.2}$$

This allows us to view $\mathcal{T}_{\mathbb{S}_h}$ as a mapping into $\mathbb{V}$.

Proof Defining $\tilde{\mathcal{T}}^h w_h$ by

$$(\tilde{\mathcal{T}}^h w_h, v)_{\mathbb{V}} = B(w_h, P_{\mathbb{S}_h} v), \quad v \in \mathbb{V}, \tag{4.6.3}$$

(where $P_{\mathbb{S}_h} : \mathbb{V} \to \mathbb{S}_h$ is again the $\mathbb{V}$-orthogonal projection) one has by (4.5.5) and self-adjointness of the orthogonal projection

$$B(w_h, P_{\mathbb{S}_h} v) = (\mathcal{T} w_h, P_{\mathbb{S}_h} v)_{\mathbb{V}} = (P_{\mathbb{S}_h} \mathcal{T} w_h, v)_{\mathbb{V}}, \quad v \in \mathbb{V},$$

In particular, this holds for $v \in \mathbb{S}_h \subset \mathbb{V}$ which by (4.6.3) confirms the first relation in (4.6.2). The second relation is just (4.5.6). ■

In a similar fashion as before one can show the following analogue to Remark 4.5.3

Remark 4.6.2 For $\mathcal{T}^h$, defined by (4.6.2), one has for any $w_h \in \mathbb{U}_h$

$$\|\mathcal{T}^h w_h\|_{\mathbb{V}} = \sup_{v_h \in \mathcal{T}^h(\mathbb{U}_h)} \frac{B(w_h, v_h)}{\|v_h\|_{\mathbb{V}}} = \sup_{v_h \in \mathbb{S}_h} \frac{B(w_h, v_h)}{\|v_h\|_{\mathbb{V}}},. \tag{4.6.4}$$

Thus, $\mathcal{T}^h = \mathcal{T}_{\mathbb{S}_h}$ produces the supremizer over $\mathcal{T}^h(\mathbb{U}_h)$ *and* over the larger space $\mathbb{S}_h$. □

Proof As before one argues

$$\begin{aligned}\sup_{v_h \in \mathbb{S}_h} \frac{B(w_h, v_h)}{\|v_h\|_{\mathbb{V}}} &\overset{(4.6.1)}{=} \sup_{v_h \in \mathbb{S}_h} \frac{(\mathcal{T}^h w_h, v_h)_{\mathbb{V}}}{\|v_h\|_{\mathbb{V}}} \overset{\mathcal{T}^h w_h \in \mathbb{S}_h}{=} \|\mathcal{T}^h w_h\|_{\mathbb{V}} \\ &= \sup_{v_h \in \mathcal{T}^h(\mathbb{U}_h)} \frac{(\mathcal{T}^h w_h, v_h)_{\mathbb{V}}}{\|v_h\|_{\mathbb{V}}} \overset{(4.6.1)}{=} \sup_{v_h \in \mathcal{T}^h(\mathbb{U}_h)} \frac{B(w_h, v_h)}{\|v_h\|_{\mathbb{V}}},\end{aligned}$$

which confirms the claim. ■

The idea is now to take $\mathcal{T}^h(\mathbb{U}_h) \subset \mathbb{S}_h \subset \mathbb{V}$ as the test space in the

Petrov-Galerkin (PG) scheme: find $u_h \in \mathbb{U}_h$ such that

$$B(u_h, v_h) = f(v_h), \quad v_h \in \mathcal{T}^h(\mathbb{U}_h) = P_{\mathbb{S}_h}(\mathcal{T}(\mathbb{U}_h)). \tag{PGh}$$

Remark 4.6.3

i) For the test-space to have the same dimension as $\mathbb{U}_h$ one must have at least
$$\dim \mathbb{S}_h \geq \dim \mathbb{U}_h.$$

ii) It is in general not clear beforehand, how to choose the search-test-space $\mathbb{S}_h$. The rationale is that under Assumption 4.0.3, the choice $\mathbb{S}_h = \mathbb{V}$ would ensure *uniform inf-sup stability* for all h. In view of the rightmost equality in (4.6.4), $\mathbb{S}_h$ "large enough" should provide a positive inf-sup constant.

iii) For the computational work to scale favorably in the end, it would be good to ensure though that
$$\dim \mathbb{S}_h \leq C \dim \mathbb{U}_h \tag{4.6.5}$$
for some uniform constant C independent of h. Whether this is possible will have to be seen.

iv) Although $\mathbb{V}_h = \mathcal{T}^h(\mathbb{U}_h) = P_{\mathbb{S}_h}(\mathcal{T}(\mathbb{U}_h))$ is the $\mathbb{V}$-orthogonal projection of the optimal (impractical) test-space $\mathcal{T}(\mathbb{U}_h)$, one does **not** need to know $\mathcal{T}(\mathbb{U}_h)$. Instead one now solves the finite-dimensional problems: find $\psi_{h,k} \in \mathbb{S}_h$ such that
$$(\psi_{h,k}, z_h)_{\mathbb{V}} = B(\phi_{h,k}, z_h), \quad z_h \in \mathbb{S}_h. \quad k = 1, \ldots, n, \tag{4.6.6}$$
to compute the test-basis-functions $\psi_{h,k}$. This requires solving $\dim \mathbb{U}_h$ linear systems of size $\dim \mathbb{S}_h$.

v) The new system matrix
$$\mathbf{B}_h = \big(B(\phi_{h,i}, \psi_{h,k})\big)_{i,k=1}^n = \big((\mathcal{T}^h \phi_{h,i}, \mathcal{T}^h \phi_{h,k})_{\mathbb{V}}\big)_{i,k=1}^n$$
is still symmetric and at least *positive semi-definite.*

vi) $\mathbf{B}_h$ is invertible, i.e., (PGh) has a unique solution, if and only if there exists a $c_{\mathcal{B}_h} > 0$ such that
$$\inf_{w_h \in \mathbb{U}_h} \sup_{v_h \in \mathcal{T}^h(\mathbb{U}_h)} \frac{B(w_h, v_h)}{\|w_h\|_{\mathbb{U}} \|v_h\|_{\mathbb{V}}} \geq c_{\mathcal{B}_h}. \tag{4.6.7}$$

Remark 4.6.4 By continuity of the bilinear form $B(\cdot,\cdot)$ and of the $\mathbb{V}$-orthogon al projection $P_{\mathbb{S}_h}$,

$$(\mathcal{B}_h w_h)(v) := B(w_h, P_{\mathbb{S}_h} v), \quad v \in \mathbb{V}, \tag{4.6.8}$$

is a bounded linear functional over $\mathbb{V}$. Thus, $\mathcal{B}_h : \mathbb{U}_h \to \mathbb{V}'$, defined by (4.6.8) is a bounded linear mapping having a natural extension to $\mathbb{U}$. By Exercise 4.5.2, its range is $\mathcal{R}_{\mathbb{V}'}\mathbb{S}_h \subset \mathbb{V}'$. □

Remark 4.6.5 Since trivially $\|\mathcal{B}_h\|_{\mathcal{L}(\mathbb{U}_h,\mathbb{V}')} \leq \|\mathcal{B}\|_{\mathcal{L}(\mathbb{U},\mathbb{V}')}$ one has

$$\mathcal{B}_h \in \mathcal{L}(\mathbb{U}, \mathbb{V}'), \tag{4.6.9}$$

see Remark 4.3.3. Moreover, for $v_h = P_{\mathbb{S}_h}\mathcal{R}_{\mathbb{V}}\mathcal{B}z_h = \mathcal{T}^h z_h \in \mathcal{T}^h(\mathbb{U}_h)$, (4.6.8) becomes

$$(\mathcal{B}_h w_h)(v_h) = B(w_h, v_h), \quad w_h \in \mathbb{U}_h,\, v_h \in \mathcal{T}^h(\mathbb{U}_h), \tag{4.6.10}$$

which (see also Exercise 4.3.1) is invertible if and only if the inf-sup condition (4.6.7) holds. According to (4.1.5) one then has

$$\kappa_{\mathbb{U},\mathbb{V}'}(\mathcal{B}_h) \leq \frac{C_{\mathcal{B}}}{c_{\mathcal{B}_h}},$$

where $C_{\mathcal{B}}$ is the continuity constant for the infinite-dimensional problem, see Assumption 4.0.3. Thus, if one manages to keep the discrete inf-sup constants $c_{\mathcal{B}_h}$ uniformly bounded away from zero, $c_{\mathcal{B}_h} \geq \beta > 0$, independently of the dimension of $\mathbb{U}_h$, the operators $\mathcal{B}_h$ would have uniformly bounded condition as mappings from $\mathbb{U}$ into $\mathbb{V}$. This does **not** imply that the matrices $\mathbf{B}_h$ have uniformly bounded spectral condition numbers $\kappa_2(\mathbf{B}_h) = \kappa_{\ell_2,\ell_2}(\mathbf{B}_h)$, why? □

The concrete choice of the test-search-space depends on the concrete weak formulation. In this context the following general criterion is often useful.

Definition 4.6.1 Let $\delta \in (0, 1)$. The space $\mathbb{V}^\delta \subset \mathbb{V}$ is called *δ-proximal for* $\mathbb{U}_h \subset \mathbb{U}$ if

$$\|v - P_{\mathbb{V}^\delta} v\|_{\mathbb{V}} \leq \delta \|v\|_{\mathbb{V}}, \quad v \in \mathcal{T}(\mathbb{U}_h). \tag{4.6.11}$$

Thus, δ-proximality of $\mathbb{V}^\delta$ means that all elements of the optimal test-space $\mathcal{T}(\mathbb{U}_h)$ can be approximated by elements in $\mathbb{V}^\delta$ uniformly with *relative* accuracy $\delta < 1$.

The next result says that the Petrov-Galerkin scheme (PGh) with respect to the spaces $\mathbb{U}_h, \mathcal{T}_{\mathbb{S}_h}(\mathbb{U}_h)$ is inf-sup stable if and only if $\mathbb{S}_h$ is δ-proximal for $\mathbb{U}_h$ for some $\delta \in [0,1)$.

Proposition 4.6.1 *If $\mathbb{V}^\delta \subset \mathbb{V}$ is δ-proximal for $\mathbb{U}_h$ then*

$$\inf_{w_h \in \mathbb{U}_h} \sup_{v_h \in \mathbb{V}^\delta} \frac{B(w_h, v_h)}{\|w_h\|_{\mathbb{U}} \|v_h\|_{\mathbb{V}}} \geq c_{\mathcal{B}} \sqrt{1-\delta}, \tag{4.6.12}$$

where $c_{\mathcal{B}}$ is the inf-sup constant of the infinite-dimensional problem (**??**)*.*

Conversely, if for some $\beta > 0$

$$\inf_{w_h \in \mathbb{U}_h} \sup_{v_h \in \mathbb{V}^\delta} \frac{B(w_h, v_h)}{\|w_h\|_{\mathbb{U}} \|v_h\|_{\mathbb{V}}} \geq \beta, \tag{4.6.13}$$

then (4.6.11) *holds with $\delta = \sqrt{1 - (\beta / C_{\mathcal{B}})^2}$, where $C_{\mathcal{B}}$ is the continuity constant in Assumption 4.0.3.* □

Proof Suppose that (4.6.11) holds. By orthogonality (4.6.2), we have

$$\begin{aligned} \|\mathcal{T}^h w_h\|_{\mathbb{V}}^2 &= \|\mathcal{T} w_h\|_{\mathbb{V}}^2 - \|(I - P_{\mathbb{S}_h})\mathcal{T} w_h\|_{\mathbb{V}}^2 \\ &\overset{(4.6.2)}{\geq} \|\mathcal{T} w_h\|_{\mathbb{V}}^2 (1-\delta^2) \\ &\overset{(4.5.14)}{\geq} c_{\mathcal{B}}^2 (1-\delta^2) \|w_h\|_{\mathbb{U}}^2. \end{aligned} \tag{4.6.14}$$

Hence

$$\sup_{v_h \in \mathbb{V}^\delta} \frac{B(w_h, v_h)}{\|v_h\|_{\mathbb{V}}} \overset{(4.6.4)}{=} \|\mathcal{T}^h w_h\|_{\mathbb{V}} \overset{(4.6.14)}{\geq} c_{\mathcal{B}} (1-\delta^2)^{1/2} \|w_h\|_{\mathbb{U}}, \tag{4.6.15}$$

confirming (4.6.12).

Conversely, by (4.5.6) and boundedness of $\mathcal{B}$ (see Assumption 4.0.3), we have

$$\|\mathcal{T} w_h\|_{\mathcal{L}(\mathbb{U},\mathbb{V})} \leq \|\mathcal{B} w_h\|_{\mathcal{L}(\mathbb{U},\mathbb{V}')} \leq C_{\mathcal{B}} \|w_h\|_{\mathbb{U}}, \quad w_h \in \mathbb{U}_h. \tag{4.6.16}$$

Moreover, combining (4.6.13) and (4.6.4) yields

$$\|\mathcal{T}^h w_h\|_{\mathbb{V}} \geq \beta \|w_h\|_{\mathbb{U}}, \quad w_h \in \mathbb{U}_h. \tag{4.6.17}$$

Now, by the first line in (4.6.14), we have, ,

$$\begin{aligned}\|(I - P_{\mathbb{S}_h})\mathcal{T} w_h\|_{\mathbb{V}}^2 &= \|\mathcal{T} w_h\|_{\mathbb{V}}^2 - \|\mathcal{T}^h w_h\|_{\mathbb{V}}^2 \\ &\overset{(4.6.17)}{\leq} \|\mathcal{T} w_h\|_{\mathbb{V}}^2 - \beta^2 \|w_h\|_{\mathbb{U}}^2 \\ &= \left\{1 - \beta^2 \Big(\frac{\|w_h\|_{\mathbb{U}}}{\|\mathcal{T} w_h\|_{\mathbb{V}}}\Big)^2\right\} \|\mathcal{T} w_h\|_{\mathbb{V}}^2 \\ &\overset{(4.6.16)}{\leq} \left\{1 - \Big(\frac{\beta}{C_{\mathcal{B}}}\Big)^2\right\} \|\mathcal{T} w_h\|_{\mathbb{V}}^2,\end{aligned}$$

which completes the proof. ■

Remark 4.6.6 If $\mathbb{V}^\delta \subset \mathbb{V}$ is δ-proximal for $\mathbb{U}_h \subset \mathbb{U}$, taking $\mathbb{V}^\delta = \mathbb{S}_h$ as the search-test-space, the corresponding PG-scheme with test-space $\mathcal{T}_{\mathbb{V}^\delta}(\mathbb{U}_h)$ is well posed and, by Remark 4.6.2, one has

$$\inf_{w_h \in \mathbb{U}_h} \sup_{v_h \in \mathcal{T}_{\mathbb{V}^\delta}(\mathbb{U}_h)} \frac{B(w_h, v_h)}{\|w_h\|_{\mathbb{U}} \|v_h\|_{\mathbb{V}}} \geq c_{\mathcal{B}} \sqrt{1 - \delta}.$$

Remark 4.6.7 When $\mathcal{B} : \mathbb{U} \to \mathbb{V}'$ is an isometry, i.e., $c_{\mathcal{B}} = C_{\mathcal{B}} = 1$, inf-sup stability with constant $\sqrt{1 - \delta^2}$ and δ-proximality are equivalent. □

In general, Proposition 4.6.1 says that inf-sup stability of a Petrov-Galerkin formulation is equivalent to the approximability of the optimal test-space by the search-test space within some uniform *relative accuracy* strictly less than one. This is best explained by the following general setting.

Exercise 4.6.1 *Let $\mathbb{X}, \mathbb{Y}$ be finite-dimensional subspaces of a Hilbert space $\mathbb{H}$ with inner product $(\cdot, \cdot)$ and norm $\|\cdot\|^2 = (\cdot, \cdot)$. Define*

$$\beta(\mathbb{X}, \mathbb{Y}) := \inf_{v \in \mathbb{X}} \sup_{z \in \mathbb{Y}} \frac{(v, z)_{\mathbb{H}}}{\|v\|_{\mathbb{H}} \|z\|_{\mathbb{H}}}.$$

Prove the following properties:

i) Let $P_{\mathbb{Y}}$ denote the $\mathbb{H}$-orthogonal projector onto $\mathbb{Y}$, then

$$\beta(\mathbb{X},\mathbb{Y}) = \inf_{v\in\mathbb{X}} \frac{\|P_{\mathbb{Y}}v\|}{\|v\|}.$$

ii) One has

$$\sup_{0\neq v\in\mathbb{X}} \inf_{z\in\mathbb{Y}} \frac{\|v-z\|}{\|v\|} \leq \sqrt{1-\beta(\mathbb{X},\mathbb{Y})^2}.$$

iii) Let $\Phi = \{x_1,\dots,x_n\}$, $\quad \Psi = \{y_1,\dots,y_m\}$ be bases of $\mathbb{X},\mathbb{Y}$ (formally viewed as column vectors), respectively and let

$$\mathbf{G} := (\Phi, \Psi^T) := \big((x_i, y_k)\big)_{i=1,k=1}^{n,m}$$

be the corresponding cross-Gramian *of the two bases. Then*

$$\beta(\mathbb{X},\mathbb{Y}) = \sigma_{\min}(\mathbf{G})$$

is the smallest singular value of $\mathbf{G} := (\Phi, \Psi^T)$.

iv) Interpret $\beta(\mathbb{X},\mathbb{Y})$ geometrically in terms of angles.

Exercise 4.6.2 (Fortin's Criterion) *Assume that there exists a projector $\Pi_h : \mathbb{V} \to \mathbb{S}_h$ such that*

$$B(w_h, \Pi_h v) = B(w_h, v), \quad v \in \mathbb{V}.$$

Show that then

$$\inf_{w_h\in\mathbb{U}_h} \sup_{v_h\in\mathbb{S}_h} \frac{B(w_h, v_h)}{\|w_h\|_{\mathbb{U}}\|v_h\|_{\mathbb{V}}} \geq \frac{c_{\mathcal{B}}}{\|\Pi_h\|_{\mathcal{L}(\mathbb{V},\mathbb{V})}}.$$

So far we have only acquired some information concerning issue (I), namely what is relevant when replacing an infinite-dimensional variational problem for the computation of the optimal test-functions by finite-dimensional problems. However, as it stands, this would still require solving for each basis function in $\mathbb{U}_h$ a linear problem of size $\dim \mathbb{S}_h \geq \dim \mathbb{U}_h$, see Remark 4.6.3, (i). So, even if one manages the validity of Remark 4.6.3, (ii), the total work needed to compute the solution of the Galerkin scheme (PGh) scales at best as $(\dim \mathbb{U}_h)^2$, which for large scale problems is not acceptable. Therefore, we discuss next *principal strategies* for coping with this essential obstruction.

4.7 A Minimum Residual Scheme

A key step is to reinterpret the Petrov-Galerkin (PG) scheme (PGh) as a *least squares* problem in $\mathbb{V}'$.

Theorem 4.7.1 *Assume that* (4.6.7) *holds for the test-search-sspace* $\mathbb{S}_h$. *Then, the* $u_h \in \mathbb{U}_h$ *solves*

$$B(u_h, v_h) = f(v_h), \quad v_h \in \mathcal{T}_{\mathbb{S}_h}(\mathbb{U}_h) = P_{\mathbb{S}_h}(\mathcal{T}(\mathbb{U}_h)), \tag{4.7.1}$$

if and only if

$$u_h = \operatorname*{argmin}_{w_h \in \mathbb{U}_h} \left\{ \sup_{v_h \in \mathbb{S}_h} \frac{f(v_h) - B(w_h, v_h)}{\|v_h\|_{\mathbb{V}}} \right\}. \tag{4.7.2}$$

Proof Rewrite

$$\begin{aligned} \sup_{v_h \in \mathbb{S}_h} \frac{f(v_h) - B(w_h, v_h)}{\|v_h\|_{\mathbb{V}}} \sup_{v_h \in \mathbb{S}_h} & \frac{(P_{\mathbb{S}_h} \mathcal{R}_{\mathbb{V}} f, v_h)_{\mathbb{V}} - (\mathcal{T}^h w_h, v_h)_{\mathbb{V}}}{\|v_h\|_{\mathbb{V}}} \\ &= \|P_{\mathbb{S}_h} \mathcal{R}_{\mathbb{V}} f - \mathcal{T}^h w_h\|_{\mathbb{V}}. \end{aligned} \tag{4.7.3}$$

Next note that minimizing the quadratic functional $\|P_{\mathbb{S}_h} \mathcal{R}_{\mathbb{V}} f - \mathcal{T}^h w_h\|_{\mathbb{V}}^2$ over $\mathbb{U}_h$ is equivalent to minimizing

$$J(w_h) := \frac{1}{2}(\mathcal{T}^h w_h, \mathcal{T}^h w_h)_{\mathbb{V}} - (P_{\mathbb{S}_h} \mathcal{R}_{\mathbb{V}} f, \mathcal{T}^h w_h)_{\mathbb{V}}$$

over $w_h \in \mathbb{U}_h$. Since J is a quadratic functional and hence strictly convex we need to find the critical points, i.e., the zeroes of its Frechét derivative. To that end, note that for any $\bar{u}_h, w_h \in \mathbb{U}_h$

$$\begin{aligned} (DJ)(\bar{u}_h)(w_h) &:= \lim_{t \to 0} \frac{1}{t}\big(J(\bar{u}_h + t w_h) - J(\bar{u}_h)\big) \\ &= \lim_{t \to 0} \big\{(\mathcal{T}^h \bar{u}_h, \mathcal{T}^h w_h)_{\mathbb{V}} - (P_{\mathbb{S}_h} \mathcal{R}_{\mathbb{V}} f, \mathcal{T}^h w_h)_{\mathbb{V}} + t\|\mathcal{T}^h w_h\|_{\mathbb{V}}^2\big\} \\ &= (\mathcal{T}^h \bar{u}_h, \mathcal{T}^h w_h)_{\mathbb{V}} - (P_{\mathbb{S}_h} \mathcal{R}_{\mathbb{V}} f, \mathcal{T}^h w_h)_{\mathbb{V}}. \end{aligned} \tag{4.7.4}$$

By Exercise (4.7.1), $DJ(u_h)(w_h) = 0$ for all $w_h \in \mathbb{U}_h$ holds if and only if u_h solves (4.7.1). The assertion follows now from Remark 4.6.3. ∎

Remark 4.7.1 (i) Note that, on the one hand

$$\sup_{v\in\mathbb{V}} \frac{f(P_{\mathbb{S}_h}v) - B(w_h, P_{\mathbb{S}_h}v)}{\|v\|_{\mathbb{V}}} = \sup_{v\in\mathbb{V}} \frac{(P_{\mathbb{S}_h}\mathcal{R}_{\mathbb{V}}f - \mathcal{T}^h w_h, v)_{\mathbb{V}}}{\|v\|_{\mathbb{V}}} \tag{4.7.5}$$

$$= \|P_{\mathbb{S}_h}\mathcal{R}_{\mathbb{V}}f - \mathcal{T}^h w_h\|_{\mathbb{V}}. \tag{4.7.6}$$

On the other hand,

$$\sup_{v\in\mathbb{V}} \frac{f(P_{\mathbb{S}_h}v) - B(w_h, P_{\mathbb{S}_h}v)}{\|v\|_{\mathbb{V}}} = \|Q_h f - \mathcal{B}_h w_h\|_{\mathbb{V}'} \tag{4.7.7}$$

In view of (4.7.3) the solution u_h of (4.7.1) is also given by

$$u_h = \underset{w_h\in\mathbb{U}_h}{\operatorname{argmin}} \|Q_h f - \mathcal{B}_h w_h\|_{\mathbb{V}'}. \tag{4.7.8}$$

(ii) Thus the condition

$$0 = (\mathcal{T}^h \bar{u}_h, \mathcal{T}^h w_h)_{\mathbb{V}} - (P_{\mathbb{S}_h}\mathcal{R}_{\mathbb{V}}f, \mathcal{T}^h w_h)_{\mathbb{V}} \tag{4.7.9}$$

are just the *normal equations* for the least squares problem (4.7.8).

Exercise 4.7.1 *(i) Show that* (4.7.1) *is equivalent to*

$$(\mathcal{T}_{\mathbb{S}_h}u_h, \mathcal{T}_{\mathbb{S}_h}w_h)_{\mathbb{V}} = f(\mathcal{T}_{\mathbb{S}_h}w_h), \quad w_h \in \mathbb{U}_h. \tag{4.7.10}$$

(ii) What does this mean for the PG system matrix?

(iii) Formulate the least-squares problem which is equivalent to the optimal PG-scheme (4.5.11)

$$B(u_h, v_h) = f(v_h), \quad v_h \in \mathcal{T}(\mathbb{U}_h).$$

4.8 A Mixed Formulation

The above interpretation of the PG-scheme (PGh) as a *minimum residual* (least-squares) problem provides the basis for yet a third reinterpretation as a *saddle-point* problem which avoids the necessity of computing the Petrov-Galerkin test-functions explicitly, thereby addressing obstruction (II) at the end of Section 4.5. The main result reads as follows.

Theorem 4.8.1 *Assume that* $\mathbb{S}_h \subset \mathbb{V}$ *is* δ*-proximal for* $\mathbb{U}_h \subset \mathbb{U}$ *for some* $\delta \in (0,1)$ *(see* (4.6.1)*). Then* $u_h \in \mathbb{U}_h$ *is the unique solution of the PG-scheme* (4.7.1) *(see also* (PGh)*) if and only if* u_h *solves the problem: find* $(u_h, r_h) \in \mathbb{U}_h \times \mathbb{S}_h$ *such that for* $f \in \mathbb{V}'$

$$\begin{array}{lcll} (r_h, v_h)_{\mathbb{V}} + B(u_h, v_h) & = & f(v_h), & v_h \in \mathbb{S}_h, \\ B(w_h, r_h) & = & 0, & w_h \in \mathbb{U}_h. \end{array} \tag{4.8.1}$$

Proof By Proposition 4.6.1, δ-proximality of $\mathbb{S}_h$ for $\mathbb{U}_h$ is equivalent to the existence of a positive inf-sup constant $c_{\mathcal{B}_h}$ for the PG-scheme (PGh). As shown above, (4.7.9) is equivalent to (PGh). The key idea is to introduce an *auxiliary* unknown representing the *projected Riesz-lifted residual.* Specifically, for $u_h \in \mathbb{U}_h$ there exists a unique $r_h \in \mathbb{S}_h$ such that

$$(r_h, v_h)_{\mathbb{V}} = (P_{\mathbb{S}_h}(\mathcal{T} u_h - \mathcal{R}_{\mathbb{V}} f), v_h)_{\mathbb{V}}, \quad v_h \in \mathbb{S}_h. \tag{4.8.2}$$

This can be rewritten as

$$(r_h, v_h)_{\mathbb{V}} + (P_{\mathbb{S}_h}\mathcal{T} u_h, v_h)_{\mathbb{V}} = (P_{\mathbb{S}_h}\mathcal{R}_{\mathbb{V}} f, v_h)_{\mathbb{V}}, \quad v_h \in \mathbb{S}_h. \tag{4.8.3}$$

Noting that for $v_h \in \mathbb{S}_h$

$$\begin{aligned} (P_{\mathbb{S}_h}\mathcal{T} u_h, v_h)_{\mathbb{V}} &= (\mathcal{T} u_h, P_{\mathbb{S}_h} v_h)_{\mathbb{V}} = (\mathcal{T} u_h, v_h)_{\mathbb{V}} \overset{(4.6.1)}{=} B(u_h, v_h) \\ (P_{\mathbb{S}_h}\mathcal{R}_{\mathbb{V}} f, v_h)_{\mathbb{V}} &= (\mathcal{R}_{\mathbb{V}} f, P_{\mathbb{S}_h} v_h)_{\mathbb{V}} = (\mathcal{R}_{\mathbb{V}} f, v_h)_{\mathbb{V}} = f(v_h), \end{aligned}$$

(4.8.11) is equivalent to the first line in (4.8.1). By (4.6.1) we have

$$\begin{aligned} B(w_h, r_h) &= (\mathcal{T}^h w_h, r_h)_{\mathbb{V}} = (\mathcal{T}^h w_h, P_{\mathbb{S}_h}(\mathcal{R}_{\mathbb{V}} f - \mathcal{T} u_h))_{\mathbb{V}} \\ &= f(\mathcal{T}^h w_h) - B(u_h, \mathcal{T}^h w_h). \end{aligned}$$

Thus, the second line of (4.8.1) is equivalent to the validity of the PG-conditions (PGh) whenever r_h satisfies the first line of (4.8.1). This finishes the proof. ■

Remark 4.8.1 The idea behind Theorem 4.8.1 is simple: one introduces an *auxiliary* unknown r_h as the *projected lifted residual.* The defining relation is the first line in (4.8.1). The second line says that this lifted projected residual vanishes under testing by $\mathbb{S}_h$ which are the normal equations (4.7.9) and hence the PG-relations (PGh). The term “mixed formulation” stems from the introduction of the auxiliary variable as in the mixed formulation of second order elliptic problems where the fluxes are the auxiliary variables. □

Remark 4.8.2 (1) The advantage of the reformulation of (PGh) as the saddle-point problem (4.8.1) is that all involved spaces $\mathbb{U}_h \subset \mathbb{U}, \mathbb{S}_h \subset \mathbb{V}$ can be chosen directly, for instance, as finite element spaces. Thus, a numerical treatment of (4.8.1) can be based on standard finite element techniques for saddle-point problems. In particular, one does not have to compute a basis for $\mathcal{T}^h(\mathbb{U}_h)$ which requires solving $\dim \mathbb{U}_h$ variational problems of size $\dim \mathbb{S}_h \geq \dim \mathbb{U}_h$. Instead, one has traded these $\dim \mathbb{S}_h$ solves against a single variational problem (4.8.2) for the projected lifted residual.

(2) The disadvantage of the formulation (4.8.1) is that the finite-dimensional problem is now larger, due to the additional unknown r_h. □

Exercise 4.8.1 *Formulate a saddle-point problem which is equivalent to the PG-scheme* (4.5.11) *with the optimal test-space* $\mathcal{T}(\mathbb{U}_h)$*.*

Stability of (4.8.1)**:** The problem (4.8.1) is formally a different variational problem. We have shown that the solution u_h of the PG-scheme is a solution component of (4.8.1). It remains to show that (4.8.1) is also stable. In principle, one can apply Brezzi's theory for saddle-point problems to assert that (4.8.1) is indeed stable if $B(\cdot,\cdot)$ is inf-sup stable over $\mathbb{U}_h \times \mathcal{T}^h(\mathbb{U}_h)$. In fact, it requires that the bilinear form in the upper left corner of (4.8.1) and if $B(\cdot,\cdot)$ satisfies an inf-sup condition.

One can also argue directly as will be sketched next. Suppose $\mathbb{S}_h$ is δ-proximal for $\mathbb{U}_h$. Consider the bilinear form

$$A([r_h, u_h], [v_h, w_h]) := (r_h, v_h)_{\mathbb{V}} + B(u_h, v_h) + B(w_h, r_h), \qquad (4.8.4)$$

i.e., A is symmetric and, defining $\|[v_h, w_h]\|_{\mathbb{H}} := \left(\|v_h\|_{\mathbb{V}}^2 + \|w_h\|_{\mathbb{U}}^2\right)^{1/2}$,

$$A(\cdot,\cdot) : (\mathbb{H} := \mathbb{S}_h \times \mathbb{U}_h) \times \mathbb{H} \to \mathbb{R}. \qquad (4.8.5)$$

Continuity of $A(\cdot,\cdot)$: Notice first that $A(\cdot,\cdot)$ is continuous. In fact,

$$\begin{aligned}|A([r_h,u_h],[v_h,w_h])| &\leq \|r_h\|_{\mathbb{V}}\|v_h\|_{\mathbb{V}} + C_{\mathcal{B}}\big(\|u_h\|_{\mathbb{U}}\|v_h\|_{\mathbb{V}} + \|w_h\|_{\mathbb{U}}\|r_h\|_{\mathbb{V}}\big)\\ &\leq \big(\|r_h\|_{\mathbb{V}}^2 + \|u_h\|_{\mathbb{U}}^2\big)^{1/2}\big\{(\|v_h\|_{\mathbb{V}} + C_{\mathcal{B}}\|w_h\|_{\mathbb{U}})^2 + C_{\mathcal{B}}^2\|v_h\|_{\mathbb{V}}^2\big\}^{1/2}\\ &\leq \|[r_h,u_h]\|_{\mathbb{H}}\big\{(1+C_{\mathcal{B}}+C_{\mathcal{B}}^2)\|v_h\|_{\mathbb{V}}^2 + (C_{\mathcal{B}}+C_{\mathcal{B}}^2)\|w_h\|_{\mathbb{U}}^2\big\}^{1/2}\\ &\leq (1+C_{\mathcal{B}}+C_{\mathcal{B}}^2)^{1/2}\|[r_h,u_h]\|_{\mathbb{H}}\|[v_h,w_h]\|_{\mathbb{H}},\end{aligned}$$

which says that $A(\cdot,\cdot)$ is continuous with a continuity constant depending on $C_{\mathcal{B}}$.

Inf-sup-stability: Next, we verify an inf-sup condition. We do this by an educated guess for the test-function $[v_h,w_h]$ given $[r_h,u_h]$. Specifically, let

$$g_h := (I - P_{\mathcal{T}^h(\mathbb{U}_h)})r_h, \quad v_h := \mathcal{T}^h u_h + g_h, \tag{4.8.6}$$

and

$$w_h := z_h - u_h, \text{ where } z_h \in \mathbb{U}_h \text{ such that } \mathcal{T}^h(z_h) = P_{\mathcal{T}^h(\mathbb{U}_h)}r_h. \tag{4.8.7}$$

Then, since $(\mathcal{T}^h u_h, g_h)_{\mathbb{V}} = 0$, for $[v_h,w_h]$ as in (4.8.6), (4.8.7) we have

$$\begin{aligned}A([r_h,u_h],[v_h,w_h]) &= (r_h, \mathcal{T}^h u_h + g_h)_{\mathbb{V}} + B(u_h, \mathcal{T}^h u_h + g_h) + B(z_h - u_h, r_h)\\ &= (r_h, \mathcal{T}^h u_h + (I - P_{\mathcal{T}^h(\mathbb{U}_h)})r_h)_{\mathbb{V}} + (\mathcal{T}^h u_h, \mathcal{T}^h u_h + g_h)_{\mathbb{V}}\\ &\quad + (\mathcal{T}^h(z_h - u_h), r_h)_{\mathbb{V}}\\ &= (r_h, \mathcal{T}^h z_h + (I - P_{\mathcal{T}^h(\mathbb{U}_h)})r_h)_{\mathbb{V}} + (\mathcal{T}^h u_h, \mathcal{T}^h u_h)_{\mathbb{V}}\\ &= (r_h, P_{\mathcal{T}^h(\mathbb{U}_h)}r_h + (I - P_{\mathcal{T}^h(\mathbb{U}_h)})r_h)_{\mathbb{V}} + \|\mathcal{T}^h u_h\|_{\mathbb{V}}^2\\ &= \|r_h\|_{\mathbb{V}}^2 + \|\mathcal{T}^h u_h\|_{\mathbb{V}}^2.\end{aligned} \tag{4.8.8}$$

On the other hand,

$$\begin{aligned}\|v_h\|_{\mathbb{V}}^2 + \|w_h\|_{\mathbb{U}}^2 &\leq \|\mathcal{T}^h u_h\|_{\mathbb{V}}^2 + \|g_h\|_{\mathbb{V}}^2 + \|z_h - u_h\|_{\mathbb{U}}^2\\ &\leq \|\mathcal{T}^h u_h\|_{\mathbb{V}}^2 + \|r_h\|_{\mathbb{V}}^2 + (\|z_h\|_{\mathbb{V}} + \|u_h\|_{\mathbb{U}})^2\\ &\leq (1 + c_{\mathcal{B}_h}^{-2})\big(\|r_h\|_{\mathbb{V}}^2 + \|\mathcal{T}^h u_h\|_{\mathbb{V}}^2\big),\end{aligned} \tag{4.8.9}$$

where we have used

$$\|\mathcal{T}^h z_h\|_{\mathbb{V}} \overset{(4.6.17)}{\geq} c_{\mathcal{B}_h}\|z_h\|_{\mathbb{U}} \;\Rightarrow\; \|z_h\|_{\mathbb{U}} \leq c_{\mathcal{B}_h}^{-1}\|\mathcal{T}^h z_h\|_{\mathbb{V}} \overset{(4.8.7)}{\leq} c_{\mathcal{B}_h}^{-1}\|r_h\|_{\mathbb{V}},$$

and $\|u_h\|_{\mathbb{U}} \le c_{\mathcal{B}_h}^{-1} \|\mathcal{T}^h u_h\|_{\mathbb{V}}$. Hence, (4.8.8) and (4.8.9) provide

$$\begin{aligned} \frac{A([r_h, u_h], [v_h, w_h])}{\|[v_h, w_h]\|_{\mathbb{V}}} &\ge (1 + c_{\mathcal{B}_h}^{-2})^{-1/2} \big(\|r_h\|_{\mathbb{V}}^2 + \|\mathcal{T}^h u_h\|_{\mathbb{V}}^2 \big)^{1/2} \\ &\ge c_{\mathcal{B}_h} (1 + c_{\mathcal{B}_h}^{-2})^{-1/2} \|[r_h, u_h]\|_{\mathbb{H}}. \end{aligned} \tag{4.8.10}$$

Thus, when the PG-inf-sup constants stay bounded away from zero, uniformly in h, so do the inf-sup constants of the mixed formulation.

The Projected Lifted Residual: While (4.8.1) can be treated by standard techniques, it involves an additional unknown r_h and thus increases the discrete problem. However, the additional unknown r_h provides important information which will be discussed next.

Recall that the solution component $r_h = r_h(u_h, f)$ of (4.8.1) is defined by

$$(r_h(u_h, f), v_h)_{\mathbb{V}} = (P_{\mathbb{S}_h}(\mathcal{T} u_h - \mathcal{R}_{\mathbb{V}} f), v_h)_{\mathbb{V}}, \quad v_h \in \mathbb{S}_h, \tag{4.8.11}$$

which, by self-adjointness of $P_{\mathbb{S}_h}$ can be restated as

$$(r_h(u_h, f), v)_{\mathbb{V}} = (P_{\mathbb{S}_h}(\mathcal{T} u_h - \mathcal{R}_{\mathbb{V}} f), v)_{\mathbb{V}}, \quad v \in \mathbb{V}, \tag{4.8.12}$$

i.e.,

$$r_h(u_h, f) = P_{\mathbb{S}_h}(\mathcal{T} u_h - \mathcal{R}_{\mathbb{V}} f) \in \mathbb{S}_h \subset \mathbb{V}, \tag{4.8.13}$$

is the *projected lifted residual* of the PG-solution u_h of (PGh), while $r(u_h, f)$ defined by

$$(r, v)_{\mathbb{V}} = (\mathcal{T} u_h - \mathcal{R}_{\mathbb{V}} f, v)_{\mathbb{V}}, \quad v \in \mathbb{V}, \tag{4.8.14}$$

defines the corresponding *lifted "full" residual.* More generally, $r_h(w_h, f), r(w_h, f)$, defined in the same way are the projected, respectively full lifted residual for any $w_h \in \mathbb{U}_h$. On account of (4.5.3), one has $\|r(u_h, f)\|_{\mathbb{V}} = \|f - \mathcal{B} u_h\|_{\mathbb{V}'}$ which, in view of (4.0.2) gives

$$C_{\mathcal{B}}^{-1} \|r(u_h, f)\|_{\mathbb{V}} \le \|u - u_h\|_{\mathbb{U}} \le c_{\mathcal{B}}^{-1} \|r(u_h, f)\|_{\mathbb{V}}. \tag{4.8.15}$$

The full residual requires a maximization over all of $\mathbb{V}$, respectively, the solution of an infinite variational problem (4.5.5) and hence cannot be computed exactly.

Question: *Can one use* $r_h = r_h(u_h, f)$ *defined by* (4.8.12)*, as an* error indicator/bound *for* $\|u - u_h\|_{\mathbb{U}}$*, where* u_h *is the solution component of* (4.8.1) *and hence of the PG-scheme with projected test-space* $\mathcal{T}_{\mathbb{S}_h}(\mathbb{U}_h)$*?*

Since by (4.8.12) and (4.8.14)

$$r_h(w_h, f) = P_{\mathbb{S}_h} r(w_h, f), \tag{4.8.16}$$

we always have

$$\|r_h(w_h, f)\|_{\mathbb{V}} \leq \|r(w_h, f)\|_{\mathbb{V}}, \quad w_h \in \mathbb{U}_h, \tag{4.8.17}$$

so we can trivially replace $\|r(u_h, f)\|_{\mathbb{V}}$ in the lower estimate in (4.8.15) by the computed quantity $\|r_h(u_h, f)\|_{\mathbb{V}}$.

Remark 4.8.3 If the space $\mathbb{S}_h \subset \mathbb{V}$ is even δ-proximal for the *extended trial space*

$$\hat{\mathbb{U}}_h := \mathbb{U}_h + \mathcal{B}^{-1} f, \tag{4.8.18}$$

i.e.,

$$\sup_{v \in \mathcal{T}(\hat{\mathbb{U}}_h)} \frac{\|v - P_{\mathbb{S}_h} v\|_{\mathbb{V}}}{\|v\|_{\mathbb{V}}} \leq \delta \quad \text{holds for some } \delta \in [0, 1), \tag{4.8.19}$$

one has

$$C_{\mathcal{B}}^{-1} \|r_h(u_h, f)\|_{\mathbb{V}} \leq \|u - u_h\|_{\mathbb{U}} \leq c_{\mathcal{B}_h} \|r_h(u_h, f)\|_{\mathbb{V}}, \tag{4.8.20}$$

where

$$c_{\mathcal{B}_h} \geq c_{\mathcal{B}} (1 - \delta^2)^{1/2}.$$

Thus, if $\mathbb{S}_h$ is δ-proximal for the extended space $\hat{\mathbb{U}}_h = \mathbb{U}_h + \mathcal{B}^{-1} f$ the solution component r_h of (4.8.1) provides a lower and upper *a-posteriori* error for the PG-solution. This can be used to drive adaptive solution strategies, as shown later in specific applications. □

In fact, (4.8.19) implies that

$$\|r_h(u_h, f)\|_{\mathbb{V}} = \|P_{\mathbb{S}_h} r(u_h, f)\|_{\mathbb{V}} \geq (1 - \delta^2)^{1/2} \|r(u_h, f)\|_{\mathbb{V}}. \tag{4.8.21}$$

4.8.0.1 Uzawa Iteration

The saddle-point problem (4.8.1) is symmetric but indefinite. Therefore, a straightforward application of the (preconditioned) conjugate gradient method is not appropriate. One finds many investigations of iterative solvers for large scale saddle-point problems which in principle apply here as well.

We discuss here only the so called *Uzawa iteration* that avoids solving the full saddle point problems by reducing it to two smaller elliptic problems at each iteration stage.

Uzawa Iteration:

i) Choose an initial guess $u_h^0 \in \mathbb{U}_h$, and a suitable damping parameter $\omega > 0$; suppose that we have a bound

$$\|u_h - u_h^0\|_{\mathbb{U}} \leq \eta_0. \tag{4.8.22}$$

ii) Given $u_h^k \in \mathbb{U}_h$, solve for $r_h^k \in \mathbb{S}_h$

$$(r_h^k, v_h)_{\mathbb{V}} = -B(u_h^k, v_h) + f(v_h), \quad v_h \in \mathbb{S}_h; \tag{4.8.23}$$

iii) solve for $u_h^{k+1} \in \mathbb{U}_h$:

$$(u_h^{k+1}, w_h)_{\mathbb{U}} = (u_h^k, w_h)_{\mathbb{U}} + \omega B(w_h, r_h^k), \quad w_h \in \mathbb{U}_h; \tag{4.8.24}$$

iv) if

$$C_{\mathcal{B}} \rho^k \eta_0 > \frac{1}{8} \|r_h^k\|_{\mathbb{V}}, \tag{4.8.25}$$

set $k+1 \to k$ and go to (i), otherwise, stop.

Step (ii) and step (iii) consist of one $\mathbb{V}$-, $\mathbb{U}$-orthogonal projection into $\mathbb{S}_h, \mathbb{U}_h$, respectively, i.e., standard Galerkin projections in these spaces.

Proposition 4.8.1 *Assume that* $\mathbb{S}_h$ *satisfies* (4.8.19). *Then the above iteration converges. Specifically, one has*

$$\|u_h^k - u_h\|_{\mathbb{U}} \leq \rho^k \eta_0, \quad k \in \mathbb{N}, \tag{4.8.26}$$

and

$$\|r_h^k - r_h(u_h, f)\|_{\mathbb{V}} \leq C_{\mathcal{B}} \rho^k \eta_0, \quad k \in \mathbb{N}. \tag{4.8.27}$$

Moreover, the output $\bar{u}_h = u_h^{\bar{k}}$ where $\bar{k}$ is the terminating iteration index, satisfies

$$c_1 \|r_h^{\bar{k}}\|_{\mathbb{V}} \leq \|u - \bar{u}_h\|_{\mathbb{U}} \leq c_2 \|r_h^{\bar{k}}\|_{\mathbb{V}}, \tag{4.8.28}$$

where $c_1 \geq \frac{3}{4C_{\mathcal{B}}}$, $c_2 \leq \frac{1}{8}\Big(\frac{9}{c_{\mathcal{B}_h}} + \frac{1}{C_{\mathcal{B}}}\Big)$. Likewise,

$$\bar{c}_1 \|r(u_h, f)\|_{\mathbb{V}} \leq \|u - \bar{u}_h\|_{\mathbb{U}} \leq \bar{c}_2 \|r(u_h, f)\|_{\mathbb{V}}, \tag{4.8.29}$$

where $\bar{c}_1 \geq c_1 \frac{3(1-\delta^2)^{1/2}}{9}$ and $\bar{c}_2 \leq \frac{8c_2}{7}$. □

Proof Define the mapping $\mathcal{T}^{*,h} : \mathbb{S}_h \to \mathbb{U}_h$ by

$$(\mathcal{T}^{*,h} v_h, w_h)_{\mathbb{U}} = B(w_h, v_h), \quad v_h \in \mathbb{S}_h,\, w_h \in \mathbb{U}_h. \tag{4.8.30}$$

As earlier we see that

$$\mathcal{T}^{*,h} = P_{\mathbb{U}_h} \circ \mathcal{R}_{\mathbb{U}} \circ \mathcal{B}^*,$$

where $\mathcal{B}^*$ is the adjoint of $\mathcal{B}$. In these terms we have

$$u_h^{k+1} = u_h^k + \omega \mathcal{T}^{*,h} r_h^k, \quad r_h^k = \mathcal{T}^h u_h^k - P_{\mathbb{S}_h} \mathcal{R}_{\mathbb{V}} f,$$

so that

$$\begin{aligned} u_h^{k+1} - u_h &= u_h^k - u_h + \omega \mathcal{T}^{*,h} r_h^k & (4.8.31)\\ &= u_h^k - u_h + \omega \mathcal{T}^{*,h}\big(\mathcal{T}^h u_h^k - P_{\mathbb{S}_h} \mathcal{R}_{\mathbb{V}} f\big). & (4.8.32) \end{aligned}$$

Now use *Petrov-Galerkin orthogonality*

$$B(u - u_h, \mathcal{T}^h(w_h)) = f(\mathcal{T}^h(w_h)) - f(\mathcal{T}^h(w_h)) = 0, \quad w_h \in \mathbb{U}_h. \tag{4.8.33}$$

Therefore,

$$\begin{aligned} B(w_h, P_{\mathbb{S}_h} \mathcal{R}_{\mathbb{V}} f - \mathcal{T}^h u_h) &= (\mathcal{T}^h w_h, P_{\mathbb{S}_h} \mathcal{R}_{\mathbb{V}} f - \mathcal{T}^h u_h)_{\mathbb{V}} \\ &= (\mathcal{T}(u - u_h), P_{\mathbb{S}_h} \mathcal{T}^h w_h)_{\mathbb{V}} \\ &= B(u - u_h, \mathcal{T}^h w_h) = 0, \end{aligned}$$

so that (4.8.31) becomes

$$\begin{aligned} u_h^{k+1} - u_h &= u_h^k - u_h + \omega \mathcal{T}^{*,h}\big(\mathcal{T}^h(u_h^k - u_h)\big) \\ &= (I - \omega \mathcal{T}^{*,h}\mathcal{T}^h)(u_h^k - u_h). \end{aligned} \tag{4.8.34}$$

Since

$$\begin{aligned} \|\mathcal{T}^{*,h}\mathcal{T}^h w_h\|_{\mathbb{U}} &= \sup_{z_h \in \mathbb{U}_h} \frac{(\mathcal{T}^h w_h, \mathcal{T}^h z_h)_{\mathbb{V}}}{\|z_h\| z_{\mathbb{U}}} \geq c_{\mathcal{B}_h} \sup_{z_h \in \mathbb{U}_h} \frac{(\mathcal{T}^h w_h, \mathcal{T}^h z_h)_{\mathbb{V}}}{\|\mathcal{T}^h z_h\|_{\mathbb{U}}} \\ &\geq c_{\mathcal{B}_h}^2 \|w_h\|_{\mathbb{U}} \end{aligned}$$

and similarly

$$\|\mathcal{T}^{*,h}\mathcal{T}^h w_h\|_{\mathbb{U}} \leq C_{\mathcal{B}}^2 \|w_h\|_{\mathbb{U}},$$

and since $\mathcal{T}^{*,h}\mathcal{T}^h w_h$ is positive definite, we conclude that

$$\|I - \omega \mathcal{T}^{*,h}\mathcal{T}^h w_h\|_{\mathcal{L}(\mathbb{U},\mathbb{U})} \leq \rho = \rho(\omega, c_{\mathcal{B}_h}, C_{\mathcal{B}}) < 1, \tag{4.8.35}$$

when ω is chosen properly, e.g. for

$$\omega = \frac{2}{C_{\mathcal{B}}^2 + c_{\mathcal{B}_h}^2} \quad \text{one obtains} \quad \rho = \frac{C_{\mathcal{B}}^2 - c_{\mathcal{B}_h}^2}{C_{\mathcal{B}}^2 + c_{\mathcal{B}_h}^2}.$$

Hence

$$\|u_h^k - u_h\|_{\mathbb{U}} \leq \rho^k \|u_h^0 - u_h\|_{\mathbb{U}} \overset{(4.8.22)}{\leq} \rho^k \eta_0, \quad k \in \mathbb{N}, \tag{4.8.36}$$

which is (4.8.26). Moreover, since

$$r_h^k = \mathcal{T}^h u_h^k - P_{\mathbb{S}_h}\mathcal{R}_{\mathbb{V}} f, \quad r_h = r_h(u_h, f) = \mathcal{T}^h u_h - P_{\mathbb{S}_h}\mathcal{R}_{\mathbb{V}} f,$$

one obtains

$$\begin{aligned} \|r_h^k - r_h\|_{\mathbb{V}} &= \|\mathcal{T}^h(u_h^k - u_h)\|_{\mathbb{V}} \leq C_{\mathcal{B}} \|u_h^k - u_h\|_{\mathbb{U}} \\ &\leq C_{\mathcal{B}}\rho^k \|u_h^0 - u_h\|_{\mathbb{U}} \overset{(4.8.22)}{\leq} C_{\mathcal{B}}\rho^k \eta_0, \quad k \in \mathbb{N}, \end{aligned} \tag{4.8.37}$$

confirming (4.8.27). Hence, the iteration converges at a rate depending on the discrete condition bound $C_{\mathcal{B}}/c_{\mathcal{B}_h}$. By (4.8.19), we know that $\|r_h(u_h, f)\|_{\mathbb{V}} \sim \|r_h(u_h, f)\|_{\mathbb{V}} > 0$. Thus, the termination criterion is met for some finite $\bar{k}$ for which

$$\|u_h - u_h^{\bar{k}}\|_{\mathbb{U}} \leq \rho^{\bar{k}} \eta_0 \leq \frac{1}{8C_{\mathcal{B}}} \|r_h^{\bar{k}}\|_{\mathbb{V}}. \tag{4.8.38}$$

By (4.8.37), we have

$$\big|\|r_h(u_h,f)\|_{\mathbb{V}} - \|r_h^{\bar{k}}\|_{\mathbb{V}}\big| \le C_{\mathcal{B}}\rho^{\bar{k}}\eta_0 \le \frac{1}{8}\|r_h^{\bar{k}}\|_{\mathbb{V}}$$

and thus

$$\frac{7}{8}\|r_h^{\bar{k}}\|_{\mathbb{V}} \le \|r_h(u_h,f)\|_{\mathbb{V}} \le \frac{9}{8}\|r_h^{\bar{k}}\|_{\mathbb{V}}. \tag{4.8.39}$$

Hence

$$\begin{aligned}\|r_h^{\bar{k}}\|_{\mathbb{V}} &\le \frac{8}{7}\|r_h(u_h,f)\|_{\mathbb{V}} \le \frac{8}{7}\|r(u_h,f)\|_{\mathbb{V}} \le \frac{8C_{\mathcal{B}}}{7}\|u-u_h\|_{\mathbb{U}}\\ &\le \frac{8C_{\mathcal{B}}}{7}\big\{\|u-u_h^{\bar{k}}\|_{\mathbb{U}} + \|u_h-u_h^{\bar{k}}\|_{\mathbb{U}}\big\}\\ &\le \frac{8C_{\mathcal{B}}}{7}\big\{\|u-u_h^{\bar{k}}\|_{\mathbb{U}} + \frac{1}{8C_{\mathcal{B}}}\|r_h^{\bar{k}}\|_{\mathbb{V}}\big\}\\ &\le \frac{8C_{\mathcal{B}}}{7}\|u-u_h^{\bar{k}}\|_{\mathbb{U}} + \frac{1}{7}\|r_h^{\bar{k}}\|_{\mathbb{V}}.\end{aligned}$$

Therefore,

$$\frac{3}{4C_{\mathcal{B}}}\|r_h^{\bar{k}}\|_{\mathbb{V}} \le \|u-u_h^{\bar{k}}\|_{\mathbb{U}}, \tag{4.8.40}$$

which confirms the lower bound in (4.8.28). Likewise, by (4.8.20) and (4.8.38),

$$\begin{aligned}\|u-u_h^{\bar{k}}\|_{\mathbb{U}} &\le \|u-u_h\|_{\mathbb{U}} + \|u_h-u_h^{\bar{k}}\|_{\mathbb{U}} \le c_{\mathcal{B}_h}^{-1}\|r_h(u_h,f)\|_{\mathbb{V}} + \frac{1}{8C_{\mathcal{B}}}\|r_h^{\bar{k}}\|_{\mathbb{V}}\\ &\le \Big(\frac{9}{8c_{\mathcal{B}_h}} + \frac{1}{8C_{\mathcal{B}}}\Big)\|r_h^{\bar{k}}\|_{\mathbb{V}},\end{aligned}$$

showing the upper bound in (4.8.28).

Likewise, combining (4.8.39) and (4.8.21), yields

$$\|r(u_h,f)\|_{\mathbb{V}} \le (1-\delta^2)^{-1/2}\|r_h(u_h,f)\|_{\mathbb{V}} \overset{(4.8.39)}{\le} \frac{9}{3(1-\delta^2)^{1/2}}\|r_h^{\bar{k}}\|_{\mathbb{V}}, \tag{4.8.41}$$

which together with the lower bound in (4.8.28) confirms the lower bound in (4.8.29). Similarly,

$$\|r_h^{\bar{k}}\|_{\mathbb{V}} \le \frac{8}{7}\|r_h(u_h,f)\|_{\mathbb{V}} \le \frac{8}{7}\|r(u_h,f)\|_{\mathbb{V}},$$

which completes the proof. ■

Since $r(u_h, f)\|_{\mathbb{V}} \sim \|u - u_h\|_{\mathbb{U}}$ the output approximation performs up to uniform constants as the PG-approximation, which in turn behaves like the best approximation to the exact solution u from the trial space $\mathbb{U}_h$.

4.8.0.2 Nested Iteration

Suppose that the following routines are available:

$\mathbf{Res}[u_h, \mathbb{U}_h, \mathbb{S}_h] \to r_h \in \mathbb{S}_h$ such that for $w_h \in \mathbb{U}_h$

$$(r_h, v_h)_{\mathbb{V}} = B(u_h, v_h) - f(v_h), \quad v_h \in \mathbb{S}_h. \tag{4.8.42}$$

$\mathbf{Proj}[\tilde{u}_h, z_h, \mathbb{U}_h, \mathbb{S}_h] \to u_h$ such that for $\tilde{u}_h \in \mathbb{U}_h$, $z_h \in \mathbb{S}_h$,

$$(u_h, w_h)_{\mathbb{U}} = (\tilde{u}_h, w_h)_{\mathbb{U}} + B(w_h, z_h), \quad w_h \in \mathbb{U}_h. \tag{4.8.43}$$

$\mathbf{Exp}[\mathbb{U}_h, \mathbb{S}_h] \to (\tilde{\mathbb{U}}_h, \tilde{\mathbb{S}_h})$ such that $\mathbb{U}_h \subset \tilde{\mathbb{U}}_h$, $\mathbb{S}_h \subset \tilde{\mathbb{S}_h} \subset \mathbb{V}$ such that for some fixed $\delta < 1$ the space $\tilde{\mathbb{S}_h}$ is δ-proximal for $\tilde{\mathbb{U}}_h$ and for some constant $\zeta < 1$ one has

$$\|r_h(\tilde{u}_h, f)\|_{\mathbb{V}} \leq \zeta \|r_h(u_h, f)\|_{\mathbb{V}}, \tag{4.8.44}$$

where $r_h(u_h, f)$, $r_h(\tilde{u}_h, f)$ are the solution components in $\mathbb{V}$ of the saddle-point problems (4.8.1) with respect to $\mathbb{U}_h, \mathbb{S}_h$, $\tilde{\mathbb{U}}_h, \tilde{\mathbb{S}}_h$, respectively.

The rationale of **Exp** is the following: for δ-proximal test-search-spaces the computabel quantities $r_h(u_h, f)$ are equivelant to the error of the PG solution u_h. As shown later, one can derive from those quantities not only the current accuracy provided by a pair of trial and test-search spacees but also indications about how to expand the space $\mathbb{U}_h$ to a larger space $\tilde{\mathbb{U}}_h$ in a way that the current residual is decreased by a fixed factor. So, several subsequent expansions will decrease the errors by a corresponding fixed factor. This is done in *adaptive* methods which in the current situation can be based naturally on the computed projected lifted residuals.

Nested Iteration is a strategy to solve an infinite-dimensional problem

$$B(u, v) = f(v), \quad v \in \mathbb{V}, \tag{4.8.45}$$

approximately within a desired *target accuracy*. Roughly speaking, one proceeds as follows:

- Solve a small-dimensional discretization (D_0) exactly or very accurately;
- use the approximate solution for the preceding smaller-dimensional discretization (D_{n-1}) as an initial guess; in a suitably chosen extended trial space.
- the discrete problem (D_n) is solved approximately with an inner iteration (e.g. the above Uzawa scheme) until one finds an approximate solution within an updated target accuracy. This is repeated until the final accuracy tolerance is met.

This strategy uses an *outer iteration*, where at each step the trial space is enlarged. An *inner iteration* can be used to find an approximate solution for the current (fixed) discrete problem. For such a strategy to work one needs a-posteriori error bounds that allow one to decide when to terminate the inner iteration and whether the final target accuracy is reached.

Exercise 4.8.2 *(i) Employ the above routines to formulate a skeleton of an algorithm*

Solve$[B, f, \epsilon] \to (\mathbb{U}_\epsilon, \mathbb{V}_\epsilon, u_\epsilon)$ *such that* $\mathbb{U}_\epsilon \subset \mathbb{U}$, $\mathbb{V}_\epsilon \subset \mathbb{V}$ *are finite dimensional trial- and test-search-spaces with the following properties:*

i) $\mathbb{V}_\epsilon$ *is* δ*-proximal for* $\mathbb{U}_\epsilon$*;*

ii) the corresponding PG solution u_ϵ *satisfies*

$$\|u - u_\epsilon\|_{\mathbb{U}} \leq \epsilon,$$

where u *is the exact solution of*

$$B(u, v) = f(v), \quad v \in \mathbb{V}.$$

(ii) Estimate the computational work in terms of the number of calls of the routines **Res**, **Proj***;*

(iii) Let $\#$**Proj**$[\cdot,\cdot,\cdot,\cdot]$ *denote the number flops required for executing* **Proj** *for the given parameters (and analogously for the other routines).*

Estimate the total number $N = N(\epsilon) = \#\mathbf{Solve}[B, f, \epsilon] \to (\mathbb{U}_\epsilon, \mathbb{V}_\epsilon, u_\epsilon)$ *under the following assumptions:*

- $\#\mathbf{Proj}[\tilde{u}_h, z_h, \mathbb{U}_h, \mathbb{S}_h] \leq C(\dim \mathbb{U}_h + \dim \mathbb{S}_h)$*;*
- $\#\mathbf{Res}[u_h, \mathbb{U}_h, \mathbb{S}_h] \leq C(\dim \mathbb{U}_h + \dim \mathbb{S}_h)$*;*
- *for* $\mathbf{Exp}[\mathbb{U}_h, \mathbb{S}_h] \to (\tilde{\mathbb{U}}_h, \tilde{\mathbb{S}_h})$ *one has* $\dim \tilde{\mathbb{U}}_h \leq C \dim \mathbb{U}_h$*,*

where C *is an abslute constant. Assume that the constants* $C_{\mathcal{B}}, c_{\mathcal{B}}, \delta$ *are known to you.*

4.8.1 Optimal Norms

The tightness of residual error bounds and the quantitative performance of the above iterative schemes depend in an essential way on the quotient $C_{\mathcal{B}}/c_{\mathcal{B}}$ (the bound on $\kappa_{\mathbb{U},\mathbb{V}'}(\mathcal{B})$. We discuss next a principal way of improving the condition by modifying the norms.

Suppose that Assumption 4.0.3 applies to

$$B(u, v) = f(v), \quad v \in \mathbb{V}, \tag{4.8.46}$$

But

$$\kappa_{\mathbb{U},\mathbb{V}}(\mathcal{B}) \gg 1 \quad \text{(the bounds } C_{\mathcal{B}}, c_{\mathcal{B}} \text{ satisfy } \tfrac{C_{\mathcal{B}}}{c_{\mathcal{B}}} \gg 1). \tag{4.8.47}$$

Example 4.8.1 Convection-diffusion equation (see Section 3.1.1): let $c \in L_\infty(\Omega)$, $b : \Omega \to \mathbb{R}^d$ such that

$$c - \frac{1}{2}\operatorname{div} b \geq 0, \quad \text{ in } \Omega, \quad \epsilon > 0, \quad \operatorname{div} b \in L_\infty(\Omega). \tag{4.8.48}$$

A possible weak formulation of the boundary value problem

$$-\epsilon\Delta u + b \cdot \nabla u + cu = f \quad \text{in } \Omega,\ u|_{\partial\Omega} = 0, \tag{4.8.49}$$

is based on

$$\mathbb{U} = \mathbb{V} = H_0^1(\Omega), \tag{4.8.50}$$

and the bilinear form

$$B(u, v) = \int_\Omega \epsilon \nabla u \cdot \nabla v + b \cdot \nabla u + cuv dx. \tag{4.8.51}$$

Exercise 4.8.3 *(i) Under the above assumptions the problem*

$$B(u,f) = f(v), \quad v \in H_0^1(\Omega) \tag{4.8.52}$$

is well posed.
(ii) One has

$$\kappa_{H_0^1(\Omega),H^{-1}(\Omega)} \le \frac{\|b\|_{L_\infty(\Omega)^d}}{\epsilon}, \tag{4.8.53}$$

see the discussion in [CDW12].

The above example addresses the following situation:

- In principle, the space $H_1^0(\Omega)$ as a *set* is appropriate for trial- and test-space because the leading part of the PDE is the Laplacian and hence symmetric.
- However, when the strength of the diffusion ϵ gets small compared with the convection, the quantitative structure of the equation starts changing its type, namely approaching a pure convection - and hence hyperbolic - problem, causing the well-known serious numerical problems. This suggests, improving the condition by endowing $H_0^1(\Omega)$ with *different (problem adapted)* norms for the trial- and test-side, respectively. This principle is discussed next on an abstract level but will later be applied to examples.

A renormation principle: Assume that (4.8.46) is well-posed but ill-conditioned, i.e., (4.8.47) holds. One remedy is to fix either the trial- or the test-norm and adjust the other one.

Fixing the trial norm $\|\cdot\|_{\mathbb{U}}$: Define the *modified test-norm*

$$\|v\|_{\mathbb{V}_{\rm opt}} := \|\mathcal{B}^* v\|_{\mathbb{U}'} = \|\mathcal{R}_{\mathbb{U}}\mathcal{B}^* v\|_{\mathbb{U}}, \quad v \in \mathbb{V}, \tag{4.8.54}$$

where $\mathcal{B}^*$ is the adjoint of $\mathcal{B}$, defined by $(\mathcal{B}^*v)(u) = (\mathcal{B}u)(v)$, $u \in \mathbb{U}, v \in \mathbb{V}$. We denote by $\mathbb{V}_{\rm opt}$ the Hilbert space which agrees with $\mathbb{V}$ as a set but is endowed with the norm $\|\cdot\|_{\mathbb{V}_{\rm opt}}$.

This makes sense because $\mathcal{B} \in \mathcal{L}(\mathbb{U},\mathbb{V}')$ says that $\mathcal{B}^* \in \mathcal{L}(\mathbb{V},\mathbb{U}')$ so that the right hand side of (4.8.54) is well-defined. It is indeed a norm because $\mathcal{B}^*$

is an isomorphism and $\|v\|_{\mathbb{V}_{\rm opt}} = 0$ implies $v = 0$. Moreover, interchanging the roles of $\mathcal{B}$ and $\mathcal{B}^*$, by the Banach-Nečas Theorem 3.2.2, $\|\mathcal{B}^*v\|_{\mathbb{U}'} \leq C_{\mathcal{B}}\|v\|_{\mathbb{V}}$ and $c_{\mathcal{B}}\|v\|_{\mathbb{V}} \leq \|\mathcal{B}^*v\|_{\mathbb{U}'}$, so that both norms $\|\cdot\|_{\mathbb{V}}$ and $\|\cdot\|_{\mathbb{V}_{\rm opt}}$ are indeed equivalent.

Let us denote by $\bar{c}_{\mathcal{B}}, \bar{C}_{\mathcal{B}}$ the inf-sup- and continuity constants when considering

$$B(\cdot,\cdot) : \mathbb{U} \times \mathbb{V}_{\rm opt} \to \mathbb{R}.$$

Continuity:

$$|B(w,v)| = |(\mathcal{B}^*v)(w)| \leq \|\mathcal{B}^*v\|_{\mathbb{U}'}\|w\|_{\mathbb{U}} = \|w\|_{\mathbb{U}}\|v\|_{\mathbb{V}_{\rm opt}}, \qquad (4.8.55)$$

which means

$$\bar{C}_{\mathcal{B}} = 1. \qquad (4.8.56)$$

inf-sup constant:

$$\begin{aligned} \sup_{w\in\mathbb{U}} \frac{B(w,v)}{\|w\|_{\mathbb{U}}} &= \|\mathcal{B}^*v\|_{\mathbb{U}'} = \|v\|_{\mathbb{V}_{\rm opt}} \\ \Rightarrow \inf_{v\in\mathbb{V}_{\rm opt}} \sup_{w\in\mathbb{U}} \frac{B(w,v)}{\|v\|_{\mathbb{V}_{\rm opt}}\|w\|_{\mathbb{U}}} &= 1, \end{aligned} \qquad (4.8.57)$$

i.e.,

$$\bar{c}_{\mathcal{B}} = 1. \qquad (4.8.58)$$

Hence, the variational problem becomes perfectly conditioned.

Exercise 4.8.4 *(i) Show that $\mathcal{B} \in \mathcal{L}(\mathbb{U}, \mathbb{V}'_{\rm opt})$ is an isometry, i.e.,*

$$\|\mathcal{B}\|_{\mathcal{L}(\mathbb{U},\mathbb{V}'_{\rm opt})} = 1 = \|\mathcal{B}^{-1}\|_{\mathcal{L}(\mathbb{V}'_{\rm opt},\mathbb{U})}.$$

(see also [DHSW12]).
(ii) Represent the Riesz-map $\mathcal{R}_{\mathbb{V}_{\rm opt}}$ in terms of $\mathcal{B}$.

Although it seems to be most natural to adjust the test-norm and keep the trial norm, i.e., measure accuracy in the original metric, this is not always the case. In particular, for *singularly perturbed* problems like (4.8.49) with arbitrarily small ϵ, the original metric may be inappropriate. In fact, the convection-diffusion equation may cause very thin boundary layers and the plain H^1-error would then be completely determined by the boundary layer. In fact, it would not decrease until the boundary layer is resolved (see [CDW12]). Therefore, a problem-dependent norm which penalizes the singular behavior in the layer region less strongly, may be more appropriate.

Fixing the test norm $\|\cdot\|_{\mathbb{V}}$: one can modify the trial-norm as follows:

$$\|w\|_{\mathbb{U}_{\text{opt}}} := \|\mathcal{R}_{\mathbb{V}}\mathcal{B}w\|_{\mathbb{V}} = \|\mathcal{T}w\|_{\mathbb{V}} = \|\mathcal{B}w\|_{\mathbb{V}'}. \tag{4.8.59}$$

Since the trial-to-test map $\mathcal{T}$ from (4.5.5) is an isomorphism, $\|\cdot\|_{\mathbb{U}_{\text{opt}}}$ is a well-defined norm.

Remark 4.8.4 Roughly speaking the isomorphisms $\mathcal{T}^*$ and $\mathcal{T}$, defining the modified norms (4.8.54), (4.8.59), *absorb* the bad condition of the original formulation. □

Exercise 4.8.5 *Considering $B(\cdot,\cdot) : \mathbb{U}_{\text{opt}} \times \mathbb{V} \to \mathbb{R}$, show that one has again*

$$\inf_{w\in\mathbb{U}_{\text{opt}}} \sup_{v\in\mathbb{V}} \frac{B(w,v)}{\|w\|_{\mathbb{U}_{\text{opt}}}\|v\|_{\mathbb{V}}} = 1 = \sup_{w\in\mathbb{U}_{\text{opt}}} \sup_{v\in\mathbb{V}} \frac{B(w,v)}{\|w\|_{\mathbb{U}_{\text{opt}}}\|v\|_{\mathbb{V}}}, \tag{4.8.60}$$

i.e., we have again that

$$\bar{C}_{\mathcal{B}} = \bar{c}_{\mathcal{B}} = 1 \tag{4.8.61}$$

and $\mathcal{B} \in \mathcal{L}(\mathbb{U}_{\text{opt}}, \mathbb{V}')$ is an isometry.

Remark 4.8.5 (i) As earlier in connection with the trial-to-test map $\mathcal{T}$ the above renormation strategy is primarily a *guiding principle* that should help finding appropriate topologies for the trial- and test space, in the sense that the corresponding variational formulation gives rise to a well-conditioned operator. However, such norms could be very difficult or expensive to realize numerically because they involve in general nontrivial mappings like $\mathcal{T}$ or $\mathcal{T}^*$.

(ii) The above modification of the test-norm (4.8.54), however becomes feasible if the Riesz-map $\mathcal{R}_{\mathbb{U}}$ is trivial (the identity) which means $\mathbb{U} = \mathbb{U}'$, i.e., $\mathbb{U} = L_2(\Omega)$. This arises indeed in at least two scenarios:

(a) considering so called *ultra-weak* formulations, seeking the solution in a low-regularity space such as $L_2(\Omega)$; this may be of interest even for Poisson's equation when the right hand side involves Dirac-distributions and hence do not belong to $H^{-1}(\Omega)$ when $d > 1$;

(b) writing a higher order PDE as a system of first order PDEs where just a single integration by parts frees the unknown from any derivative. This is of interest when the solution may have layers or even discontinuities along curves/lower-dimensional manifolds (shear layers) as in the case of transport problems.

In both cases the choice of norms can be viewed as being guided by the above extreme cases of optimal pairs of norms. □

4.9 Localizing the Test-Search-Space - the Discontinuous Petrov Galerkin Method

To avoid the solution of $\dim \mathbb{U}_h$ global elliptic problems for the computation of the Petrov-Galerkin test-functions we have used in the previous section the mixed formulation (4.8.1). In this section we discuss an alternate strategy for reducing the cost of computing good test-functions.

Central Idea: use a *mesh-dependent* (infinite-dimensional) variational formulation; this offers the possibility of employing so called *broken* test-search-spaces which. This, in turn, will be seen to lead to test-functions which are localized to the cells in the underlying mesh.

The basic idea goes back already to [BM84]. It has been further developed for a wide range of PDE classes by Demkowicz and Gopalakrishnan, see e.g. [DG11].

4.9.1 A Guiding Example

One typically rewrites the original PDE as a first order system. As a motivating example concerning Remark 4.8.5, (a), (b), consider

$$\begin{aligned} & -\operatorname{div}(\epsilon\nabla u) + b\cdot\nabla u + cu \\ & \qquad = -\operatorname{div}(\epsilon\nabla u) + \operatorname{div}(ub) + (c - (\operatorname{div} b)u = f, \quad u|_{\partial\Omega} = 0, \quad (4.9.1)\end{aligned}$$

where we assume for simplicity that ϵ is a positive constant when scalar-valued or a constant symmetric positive definite matrix. Introducing the "weighted flux" $\epsilon^{1/2}\nabla u$ as a new unknown, (4.9.1) is equivalent to the *system*

$$\sigma = \epsilon^{1/2}\nabla u, \quad -\operatorname{div}(\epsilon^{1/2}\sigma - ub) + (c - (\operatorname{div} b)u = f, \qquad (4.9.2)$$

with unknowns $[u,\sigma]$ (u scalar-valued, σ vector-valued). A natural variational formulation reads

$$\begin{aligned} B([u,\sigma],[v,\tau]) = \int_\Omega & (\sigma - \epsilon^{1/2}\nabla u)\cdot\tau + (\operatorname{div}(ub - \epsilon^{1/2}\sigma)v \\ & + (c - \operatorname{div} b)uvdx. \qquad (4.9.3)\end{aligned}$$

Remark 4.9.1 The above way of splitting the diffusion coefficient (which is also well-defined, when ϵ is a positve definite symmetric matrix) is not mandatory. We could have defined $\sigma = \epsilon\nabla u$. The above way better preserves symmetry. More importantly it is more appropriate for studying the vicuous limit $\epsilon \to 0$, as seen later. □

Choosing as trial- and test-space

$$\mathbb{U}_1 := H_0^1(\Omega) \times H(\operatorname{div};\Omega), \quad \mathbb{V}_1 := L_2(\Omega) \times (L_2(\Omega))^d, \qquad (4.9.4)$$

where

$$\begin{aligned} H(\operatorname{div};\Omega) &:= \{\tau \in (L_2(\Omega))^d : \operatorname{div}\tau \in L_2(\Omega)\}, \\ \|\tau\|^2_{H(\operatorname{div};\Omega)} &:= \|\tau\|^2_{(L_2(\Omega))^d} + \|\operatorname{div}\tau\|^2_{L_2(\Omega)}. \end{aligned}$$

Cauchy-Schwarz' inequality readily confirms that

$$B(\cdot,\cdot) : \mathbb{U}_1 \times \mathbb{V}_1 \to \mathbb{R}, \qquad (4.9.5)$$

is continuous, i.e., the induced operator

$$(\mathcal{B}_1[u,\sigma])([v,\tau]) = B([u,\sigma],[v,\tau]), \quad [u,\sigma] \in \mathbb{U}_1, \tag{4.9.6}$$

satisfies

$$\mathcal{B}_1 \in \mathcal{L}(\mathbb{U}_1, \mathbb{V}_1'). \tag{4.9.7}$$

Remark 4.9.2 Since the test-space $\mathbb{V}_1$ is just a product of L_2-spaces it agrees with its dual and the Riesz-map $\mathcal{R}_{\mathbb{V}_1}$ becomes trivial. Thus, the optimal trial-norm $\|\cdot\|_{\mathbb{U}_{\rm opt}}$ is feasible. □

An alternate variational formulation is obtained as follows. Applying integration by parts, (4.9.3) takes the form

$$\begin{aligned} B([u,\sigma],[v,\tau]) = \int_\Omega & \sigma\cdot\tau + u\mathrm{div}\,(\epsilon^{1/2}\tau) - (ub - \epsilon^{1/2}\sigma)\cdot\nabla v \\ & + (c - \mathrm{div}\, b)uvdx \\ & + \int_{\partial\Omega} n\cdot(bu - \epsilon^{1/2}\sigma)vds - \int_{\partial\Omega} n\cdot\epsilon^{1/2}\tau uds. \end{aligned} \tag{4.9.8}$$

Choosing as trial- and test-space

$$\mathbb{U}_2 := L_2(\Omega)\times(L_2(\Omega))^d, \quad \mathbb{V}_2 := H_0^1(\Omega)\times H(\mathrm{div};\Omega), \tag{4.9.9}$$

by the same reasoning

$$B(\cdot,\cdot) : \mathbb{U}_2 \times \mathbb{V}_2 \to \mathbb{R}, \tag{4.9.10}$$

is also continuous. Thus, the induced operator

$$(\mathcal{B}_2[u,\sigma])([v,\tau]) = B([u,\sigma],[v,\tau]), \quad [u,\sigma] \in \mathbb{U}_2, \tag{4.9.11}$$

satisfies

$$\mathcal{B}_2 \in \mathcal{L}(\mathbb{U}_2, \mathbb{V}_2'). \tag{4.9.12}$$

Remark 4.9.3 Now the trial space $\mathbb{U}_2$ is a product of L_2-spaces so that the optimal test-norm (4.8.54) is now practically feasible. In this formulation essentially no regularity is required of the solution $[u,\sigma]$. This is sometimes referred to as *ultra-weak* formulation. □

$\mathcal{B}_1$ and $\mathcal{B}_2$ both represent the same PDE. They differ only through their respective domains and ranges. A few more remarks on the trace integrals in (4.9.8) are in order.

Remark 4.9.4 If we assume that $u, v \in H_0^1(\Omega)$ the trace integrals in the last line of (4.9.8) would vanish. However, in this last formulation no regularity is imposed on u, σ. They just need to be square integrable. But then the trace of u does not exist. However, if (non-homogeneous) boundary conditions are imposed u which belong to $H^{1/2}(\partial\Omega)$, i.e., $u = g$ on $\partial\Omega$, then $\int_{\partial\Omega} n \cdot \epsilon^{1/2}\tau g ds$ is defined as a functional acting on the *normal traces* $n \cdot \tau$ which are known to belong to $H^{-1/2}(\partial\Omega) = (H^{1/2}(\partial\Omega))'$. Thus

$$g \mapsto \int_{\partial\Omega} n \cdot \epsilon^{1/2}\tau g ds$$

is a bounded linear functional on $\mathbb{V}_2$ and thus belongs to $\mathbb{V}_2'$. Recall that boundary conditions which are incorporated in the variational formulation through a bounded linear functional are called *natural boundary conditions*. Thus, in this formulation Dirichlet boundary conditions are natural ones (in contrast to the standard second order formulation). The issue of traces plays an important role in what follows. □

Remark 4.9.5 Thus, considering (4.9.1) with (possibly) *inhomogeneous* boundary conditions $u|_{\partial\Omega} = g \in H^{1/2}(\partial\Omega)$ (in the sense of traces), the variational formulation (4.9.8) becomes

$$\begin{aligned}\int_\Omega \sigma \cdot \tau + u\mathrm{div}\,(\epsilon^{1/2}\tau) - (ub - \epsilon^{1/2}\sigma) \cdot \nabla v + (c - \mathrm{div}\, b)uvdx \\ = \int_{\partial\Omega} n \cdot \epsilon^{1/2}\tau g ds - \int_{\partial\Omega} n \cdot (bg - \epsilon^{1/2}\sigma)v ds + f(v), \\ = \int_{\partial\Omega} n \cdot \epsilon^{1/2}\tau g ds + f(v), \quad [v, \tau] \in H_0^1(\Omega) \times H(\mathrm{div}; \Omega). \qquad (4.9.13)\end{aligned}$$

Exact and formal adjoints: To understand the essential mechanisms behind the formulations (4.9.3) and (4.9.8) it is helpful to have a closer look at the notion of the adjoint of an operator. The operator $\mathcal{B}_1$ from (4.9.7) is given by

$$\mathcal{B}_1([u, \sigma]) = \begin{pmatrix} \mathrm{div}\,(ub - \epsilon^{1/2}\sigma) + (c - \mathrm{div}\, b) \\ \sigma - \epsilon^{1/2}\nabla u \end{pmatrix}. \qquad (4.9.14)$$

Remark 4.9.6 The adjoint $\mathcal{B}_1^*$ of $\mathcal{B}_1$, defined by

$$(\mathcal{B}_1^*[v,\tau])([u,\sigma]) = B([u,\sigma],[v,\tau]), \quad [v,\tau] \in \mathbb{V}_1,\ [u,\sigma] \in \mathbb{U}_1, \tag{4.9.15}$$

belongs to

$$\mathcal{L}(\mathbb{V}_1, \mathbb{U}_1') = \mathcal{L}(L_2(\Omega) \times (L_2(\Omega))^d, H^{-1}(\Omega) \times H(\mathrm{div};\Omega)').$$

While the domain of the *exact* adjoint $\mathcal{B}_1^*$ is $\mathbb{V}_1 = L_2(\Omega) \times (L_2(\Omega))^d$, when restricted to smooth functions with vanishing boundary traces, it is, in view of (4.9.8) given by

$$\mathcal{B}_1'([v,\tau]) = \begin{pmatrix} \mathrm{div}\,(\epsilon^{1/2}\tau) - b \cdot \nabla v + (c - \mathrm{div}\, b)v \\ \tau + \epsilon^{1/2}\nabla v \end{pmatrix}. \tag{4.9.16}$$

We call $\mathcal{B}_1'$ the *formal adjoint* of $\mathcal{B}_1$ and record

$$\mathcal{B}_1'([u,\sigma]) = \mathcal{B}_1^*([u,\sigma]), \quad [u,\sigma] \in C_0^\infty(\Omega) \times (C_0^\infty(\Omega))^d. \tag{4.9.17}$$

In fact, they agree by density on $H_0^1(\Omega) \times H(\mathrm{div};\Omega)$. □

Remark 4.9.7 In the above terms we have

$$(\mathcal{B}_1')^* = \mathcal{B}_2. \tag{4.9.18}$$

4.9.2 The Discontinuous Petrov-Galerkin Concept (DPG)

For convenience, abbreviate for any subdomain $D \subseteq \Omega$

$$(w,v)_D = \int_D wv dx, \quad \langle w,v\rangle_{\partial D} = \int_{\partial D} wv ds$$

where the second boundary integral is interpreted as a functional acting on one of the arguments whenever the other argument belongs to the corresponding dual trace space. Only when $w,v \in L_2(\partial D)$ the right intgral representation is in strict terms justified.

The above discussion can then be summarized as follows:

(I) write the original PDE as a first order system whose weak formulation is

$$B(u,v) = (\mathcal{B}u, v)_\Omega = f(v), \quad v \in \mathbb{V}.$$

(II) Rewrite

$$(\mathcal{B}u, v)_\Omega = (u, \mathcal{B}'v)_\Omega = f(v) + g(v), \quad v \in \mathbb{V},$$

where $\mathbb{V}$ is chosen in a way that $\mathcal{B}'v \in L_2(\Omega)$ and $g, f \in \mathbb{V}'$, where g represents the boundary conditions which have now become natural ones.

The DPG concept works by applying step (II) cellwise on a given partition $\mathcal{P}$ of Ω, using that

$$(w, v)_\Omega = \sum_{K \in \mathcal{P}} (w, v)_K,$$

and that differential operators localize. This provides formally for smooth functions u, v

$$\begin{aligned} (\mathcal{B}u, v)_\Omega &= \sum_{K \in \mathcal{P}} (\mathcal{B}u, v)_K \\ &= \sum_{K \in \mathcal{P}} (u, \mathcal{B}'v)_K + \{(\mathcal{B}u, v)_K - (u, \mathcal{B}'v)_K\} \\ &= \sum_{K \in \mathcal{P}} (u, \mathcal{B}'v)_K + J_{\partial K}(u, v). \end{aligned} \tag{4.9.19}$$

The term $J_{\partial K}(u,v)$ indicates that it “lives” on the cell boundary ∂K. In fact, it represents the trace-integral contributions when performing integration by parts on K to go from $(\mathcal{B}u, v)_K$ to $(u, \mathcal{B}'v)_K$ where $\mathcal{B}'$ is the local formal adjoint.

We exemplify this for the previous example (4.9.3).

$$\begin{aligned}(\mathcal{B}[u,\sigma],[v,\tau])_K &= \int_K (\sigma - \epsilon^{1/2}\nabla u)\cdot\tau + (\operatorname{div}(ub - \epsilon^{1/2}\sigma)v \\ &\quad + (c - \operatorname{div} b)uvdx \\ &= \int_K \sigma\cdot\tau + u\operatorname{div}(\epsilon^{1/2}\tau) - (ub - \epsilon^{1/2}\sigma)\cdot\nabla v \\ &\quad + (c - \operatorname{div} b)uvdx \\ &\quad + \int_{\partial K} n\cdot(bu - \epsilon^{1/2}\sigma)vds - \int_{\partial K} n\cdot\epsilon^{1/2}\tau uds,\end{aligned} \tag{4.9.20}$$

i.e., in this case we have

$$\begin{aligned}(\mathcal{B}[u,\sigma],[v,\tau])_K &= ([u,\sigma],\mathcal{B}'[v,\tau])_K \\ &\quad + \int_{\partial K} n\cdot(bu - \epsilon^{1/2}\sigma)vds - \int_{\partial K} n\cdot\epsilon^{1/2}\tau uds.\end{aligned} \tag{4.9.21}$$

which means

$$J_{\partial K}([u,\sigma],[v,\tau]) = \int_{\partial K} n\cdot(bu - \epsilon^{1/2}\sigma)vds - \int_{\partial K} n\cdot\epsilon^{1/2}\tau uds.$$

Remark 4.9.8 (a) Remark 4.9.4 already indicates a fundamental difficulty encountered with this procedure. As long as $[u,\sigma]$, $[v,\tau]$ are smooth all terms in (4.9.21) are well defined. But then one would like to extend these expression to spaces imposing possible weak regularity conditions on the unknown $[u,\sigma]$, namely that they need only belong to $L_2(K)\times(L_2(K))^d$ for each $K\in\mathcal{P}$. Since we have to deal with such traces for each cell K we cannot argue with boundary conditions.

(b) One could avoid this difficulty by formulating a resulting variational problem only for finite-dimensional trial- and test-spaces comprised of piecewise polynomials (which is done for standard DG-methods). But one would then give up on the rule:

- identify first a well-conditioned variational formulation of the given problem, which in particular means to identify the right (infinite-dimensional) trial- and test-spaces;

- then formulate finite-dimensional inf-sup stable formulations based on the norms for the infinite-dimensional case. □

(c) The way out chosen in the DPG concept is: since the unknowns may in general be sought in a function space for which traces on the cell boundaries are not defined, one introduces *new unknowns* $\hat{u}$ living on the *skeleton*

$$\partial\mathcal{P} := \bigcup_{K\in\mathcal{P}} \partial K. \tag{4.9.22}$$

Therefore, (4.9.19) suggests the following: find $[u,\hat{u}] \in \mathbb{U}_{\mathcal{P}}$ such that

$$B_{\mathcal{P}}([u,\hat{u}],v) = \sum_{K\in\mathcal{P}} (u,\mathcal{B}'v)_K + J_{\partial K}(\hat{u},v) = f(v), \quad v \in \mathbb{V}_{\mathcal{P}}. \tag{4.9.23}$$

Here u could have several components such as in the above example with the correspondence $u \leftrightarrow [u,\sigma]$, i.e., one has to introduce also new trace unknowns $\hat{\sigma}$ for the component σ.

The trial-space $\mathbb{U}_{\mathcal{P}}$ is a product-space with bulk-factors just being (products of) $L_2(\Omega)$. It remains to choose a suitable space for the skeleton-component. The choice of the test-space $\mathbb{V}_{\mathcal{P}}$ is again dictated by ensuring continuity of $B_{\mathcal{P}}([u,\hat{u}],v)$.

By (4.8.54) the ideal test-space would be comprised of all those v for which

$$\|v\|_{\mathbb{V}_{\rm opt}} := \|\mathcal{B}^*v\|_{\mathbb{U}'_{\mathcal{P}}} = \sup_{[u,\hat{u}]\in\mathbb{U}_{\mathcal{P}}} \frac{B_{\mathcal{P}}([u,\hat{u}],v)}{\|[u,\hat{u}]\|_{\mathbb{U}_{\mathcal{P}}}} \tag{4.9.24}$$

is finite. Although $\mathbb{U}_{\mathcal{P}}$ has $L_2(\Omega)$ as a factor the skeleton component still renders the Riesz-map $\mathcal{R}_{\mathbb{U}_{\mathcal{P}}}$ non-trivial. The typical DPG realization therefore does not use $\mathbb{V}_{\rm opt}$ but settles on a norm that is in some sense "close" but defines a very simply structured test-space, namely a so-called "broken" space

$$\begin{aligned} \mathbb{V}_{\mathcal{P}} &:= \textstyle\prod_{K\in\mathcal{P}} H(\mathcal{B};K), \\ H(\mathcal{B};K) &:= \{v \in L_2(K) : \|v\|^2_{L_2(K)} + \|\mathcal{B}'v\|^2_{L_2(K)} < \infty\}, \end{aligned} \tag{4.9.25}$$

which is a product of local spaces, endowed with the norm

$$\|v\|_{\mathbb{V}_{\mathcal{P}}}^2 = \sum_{K\in\mathcal{P}} \|v\|_{H(\mathcal{B};K)}^2. \tag{4.9.26}$$

It is not clear beforehand whether this choice renders (4.9.23) well-posed or even well-conditioned. Clearly, for localizable $\mathcal{B}$

$$\begin{aligned}|B_{\mathcal{P}}([u,\hat{u}],v)| &\le \sum_{K\in\mathcal{P}} \|u\|_{L_2(K)}\|v\|_{H(\mathcal{B};K)} + \Big|\sum_{K\in\mathcal{P}} J_{\partial K}(\hat{u},v)\Big| \\ &\le \Big(\sum_{K\in\mathcal{P}} \|u\|_{L_2(K)}^2\Big)^{1/2}\Big(\sum_{K\in\mathcal{P}} \|v\|_{H(\mathcal{B};K)}^2\Big)^{1/2} \\ &\qquad + \Big|\sum_{K\in\mathcal{P}} J_{\partial K}(\hat{u},v)\Big| \\ &= \|u\|_{L_2(\Omega)}\|v\|_{\mathbb{V}_{\mathcal{P}}} + \Big|\sum_{K\in\mathcal{P}} J_{\partial K}(\hat{u},v)\Big|. \end{aligned} \tag{4.9.27}$$

Thus, continuity now hinges on the proper choice of the skeleton space $\mathbb{U}_{\partial\mathcal{P}}$ that eventually allows one to conclude that the trace part is controled by $\|\hat{u}\|_{\mathbb{U}_{\partial\mathcal{P}}}\|v\|_{\mathbb{V}_{\mathcal{P}}}$:

$$\Big|\sum_{K\in\mathcal{P}} J_{\partial K}(\hat{u},v)\Big| \lesssim \|\hat{u}\|_{\mathbb{U}_{\partial\mathcal{P}}}\|v\|_{\mathbb{V}_{\mathcal{P}}}, \tag{4.9.28}$$

which together with (4.9.27) gives

$$\begin{aligned}|B_{\mathcal{P}}([u,\hat{u}],v)| &\lesssim \big(\|u\|_{L_2(\Omega)} + \|\hat{u}\|_{\mathbb{U}_{\partial\mathcal{P}}}\big)\|v\|_{\mathbb{V}_{\mathcal{P}}} \\ &\sim \|[u,\hat{u}]\|_{\mathbb{U}_{\mathcal{P}}}\|v\|_{\mathcal{P}}, \quad [u,\hat{u}]\in\mathbb{U}_{\mathcal{P}},\ v\in\mathbb{V}_{\mathcal{P}}.\end{aligned} \tag{4.9.29}$$

Let us suppose for the moment that we have established (4.9.29) and also the validity of an inf-sup condition. The key point that motivates the above ansatz is the fact that the trial-test-map now localizes.

Proposition 4.9.1 *Adhering to the above setting consider the local mappings $\mathcal{T}_K : \mathbb{U}_{\mathcal{P}} \to \mathbb{V}_K := H(\mathcal{B};K)$, defined by*

$$(\mathcal{T}_K[w,\hat{w}],v)_{H(\mathcal{B};K)} = B_K([w,\hat{w}],v), \quad v\in H(\mathcal{B};K), \quad K\in\mathcal{P}, \tag{4.9.30}$$

where

$$B_K([w,\hat{w}],v) := (w,\mathcal{B}'v)_K + J_{\partial K}(\hat{w},v), \quad K \in \mathcal{P}, \tag{4.9.31}$$

and $(\cdot,\cdot)_{H(\mathcal{B};K)}$ *is the inner product inducing the norm* $\|\cdot\|_{H(\mathcal{B};K)}$. *Then, the global trial-to-test map* $\mathcal{T} : \mathbb{U}_{\mathcal{P}} \to \mathbb{V}_{\mathcal{P}}$ *is given by*

$$(\mathcal{T}[w,\hat{w}],v)_{\mathbb{V}_{\mathcal{P}}} = \sum_{K\in\mathcal{P}} (\mathcal{T}_K[w,\hat{w}],v_K)_{H(\mathcal{B}';K)}. \tag{4.9.32}$$

Proof Every element of $\mathbb{V}_{\mathcal{P}}$ is of the form $v = (v_K)_{K\in\mathcal{P}}$, $v_K \in H(\mathcal{B};K)$, $K \in \mathcal{P}$ and

$$(v,z)_{\mathbb{V}_{\mathcal{P}}} = \sum_{K\in\mathcal{P}} (v_K,z_K)_{H(\mathcal{B};K)},$$

from which the assertion easily follows. ■

Remark 4.9.9 (a) The trial-to-test map is assembled by solving for each cell $K \in \mathcal{P}$ a *local* variational problem (4.9.30).

(b) A finite-dimensional problem is obtained by choosing a finite-dimensional trial space $\mathbb{U}^h_{\mathcal{P}} \subset \mathbb{U}_{\mathcal{P}}$.

(c) The optimal test-space for $\mathbb{U}^h_{\mathcal{P}}$ is then given by

$$\mathcal{T}(\mathbb{U}^h_{\mathcal{P}}) = \{\mathcal{T}([u_h,\hat{u}_h]) : [u_h,\hat{u}_h] \in \mathbb{U}^h_{\mathcal{P}}\}. \tag{4.9.33}$$

(d) Each local problem (4.9.30) is still an *infinite-dimensional* problem. A practicable version can be obtained by choosing for each K a *finite-dimensional* test-search-space $\mathbb{S}_K \subset H(\mathcal{B};K)$, large enough so that $\prod_{K\in\mathcal{P}} \mathbb{S}_K$ is δ-proximal for $\mathbb{U}^h_{\mathcal{P}}$.

(e) The projected trial-to-test-map $\mathcal{T}^h : \mathbb{U}^h_{\mathcal{P}} \to \prod_{K\in\mathcal{P}} \mathbb{S}_K$ can again be assembled from the local components

$$(\mathcal{T}^h_K[w_h,\hat{w}_h],v_K)_{H(\mathcal{B};K)} = B_K([w_h,\hat{w}_h],v_K), \quad v_K \in \mathbb{S}_K, \quad K \in \mathcal{P}. \tag{4.9.34}$$

i.e.,

$$\mathcal{T}^h_K = P_{\mathbb{S}_K} \circ \mathcal{T}_K, \quad K \in \mathcal{P}.$$

(e) The practical DPG-scheme is then of the form: find $[u_h, \hat{u}_h] \in \mathbb{U}^h_{\mathcal{P}}$ such that

$$B_{\mathcal{P}}([u_h, \hat{u}_h], v_h) = f(v_h), \quad v_h \in \mathcal{T}^h(\mathbb{U}^h_{\mathcal{P}}). \tag{4.9.35}$$

(f) Suppose that the bulk- and skeleton component of $\mathbb{U}^h_{\mathcal{P}}$ are comprised of *piecewise polynomials* of some fixed degree, and that $\dim \mathbb{S}_K = M_h$, $K \in \mathcal{P}$. Then the total computational cost of (4.9.35) scales at least as

$$\#(\mathcal{P}) M_h \sim M_h \mathrm{dim}\, \mathbb{U}^h_{\mathcal{P}},$$

i.e., it remains proportional to the size of the trial-space. However, it is not clear whether M_h can be kept uniformly bounded in $\mathcal{P}$.

(g) One can transform (4.9.35) into an equivalent mixed formulation (4.8.1) avoiding the computation of the test-functions. The system (4.8.23) is now **block-diagonal** and thus easy to solve. □

In summary, this leads to the following

DPG-strategy:

i) The DPG-method is based on a *family* of mesh-dependent infinite-dimensional variational formulations based on a hierarchy of partitions $\mathfrak{P} = \{\mathcal{P}\}$;

ii) for each of the underlying partitions $\mathcal{P}$ one selects a pair of infinite-dimensional trial- and test-spaces $\mathbb{U}_{\mathcal{P}}$, $\mathbb{V}_{\mathcal{P}}$;

iii) the spaces $\mathbb{V}_{\mathcal{P}}$ should be "broken" spaces of the form (4.9.25);

iv) one then needs to establish that the corresponsing mesh-dependent formulations are *uniformly* well-posed (or better well-conditioned) with respect to $\mathcal{P}$.

v) for given finite-dimensional trial spaces $\mathbb{U}^h_{\mathcal{P}} \subset \mathbb{U}_{\mathcal{P}}$ find *finite-dimensional* test-search spaces $\mathbb{S}_K$, $K \in \mathcal{P}$, so that the inf-sup constants of the corresponding finite-dimensional Petrov-Galerkin schemes remain bounded away from zero.

Ideal Scenario: show that uniform inf-sup stability is achieved through local test-serch spaces $\mathbb{S}_K$ of *uniformly bounded finite dimension*

$$\dim \mathbb{S}_K \le M, \quad \forall\, K \in \mathcal{P}, \forall\, \mathcal{P} \in \mathfrak{P}. \tag{4.9.36}$$

The numerical cost would then scale like $\dim \mathbb{U}^h_{\mathcal{P}}$, $\mathcal{P} \in \mathfrak{P}$.

While DPG methods have been applied to a wide spectrum of problems a rigorous justification in the sense of the *ideal scenario* is so far only available for elliptic problems and Maxwell's equation [GQ14, CDG16], and for transport equations [BDS].

5 Transport Equations

5.1 Why Transport Equations

Recall from Section 3.1.6 the pure transport equation

$$b \cdot \nabla u + cu \;=\; f \quad \text{in } \Omega,\; u = 0 \quad \text{on } \Gamma_-, \tag{5.1.1}$$

where b is a possibly variable vector field representing convection and c is (als a possibly x-dependent) reaction coefficient. For the problem to be well-posed boundary conditions can only be prescribed on the *inflow-boundary* Γ_- where

$$\Gamma_\pm := \{x \in \partial\Omega : \operatorname{sgn}\left(n(x) \cdot b(x)\right) = \pm\}.$$

There may be a non-trivial remaining boundary portion

$$\Gamma_0 := \partial\Omega \setminus (\Gamma_- \cup \Gamma_+),$$

which is called the *characteristic boundary* on which $b \cdot n = 0$, i.e., it is parallel to the flow direction b.

Remark 5.1.1 The time-dependent analog $\partial_t u + b \cdot \nabla u + cu = f$ can be treated in exactly the same manner by replacing x by $\hat{x} := (x, t)$ and b by $(b^T, 1)^t$ with a corresponding inflow-boundary of the space-time cylinder $\hat{\Omega} := \Omega \times [0, T)$. □

The interest in such simple transport equations has several sources:

- (5.1.1) arises as the singular limit for $\epsilon \to 0$ for our guiding example in the previous section.
- There exist so far no variational formulations for transport equations with tight error-residual bounds.

- This latter fact is particularly interesting in other problem classes where linear transport equations appear as core constituent. This is the case for a variety of *kinetic models* of Boltzmann-type. One such example is (the simplified) *neutron transport equation*

$$y \cdot \nabla u(x,y) + \sigma(x,y)u(x,y) - \int_{\mathcal{Y}} K(x,y,y')u(x,y')dy' = f(x) \quad \text{in}\, \Omega, \tag{5.1.2}$$

  supplemented by inflow-boundary conditions $u|_{\Gamma_-(y)} = g$. where for instance $\mathcal{Y}$ is the $(d-1)$-sphere. This describes the neutron density in the domain Ω which results from particle transport in direction y, absorbtion of particals, modeled by part of σ, and scattering, modeled by a global kernel $K(x,y,y')$. These problems are challenging since the solution is not only a function of space (and possible time) but also of the transport directions y ranging over the whole sphere. Moreover, a discretization will have densely populated matrices because of the the global kernel. To avoid the inversion of such matrices one can formulate an (infinite-dimensional) iteration where at each stage the integral operator is only applied to a current approximation and only pure transport problems need to be solved within a certain accuracy tolerance (nested iteration). This requires a-posteriori bounds.

- Last but not least, the techniques are completely different from the usual elliptic settings.

5.2 A Well-Posed Variational Formulation

As in the case of a convection-diffusion equation we have two possible variational formulations, namley multiplying the equation by test-functions and either keep it as it is which gives

$$B(u,v) := \int_{\Omega} (b \cdot \nabla u + cu)v dx, \tag{5.2.1}$$

or to apply integration by parts to obtain in analogy to (4.9.8)

$$\begin{aligned} B(u,v) &:= \int_\Omega u(-\mathrm{div}(bv)) + cuvdx + \int_{\partial\Omega} b\cdot nuvds \\ &= \int_\Omega -u(b\cdot\nabla v) + (c - \operatorname{div} b)uvdx \\ &\qquad + \int_{\partial\Omega} b\cdot nuvds, \end{aligned} \tag{5.2.2}$$

Remark 5.2.1 When setting ϵ to zero in (4.9.8) one obtains exactly (5.2.2). In fact, the only component with the test-function τ which is left is the quation $\int_\Omega \sigma\cdot\tau dx = 0$ for all $\tau \in (L_2(\Omega))^d$ which means $\sigma = 0$.□

To identify good trial- and test-spaces for (5.2.1) or (5.2.2) we are guided by making the bilinear form B continuous. It obviously matters whether directional derivatives of the test-functions are in L_2. This suggests considering the spaces

$$H(b;\Omega) := \{v \in L_2(\Omega) : \|v\|^2_{H(b;\Omega)} := \|v\|^2_{L_2(\Omega)} + \|b\cdot\nabla v\|^2_{L_2(\Omega)} < \infty\}, \tag{5.2.3}$$

and

$$H_{0,\Gamma_\pm}(b;\Omega) := \mathrm{clos}_{\|\cdot\|_{H(b;\Omega)}}\big(\{v \in C^1(\Omega) : v|_{\Gamma_\pm} = 0\}\big). \tag{5.2.4}$$

Obviously, one has for (5.2.1)

$$|B(u,v)| \le \big(1 + \|c\|^2_{L_\infty(\Omega)}\big)^{1/2}\|u\|_{H(b;\Omega)}\|v\|_{L_2(\Omega)}. \tag{5.2.5}$$

Hence, the operator $(\mathcal{B}_1 u)(v) = B(u,v)$ induced when B is considered as a bilinear form on

$$\mathbb{U}_1 = H_{0,\Gamma_-}(b;\Omega), \quad \mathbb{V}_1 = L_2(\Omega), \tag{5.2.6}$$

is obviously bounded, $\mathcal{B}_1 \in \mathcal{L}(\mathbb{U}_1, \mathbb{V}_1')$ and leads to the variational formulation: find $u \in \mathbb{U}_1 = H_{0,\Gamma_-}(b;\Omega)$ such that

$$B(u,v) = f(v), \quad v \in L_2(\Omega) = \mathbb{V}_1. \tag{5.2.7}$$

Likewise for (5.2.2)

$$\begin{aligned}|B(u,v)| &\le \|u\|_{L_2(\Omega)}\| - b\nabla v + (c - \operatorname{div} b)v\|_{L_2(\Omega)}\\ &= \|u\|_{L_2(\Omega)}\|\mathcal{B}^* v\|_{L_2(\Omega)}\\ &\le \|u\|_{L_2(\Omega)}\big(1 + (\|c\|_{L_\infty(\Omega)} + \|\operatorname{div} b\|_{L_\infty(\Omega)})^2\big)^{1/2}\|v\|_{H(b;\Omega)}, \quad (5.2.8)\end{aligned}$$

see (4.8.54).

In the case (5.2.2) one has, of course, still the same problem with the trace of $u \in L_2(\Omega)$ on $\partial\Omega$. We know that we can only prescribe boundary conditions on the inflow-boundarey Γ_-. Splitting the boundary integral in (5.2.2) as

$$\int_{\partial\Omega} b\cdot nuvds = \int_{\Gamma_-} b\cdot nuvds + \int_{\Gamma_+} b\cdot nuvds,$$

we could replace u on Γ_- in the first summand by the given boundary conditions. The second summand vanishes if we choose as test-functions only elements that vanish on the outflow boundary Γ_+. This suggests taking

$$\mathbb{U}_2 = L_2(\Omega), \quad \mathbb{V}_2 = H_{0,\Gamma_+}(b;\Omega). \qquad (5.2.9)$$

The corresponding variational formulation of (5.1.1) then becomes: find $u \in \mathbb{U}_2 = L_2(\Omega)$ such that

$$B(u,v) = -\int_{\Gamma_-} n\cdot bgvds + f(v), \quad v \in \mathbb{V}_2 = H_{0,\Gamma_+}(b;\Omega). \qquad (5.2.10)$$

Dirichlet boundary conditions become part of the variational formulation as functional on the right hand side. For well-posedness

$$g(v) := -\int_{\Gamma_-} n\cdot bgvds = \int_{\Gamma_-} |n\cdot b|gvds$$

has to belong to $\mathbb{V}_2'$. This indeed the case, due to a trace theorem that roughly says (see [DHSW12] and the literature cited there):

Trace Theorem: *If an element $v \in H(b;\Omega)$ has a a trace in the weighted L_2-space*

$$L_{2,|n\cdot b|}(\Gamma_\pm) = \{z \in L_{1,\mathrm{loc}}(\Gamma_\pm) : \int_{\Gamma_\pm} |b\cdot n|z^2 ds = \|z\|^2_{L_{2,|n\cdot b|}(\Gamma_\pm)} < \infty\}, \qquad (5.2.11)$$

then it also has a trace in $L_{2,|n\cdot b|}(\Gamma_\mp)$ *and* $\|z\|_{L_{2,|n\cdot b|}(\Gamma_\pm)} \lesssim \|z\|_{H(b;\Omega)}$.

Hence $\mathcal{B}_2$ induced by

$$B(u,v) := \int_\Omega -u(b\cdot\nabla v)+(c-\operatorname{div} b)uvdx = \int_\Omega (cv-\operatorname{div}(vb))udx \quad (5.2.12)$$

over $\mathbb{U}_2 \times \mathbb{V}_2$ belongs to $\mathcal{L}(\mathbb{U}_2, \mathbb{V}_2')$.

Inf-sup stability depends somewhat on the coefficients b, c. Here is the typical argument for conforming:

i) Show that $\mathcal{B}$ and $\mathcal{B}'$ are injective on dense subsets of their respective ranges. For instance, when $b(x)$ is a regular C^1-field (e.g. $b =$ constant would do), this can be easily shown by the method of characteristics which provides an explicit representation. Another case is that $c \geq c_0 > 0$ in Ω, see the discussion in [DHSW12, BDS].

ii) Let $\mathbb{U} = L_2(\Omega)$ and $\mathbb{V} := \operatorname{clos}_{\|\mathcal{B}^*\cdot\|_{\mathbb{U}'}}\big(\{v \in C^1(\Omega) : v|_{\Gamma_+} = 0\}\big)$ and $\|\mathcal{B}^* v\|_{\mathbb{U}'} = \|\mathcal{B}^* v\|_{L_2(\Omega)}$ is a norm on $\mathbb{V}$.

iii) Verify the Poincarè-type estimate

$$\|v\|_{L_2(\Omega)} \lesssim \|b\cdot\nabla v\|_{L_2(\Omega)}, \quad v \in H_{0,\Gamma_\pm}(b;\Omega). \quad (5.2.13)$$

iv) From (5.2.13) one easily derives that

$$\|\mathcal{B}^* v\|_{L_2(\mathbb{U})} = \| - b\cdot\nabla v + (c - \operatorname{div} b)v\|_{L_2(\Omega)} \sim \|v\|_{H(b;\Omega)} \quad (5.2.14)$$

are equivalent norms, i.e., $\mathbb{V} = H_{0,\Gamma_+}(b;\Omega)$.

v) For the optimal test-norm $\|\mathcal{B}^* v\|_{L_2(\Omega)}$ one has an inf-sup constant equal to one (see (4.8.60)). By (5.2.8), the continuity constant equals one as well. Thus, the problem is well-posed and because of (iv), it is still well-posed also when $\mathbb{V}$ is endowed with the norm $\|\cdot\|_{H(b;\Omega)}$.

vi) The argument for (5.2.7) is similar, use e.g. (4.9.18).

In summary, (5.2.10) is a well-posed conforming weak formulation of (5.1.1).

5.3 An Infinite-dimensional DPG Formulation

The DPG-formulation builds on the well-posed conforming formulations (5.2.7), (5.2.10), which we will now assume to be the case (i.e., b, c are such that the previous conclusions apply). The details of the following results can be found in [BDS]

Given a partition $\mathcal{P}$ of Ω, as in (4.9.20) we do (5.2.12) elementwise and then replace the traces of u by new unknowns $\hat{u}$ living only on the skeleton

$$\partial\mathcal{P} = \bigcup_{K\in\mathcal{P}} (\partial K_- \cup \partial K_+).$$

This yields the DPG bilinear form

$$B_{\mathcal{P}}([u,\hat{u}],v) = \int_\Omega (cv - b\cdot\nabla_{\mathcal{P}}v - v\mathrm{div}\, b)u dx + \int_{\partial\mathcal{P}} [[vb]]\hat{u} ds, \qquad (5.3.1)$$

where $\nabla_{\mathcal{P}}$ is the piecewise gradient, i.e.,

$$\int_\Omega b\cdot\nabla_{\mathcal{P}}v dx = \sum_{K\in\mathcal{P}} b\cdot\nabla v dx,$$

and where for $x \in K \cap K'$ (interface f neighboring closed cells K, K')

$$[[vb]](x) := (vb|_K \cdot n_K + vb|_{K'} \cdot n_{K'})(x)$$

are jump terms.

Choice of trial- and test space: As in (4.9.26) as a test-space we take the "broken analog" to (5.2.14), i.e.,

$$\|v\|_{\mathbb{V}_{\mathcal{P}}} = \|v\|_{H(b;\mathcal{P})} := \Big(\sum_{K\in\mathcal{P}} \|v_K\|^2_{H(b;K)}\Big)^{1/2}. \qquad (5.3.2)$$

Clearly, the bulk-unknown u should belong to $L_2(\Omega)$. As indicated earlier, it is important to choose the right norm for the skeleton component $\hat{u}$ to ensure continuity of $B_{\mathcal{P}}$. To this end, consider the space

$$H_{0,\Gamma_-}(b;\partial\mathcal{P}) := \{w|_{\partial\mathcal{P}} : w \in H_{0,\Gamma_-}(b;\Omega)\}, \qquad (5.3.3)$$

i.e., the skeleton functions is viewed as restrictions of the elements in the conforming infinite-dimensional trial space $\mathbb{U}_1 = H_{0,\Gamma_-}(b;\Omega)$ while the DPG-formulation for the bulk-unknown u is inspired by the weak formulation (5.2.10). In this sense the DPG-formulation draws on both types of conforming formulations. The space $H_{0,\Gamma_-}(b;\partial\mathcal{P})$ is endowed with the *factor-norm*

$$\|\hat{u}\|_{H_{0,\Gamma_-}(b;\partial\mathcal{P})} := \inf\{\|w\|_{H(b;\Omega)} : w|_{\partial\mathcal{P}} = \hat{u},\, w \in H_{0,\Gamma_-}(b;\Omega)\}. \tag{5.3.4}$$

Thus, we arrive at

$$\mathbb{U}_{\mathcal{P}} = L_2(\Omega) \times H_{0,\Gamma_-}(b;\partial\mathcal{P}), \quad \mathbb{V}_{\mathcal{P}} = H(b;\mathcal{P}), \tag{5.3.5}$$

and

$$\mathcal{B}_{\mathcal{P}} : \mathbb{U}_{\mathcal{P}} \to \mathbb{V}_{\mathcal{P}} \text{ defined by } (\mathcal{B}_{\mathcal{P}}[u,\hat{u}])(v) = B_{\mathcal{P}}([u,\hat{u}],v) \tag{5.3.6}$$

is clearly bounded.

Well-posedness:

Theorem 5.3.1 ([BDS, Theorem 3.1]) *Assume that* $c \in L_\infty(\Omega), \operatorname{div} b \in L_\infty(\Omega)$ *and that for these coefficients* (5.2.7) *and* (5.2.10) *are well-posed. Then* $\mathcal{B}_{\mathcal{P}}$ *defined by* (**??**) *is an isomorphism, i.e.,*

$$\|\mathcal{B}_{\mathcal{P}}\|_{\mathcal{L}(\mathbb{U}_{\mathcal{P}},\mathbb{V}'_{\mathcal{P}})} \leq C_{\mathcal{B}_{\mathcal{P}}}, \quad \|\mathcal{B}_{\mathcal{P}}^{-1}\|_{\mathcal{L}(\mathbb{V}'_{\mathcal{P}},\mathbb{U}_{\mathcal{P}})} \leq c_{\mathcal{B}_{\mathcal{P}}}^{-1}, \tag{5.3.7}$$

where $C_{\mathcal{B}_{\mathcal{P}}}$ *depends on* c, b *while* $c_{\mathcal{B}_{\mathcal{P}}}$ *depends in addition on* $\|(\mathcal{B}_2^*)^{-1}\|_{\mathcal{L}(\mathbb{U}'_2,\mathbb{V}_2)}$ *and* $\|\mathcal{B}_1^{-1}\|_{\mathcal{L}(\mathbb{V}'_1,\mathbb{U}_1)}$. □

The theorem says that the *infinite-dimensional* DPG formulations are inf-.sup stable *uniformly* in a given hierarchy of shape-regular partitions $\mathcal{P} \in \mathfrak{P}$.

An important ingredient in the proof is to show that (see [BDS, Lemma 3.4])

$$\begin{gathered}[[vb]] \in (H_{0,\Gamma_-}(b;\partial\mathcal{P}))', \\ \|[[vb]]\|_{(H_{0,\Gamma_-}(b;\partial\mathcal{P}))'} \sim \inf_{z \in H_{0,\Gamma_+}(b;\Omega)} \|v - z\|_{H(b;\mathcal{P})},\end{gathered} \tag{5.3.8}$$

where the equivalence constants depend only on c, b and $\|\mathcal{B}_1^{-1}\|_{\mathcal{L}(\mathbb{V}'_1,\mathbb{U}_1)}$.

5.4 A Fully Discrete DPG Scheme

A practical DPG-scheme is obtained by choosing suitable finite dimensional trial spaces $\mathbb{U}_{\mathcal{P}}^m \subset \mathbb{U}_{\mathcal{P}}$ and appropriate test-search spaces $\mathbb{S}_{\mathcal{P}}^m \subset \mathbb{V}_{\mathcal{P}}$. Specifically, take

$$\mathbb{U}_{\mathcal{P}}^m := \prod_{K\in\mathcal{P}} \mathbb{P}_m(K) \times H_{0,\Gamma_-}(b;\Omega) \cap \prod_{K\in\mathcal{P}} \mathbb{P}_m(K). \tag{5.4.1}$$

For the test-search space consider a refinement $\mathcal{P}_r$ of $\mathcal{P}$ of fixed depth $r \in \mathbb{N}$, i.e., each cell in $\mathcal{P}$ is subdivided r times according to the refinement rule in $\mathfrak{P}$. Then take

$$\mathbb{S}_{\mathcal{P}}^m := \prod_{K'\in\mathcal{P}_r} \mathbb{P}_{m+1}(K') \subset \mathbb{V}_{\mathcal{P}_r}. \tag{5.4.2}$$

The test-space for the corresponding Petrov-Galerkin scheme is then given by

$$\mathcal{T}_{\mathbb{S}_{\mathcal{P}}^m}(\mathbb{U}_{\mathcal{P}}^m) = \prod_{K'\in\mathcal{P}_r} \mathcal{T}_{K,\mathbb{S}_{\mathcal{P}}^m}(\mathbb{U}_{\mathcal{P}}^m), \tag{5.4.3}$$

i.e., the test-functions for any local basis elements in $\mathbb{U}_{\mathcal{P}}^m$ are obtained by projection from a local test-search space of uniformly bounded dimension depending on the subgrid-depth r and the polynomial degree m of the trial functions. Thus the overall problem size scales like $\dim \mathbb{U}_{\mathcal{P}}^m$.

Theorem 5.4.1 ([BDS, Theorem 4.8]) *Assume that $c \in W^1(L_\infty(\Omega))$, $b \in W^1(L_\infty(\Omega; \mathrm{div})$, $|b|^{-1} \in L_\infty(\Omega)$. Let $\mathfrak{P}$ be a hierarchy of shape-regular partitions. Then for a fixed by sufficiently large subgrid-depth r the DPG scheme: find $[u_m, \hat{u}_m] \in \mathbb{U}_{\mathcal{P}}^m$ such that*

$$B_{\mathcal{P}_r}([u_m, \hat{u}_m], v_m) = f(v_m), \quad v_m \in \mathcal{T}_{\mathbb{S}_{\mathcal{P}}^m}(\mathbb{U}_{\mathcal{P}}^m),$$

is uniformly inf-sup-stable. □

Roadmap for the proof:

- When b is a constant field and c is constant, one can determine the exact optimal test-functions explicitly when using a somewhat modified but equivalent inner product on $H(b; K)$.

- One then proceeds to show that these test-functions can be approximated in the δ-proximal sense well enough by piecewise polynomials of degree one higher that the trial degree on a refined subgrid.

- To treat variable convection fields, one uses an elaborate perturbation argument, replacing b cell-wise by a constrant field. The difficulty then is that on the infinite-dimensional level a piecewise constant b does not give rise to a well-posed problem. The fact that the relevant function spaces depend on b makes things rather delicate.

Bibliography

[Ada75] Robert Alexander Adams. *Sobolev Spaces*, volume 65 of *Pure and applied mathematics.* Academic Press, New York, 1975.

[Alt85] Hans Wilhelm Alt. *Lineare Funktionalanalysis, eine anwendungsorientierte Einführung.* Springer, Berlin, 1985.

[AU10] Wolfgang Arendt and Karsten Urban. *Partielle Differentialgleichungen. Eine Einfürung in analytische und numerische Methoden.* Spektrum Akademischer Verlag, Heidelberg, 2010.

[BDS] D. Broersen, W. Dahmen, and R.P. Stevenson. On the stability of DPG formulations of transport equations. *To appear in Math. Comp.*

[BF91] Franco Brezzi and Michel Fortin. *Mixed and Hybrid Finite Element Methods*, volume 15 of *Series in Computational Mathematics.* Springer, Berlin, 1991.

[BL76] Jöran Bergh and Jörgen Löfström. *Interpolation Spaces. An Introduction.*, volume 223 of *Grundlehren der mathematischen Wissenschaften.* Springer, Berlin, 1976.

[BM84] J.W. Barrett and K.W. Morton. Approximate symmetrization and petrov-galerkin methods for diffusion-convection problems. *Comput. Method. Appl. M.*, 45:97–12, 1984.

[CDG16] C Carstensen, L. Demkowicz, and J. Gopalakrishnan. Breaking spaces and forms for the DPG method and applications including Maxwell equations. *Comput. Math. Appl.*, 72(3):494–522, 2016.

[CDW12] Albert Cohen, Wolfgang Dahmen, and Gerrit Welper. Adaptivity and variational stabilization for convection-diffusion equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46(5):1247–1273, 2012.

[DG11] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. II Optimal test functions. *Numer. Methods Partial Differential Equations*, 27(1):70–105, 2011.

[DHSW12] Wolfgang Dahmen, Chunyan Huang, Christoph Schwab, and Gerrit Welper. Adaptive petrov-galerkin methods for first order transport equations. *SIAM J. Numer. Anal.*, 50(5):24202445, 2012.

[Eva98] Lawrence Craig Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Matematics*. American Mathematical Society, Providence, RI, 1998.

[GQ14] J. Gopalakrishnan and W. Qiu. An analysis of the practical DPG method. *Math. Comp.*, 83(286):537–552, 2014.

[Ha06] Dzung Minh Ha. *Functional Analysis, A Gentle Introduction.* Matrix Editions, Ithaca, New York, 2006.

[Szy06] Daniel B. Szyld. The many proofs of an identity on the norm of oblique projections. *Numer. Algor.*, 42:309–323, 2006.