

## Weak Solutions, Elliptic Problems and Sobolev Spaces

### 3.1 Introduction

In Chapter 2 we discussed difference methods for the numerical treatment of partial differential equations. The basic idea of these methods was to use information from a discrete set of points to approximate derivatives by difference quotients.

Now we start to discuss a different class of discretization methods: the so-called *ansatz methods*. An ansatz method is characterized by prescribing some approximate solution in a certain form. In general, this is done by determining the coefficients in a linear combination of a set of functions chosen by the numerical analyst. One cannot then expect to get an exact solution of the differential equation in all cases. Thus a possible strategy is to determine the coefficients in a way that approximately satisfies the differential equation (and perhaps some additional conditions).

For instance, one can require that the differential equation be satisfied at a specified discrete set of points; this method is called collocation. It is, however, much more popular to use methods that are based on a weak formulation of the given problem. Methods of this type do not assume that the differential equation holds at every point. Instead, they are based on a related variational problem or variational equation. The linear forms defined by the integrals in the variational formulation require the use of appropriate function spaces to guarantee, for instance, the existence of weak solutions. It turns out that existence theorems for weak solutions are valid under assumptions that are much more realistic than in the corresponding theorems for classical solutions. Moreover, ansatz functions can have much less smoothness than, e.g., functions used in collocation methods where the pointwise validity of the differential equation is required.

As a first simple example let us consider the two-point boundary value problem

$$-u''(x) + b(x)u'(x) + c(x)u(x) = f(x) \quad \text{in} \quad \Omega := (0, 1), \quad (1.1)$$

$$u(0) = u(1) = 0. \quad (1.2)$$

Let  $b$ ,  $c$  and  $f$  be given continuous functions. Assume that a classical solution exists, i.e., a twice continuously differentiable function  $u$  that satisfies (1.1) and (1.2). Then for an arbitrary continuous function  $v$  we have

$$\int_{\Omega} (-u'' + bu' + cu)v \, dx = \int_{\Omega} f v \, dx. \quad (1.3)$$

The reverse implication is also valid: if a function  $u \in C^2(\bar{\Omega})$  satisfies equation (1.3) for all  $v \in C(\bar{\Omega})$ , then  $u$  is a classical solution of the differential equation (1.1).

If  $v \in C^1(\bar{\Omega})$ , then we can integrate by parts in (1.3) and obtain

$$-u'v \Big|_{x=0} + \int_{\Omega} u'v' \, dx + \int_{\Omega} (bu' + cu)v \, dx = \int_{\Omega} f v \, dx.$$

Under the additional condition  $v(0) = v(1) = 0$  this is equivalent to

$$\int_{\Omega} u'v' \, dx + \int_{\Omega} (bu' + cu)v \, dx = \int_{\Omega} f v \, dx. \quad (1.4)$$

Unlike (1.1) or (1.3), equation (1.4) still makes sense if we know only that  $u \in C^1(\bar{\Omega})$ . But we have not yet specified a topological space in which mappings implicitly defined by a weak form of (1.1) such as (1.4) have desirable properties like continuity, boundedness, etc. It turns out that Sobolev spaces, which generalize  $L_p$  spaces to spaces of functions whose generalized derivatives also lie in  $L_p$ , are the correct setting in which to examine weak formulations of differential equations. The book [Ada75] presents an excellent general survey of Sobolev spaces. In Section 3.2 we shall give some basic properties of Sobolev spaces that will allow us to analyse discretization methods—at least in standard situations.

But first we explain the relationship of the simple model problem (1.1), (1.2) to variational problems. Assume that  $b(x) \equiv 0$  and  $c(x) \geq 0$ . Define a functional  $J$  by

$$J(u) := \frac{1}{2} \int_{\Omega} (u'^2 + cu^2) \, dx - \int_{\Omega} f u \, dx. \quad (1.5)$$

Consider now the following problem: Find a function  $u \in C^1(\bar{\Omega})$  with  $u(0) = u(1) = 0$  such that

$$J(u) \leq J(v) \quad \text{for all } v \in C^1(\bar{\Omega}) \text{ with } v(0) = v(1) = 0. \quad (1.6)$$

For such problems a necessary condition for optimality is well known: the first variation  $\delta J(u, v)$  must vanish for arbitrarily admissible directions  $v$  (see, for

instance, [Zei90]). This first variation is defined by  $\delta J(u, v) := \Phi'(0)$  where  $\Phi(t) := J(u + tv)$  for fixed  $u, v$  and real  $t$ .

For the functional  $J(\cdot)$  defined by (1.5), one has

$$J(u + tv) = \frac{1}{2} \int_{\Omega} [(u' + tv')^2 + c(u + tv)^2] dx - \int_{\Omega} f \cdot (u + tv) dx,$$

and consequently

$$\Phi'(0) = \int_{\Omega} u' v' dx + \int_{\Omega} c uv dx - \int_{\Omega} f v dx.$$

Thus in the case  $b(x) \equiv 0$ , the condition  $\delta J(u, v) = 0$  necessary for optimality in (1.6) is equivalent to the *variational equation* (1.4). This equivalence establishes a close connection between boundary value problems and variational problems. The differential equation (1.1) is described as the Euler equation of the variational problem (1.6). The derivation of Euler equations for general variational problems that are related to boundary value problems is discussed in [Zei90]. Later we shall discuss in more detail the role played by the condition  $v(0) = v(1) = 0$  in the formulation of (1.4).

Variational problems often appear when modelling applied problems in the natural and technical sciences because in many situations nature follows minimum or maximum laws such as the principle of minimum energy.

Next we consider a simple elliptic model problem in two dimensions. Let  $\Omega \subset \mathbb{R}^2$  be a simply connected open set with a (piecewise) smooth boundary  $\Gamma$ . Let  $f : \Omega \rightarrow \mathbb{R}$  be a given function. We seek a twice differentiable function  $u$  that satisfies

$$-\Delta u(\xi, \eta) = f(\xi, \eta) \quad \text{in } \Omega, \quad (1.7)$$

$$u|_{\Gamma} = 0. \quad (1.8)$$

To derive a variational equation, we again take a continuous function  $v$ , multiply (1.7) by  $v$  and integrate:

$$-\int_{\Omega} \Delta u v dx = \int_{\Omega} f v dx.$$

If  $v \in C^1(\bar{\Omega})$  with  $v|_{\Gamma} = 0$ , the application of an integral theorem (the two-dimensional analogue of integration by parts—see the next section for details) yields

$$\int_{\Omega} \left( \frac{\partial u}{\partial \xi} \frac{\partial v}{\partial \xi} + \frac{\partial u}{\partial \eta} \frac{\partial v}{\partial \eta} \right) dx = \int_{\Omega} f v dx. \quad (1.9)$$

This is the variational equation derived from the boundary value problem (1.7), (1.8). In the opposite direction, assuming  $u \in C^2(\bar{\Omega})$ , we can infer from (1.9) that  $u$  satisfies the Poisson equation (1.7).

So far we have said nothing about the existence and uniqueness of solutions in our new variational formulation of boundary value problems, because to deal adequately with these topics it is necessary to work in the framework of Sobolev spaces. In the following sections we introduce these spaces and discuss not only existence and uniqueness of solutions to variational problems but also the numerical approximation of these solutions by means of certain ansatz functions.

### 3.2 Function Spaces for the Variational Formulation of Boundary Value Problems

In the classical treatment of differential equations, the solution and certain of its derivatives are required to be continuous functions. One therefore works in the spaces  $C^k(\bar{\Omega})$  that contain functions with continuous derivatives up to order  $k$  on the given domain  $\Omega$ , or in spaces where these derivatives are Hölder continuous.

When the strong form (e.g. (1.7)) of a differential equation is replaced by a variational formulation, then instead of pointwise differentiability we need only ensure the existence of some integrals that contain the unknown function as certain derivatives. Thus it makes sense to use function spaces that are specially suited to this situation.

We start with some basic facts from functional analysis.

Let  $U$  be a linear (vector) space. A mapping  $\|\cdot\| : U \rightarrow \mathbb{R}$  is called a *norm* if it has the following properties:

- i)  $\|u\| \geq 0$  for all  $u \in U$ ,  $\|u\| = 0 \Leftrightarrow u = 0$ ,
- ii)  $\|\lambda u\| = |\lambda| \|u\|$  for all  $u \in U$ ,  $\lambda \in \mathbb{R}$ ,
- iii)  $\|u + v\| \leq \|u\| + \|v\|$  for all  $u, v \in U$ .

A linear space  $U$  endowed with a norm is called a *normed space*. A sequence  $\{u^k\}$  in a normed space is a *Cauchy sequence* if for each  $\varepsilon > 0$  there exists a number  $N(\varepsilon)$  such that

$$\|u^k - u^l\| \leq \varepsilon \quad \text{for all } k, l \geq N(\varepsilon).$$

The next property is of fundamental importance both in existence theorems for solutions of variational problems and in proofs of convergence of numerical methods. A normed space is called *complete* if every Cauchy sequence  $\{u^k\} \subset U$  converges in  $U$ , i.e., there exists a  $u \in U$  with

$$\lim_{k \rightarrow \infty} \|u^k - u\| = 0.$$

Equivalently,

$$u = \lim_{k \rightarrow \infty} u^k.$$

Complete normed spaces are often called *Banach spaces*.

Let  $U, V$  be two normed spaces with norms  $\|\cdot\|_U$  and  $\|\cdot\|_V$  respectively. A mapping  $P : U \rightarrow V$  is continuous at  $u \in U$  if for any sequence  $\{u^k\} \subset U$  converging to  $u$  one has

$$\lim_{k \rightarrow \infty} Pu^k = Pu,$$

i.e.,

$$\lim_{k \rightarrow \infty} \|u^k - u\|_U = 0 \quad \Rightarrow \quad \lim_{k \rightarrow \infty} \|Pu^k - Pu\|_V = 0.$$

A mapping is continuous if it is continuous at every point  $u \in U$ .

A mapping  $P : U \rightarrow V$  is called *linear* if

$$P(\lambda u + \mu v) = \lambda Pu + \mu Pv \text{ for all } u, v \in U, \quad \lambda, \mu \in \mathbb{R}.$$

A linear mapping is *continuous* if there exists a constant  $M \geq 0$  such that

$$\|Pu\| \leq M\|u\| \text{ for all } u \in U.$$

A mapping  $f : U \rightarrow \mathbb{R}$  is usually called a functional. Consider the set of all continuous linear functionals  $f : U \rightarrow \mathbb{R}$ . These form a normed space with norm defined by

$$\|f\|_* := \sup_{v \neq 0} \frac{|f(v)|}{\|v\|}.$$

This space is in fact a Banach space. It is the *dual space*  $U^*$  of  $U$ . When  $f \in U^*$  and  $u \in U$  we shall sometimes write  $\langle f, u \rangle$  instead of  $f(u)$ .

Occasionally it is useful to replace convergence in the normed space by convergence in a weaker sense: if

$$\lim_{k \rightarrow \infty} \langle f, u^k \rangle = \langle f, u \rangle \text{ for all } f \in U^*$$

for a sequence  $\{u^k\} \subset U$  and  $u \in U$ , then we say that the sequence  $\{u^k\}$  *converges weakly* to  $u$ . It is standard to use the notation

$$u^k \rightharpoonup u \quad \text{for } k \rightarrow \infty$$

to denote weak convergence. If  $u = \lim_{k \rightarrow \infty} u^k$  then  $u^k \rightharpoonup u$ , i.e. convergence implies weak convergence, but the converse is false: a weakly convergent sequence is not necessarily convergent.

It is particularly convenient to work in linear spaces that are endowed with a scalar product. A mapping  $(\cdot, \cdot) : U \times U \rightarrow \mathbb{R}$  is called a (real-valued) *scalar product* if  $U$  has the following properties:

- i)  $(u, u) \geq 0$  for all  $u \in U$ ,  $(u, u) = 0 \Leftrightarrow u = 0$ ,
- ii)  $(\lambda u, v) = \lambda(u, v)$  for all  $u, v \in U$ ,  $\lambda \in \mathbb{R}$ ,
- iii)  $(u, v) = (v, u)$  for all  $u, v \in U$ ,
- iv)  $(u + v, w) = (u, w) + (v, w)$  for all  $u, v, w \in U$ .

Given a scalar product, one can define an induced norm by  $\|u\| := \sqrt{(u, u)}$ . But not all norms are induced by related scalar products.

A Banach space in which the norm is induced by a scalar product is called a (real) *Hilbert space*. From the properties of the scalar product one can deduce the useful *Cauchy-Schwarz inequality*:

$$|(u, v)| \leq \|u\| \|v\| \quad \text{for all } u, v \in U.$$

Continuous linear functionals on Hilbert spaces have a relatively simple structure that is important in many applications. It is stated in the next result.

**Theorem 3.1 (Riesz).** *Let  $f : V \rightarrow \mathbb{R}$  be a continuous linear functional on a Hilbert space  $V$ . Then there exists a unique  $w \in V$  such that*

$$(w, v) = f(v) \quad \text{for all } v \in V.$$

Moreover, one has  $\|f\|_* = \|w\|$ .

The Lebesgue spaces of integrable functions are the starting point for the construction of the Sobolev spaces. Let  $\Omega \subset \mathbb{R}^n$  (for  $n = 1, 2, 3$ ) be a bounded domain (i.e., open and connected) with boundary  $\Gamma := \partial\Omega$ . Let  $p \in [1, +\infty)$ . The class of all functions whose  $p$ -th power is integrable on  $\Omega$  is denoted by

$$L_p(\Omega) := \left\{ v : \int_{\Omega} |v(x)|^p dx < +\infty \right\}.$$

Furthermore

$$\|v\|_{L_p(\Omega)} := \left[ \int_{\Omega} |v(x)|^p dx \right]^{1/p}$$

is a norm on  $L_p$ . It is important to remember that we work with Lebesgue integrals (see, e.g., [Wlo87]), so all functions that differ only on a set of measure zero are identified. It is in this sense that  $\|v\| = 0$  implies  $v = 0$ . Moreover, the space  $L_p(\Omega)$  is complete, i.e., is a Banach space.

In the case  $p = 2$  the integral

$$(u, v) := \int_{\Omega} u(x) v(x) dx$$

defines a scalar product, so  $L_2(\Omega)$  is a Hilbert space.

The definition of these spaces can be extended to the case  $p = \infty$  with

$$L_{\infty}(\Omega) := \left\{ v : \operatorname{ess\,sup}_{x \in \Omega} |v(x)| < +\infty \right\}$$

and associated norm

$$\|v\|_{L_{\infty}(\Omega)} := \operatorname{ess\,sup}_{x \in \Omega} |v(x)|.$$

Here  $\text{ess sup}$  denotes the essential supremum, i.e., the lowest upper bound over  $\Omega$  excluding subsets of  $\Omega$  of Lebesgue measure zero.

To treat differential equations, the next step is to introduce derivatives into the definitions of suitable spaces. In extending the Lebesgue spaces to Sobolev spaces one needs generalized derivatives, which we now describe. For the reader familiar with derivatives in the sense of distributions this introduction will be straightforward.

Denote by  $cl_V A$  the closure of a subset  $A \subset V$  with respect to the topology of the space  $V$ . For  $v \in C(\bar{\Omega})$  the *support* of  $v$  is then defined by

$$\text{supp } v := cl_{\mathbb{R}^n} \{x \in \Omega : v(x) \neq 0\}.$$

For our bounded domain  $\Omega$ , set

$$C_0^\infty(\Omega) := \{v \in C^\infty(\Omega) : \text{supp } v \subset \Omega\}.$$

In our further considerations the role of integration by parts in several dimensions is very important. For instance, for arbitrary  $u \in C^1(\bar{\Omega})$  and  $v \in C_0^\infty(\Omega)$  one has

$$\int_{\Omega} \frac{\partial u}{\partial x_i} v \, dx = \int_{\Gamma} uv \cos(n, e^i) \, ds - \int_{\Omega} u \frac{\partial v}{\partial x_i} \, dx$$

where  $e^i$  is the unit vector in the  $i$ th coordinate direction and  $n$  is the outward-pointing unit vector normal to  $\Gamma$ . Taking into account that  $v|_{\Gamma} = 0$  we get

$$\int_{\Omega} u \frac{\partial v}{\partial x_i} \, dx = - \int_{\Omega} \frac{\partial u}{\partial x_i} v \, dx. \quad (2.1)$$

This identity is the starting point for the generalization of standard derivatives on Lebesgue spaces.

First we need more notation. To describe partial derivatives one uses a multi-index  $\alpha := (\alpha_1, \dots, \alpha_n)$  where each  $\alpha_i$  is a non-negative integer. Set  $|\alpha| = \sum_i \alpha_i$ . We introduce

$$D^\alpha u := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}} u$$

for the derivative of order  $|\alpha|$  with respect to the multi-index  $\alpha$ .

Now, recalling (2.1), we say that an integrable function  $u$  is in a generalized sense differentiable with respect to the multi-index  $\alpha$  if there exists an integrable function  $w$  with

$$\int_{\Omega} u D^\alpha v \, dx = (-1)^{|\alpha|} \int_{\Omega} w v \, dx \quad \text{for all } v \in C_0^\infty(\Omega). \quad (2.2)$$

The function  $D^\alpha u := w$  is called the *generalized derivative* of  $u$  with respect to the multi-index  $\alpha$ .

Applying this definition to each first-order coordinate derivative, we obtain a generalized gradient  $\nabla u$ . Furthermore, if for a componentwise integrable vector-valued function  $\underline{u}$  there exists a integrable function  $z$  with

$$\int_{\Omega} \underline{u} \nabla v \, dx = - \int_{\Omega} z v \, dx \quad \text{for all } v \in C_0^\infty(\Omega),$$

then we call  $z$  the *generalized divergence* of  $\underline{u}$  and we write  $\operatorname{div} \underline{u} := z$ .

Now we are ready to define the Sobolev spaces. Let  $l$  be a non-negative integer. Let  $p \in [2, \infty)$ . Consider the subspace of all functions from  $L_p(\Omega)$  whose generalized derivatives up to order  $l$  exist and belong to  $L_p(\Omega)$ . This subspace is called the *Sobolev space*  $W_p^l(\Omega)$  (Sobolev, 1938). The norm in  $W_p^l(\Omega)$  is chosen to be

$$\|u\|_{W_p^l(\Omega)} := \left[ \int_{\Omega} \sum_{|\alpha| \leq l} |[D^\alpha u](x)|^p \, dx \right]^{1/p}. \quad (2.3)$$

Starting from  $L_\infty(\Omega)$ , the Sobolev space  $W_\infty^l(\Omega)$  is defined analogously.

Today it is known that Sobolev spaces can be defined in several other equivalent ways. For instance, Meyers and Serrin (1964) proved the following (see [Ada75]):

$$\text{For } 1 \leq p < \infty \text{ the space } C^\infty(\Omega) \cap W_p^l(\Omega) \text{ is dense in } W_p^l(\Omega). \quad (2.4)$$

That is, for these values of  $p$  the space  $W_p^l(\Omega)$  can be generated by completing the space  $C^\infty(\Omega)$  with respect to the norm defined by (2.3). In other words,

$$W_p^l(\Omega) = \operatorname{cl}_{W_p^l(\Omega)} C^\infty(\Omega).$$

This result makes clear that one can approximate functions in Sobolev spaces by functions that are differentiable in the classical sense. Hence, various desirable properties of Sobolev spaces can be proved by first verifying them for classical functions and then using the above density identity to extend them to Sobolev spaces.

When  $p = 2$  the spaces  $W_p^l(\Omega)$  are Hilbert spaces with scalar product

$$(u, v) = \int_{\Omega} \left( \sum_{|\alpha| \leq l} D^\alpha u D^\alpha v \right) \, dx. \quad (2.5)$$

It is standard to use the notation  $H^l(\Omega)$  in this case, i.e.,  $H^l(\Omega) = W_2^l(\Omega)$ . In the treatment of second-order elliptic boundary value problems the Sobolev spaces  $H^1(\Omega)$  play a fundamental role, while for fourth-order elliptic problems one uses the spaces  $H^2(\Omega)$ .

If additional boundary conditions come into the game, then additional information concerning certain subspaces of these Sobolev spaces is required. Let us first introduce the spaces



$$\mathring{W}_p^l(\Omega) := cl_{W_p^l(\Omega)} C_0^\infty(\Omega).$$

In the case  $p = 2$  these spaces are Hilbert spaces with the same scalar product as in (2.5), and they are denoted by  $H_0^l(\Omega)$ . When  $l = 1$  this space can be considered as a subspace of  $H^1(\Omega)$  comprising those functions that vanish (in a certain sense) on the boundary  $\Gamma$ ; we shall explain this in detail later in our discussion of traces following Lemma 3.3. The standard notation for the dual spaces of the Sobolev spaces  $H_0^l(\Omega)$  is

$$H^{-l}(\Omega) := (H_0^l(\Omega))^*. \quad (2.6)$$

In some types of variational inequalities—for instance, in mixed formulations of numerical methods—we shall also need special spaces of vector-valued functions. As an example we introduce

$$H(\operatorname{div}; \Omega) := \{ \underline{u} \in L_2(\Omega)^n : \operatorname{div} \underline{u} \in L_2(\Omega) \} \quad (2.7)$$

with

$$\|\underline{u}\|_{\operatorname{div}, \Omega}^2 := \|\underline{u}\|_{H(\operatorname{div}; \Omega)}^2 := \sum_{i=1}^n \|u_i\|_{L_2(\Omega)}^2 + \|\operatorname{div} \underline{u}\|_{L_2(\Omega)}^2. \quad (2.8)$$

Now we begin to use Sobolev spaces in the weak formulation of boundary value problems. Our first example is the Poisson equation (1.7) with homogeneous Dirichlet boundary conditions (1.8). The *variational problem* related to that example can be stated precisely in the following way:

Find  $u \in H_0^1(\Omega)$  such that

$$\int_{\Omega} \left( \frac{\partial u}{\partial x_1} \frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2} \frac{\partial v}{\partial x_2} \right) dx = \int_{\Omega} f v dx \quad \text{for all } v \in H_0^1(\Omega). \quad (2.9)$$

The derivatives here are generalized derivatives and the choice of spaces is made to ensure the existence of all integrals. Every classical solution of the Dirichlet problem (1.7), (1.8) satisfies the variational equation (2.9), as we already saw in (1.9), using integration by parts. But is a weak solution in the sense of (2.9) also a classical solution? To answer this question we need further properties of Sobolev spaces. In particular we need to investigate the following:

- What classical differentiability properties does the *weak solution* of the variational problem (2.9) possess?
- In what sense does the weak solution satisfy the boundary conditions?

To address these issues one needs theorems on regularity, embedding and traces for Sobolev spaces, which we now discuss.

The validity of embedding and trace theorems depends strongly on the properties of the boundary of the given domain. It is not our aim to discuss

here minimal boundary assumptions for these theorems, as this is a delicate task; results in many cases can be found in [Ada75].

*Here we assume generally—as already described in Chapter 1.3—that for every point of the boundary  $\partial\Omega$  there exists a local coordinate system in which the boundary corresponds to some hypersurface with the domain  $\Omega$  lying on one side of that surface. The regularity class of the boundary (and domain) is defined by the smoothness of the boundary's parametrization in this coordinate system: we distinguish between Lipschitz,  $C^k$  and  $C^\infty$  boundaries and domains.*

Many other characterizations of boundaries are also possible.

In most practical applications it is sufficient to consider Lipschitz domains. In two dimensions, a polygonal domain is Lipschitz if all its interior angles are less than  $2\pi$ , i.e., if the domain contains no slits.

Let  $U, V$  be normed spaces with norms  $\|\cdot\|_U$  and  $\|\cdot\|_V$ . We say the space  $U$  is *continuously embedded* into  $V$  if  $u \in V$  for all  $u \in U$  and moreover there exists a constant  $c > 0$  such that

$$\|u\|_V \leq c \|u\|_U \quad \text{for all } u \in U. \quad (2.10)$$

Symbolically, we write  $U \hookrightarrow V$  for the continuous embedding of  $U$  into  $V$ . The constant  $c$  in inequality (2.10) is called the embedding constant.

The obvious embedding

$$W_p^l(\Omega) \hookrightarrow L_p(\Omega) \quad \text{for every integer } l \geq 0$$

is a direct consequence of the definitions of the spaces  $W_p^l(\Omega)$  and  $L_p(\Omega)$  and their norms. It is more interesting to study the imbedding relations between different Sobolev spaces or between Sobolev spaces and the classical spaces  $C^k(\bar{\Omega})$  and  $C^{k,\beta}(\bar{\Omega})$  with  $\beta \in (0, 1)$ . The corresponding norms are

$$\begin{aligned} \|v\|_{C^k(\bar{\Omega})} &= \sum_{|\alpha| \leq k} \max_{x \in \bar{\Omega}} |D^\alpha v(x)|, \\ \|v\|_{C^{k,\beta}(\bar{\Omega})} &= \|v\|_{C^k(\bar{\Omega})} + \sum_{|\alpha|=k} |D^\alpha v|_{C^\beta(\bar{\Omega})} \end{aligned}$$

with the Hölder seminorm

$$|v|_{C^\beta(\bar{\Omega})} = \inf \{ c : |v(x) - v(y)| \leq c|x - y|^\beta \text{ for all } x, y \in \bar{\Omega} \}.$$

Then one has the following important theorem (see [Ada75, Wlo87]):

**Theorem 3.2 (Embedding theorem).** *Let  $\Omega \subset \mathbb{R}^n$  be a bounded domain with Lipschitz boundary. Assume that  $0 \leq j \leq k$ ,  $1 \leq p, q < +\infty$  and  $0 < \beta < 1$ .*

*i) For  $k - j \geq n(\frac{1}{p} - \frac{1}{q})$  one has the continuous embeddings*

$$W_p^k(\Omega) \hookrightarrow W_q^j(\Omega), \quad \dot{W}_p^k(\Omega) \hookrightarrow \dot{W}_q^j(\Omega).$$

ii) For  $k - j - \beta > \frac{n}{p}$  one has the continuous embeddings

$$W_p^k(\Omega) \hookrightarrow C^{j,\beta}(\bar{\Omega}).$$

Note that the definition of the Hölder spaces  $C^{j,\beta}(\bar{\Omega})$  shows that they are continuously embedded into  $C^j(\bar{\Omega})$ , i.e.,

$$C^{j,\beta}(\bar{\Omega}) \hookrightarrow C^j(\bar{\Omega}).$$

Next we study the behaviour of restrictions of functions  $u \in W_p^l(\Omega)$  to the boundary  $\Gamma$ , which is a key step in understanding the treatment of boundary conditions in weak formulations. The following lemma from [Ada75] is the basic tool.

**Lemma 3.3 (Trace lemma).** *Let  $\Omega$  be a bounded domain with Lipschitz boundary  $\Gamma$ . Then there exists a constant  $c > 0$  such that*

$$\|u\|_{L_p(\Gamma)} \leq c \|u\|_{W_p^1(\Omega)} \text{ for all } u \in C^1(\bar{\Omega}).$$

Lemma 3.3 guarantees the existence of a linear continuous mapping

$$\gamma : W_p^1(\Omega) \rightarrow L_p(\Gamma)$$

which is called the *trace mapping*. The image of  $W_p^1(\Omega)$  under this mapping is a subspace of  $L_p(\Gamma)$  that is a new function space defined on the boundary  $\Gamma$ . For us the case  $p = 2$  is particularly important; we then obtain

$$H^{1/2}(\Gamma) := \{ w \in L_2(\Gamma) : \text{there exists a } v \in H^1(\Omega) \text{ with } w = \gamma v \}.$$

It is possible to define a norm on  $H^{1/2}(\Gamma)$  by

$$\|w\|_{H^{1/2}(\Gamma)} = \inf \{ \|v\|_{H^1(\Omega)} : v \in H^1(\Omega), w = \gamma v \}.$$

The space dual to  $H^{1/2}(\Gamma)$  is denoted by  $H^{-1/2}(\Gamma)$ , and its norm is given by

$$\|g\|_{H^{-1/2}(\Gamma)} = \sup_{w \in H^{1/2}(\Gamma)} \frac{|g(w)|}{\|w\|_{H^{1/2}(\Gamma)}}.$$

The relationship between the spaces  $H^1(\Omega)$  and  $H^{1/2}(\Gamma)$  allows a characterization of the norms in  $H^{1/2}(\Gamma)$  and  $H^{-1/2}(\Gamma)$  by means of suitably defined variational inequalities; see [BF91].

Taking into account the definition of the spaces  $\dot{W}_p^l(\Omega)$ , Lemma 3.3 implies that

$$\gamma u = 0 \text{ for all } u \in \dot{W}_p^1(\Omega)$$

and

$$\gamma D^\alpha u = 0 \text{ for all } u \in \dot{W}_p^l(\Omega) \text{ and } |\alpha| \leq l - 1.$$

In handling boundary value problems, we usually need not only the norms  $\|\cdot\|_{W_p^l(\Omega)}$  defined by (2.3) but also the associated seminorms

$$|u|_{W_p^s(\Omega)} := \left[ \int_{\Omega} \sum_{|\alpha|=s} |[D^\alpha u](x)|^p dx \right]^{1/p}.$$

It is clear that these seminorms can be estimated by the foregoing norms:

$$|v|_{W_p^s(\Omega)} \leq \|v\|_{W_p^l(\Omega)} \text{ for all } v \in W_p^l(\Omega) \text{ and } 0 \leq s \leq l. \quad (2.11)$$

Is a converse inequality true (at least for certain  $v$ )? Here the following result plays a fundamental role.

**Lemma 3.4.** *Let  $\Omega \subset \mathbb{R}^n$  be a bounded domain. Then there exists a constant  $c > 0$  such that*

$$\|v\|_{L_2(\Omega)} \leq c|v|_{W_2^1(\Omega)} \quad \text{for all } v \in H_0^1(\Omega). \quad (2.12)$$

Inequality (2.12) is known as the *Friedrichs inequality*. Once again a proof is in [Ada75].

*Remark 3.5.* The smallest constant  $c$  in Friedrichs' inequality can be characterized as the reciprocal of the minimal eigenvalue  $\lambda$  of the problem

$$-\Delta u = \lambda u \text{ on } \Omega, \quad u|_{\partial\Omega}=0.$$

For parallelepipeds  $\Omega$  the value of this eigenvalue is known. Furthermore, the eigenvalue does not increase in value if the domain is enlarged. Consequently in many cases one can compute satisfactory bounds for the constant in (2.12).

A detailed discussion of the values of constants in many fundamental inequalities related to Sobolev spaces can be found in the book [Mik86].  $\square$

From Lemma 3.4 and (2.11) it follows that

$$c_1\|v\|_{W_2^1(\Omega)} \leq |v|_{W_2^1(\Omega)} \leq \|v\|_{W_2^1(\Omega)} \quad \text{for all } v \in H_0^1(\Omega) \quad (2.13)$$

for some constant  $c_1 > 0$ . Therefore the definition

$$\|v\| := |v|_{W_2^1(\Omega)}$$

is a new norm on  $H_0^1(\Omega)$ , which is equivalent to the  $H^1$  norm. This norm is often used as the natural norm on  $H_0^1(\Omega)$ . It is induced by the scalar product

$$(u, v) := \int_{\Omega} \sum_{|\alpha|=1} D^\alpha u D^\alpha v dx.$$

Using this scalar product, the unique solvability of the weak formulation of the Poisson equation with homogeneous boundary conditions follows immediately from Riesz's theorem if the linear functional defined by

$$v \mapsto \int_{\Omega} f v \, dx$$

is continuous on  $H_0^1(\Omega)$ . This is true when  $f \in L_2(\Omega)$ , for instance.

Based on the following inequalities it is possible ([GGZ74], Lemma 1.36) to define other norms that are equivalent to  $\|\cdot\|_{W_2^1(\Omega)}$ :

**Lemma 3.6.** *Let  $\Omega \subset \mathbb{R}^n$  be a bounded Lipschitz domain. Assume also that  $\Omega_1$  is a subset of  $\Omega$  with positive measure and  $\Gamma_1$  a subset of  $\Gamma$  with positive  $(n-1)$ -dimensional measure. Then for  $u \in H^1(\Omega)$  one has*

$$\|u\|_{L_2(\Omega)}^2 \leq c \left\{ |u|_{1,\Omega}^2 + \left( \int_{\Omega_1} u \right)^2 \right\},$$

$$\|u\|_{L_2(\Omega)}^2 \leq c \left\{ |u|_{1,\Omega}^2 + \left( \int_{\Gamma_1} u \right)^2 \right\}.$$

These types of inequalities are proved for the more general  $W^{1,p}(\Omega)$  case in [GGZ74]. In the special case  $\Omega_1 = \Omega$  the first inequality is called the *Poincaré inequality*. The second inequality generalizes Friedrichs' inequality.

To simplify the notation, we shall write in future

$$|v|_{l,p,\Omega} := |v|_{W_p^l(\Omega)} \quad \text{and} \quad |v|_{l,\Omega} := |v|_{W_2^l(\Omega)}.$$

Next we consider the technique of integration by parts and study its application to the weak formulation of boundary value problems.

**Lemma 3.7 (integration by parts).** *Let  $\Omega \subset \mathbb{R}^n$  be a bounded Lipschitz domain. Then one has*

$$\int_{\Omega} \frac{\partial u}{\partial x_i} v \, dx = \int_{\Gamma} u v \cos(n, e^i) \, ds - \int_{\Omega} u \frac{\partial v}{\partial x_i} \, dx$$

for arbitrary  $u, v \in C^1(\bar{\Omega})$ . Here  $n$  is the outward-pointing unit vector normal to  $\Gamma$  and  $e^i$  is the unit vector in the  $i$ th coordinate direction.

Hence one obtains *Green's formula*:

$$\int_{\Omega} \Delta u v \, dx = \int_{\Gamma} \frac{\partial u}{\partial n} v \, ds - \int_{\Omega} \nabla u \nabla v \, dx \quad \text{for all } u \in H^2(\Omega), v \in H^1(\Omega) \quad (2.14)$$

—first apply integration by parts to classical differentiable functions then extend the result to  $u \in H^2(\Omega)$  and  $v \in H^1(\Omega)$  by a density argument based on (2.4).

Here and subsequently the term  $\nabla u \nabla v$  denotes a scalar product of two vectors; it could be written more precisely as  $(\nabla u)^T \nabla v$ . In general we tend

to use the simplified form, returning to the precise form only if the simplified version could lead to confusion.

The validity of Green's formula depends strongly on the geometry of  $\Omega$ . In general we shall consider only bounded Lipschitz domains so that (2.14) holds. For its validity on more general domains, see [Wlo87].

Now we resume our exploration of Poisson's equation with homogeneous Dirichlet boundary conditions:

$$-\Delta u = f \text{ in } \Omega, \quad u|_{\Gamma} = 0. \quad (2.15)$$

Every classical solution  $u$  of (2.15) satisfies, as we have seen, the variational equation

$$\int_{\Omega} \nabla u \nabla v \, dx = \int_{\Omega} f v \, dx \text{ for all } v \in H_0^1(\Omega). \quad (2.16)$$

If one defines the mapping  $a(\cdot, \cdot) : H_0^1(\Omega) \times H_0^1(\Omega) \rightarrow \mathbb{R}$  by

$$a(u, v) := \int_{\Omega} \nabla u \nabla v \, dx,$$

then Lemma 3.4 ensures the existence of a constant  $\gamma > 0$  such that

$$a(v, v) \geq \gamma \|v\|_{H_0^1(\Omega)}^2 \text{ for all } u, v \in H_0^1(\Omega).$$

This inequality is of fundamental importance in proving the existence of a unique solution  $u \in H_0^1(\Omega)$  of the variational equation (2.16) for each  $f \in L_2(\Omega)$ . In the next section we shall present a general existence theory for variational equations and discuss conditions sufficient for guaranteeing that weak solutions are also classical solutions.

If one reformulates a boundary value problem as a variational equation in order to define a weak solution, the type of boundary condition plays an important role. To explain this basic fact, we consider the following example:

$$\begin{aligned} -\Delta u + cu &= f \text{ in } \Omega, \\ u &= g \text{ on } \Gamma_1, \\ \frac{\partial u}{\partial n} + pu &= q \text{ on } \Gamma_2. \end{aligned} \quad (2.17)$$

Here  $\Gamma_1$  and  $\Gamma_2$  are subsets of the boundary with  $\Gamma_1 \cap \Gamma_2 = \emptyset$ ,  $\Gamma_1 \cup \Gamma_2 = \Gamma$  and the given functions  $c, f, g, p, q$  are continuous (say) with  $c \geq 0$  in  $\Omega$ . As usual, multiply the differential equation by an arbitrary function  $v \in H^1(\Omega)$ , then integrate over  $\Omega$  and apply integration by parts to get

$$\int_{\Omega} (\nabla u \nabla v + c uv) \, dx - \int_{\Gamma} \frac{\partial u}{\partial n} v \, ds = \int_{\Omega} f v \, dx.$$

Taking into account the boundary conditions for  $u$  we have

$$\int_{\Omega} (\nabla u \nabla v + c uv) dx + \int_{\Gamma_2} (pu - q)v ds - \int_{\Gamma_1} \frac{\partial u}{\partial n} v ds = \int_{\Omega} f v dx.$$

On  $\Gamma_1$  we have no information about the normal derivative of  $u$ . Therefore we restrict  $v$  to lie in  $V := \{ v \in H^1(\Omega) : v|_{\Gamma_1} = 0 \}$ . Then we obtain the variational equation

$$\int_{\Omega} (\nabla u \nabla v + c uv) dx + \int_{\Gamma_2} (pu - q)v ds = \int_{\Omega} f v dx \quad \text{for all } v \in V. \quad (2.18)$$

Of course, we require  $u$  to satisfy  $u \in H^1(\Omega)$  and  $u|_{\Gamma_1} = g$ . The variational equation (2.18) then defines weak solutions of our example (2.17). If the weak solution has some additional smoothness, then it is also a classical solution:

**Theorem 3.8.** *Let  $u \in H^1(\Omega)$  with  $u|_{\Gamma_1} = g$  be a solution of the variational equation (2.18). Moreover, let  $u$  be smooth:  $u \in C^2(\bar{\Omega})$ . Then  $u$  is a solution of the boundary value problem (2.17).*

**Proof:** Taking  $v|_{\Gamma_1} = 0$  into account, Green's formula applied to (2.18) yields

$$\int_{\Omega} (\Delta u + c u)v dx + \int_{\Gamma_2} \left( \frac{\partial u}{\partial n} + pu - q \right) v ds = \int_{\Omega} f v dx \quad \text{for all } v \in V. \quad (2.19)$$

Because  $H_0^1(\Omega) \subset V$  it follows that

$$\int_{\Omega} (\Delta u + c u)v dx = \int_{\Omega} f v dx \quad \text{for all } v \in H_0^1(\Omega).$$

Hence, using a well-known lemma of de la Vallée-Poussin, one obtains

$$-\Delta u + cu = f \quad \text{in } \Omega.$$

Now (2.19) implies that

$$\int_{\Gamma_2} \left( \frac{\partial u}{\partial n} + pu - q \right) v ds = 0 \quad \text{for all } v \in V.$$

Again we can conclude that

$$\frac{\partial u}{\partial n} + pu = q \quad \text{on } \Gamma_2. \quad (2.20)$$

The remaining condition  $u|_{\Gamma_1} = g$  is already satisfied by hypothesis. This is a so-called *essential boundary condition* that does not affect the variational equation but must be imposed directly on the solution itself. ■

**Remark 3.9.** In contrast to the essential boundary condition, the condition (2.20) follows from the variational equation (2.18) so it is not necessary to impose it explicitly on  $u$  in the variational formulation of the problem. Observe that the weak form of the boundary value problem was influenced by (2.20). Boundary conditions such as (2.20) are called *natural boundary conditions*.  $\square$

**Remark 3.10.** Let  $A$  be a positive definite matrix. Consider the differential equation

$$-\operatorname{div}(A \operatorname{grad} u) = f \quad \text{in } \Omega.$$

Integration by parts gives

$$-\int_{\Omega} \operatorname{div}(A \operatorname{grad} u) v \, dx = -\int_{\Gamma} n \cdot (A \operatorname{grad} u) v \, ds + \int_{\Omega} \operatorname{grad} v \cdot (A \operatorname{grad} u) \, dx.$$

In this case the natural boundary conditions contain the so-called *conormal derivative*  $n \cdot (A \operatorname{grad} u)$  instead of the normal derivative  $\frac{\partial u}{\partial n}$  that we met in the special case of the Laplacian (where  $A$  is the identity matrix).  $\square$

As we saw in Chapter 2, maximum principles play an important role in second-order elliptic boundary value problems. Here we mention briefly that even for weak solutions one can have maximum principles. For instance, the following *weak maximum principle* (see [GT83]) holds:

**Lemma 3.11.** *Let  $\Omega \subset \mathbb{R}^n$  be a bounded Lipschitz domain. If  $u \in H_0^1(\Omega)$  satisfies the variational inequality*

$$\int_{\Omega} \nabla u \nabla v \, dx \geq 0 \quad \text{for all } v \in H_0^1(\Omega) \text{ with } v \geq 0,$$

*then  $u \geq 0$ .*

Here  $u \geq 0$  and  $v \geq 0$  are to be understood in the  $L_2$  sense, i.e., almost everywhere in  $\Omega$ .

**Exercise 3.12.** Let  $\Omega = \{x \in \mathbb{R}^n : |x_i| < 1, i = 1, \dots, n\}$ . Prove:

a) The function defined by  $f(x) = |x_1|$  has on  $\Omega$  the generalized derivatives

$$\frac{\partial f}{\partial x_1} = \operatorname{sign}(x_1), \quad \frac{\partial f}{\partial x_j} = 0 \quad (j \neq 1).$$

b) The function defined by  $f(x) = \operatorname{sign}(x_1)$  does not have a generalized derivative  $\partial f / \partial x_1$  in  $L_2$ .

**Exercise 3.13.** Prove: If  $u : \Omega \rightarrow \mathbb{R}$  has the generalized derivatives  $v = D^\alpha u \in L_2(\Omega)$  and  $w$  the generalized derivatives  $w = D^\beta v \in L_2(\Omega)$ , then  $w = D^{\alpha+\beta} u$ .



**Exercise 3.14.** Let  $\Omega \subset \mathbb{R}^n$  be a bounded domain with  $0 \in \Omega$ . Prove that the function defined by  $u(x) = \|x\|_2^\sigma$  has first-order generalized derivatives in  $L_2(\Omega)$  if  $\sigma = 0$  or  $2\sigma + n > 2$ .

**Exercise 3.15.** Let  $\Omega = (a, b) \in \mathbb{R}$ . Prove that every function  $u \in H^1(\Omega)$  is continuous, and moreover  $u$  belongs to the Hölder space  $C^{1/2}(\Omega)$ .

**Exercise 3.16.** Let  $\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < r_0^2\}$  with  $r_0 < 1$ . Determine if the function

$$f(x, y) = \left( \ln \frac{1}{\sqrt{x^2 + y^2}} \right)^k, \text{ where } k < 1/2,$$

is continuous in  $\Omega$ . Is  $f \in H^1(\Omega)$ ?

**Exercise 3.17.** Consider the space of all continuous functions on the interval  $[a, b]$ . Prove that the norms

$$\|f\|_1 = \max_{x \in [a, b]} |f(x)| \quad \text{and} \quad \|f\|_2 = \int_a^b |f(x)| dx$$

are not equivalent.

**Exercise 3.18.** Let  $\Omega \subset [a_1, b_1] \times \cdots \times [a_n, b_n]$  be a convex domain. Let  $v \in H_0^1(\Omega)$ . Prove the Friedrichs inequality

$$\int_{\Omega} v^2 \leq \gamma \int_{\Omega} |\nabla v|^2 \quad \text{with} \quad \gamma = \sum_{k=1}^n (b_k - a_k)^2.$$

**Exercise 3.19.** Let  $\Omega \subset \mathbb{R}^n$ . Let  $u \in H^k(\Omega)$  for some integer  $k$ . For which dimensions  $n$  does Sobolev's embedding theorem guarantee that (i)  $u$  (ii)  $\nabla u$  is continuous?

**Exercise 3.20.** a) Let  $\Omega = (0, 1)$ ,  $u(x) = x^\alpha$ . Use this example to show that it is impossible to improve the continuous embedding  $H^1(\Omega) \hookrightarrow C^{1/2}(\overline{\Omega})$  to  $H^1(\Omega) \hookrightarrow C^\lambda(\overline{\Omega})$  with  $\lambda > 1/2$ .

b) Investigate for  $\Omega \subset \mathbb{R}^2$  whether or not the embedding  $H^1(\Omega) \hookrightarrow L_\infty(\Omega)$  holds.

**Exercise 3.21.** Let  $\Omega \subset \mathbb{R}^n$  with  $0 \in \Omega$ . Does the mapping

$$g \mapsto \langle f, g \rangle = g(0) \quad \text{for} \quad g \in H_0^1(\Omega)$$

define a continuous linear functional  $f$  on  $H_0^1(\Omega)$ ? If yes, determine  $\|f\|_*$ .

**Exercise 3.22.** Consider the boundary value problem

$$-u'' = f \text{ on } (0, 1), \quad u(-1) = u(1) = 0$$

with  $f$  the  $\delta$  distribution. How one can define the problem correctly in a weak sense? Determine the exact solution!

**Exercise 3.23.** Consider the boundary value problem

$$-(a(x)u')' = 0 \text{ on } (0, 1), \quad u(-1) = 3, \quad u(1) = 0$$

with

$$a(x) = \begin{cases} 1 & \text{for } -1 \leq x < 0, \\ 0.5 & \text{for } 0 \leq x \leq 1. \end{cases}$$

Formulate the related variational equation and solve the problem exactly.

### 3.3 Variational Equations and Conforming Approximation

In the previous section we described the relationship between an elliptic boundary value problem and its variational formulation in the case of the Poisson equation with homogeneous Dirichlet boundary conditions. Before we present an abstract framework for the analysis of general variational equations, we give weak formulations for some other standard model problems.

Let  $\Omega \subset \mathbb{R}^2$  with  $\Gamma = \partial\Omega$ . We consider, for a given sufficiently smooth function  $f$ , the boundary value problem

$$\begin{aligned} \frac{\partial^4}{\partial x^4} u(x, y) + 2 \frac{\partial^4}{\partial x^2 \partial y^2} u(x, y) + \frac{\partial^4}{\partial y^4} u(x, y) &= f(x, y) \quad \text{in } \Omega \\ u|_{\Gamma} &= \frac{\partial}{\partial n} u|_{\Gamma} = 0. \end{aligned} \quad (3.1)$$

This problem models the behaviour of a horizontally clamped plate under some given load distribution. Thus the differential equation in (3.1) is often called the *plate equation*. In terms of the Laplacian we have equivalently

$$\begin{aligned} \Delta^2 u &= f \quad \text{in } \Omega, \\ u|_{\Gamma} &= \frac{\partial}{\partial n} u|_{\Gamma} = 0. \end{aligned}$$

Now we formulate this problem weakly. Taking into account the boundary conditions, we apply Green's formula twice to obtain

$$\begin{aligned} \int_{\Omega} \Delta^2 u v \, dx &= \int_{\Gamma} \frac{\partial}{\partial n} (\Delta u) v \, ds - \int_{\Omega} \nabla(\Delta u) \nabla v \, dx \\ &= - \int_{\Gamma} \Delta u \frac{\partial v}{\partial n} \, ds + \int_{\Omega} \Delta u \Delta v \, dx \\ &= \int_{\Omega} \Delta u \Delta v \, dx \quad \text{for all } u \in H^4(\Omega), \, v \in H_0^2(\Omega). \end{aligned}$$

Therefore, the weak formulation of the given problem (3.1) reads as follows: Find  $u \in H_0^2(\Omega)$  such that

$$\int_{\Omega} \Delta u \Delta v \, dx = \int_{\Omega} f v \, dx \quad \text{for all } v \in H_0^2(\Omega). \quad (3.2)$$

With the abbreviations  $V := H_0^2(\Omega)$  and

$$a(u, v) := \int_{\Omega} \Delta u \Delta v \, dx, \quad (f, v) := \int_{\Omega} f v \, dx \quad \text{for all } u, v \in V,$$

the variational equation (3.2) can be written in the abstract form

$$a(u, v) = (f, v) \quad \text{for all } v \in V.$$

Using Friedrichs' inequality one can show that there exists some constant  $c > 0$  such that

$$c \|v\|_{H^2(\Omega)}^2 \leq a(v, v) \quad \text{for all } v \in H_0^2(\Omega).$$

This property is critical in the general existence theory for the weak solution of (3.2), as we shall see shortly in the Lax-Milgram lemma.

*Remark 3.24.* Up to this point in the plate problem, we considered the boundary conditions

$$u|_T = \frac{\partial}{\partial n} u|_T = 0,$$

which correspond to a clamped plate. Both of these conditions are essential boundary conditions. If instead we study a simply supported plate, whose boundary conditions are

$$u|_T = 0 \quad \text{and} \quad \Delta u|_T = \phi,$$

then the standard technique produces the weak formulation

$$a(u, v) = (f, v) + \int_T \phi \frac{\partial v}{\partial n}$$

with  $u, v \in H^2(\Omega) \cap H_0^1(\Omega)$ . This means that the first boundary condition is essential but the second is natural. Of course, in practical applications still other boundary conditions are important and in each case a careful study is required to classify each condition.  $\square$

Our last model problem plays an important role in fluid mechanics. Consider a domain  $\Omega \subset \mathbb{R}^n$  for  $n = 2$  or  $3$  and given functions  $f_i : \Omega \rightarrow \mathbb{R}$ ,  $i = 1, \dots, n$ . We seek solutions of the following system of partial differential equations with unknowns  $u_i : \overline{\Omega} \rightarrow \mathbb{R}$  for  $i = 1, \dots, n$  and  $p : \overline{\Omega} \rightarrow \mathbb{R}$ :

$$\begin{aligned} -\Delta u_i + \frac{\partial p}{\partial x_i} &= f_i & \text{in } \Omega, & \quad i = 1, \dots, n, \\ \sum_{i=1}^n \frac{\partial u_i}{\partial x_i} &= 0, \\ u_i|_T &= 0, & \quad i = 1, \dots, n. \end{aligned} \quad (3.3)$$

This is the so-called *Stokes problem*. In fluid mechanics, the quantities  $u_i$  denote the components of the velocity field while  $p$  represents the pressure.

Let  $u = (u_1, \dots, u_n)$  denote a vector-valued function. Let us choose the function space

$$V = \{u \in H_0^1(\Omega)^n : \operatorname{div} u = 0\} \subset H(\operatorname{div}; \Omega).$$

Then applying our standard technique (multiplication, integration, integration by parts) and adding all the resulting equations yields the following weak formulation of (3.3):

Find some  $u \in V$  with

$$\sum_{i=1}^n \int_{\Omega} \nabla u_i \nabla v_i \, dx = \sum_{i=1}^n \int_{\Omega} f_i v_i \, dx \quad \text{for all } v \in V. \quad (3.4)$$

It is remarkable that the pressure  $p$  has disappeared: integration by parts in the corresponding term gives 0 because  $\operatorname{div} v = 0$ . In the theory of mixed methods the pressure can be interpreted as a dual quantity; see Chapter 4.6. Alternative weak formulations of the Stokes problem are also possible.

If we introduce

$$a(u, v) := \sum_{i=1}^n \int_{\Omega} \nabla u_i \nabla v_i \, dx \quad \text{and} \quad f(v) := \sum_{i=1}^n \int_{\Omega} f_i v_i \, dx \quad \text{for all } u, v \in V,$$

then the weak formulation (3.4) of the Stokes problem can also be written in the form

$$a(u, v) = f(v) \quad \text{for all } v \in V. \quad (3.5)$$

Now we are ready to present an abstract theory encompassing (2.16), (3.2) and (3.5). The abstract setting allows us to characterize clearly those properties of variational equations that guarantee the existence of a unique solution. Then in every concrete situation one has only to check these properties.

Let  $V$  be a given Hilbert space with scalar product  $(\cdot, \cdot)$  and corresponding norm  $\|\cdot\|$ . Furthermore, let there be given a mapping  $a : V \times V \rightarrow \mathbb{R}$  with the following properties:

- i) for arbitrary  $u \in V$ , both  $a(u, \cdot)$  and  $a(\cdot, u)$  define linear functionals on  $V$ ;
- ii) there exists a constant  $M > 0$  such that

$$|a(u, v)| \leq M \|u\| \|v\| \quad \text{for all } u, v \in V;$$

- iii) there exists a constant  $\gamma > 0$  such that

$$a(u, u) \geq \gamma \|u\|^2 \quad \text{for all } u \in V.$$

A mapping  $a(\cdot, \cdot)$  satisfying i) and ii) is called a *continuous bilinear form on  $V$* . Property ii) guarantees the boundedness of the bilinear form. The essential property iii) is called  *$V$ -ellipticity*.

The existence of solutions of variational equations is ensured by the following fundamental result.

**Lemma 3.25 (Lax-Milgram).** *Let  $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$  be a continuous,  $V$ -elliptic bilinear form. Then for each  $f \in V^*$  the variational equation*

$$a(u, v) = f(v) \quad \text{for all } v \in V \quad (3.6)$$

*has a unique solution  $u \in V$ . Furthermore, the a priori estimate*

$$\|u\| \leq \frac{1}{\gamma} \|f\|_* \quad (3.7)$$

*is valid.*

**Proof:** First we show that the solution of (3.6) is unique. Suppose that  $u \in V$  and  $\tilde{u} \in V$  are both solutions. Then the linearity of  $a(\cdot, v)$  implies that

$$a(\tilde{u} - u, v) = 0 \quad \text{for all } v \in V.$$

Choosing  $v := \tilde{u} - u$  we get  $a(v, v) = 0$ , which by  $V$ -ellipticity implies that  $v = 0$ , as desired. Note that  $V$ -ellipticity, however, is stronger than the condition “ $a(v, v) = 0$  implies  $v = 0$ ”.

To prove the existence of a solution to (3.6) we use Banach’s fixed-point theorem. Therefore, we need to choose a contractive mapping that has as a fixed point a solution of (3.6).

For each  $y \in V$  the assumptions i) and ii) for the bilinear form guarantee that

$$a(y, \cdot) - f \in V^*.$$

Hence, Riesz’s theorem ensures the existence of a solution  $z \in V$  of

$$(z, v) = (y, v) - r[a(y, v) - f(v)] \quad \text{for all } v \in V \quad (3.8)$$

for each real  $r > 0$ . Now we define the mapping  $T_r : V \rightarrow V$  by

$$T_r y := z$$

and study its properties—especially contractivity. The relation (3.8) implies

$$(T_r y - T_r w, v) = (y - w, v) - r a(y - w, v) \quad \text{for all } v, w \in V. \quad (3.9)$$

Given  $p \in V$ , by applying Riesz’s theorem again we define an auxiliary linear operator  $S : V \rightarrow V$  by

$$(Sp, v) = a(p, v) \quad \text{for all } v \in V. \quad (3.10)$$

Property ii) of the bilinear form implies that

$$\|Sp\| \leq M \|p\| \quad \text{for all } p \in V. \quad (3.11)$$

The definition of the operator  $S$  means that (3.9) can be rewritten as

$$(T_r y - T_r w, v) = (y - w - rS(y - w), v) \quad \text{for all } v, w \in V.$$

This allows us to investigate whether  $T_r$  is contractive:

$$\begin{aligned} \|T_r y - T_r w\|^2 &= (T_r y - T_r w, T_r y - T_r w) \\ &= (y - w - rS(y - w), y - w - rS(y - w)) \\ &= \|y - w\|^2 - 2r(S(y - w), y - w) + r^2(S(y - w), S(y - w)). \end{aligned}$$

By (3.10) and (3.11) this yields

$$\|T_r y - T_r w\|^2 \leq \|y - w\|^2 - 2ra(y - w, y - w) + r^2 M^2 \|y - w\|^2.$$

Finally, invoking the  $V$ -ellipticity of  $a(\cdot, \cdot)$  we get

$$\|T_r y - T_r w\|^2 \leq (1 - 2r\gamma + r^2 M^2) \|y - w\|^2 \quad \text{for all } y, w \in V.$$

Consequently the operator  $T_r : V \rightarrow V$  is contractive if  $0 < r < 2\gamma/M^2$ .

Choose  $r = \gamma/M^2$ . Now Banach's fixed-point theorem tells us that there exists  $u \in V$  with  $T_r u = u$ . Since  $r > 0$ , the definition (3.8) of  $T_r$  then implies that

$$a(u, v) = f(v) \quad \text{for all } v \in V. \quad (3.12)$$

The a priori estimate (3.7) is an immediate consequence of the ellipticity of  $a(\cdot, \cdot)$ : choose  $v = u$  in (3.6). ■

We remark that in the case where  $a(\cdot, \cdot)$  is symmetric, the existence of  $u$  in the Lax-Milgram lemma follows directly from Riesz's theorem. In the symmetric case, moreover, there is a close relationship between variational equations and variational problems:

**Lemma 3.26.** *In addition to the assumptions of Lemma 3.25, suppose that  $a(\cdot, \cdot)$  is symmetric, i.e.,*

$$a(v, w) = a(w, v) \quad \text{for all } v, w \in V.$$

*Then  $u \in V$  is a solution of the variational problem*

$$\min_{v \in V} J(v), \quad \text{where } J(v) := \frac{1}{2}a(v, v) - f(v) \quad \text{for } v \in V, \quad (3.13)$$

*if and only if  $u$  is a solution of the variational equation (3.6).*

**Proof:** The symmetry of the bilinear form  $a(\cdot, \cdot)$  implies that

$$\begin{aligned} a(w, w) - a(u, u) &= a(w + u, w - u) \\ &= 2a(u, w - u) + a(w - u, w - u) \quad \text{for } u, w \in V. \end{aligned} \quad (3.14)$$

First we shall show that (3.6) is a sufficient condition for the optimality of  $u$  in the variational problem (3.13). From (3.14) one has

$$\begin{aligned} J(w) &= \frac{1}{2}a(w, w) - f(w) \\ &= \frac{1}{2}a(u, u) - f(u) + a(u, w - u) \\ &\quad - f(w - u) + \frac{1}{2}a(w - u, w - u). \end{aligned} \quad (3.15)$$

Taking  $v := w - u$  in (3.6) and using property iii) of the bilinear form  $a(\cdot, \cdot)$  leads to

$$J(w) \geq J(u) \quad \text{for all } w \in V;$$

that is,  $u$  is a solution of the variational problem (3.13).

We prove the converse implication indirectly. Given some  $u \in V$ , assume that there exists  $v \in V$  with

$$a(u, v) \neq f(v).$$

Because  $V$  is a linear space we can assume without loss of generality that in fact

$$a(u, v) < f(v). \quad (3.16)$$

Now set  $w := u + tv$  with a real parameter  $t > 0$ .

Definition (3.13) implies, using standard properties of  $a(\cdot, \cdot)$  and  $f$ , that

$$J(w) = J(u) + t[a(u, v) - f(v)] + t^2 \frac{1}{2}a(v, v).$$

By (3.16) we can choose  $t > 0$  in such a way that

$$J(w) < J(u).$$

That is,  $u$  cannot be an optimal solution of the variational problem. Consequently (3.6) is a necessary optimality condition for the variational problem (3.13). ■

Lemma 3.26 can be applied to our familiar example of the Poisson equation with homogeneous Dirichlet boundary conditions:

$$-\Delta u = f \text{ on } \Omega, \quad u|_T = 0.$$

We proved already in Section 3.1 that the corresponding bilinear form is  $V$ -elliptic:

$$a(v, v) \geq \gamma \|v\|_1^2.$$

The boundedness of the bilinear form is obvious. Therefore, the Lax-Milgram lemma tells us that there exists a unique weak solution if we assume only that

$f \in H^{-1}(\Omega)$ . Because the bilinear form is symmetric, this weak solution is also a solution of the variational problem

$$\min_{v \in H_0^1(\Omega)} \left[ \frac{1}{2} \int_{\Omega} (\nabla v)^2 - f(v) \right].$$

Next we study the nonsymmetric convection-diffusion problem

$$-\Delta u + b \cdot \nabla u + cu = f \text{ on } \Omega, \quad u|_{\Gamma} = 0.$$

The associated bilinear form

$$a(u, v) := (\nabla u, \nabla v) + (b \cdot \nabla u + cu, v)$$

is not necessarily  $H_0^1$ -elliptic. Integration by parts of the term  $(b \cdot \nabla v, v)$  shows that the condition

$$c - \frac{1}{2} \operatorname{div} b \geq 0$$

is sufficient for  $H_0^1$ -ellipticity.

*Remark 3.27 (Neumann boundary conditions).* Consider the boundary value problem

$$-\Delta u + cu = f \text{ on } \Omega, \quad \frac{\partial u}{\partial n}|_{\Gamma} = 0.$$

If  $c(x) \geq c_0 > 0$ , then the bilinear form

$$a(u, v) := (\nabla u, \nabla v) + (cu, v)$$

is  $V$ -elliptic on  $V = H^1(\Omega)$ . The Lax-Milgram lemma can now be readily applied to the weak formulation of the problem, which shows that it has a unique solution in  $H^1(\Omega)$ .

If instead  $c = 0$ , then any classical solution of the Neumann problem above has the property that adding a constant to the solution yields a new solution. How does one handle the weak formulation in this case? Is it possible to apply the Lax-Milgram lemma?

To deal with this case we set  $V = \{v \in H^1(\Omega) : \int_{\Gamma} v = 0\}$ . Then Lemma 3.6 implies that the bilinear form  $a(u, v) = (\nabla u, \nabla v)$  is  $V$ -elliptic with respect to the space  $V$ . It is easy to see that the bilinear form is bounded on  $V \times V$ . Therefore, surprisingly, our weak formulation for the Neumann problem with  $c = 0$  is

$$a(u, v) = (f, v) \quad \text{for all } v \in V, \tag{3.17}$$

and this equation has a unique solution  $u \in V$  for each  $f \in L_2(\Omega)$ .

But in the case  $c = 0$  if one wants for smooth  $u$  to return from the variational equation (3.17) to the classical formulation of the problem, then (3.17) must be valid for all  $v \in H^1(\Omega)$ . On choosing  $v = 1$ , this implies the condition



$$\int_{\Omega} f = 0.$$

In this way we get the well-known solvability condition for classical solutions, which alternatively follows from the classical formulation by invoking Gauss's integral theorem. A detailed discussion of the consequences for the finite element method applied to this case can be found in [17].  $\square$

Classical solutions of boundary value problems are also weak solutions. For the converse implication, weak solutions must have sufficient smoothness to be classical solutions. We now begin to discuss regularity theorems that give sufficient conditions for additional regularity of weak solutions. Embedding theorems are also useful in deducing smoothness in the classical sense from smoothness in Sobolev spaces.

From [Gri85] we quote

**Lemma 3.28.** *Let  $\Omega$  be a domain with  $C^k$  boundary. If  $f \in H^k(\Omega)$  for some  $k \geq 0$ , then the solution  $u$  of (2.16) has the regularity property*

$$u \in H^{k+2}(\Omega) \cap H_0^1(\Omega).$$

Furthermore, there exists a constant  $C$  such that

$$\|u\|_{k+2} \leq C \|f\|_k.$$

A result of this type—where a certain regularity of  $f$  yields a higher degree of regularity in  $u$ —is called a *shift theorem*.

**Corollary 3.29.** *Let  $\Omega \subset \mathbb{R}^n$  have  $C^k$  boundary. Let  $f \in H^k(\Omega)$  with  $k > \frac{n}{2}$ . Then the solution  $u$  of (2.16) satisfies*

$$u \in C^2(\bar{\Omega}) \cap H_0^1(\Omega).$$

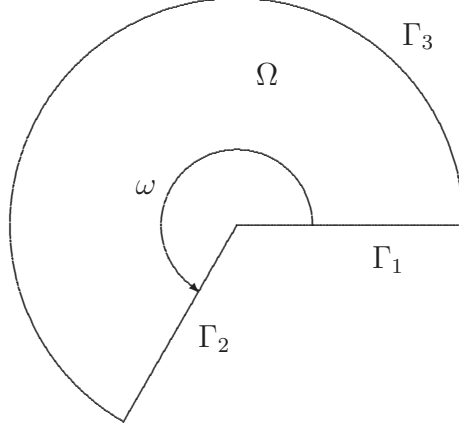
Thus  $u$  is a solution of the boundary value problem (2.15) in the classical sense.

**Proof:** Lemma 3.28 implies that  $u \in H^{k+2}(\Omega)$ . Then the continuous embedding  $W_2^{k+2}(\Omega) \hookrightarrow C^2(\bar{\Omega})$  for  $k > n/2$  yields the result.  $\blacksquare$

The assumption of Lemma 3.28 that the domain  $\Omega$  possesses a  $C^k$  boundary is very restrictive, for in many practical examples the domain has corners. Thus it is more realistic to assume only that the boundary is piecewise smooth.

What regularity does the solution have at a corner of the domain? To answer this question, we shall study the Laplace equation in the model domain

$$\Omega = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 : x = r \cos \varphi, y = r \sin \varphi, r \in (0, 1), \varphi \in (0, \omega) \right\} \quad (3.18)$$

**Figure 3.1** Example: corner singularity

for some parameter  $\omega \in (0, 2\pi)$ ; see Figure 3.1. This domain has a piecewise smooth boundary  $\Gamma$  with corners at the points  $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ ,  $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$  and  $\begin{pmatrix} \cos \omega \\ \sin \omega \end{pmatrix}$ . We decompose  $\Gamma$  into the three smooth pieces

$$\begin{aligned} \Gamma_1 &= \left\{ \begin{pmatrix} x \\ y \end{pmatrix} : x \in [0, 1], y = 0 \right\}, \\ \Gamma_2 &= \left\{ \begin{pmatrix} x \\ y \end{pmatrix} : x = r \cos \omega, y = r \sin \omega, r \in (0, 1) \right\}, \\ \Gamma_3 &= \left\{ \begin{pmatrix} x \\ y \end{pmatrix} : x = \cos \varphi, y = \sin \varphi, \varphi \in (0, \omega] \right\}. \end{aligned}$$

Then  $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$ . Now consider the following Dirichlet problem for the Laplacian:

$$\begin{aligned} -\Delta u &= 0 \quad \text{in } \Omega, \\ u|_{\Gamma_1 \cup \Gamma_2} &= 0, \\ u|_{\Gamma_3} &= \sin\left(\frac{\pi}{\omega}\varphi\right). \end{aligned} \tag{3.19}$$

The problem has the unique solution

$$u(r, \varphi) = r^{\pi/\omega} \sin\left(\frac{\pi}{\omega}\varphi\right).$$

Consequently  $u \in H^2(\Omega)$  if and only if  $\omega \in (0, \pi]$ . We infer that the solutions of Dirichlet problems in non-convex domains do not in general have the regularity property  $u \in H^2(\Omega)$ .

Next, consider instead of (3.19) the boundary value problem

$$\begin{aligned}
-\Delta u &= 0 && \text{in } \Omega, \\
u|_{\Gamma_1} &= 0, \\
\frac{\partial u}{\partial n}|_{\Gamma_2} &= 0, \\
u|_{\Gamma_3} &= \sin\left(\frac{\pi}{2\omega}\varphi\right).
\end{aligned} \tag{3.20}$$

Its solution is

$$u(r, \varphi) = r^{\pi/2\omega} \sin\left(\frac{\pi}{2\omega}\varphi\right).$$

For this problem with mixed boundary conditions one has  $u \notin H^2(\Omega)$  if  $\omega > \pi/2$ . In the case  $\omega = \pi$ , for instance, the solution has a corner singularity of the type  $r^{1/2}$ .

These examples show clearly that the regularity of the solution of a boundary value problem depends not only on the smoothness of the data but also on the geometry of the domain and on the type of boundary conditions. It is important to remember this dependence to guard against proving convergence results for discretization methods under unrealistic assumptions. Lemma 3.28, for instance, is powerful and elegant but it does treat an ideal situation because of the smoothness of the boundary and the homogeneous Dirichlet boundary conditions.

In the books of Dauge [Dau88] and Grisvard [Gri85, Gri92] the reader will find many detailed results regarding the behaviour of solutions of elliptic boundary value problems in domains with corners. We shall quote only the following theorem, which ensures  $H^2$ -regularity for convex domains.

**Theorem 3.30.** *Let  $\Omega$  be a convex domain. Set  $V = H_0^1(\Omega)$ . Let  $a(\cdot, \cdot)$  be a  $V$ -elliptic bilinear form that is generated by a second-order elliptic differential operator with smooth coefficients. Then for each  $f \in L_2(\Omega)$ , the solution  $u$  of the Dirichlet problem*

$$a(u, v) = (f, v) \quad \text{for all } v \in V$$

*lies in the space  $H^2(\Omega)$ . Furthermore, there exists a constant  $C$  such that*

$$\|u\|_2 \leq C\|f\|_0.$$

A similar result holds for elliptic second-order boundary value problems in convex domains if the boundary conditions are of a different type—but not mixed as the example above has shown us. For fourth-order boundary value problems, however, a convex domain is not sufficient in general to guarantee  $u \in H^4(\Omega)$ .

Now we start to discuss the approximation of solutions of variational equations.

First we describe *Ritz's method*. It is a technique for approximately solving variational problems such as (3.13). Instead of solving the given problem in the space  $V$ , which is in general a infinite-dimensional space, one chooses a finite-dimensional subspace  $V_h \subset V$  and solves

$$\min_{v_h \in V_h} J(v_h), \text{ where } J(v_h) = \frac{1}{2}a(v_h, v_h) - f(v_h). \quad (3.21)$$

As  $V_h$  is finite-dimensional, it is a closed subspace of  $V$  and therefore a Hilbert space endowed with the same scalar product  $(\cdot, \cdot)$ . Consequently the bilinear form  $a(\cdot, \cdot)$  has the same properties on  $V_h$  as on  $V$ . Thus our abstract theory applies to the problem (3.21). Hence (3.21) has a unique solution  $u_h \in V_h$ , and  $u_h$  satisfies the necessary and sufficient optimality condition

$$a(u_h, v_h) = f(v_h) \quad \text{for all } v_h \in V_h. \quad (3.22)$$

Ritz's method assumes that the bilinear form  $a(\cdot, \cdot)$  is symmetric. Nevertheless in the nonsymmetric case it is an obvious idea to go directly from the variational equation

$$a(u, v) = f(v) \quad \text{for all } v \in V$$

to its finite-dimensional counterpart (3.22). The discretization of the variational equation by (3.22) is called the *Galerkin method*. Because in the symmetric case the Ritz method and the Galerkin method coincide, we also use the terminology *Ritz-Galerkin method*.

The following result, often called *Cea's lemma*, is the basis for most convergence results for Ritz-Galerkin methods:

**Theorem 3.31 (Cea).** *Let  $a(\cdot, \cdot)$  be a continuous,  $V$ -elliptic bilinear form. Then for each  $f \in V^*$  the continuous problem (3.6) has a unique solution  $u \in V$  and the discrete problem (3.22) has a unique solution  $u_h \in V_h$ . The error  $u - u_h$  satisfies the inequality*

$$\|u - u_h\| \leq \frac{M}{\gamma} \inf_{v_h \in V_h} \|u - v_h\|. \quad (3.23)$$

**Proof:** Existence and uniqueness of  $u$  and  $u_h$  are immediate consequences of the Lax-Milgram lemma.

As  $V_h \subset V$ , it follows from (3.6) that

$$a(u, v_h) = f(v_h) \quad \text{for all } v_h \in V_h.$$

By the linearity of the bilinear form and (3.22) we then get

$$a(u - u_h, v_h) = 0 \quad \text{for all } v_h \in V_h.$$

This identity and linearity yield

$$a(u - u_h, u - u_h) = a(u - u_h, u - v_h) \quad \text{for all } v_h \in V_h.$$

The  $V$ -ellipticity and boundedness of  $a(\cdot, \cdot)$  now imply

$$\gamma \|u - u_h\|^2 \leq M \|u - u_h\| \|u - v_h\| \quad \text{for all } v_h \in V_h.$$

The estimate (3.23) follows since  $v_h$  is an arbitrary element of  $V_h$ . ■

The property

$$a(u - u_h, v_h) = 0 \quad \text{for all } v_h \in V_h$$

that we met in the above proof tells us that the error  $u - u_h$  is “orthogonal” in a certain sense to the space  $V_h$  of ansatz functions. Galerkin used this idea in formulating his method in 1915. We call the property *Galerkin orthogonality*.

*Remark 3.32.* Cea’s lemma relates the discretization error to the best approximation error

$$\inf_{v_h \in V_h} \|u - v_h\|. \quad (3.24)$$

Because the two errors differ only by a fixed multiplicative constant, the Ritz-Galerkin method is described as *quasi-optimal*.

If as (say)  $h \rightarrow 0$  the best approximation error goes to zero, then it follows that

$$\lim_{h \rightarrow 0} \|u - u_h\| = 0.$$

It is often difficult to compute the best approximation error. Then we choose an easily-computed projector  $\Pi_h : V \rightarrow V_h$ , e.g. an interpolation operator, and estimate the approximation error by

$$\inf_{v_h \in V_h} \|u - v_h\| \leq \|u - \Pi_h u\|.$$

In Section 4.4 we shall estimate  $\|u - \Pi_h u\|$  explicitly for specially chosen spaces  $V_h$  used in finite element methods. □

*Remark 3.33.* If the bilinear form  $a(\cdot, \cdot)$  is symmetric, then instead of (3.23) one can prove that

$$\|u - u_h\| \leq \sqrt{\frac{M}{\gamma}} \inf_{v_h \in V_h} \|u - v_h\|.$$

□

*Remark 3.34.* The assumption that  $V_h \subset V$  guarantees that certain properties valid on  $V$  remain valid on the finite-dimensional space  $V_h$ . If we do not require  $V_h \subset V$ , then we have to overcome some technical difficulties (see Chapter 4). Methods with  $V_h \subset V$  that use the same bilinear form  $a(\cdot, \cdot)$  and functional  $f(\cdot)$  in both the continuous and discrete problems are called *conforming methods*. □

*Remark 3.35.* For the practical implementation of the Galerkin method one needs a suitably chosen space of ansatz functions  $V_h \subset V$  and one must compute  $a(w, v)$  and  $f(v)$  for given  $v, w \in V_h$ . The exact computation of the

integrals involved is often impossible, so quadrature formulas are used. But the introduction of such formulas is equivalent to changing  $a(\cdot, \cdot)$  and  $f(\cdot)$ , so it makes the method nonconforming; see Chapter 4.  $\square$

*Remark 3.36.* The dimension of  $V_h$  is finite. Thus this space has a basis, i.e., a finite number of linearly independent functions  $\varphi_i \in V_h$ , for  $i = 1, \dots, N$ , that span  $V_h$ :

$$V_h = \left\{ v : v(x) = \sum_{i=1}^N d_i \varphi_i(x) \right\}.$$

Because  $a(\cdot, \cdot)$  and  $f(\cdot)$  are linear, the relation (3.22) is equivalent to

$$a(u_h, \varphi_i) = f(\varphi_i), \quad i = 1, \dots, N.$$

Writing the unknown  $u_h \in V_h$  as

$$u_h(x) = \sum_{j=1}^N s_j \varphi_j(x), \quad x \in \Omega,$$

the unknown coefficients  $s_j \in \mathbb{R}$  ( $j = 1, \dots, N$ ) satisfy the linear system of equations

$$\sum_{j=1}^N a(\varphi_j, \varphi_i) s_j = f(\varphi_i), \quad i = 1, \dots, N. \quad (3.25)$$

We call the system (3.25) the *Galerkin equations*. In Chapter 8 we shall discuss its properties in detail, including practical effective methods for its solution. For the moment we remark only that the  $V$ -ellipticity of  $a(\cdot, \cdot)$  implies that the coefficient matrix of (3.25) is invertible: let  $z = (z_1, \dots, z_N) \in \mathbb{R}^N$  be a solution of the homogeneous system

$$\sum_{j=1}^N a(\varphi_j, \varphi_i) z_j = 0, \quad i = 1, \dots, N. \quad (3.26)$$

Then

$$\sum_{i=1}^N \sum_{j=1}^N a(\varphi_j, \varphi_i) z_j z_i = 0.$$

By the linearity of  $a(\cdot, \cdot)$  this is the same as

$$a\left(\sum_{j=1}^N z_j \varphi_j, \sum_{i=1}^N z_i \varphi_i\right) = 0,$$

which by  $V$ -ellipticity forces

$$\sum_{j=1}^N z_j \varphi_j = 0.$$

Because the functions  $\varphi_j$  are linearly independent, we get  $z = \mathbf{0}$ . That is, the homogeneous system (3.26) has only the trivial solution. Consequently the coefficient matrix of (3.25) is nonsingular.  $\square$

In the derivation of the Galerkin equations (3.25) we used the same basis functions  $\{\varphi_i\}_{i=1}^N$  of  $V_h$  for both the ansatz and the test functions. This guarantees that the *stiffness matrix*  $A_h = (a_{ij}) = (a(\varphi_j, \varphi_i))$  has nice properties; in the case of a symmetric bilinear form the stiffness matrix is symmetric and positive definite.

Alternatively, one can use different spaces  $V_h$  and  $W_h$  for the ansatz and the test functions, but they must have the same dimension. Let us denote by  $\{\varphi_i\}_{i=1}^N$  and  $\{\psi_i\}_{i=1}^N$  the basis functions of  $V_h$  and  $W_h$ , i.e.,

$$V_h = \text{span}\{\varphi_i\}_{i=1}^N, \quad W_h = \text{span}\{\psi_i\}_{i=1}^N.$$

Setting

$$u_h(x) = \sum_{j=1}^N s_j \varphi_j(x),$$

the discrete variational equation

$$a(u_h, v_h) = f(v_h) \quad \text{for all } v_h \in W_h \quad (3.27)$$

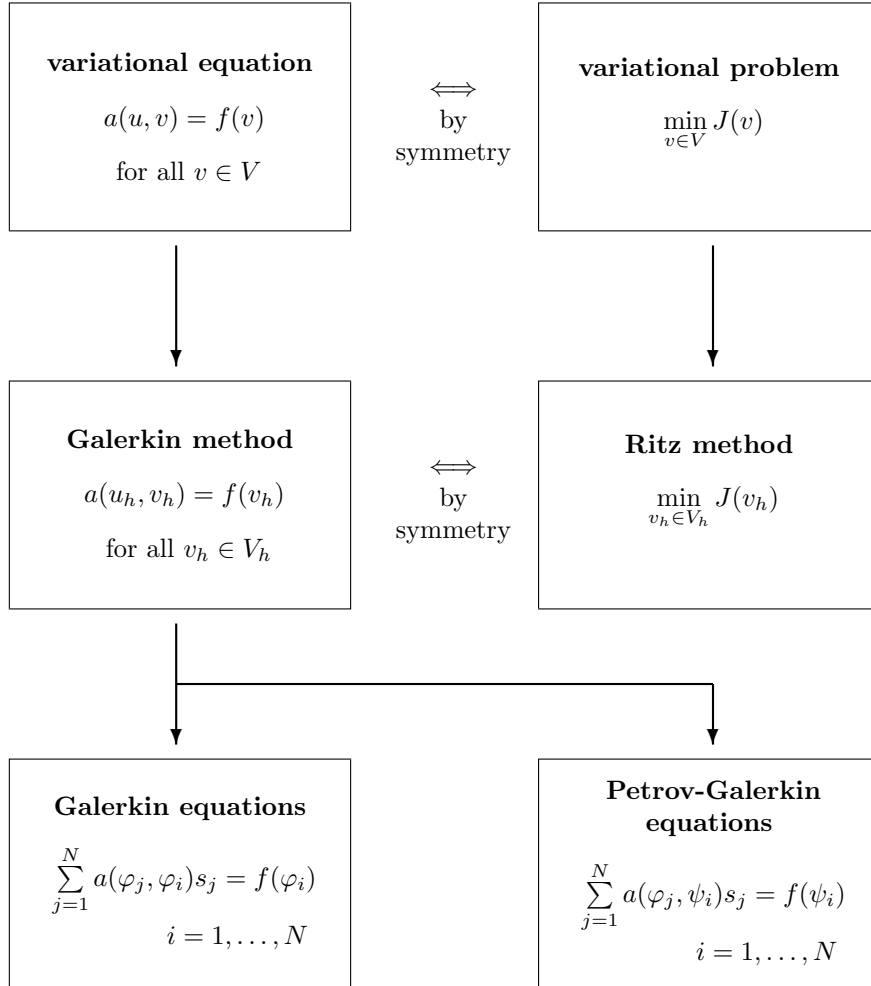
is equivalent to

$$\sum_{j=1}^N a(\varphi_j, \psi_i) s_j = f(\psi_i), \quad i = 1, \dots, N. \quad (3.28)$$

This generalization of the Galerkin method, where the *ansatz functions* differ from the *test functions*, is called the *Petrov-Galerkin method*.

One could choose  $V_h$  and  $W_h$  with the aim of imposing certain properties on the discrete problem (3.28), but Petrov-Galerkin methods are more usually the result of a weak formulation that is based on different ansatz and test spaces: see the next Section. For instance, they are often used in the treatment of first-order hyperbolic problems and singularly perturbed problems.

Setting  $J(v) = \frac{1}{2}a(v, v) - f(v)$ , here is a summary of our basic discretizations for variational equations:



Before continuing our study of the properties of the Ritz-Galerkin method, we illustrate it by some simple examples.

*Example 3.37.* Let us study the two-point boundary value problem

$$\begin{aligned} -u'' &= f && \text{in } (0, 1), \\ u(0) &= u(1) = 0. \end{aligned} \tag{3.29}$$

We choose  $V = H_0^1(0, 1)$ . As the Dirichlet boundary conditions are homogeneous, integration by parts generates the bilinear form



$$a(u, v) = \int_0^1 u'(x) v'(x) dx. \quad (3.30)$$

Next we choose as ansatz functions

$$\varphi_j(x) = \sin(j\pi x), \quad j = 1, \dots, N,$$

and set  $h := 1/N$ . Then  $V_h \subset V$  is defined by

$$V_h := \text{span}\{\varphi_j\}_{j=1}^N := \left\{ v : v(x) = \sum_{j=1}^N c_j \varphi_j(x) \right\}.$$

It is possible to show (see, e.g., [Rek80]) that

$$\lim_{h \rightarrow 0} \left[ \inf_{v \in V_h} \|u - v\| \right] = 0$$

for any given  $u \in V$ . The estimate (3.23) proves convergence of the Galerkin method in this case. Now

$$\begin{aligned} a(\varphi_i, \varphi_j) &= \pi^2 \int_0^1 ij \cos(i\pi x) \cos(j\pi x) dx \\ &= \begin{cases} \pi^2 j^2 / 2 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \end{aligned}$$

Setting

$$q_i := \int_0^1 f(x) \varphi_i(x) dx,$$

the solution of the Galerkin equations (3.25) is easily seen to be

$$s_j = \frac{2q_j}{\pi^2 j^2}, \quad j = 1, \dots, N. \quad (3.31)$$

The Galerkin approximation  $u_h$  for the solution of (3.22) is then

$$u_h(x) = \sum_{j=1}^N s_j \sin(j\pi x).$$

Why was it possible to derive an explicit formula for the Galerkin approximation? The reason is that our ansatz functions were the eigenfunctions of the differential operator of (3.29). This is an exceptional situation, since in general the differential operator's eigenfunctions are not known. Consequently, we do not usually have the orthogonality relation

$$a(\varphi_i, \varphi_j) = 0 \quad \text{for } i \neq j.$$

□

*Example 3.38.* Let us modify problem (3.29) by considering instead the problem

$$\begin{aligned} -u'' &= f && \text{in } (0, 1), \\ u(0) &= u'(1) = 0. \end{aligned} \quad (3.32)$$

The condition  $u'(1) = 0$  is a natural boundary condition. In the weak formulation we get  $a(u, v) = \int_0^1 u'(x)v'(x) dx$  as before but now the underlying function space is

$$V = \{v \in H^1(0, 1) : v(0) = 0\}.$$

We choose  $V_h$  to be the polynomial subspace

$$V_h = \text{span} \left\{ \frac{1}{i} x^i \right\}_{i=1}^N. \quad (3.33)$$

The Galerkin method generates a linear system of equations

$$As = b \quad (3.34)$$

for the unknown coefficients  $s_i \in \mathbb{R}$ ,  $i = 1, \dots, N$ , in the representation

$$u_h(x) = \sum_{i=1}^N \frac{s_i}{i} x^i.$$

The entries in the coefficient matrix  $A = (a_{ij})$  are

$$a_{ij} = a(\varphi_j, \varphi_i) = \frac{1}{i+j-1}, \quad i, j = 1, \dots, N.$$

This particular matrix  $A$  is called the Hilbert matrix. It is well known to be extremely ill-conditioned. For example, when  $N = 10$  the condition number  $\text{cond}(A)$  is approximately  $10^{13}$ .

The example reveals that the choice of the ansatz functions is very important—the Galerkin method with ansatz (3.33) for the boundary value problem (3.37) is impracticable because the linear system generated is numerically unstable and cannot be solved satisfactorily owing to rounding errors.  $\square$

*Example 3.39.* Consider again the boundary value problem (3.29) with  $V = H_0^1(0, 1)$ . Now we choose the discrete space  $V_h = \text{span}\{\varphi_j\}_{j=1}^N$  to be the span of the piecewise linear functions

$$\varphi_j(x) = \begin{cases} \frac{x - x_{j-1}}{h} & \text{if } x \in (x_{j-1}, x_j], \\ \frac{x_{j+1} - x}{h} & \text{if } x \in (x_j, x_{j+1}), \\ 0 & \text{otherwise.} \end{cases} \quad (3.35)$$

where  $j = 1, \dots, N-1$ . Here  $\{x_j\}_{j=0}^N$  is an equidistant mesh on the given interval  $(0, 1)$ , i.e.,  $x_j = j \cdot h$  for  $j = 0, 1, \dots, N$  with  $h = 1/N$ . The best approximation error (3.24) from this discrete space will be studied in detail in Chapter 4.

From (3.30) and (3.35) it follows that

$$a(\varphi_i, \varphi_j) = \begin{cases} \frac{2}{h} & \text{if } i = j, \\ -\frac{1}{h} & \text{if } |i - j| = 1, \\ 0 & \text{otherwise.} \end{cases} \quad (3.36)$$

The Galerkin equations (3.23) yield in this case the linear tridiagonal system

$$\begin{aligned} -s_{i-1} + 2s_i - s_{i+1} &= h \int_0^1 f(x) \varphi_i(x) dx, & i = 1, \dots, N-1, \\ s_0 &= s_N = 0, \end{aligned} \quad (3.37)$$

for the unknown coefficients  $s_i$  in the representation  $u_h(x) = \sum_{i=1}^{N-1} s_i \varphi_i(x)$  of the approximate solution. Because  $\varphi_i(x_j) = \delta_{ij}$ , we have the important property  $s_i = u_h(x_i)$  for all  $i$ .

For smooth  $f$  there exists a constant  $L$  such that

$$\left| f(x_i) - \frac{1}{h} \int_0^1 f(x) \varphi_i(x) dx \right| \leq \frac{2}{3} L h^2 \quad \text{for } i = 1, \dots, N-1.$$

This observation reveals the affinity of (3.37) with the standard central difference scheme for the boundary value problem (3.29). More precisely, (3.37) is a difference scheme where each function value  $f(x_i)$  is replaced by the integral mean  $\frac{1}{h} \int_0^1 f(x) \varphi_i(x) dx$ .  $\square$

We hope that the examples discussed above make clear the importance of choosing a good discrete space  $V_h$  in the Galerkin method.

The finite element method, which we shall discuss in great detail in Chapter 4, generalizes the choice of ansatz functions in Example 3.39: one uses ansatz functions—often piecewise polynomials—with a relatively small support

$$\text{supp } \varphi_i := \text{cl}_{\mathbb{R}^n} \{x \in \Omega : \varphi_i(x) \neq 0\},$$

and one aims to ensure that the quantity

$$\sum_{i=1}^N \text{card}\{j \in \{1, \dots, N\} : (\text{supp } \varphi_i \cap \text{supp } \varphi_j) \neq \emptyset\}$$

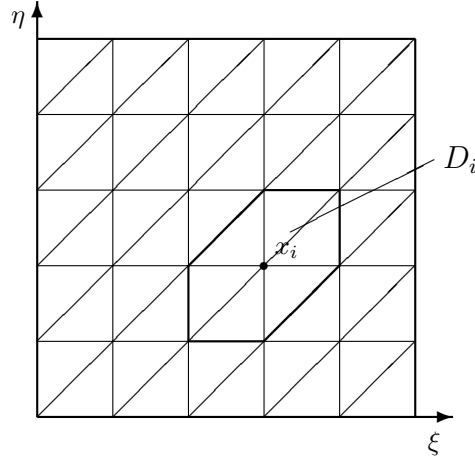
is not too large since it is an upper bound for the number of nonzero elements in the stiffness matrix  $A = (a(\varphi_j, \varphi_i))_{i,j=1}^N$  of the Galerkin system (3.25).

Let us go through the details of the method for a simple example in 2D:

*Example 3.40.* Let  $\Omega = (0,1) \times (0,1) \subset \mathbb{R}^2$ . Consider the boundary value problem

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega, \\ u|_{\Gamma} &= 0. \end{aligned} \quad (3.38)$$

We choose  $V = H_0^1(\Omega)$  for the weak formulation and decompose  $\Omega$  into a uniform triangular mesh as in Figure 3.2:



**Figure 3.2** uniform triangular mesh

The decomposition of  $\Omega$  is generated by a uniform mesh of mesh size  $h = 1/N$  in each of the coordinate directions  $\xi$  and  $\eta$ , then the resulting squares are bisected by drawing diagonals as in Figure 3.2.

Denote the inner mesh points by  $x_i = \begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix}$  for  $i = 1, \dots, M$  with  $M = (N-1)^2$ , and the points on the boundary by  $x_i$  for  $i = M+1, \dots, N^2$ . Analogously to Example 3.39 we define the piecewise linear ansatz functions  $\varphi_i \in C(\bar{\Omega})$  indirectly by the property

$$\varphi_i(x_j) := \delta_{ij}, \quad i = 1, \dots, M, \quad j = 1, \dots, N. \quad (3.39)$$

Then for the support of each basis function  $\varphi_i$  we have

$$\text{supp } \varphi_i = \left\{ \begin{pmatrix} \xi \\ \eta \end{pmatrix} \in \bar{\Omega} : |\xi - \xi_i| + |\eta - \eta_i| + |\xi - \eta - \xi_i + \eta_i| \leq 2h \right\}.$$

Using the bilinear form

$$a(u, v) = \int_{\Omega} \nabla u \nabla v \, dx,$$

the Galerkin method generates the linear system

$$As = b \quad (3.40)$$

with stiffness matrix  $A = (a_{ij})_{i,j=1}^M$  and right-hand side vector  $b = (b_i)_{i=1}^M$ . A direct computation yields

$$a_{ij} = \begin{cases} 4 & \text{if } i = j, \\ -1 & \text{if } |\xi_i - \xi_j| + |\eta_i - \eta_j| = h, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$b_i = \int_{\Omega} f(x) \varphi_i(x) dx.$$

The small support of each basis function results in only five nonzero elements in each row of the stiffness matrix. Similarly to Example 3.39, we recognize the affinity of this method with the five-point difference scheme for the boundary value problem (3.38) that appeared in Chapter 2.  $\square$

The examples and model problems we have just studied, though they are relatively simple, nevertheless demonstrate some essential features of the Galerkin method:

- It is necessary to choose a discrete space that has good approximation properties and generates linear systems that can be solved efficiently.
- When using piecewise-defined ansatz functions one has to ensure that the discrete space satisfies  $V_h \subset V$ . As we shall see, this is not a problem for elliptic second-order problems but difficulties can arise with, e.g., fourth-order problems where globally smooth functions are needed and for the Stokes problem where some care is needed to satisfy the divergence condition.
- The computation of the stiffness matrix  $A = (a_{ij})_{ij}$  with  $a_{ij} = a(\varphi_j, \varphi_i)$  and the vector  $b$  of the Galerkin equations both require, in general, the application of numerical integration.

These and other requirements have lead to intensive work on several manifestations of the Galerkin method. The most popular variants are *spectral methods*, where (usually) orthogonal polynomials are used as ansatz functions—for an excellent overview of spectral methods see [QV94] and the recent [CHQZ06]—and the *finite element method* where splines are used as ansatz functions. In Chapter 4 we shall examine the finite element method in detail; as well as presenting the basic facts and techniques, we also discuss advances in the method and its practical implementation.

**Exercise 3.41.** Approximately solve the boundary value problem

$$Lu := u'' - (1 + x^2)u = 1 \quad \text{on } (-1, 1), \quad u(-1) = u(1) = 0,$$

using the ansatz

$$\tilde{u}(x) = c_1\varphi_1(x) + c_2\varphi_2(x) \quad \text{with} \quad \varphi_1(x) = 1 - x^2, \quad \varphi_2(x) = 1 - x^4.$$

Determine  $c_1$  and  $c_2$

- a) by means of the Ritz-Galerkin technique;  
 b) using the “Galerkin equations”

$$(L\tilde{u} - 1, \varphi_1) = 0, \quad (L\tilde{u} - 1, \varphi_2) = 0;$$

- c) by computing

$$\min_{\tilde{u}} \int_{-1}^1 [(\tilde{u}')^2 + (1 + x^2)\tilde{u}^2 + 2\tilde{u}] dx.$$

**Exercise 3.42.** Consider the boundary value problem

$$\begin{aligned} -\Delta u(x, y) &= \pi^2 \cos \pi x \quad \text{in } \Omega = (0, 1) \times (0, 1), \\ \frac{\partial u}{\partial n} &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

- a) Construct the weak formulation and compute a Ritz-Galerkin approximation  $\tilde{u}$  using the basis

$$\varphi_1(x, y) = x - 1/2, \quad \varphi_2(x, y) = (x - 1/2)^3.$$

- b) Verify that the problem formulated in a) has a unique solution in

$$W = \left\{ v \in H^1(\Omega) : \int_{\Omega} v = 0 \right\}$$

and that  $\tilde{u} \in W$ .

- c) Verify that the problem formulated in a) does not have a unique classical solution in  $C^2(\Omega)$ . Determine the solution  $u$  in  $C^2(\Omega) \cap W$ . For the approximation  $\tilde{u}$ , determine

- the pointwise error at  $x = 0.25$
- the defect (i.e., amount by which it is in error) in the differential equation at  $x = 0.25$
- the defect in the boundary condition at  $x = 0$ .

**Exercise 3.43.** Let  $\Omega = \{(x, y) \in \mathbb{R}^2 : x > 0, y > 0, x + y < 1\}$ . Approximately determine the minimal eigenvalue in the eigenvalue problem

$$\Delta u + \lambda u = 0 \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega$$

by using the ansatz function  $\tilde{u}(x, y) = xy(1 - x - y)$  and computing  $\tilde{\lambda}$  from the Galerkin orthogonality property

$$\int_{\Omega} (\Delta \tilde{u} + \tilde{\lambda} \tilde{u}) \tilde{u} = 0.$$

**Exercise 3.44.** Let  $\Omega \subset \mathbb{R}^N$  be a bounded domain.

a) Verify that

$$\|u\|_{\Omega, c}^2 = \int_{\Omega} [|grad u|^2 + c(x)u^2] dx$$

defines a norm on  $V = H_0^1(\Omega)$  if the function  $c \in L_{\infty}(\Omega)$  is nonnegative almost everywhere.

b) Prove the coercivity over  $V$  of the bilinear form associated with the Laplacian and discuss the dependence of the coercivity constant on the norm used.

In the remaining sections of this chapter we shall present some generalizations of the earlier theory that include certain nonlinear features.

### 3.4 Weakening V-ellipticity

In Section 3.3 we investigated elliptic variational equations and used the Lax-Milgram lemma to ensure existence and uniqueness of solutions for both the continuous problem and its conforming Galerkin approximation. The  $V$ -ellipticity of the underlying bilinear form  $a(\cdot, \cdot)$  was a key ingredient in the proofs of the Lax-Milgram and Cea lemmas.

In the present section we weaken the  $V$ -ellipticity assumption. This is important, for example, when analysing finite element methods for first-order hyperbolic problems or mixed finite element methods.

First we study variational equations that satisfy some stability condition and hence derive results similar to the Lax-Milgram lemma.

Let  $V$  be a Hilbert space and  $a : V \times V \rightarrow \mathbb{R}$  a continuous bilinear form. Then there exists a constant  $M > 0$  such that

$$|a(u, v)| \leq M \|u\| \|v\| \quad \text{for all } u, v \in V. \quad (4.1)$$

Now we assume that the variational equation

$$a(u, v) = f(v) \quad \text{for all } v \in V \quad (4.2)$$

has for each  $f \in V^*$  a solution  $u \in V$  that satisfies the stability condition

$$\|u\| \leq \sigma \|f\|_* \quad (4.3)$$

for some constant  $\sigma > 0$ . This stability condition implies uniqueness of the solution of the variational equation (4.2): for if two elements  $\tilde{u}, \hat{u} \in V$  are solutions of (4.2), then the linearity of  $a(\cdot, \cdot)$  leads to

$$a(\tilde{u} - \hat{u}, v) = 0 \quad \text{for all } v \in V,$$

and now the estimate (4.3) yields

$$0 \leq \|\tilde{u} - \hat{u}\| \leq (\sigma)(0),$$

whence  $\tilde{u} = \hat{u}$ .

Consider a conforming Ritz-Galerkin approximation of the problem (4.2). Thus with  $V_h \subset V$  we seek  $u_h \in V_h$  such that

$$a(u_h, v_h) = f(v_h) \quad \text{for all } v_h \in V_h. \quad (4.4)$$

Analogously to the continuous problem, we require that for each  $f \in V^*$  the discrete problem (4.4) is solvable and that its solution  $u_h \in V_h$  satisfies

$$\|u_h\| \leq \sigma_h \|f\|_{*,h} \quad (4.5)$$

for some constant  $\sigma_h > 0$ . Here we used

$$\|f\|_{*,h} := \sup_{v_h \in V_h} \frac{|f(v_h)|}{\|v_h\|}.$$

Then, similarly to Cea's lemma, we obtain:

**Lemma 3.45.** *Assume that the bilinear form  $a(\cdot, \cdot)$  is continuous on  $V \times V$ , with  $M$  defined in (4.1). Assume that both the continuous problem (4.2) and the discrete problem (4.4) have solutions, and that the solution  $u_h$  of the discrete problem satisfies the stability estimate (4.5). Then the error of the Ritz-Galerkin approximation satisfies the inequality*

$$\|u - u_h\| \leq (1 + \sigma_h M) \inf_{v_h \in V_h} \|u - v_h\|.$$

**Proof:** Since  $u \in V$  and  $u_h \in V_h$  satisfy (4.2) and (4.4) respectively and  $V_h \subset V$ , we get

$$a(u - u_h, v_h) = 0 \quad \text{for all } v_h \in V_h.$$

Hence, for arbitrary  $y_h \in V_h$  one has

$$a(u_h - y_h, v_h) = a(u - y_h, v_h) \quad \text{for all } v_h \in V_h.$$

But  $a(u - y_h, \cdot) \in V^*$  so the stability estimate (4.5) implies that

$$\|u_h - y_h\| \leq \sigma_h \|a(u - y_h, \cdot)\|_{*,h}.$$

The continuity of  $a(\cdot, \cdot)$  and the property  $V_h \subset V$  then lead to

$$\|u_h - y_h\| \leq \sigma_h M \|u - y_h\|.$$

An application of the triangle inequality yields

$$\|u - u_h\| \leq \|u - y_h\| + \|y_h - u_h\| \leq (1 + \sigma_h M) \|u - y_h\|.$$

As  $y_h \in V_h$  is arbitrary, the statement of the lemma follows. ■



*Remark 3.46.* If for a family of discretizations the variational equations (4.4) are uniformly stable, i.e., there exists a constant  $\tilde{\sigma} > 0$  with

$$\sigma_h \leq \tilde{\sigma} \quad \text{for all } h < h_0,$$

with some  $h_0 > 0$  then, like Cea's lemma in the case of  $V$ -ellipticity, our Lemma 3.45 guarantees the quasi-optimality of the Ritz-Galerkin method.  $\square$

In Section 4.6 we shall apply these results to extended variational equations that correspond to so-called mixed formulations. Special conditions there will ensure existence of solutions and the uniform stability of the discrete problem.

Next we consider a different weakening of  $V$ -ellipticity. Recall that in Section 3.3 we already met Petrov-Galerkin methods, where it can be useful to choose differing ansatz and test spaces. This is of interest in various situations such as first-order hyperbolic problems, singularly perturbed problems, and error estimates in norms other than the norm on  $V$  (e.g. for second-order problems the norm on  $V$  is typically an “energy norm”, but one might desire an error estimate in the  $L_\infty$  norm).

*Example 3.47.* Let us consider the first-order hyperbolic convection problem

$$b \cdot \nabla u + cu = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \Gamma^-.$$

Here the inflow boundary of  $\Omega$  is defined by  $\Gamma^- = \{x \in \Gamma : b \cdot n < 0\}$ , where  $n$  is as usual an outer-pointing unit vector that is normal to the boundary  $\Gamma$ . Setting

$$W = L_2(\Omega), \quad V = H^1(\Omega)$$

and

$$a(u, v) = - \int_{\Omega} u \operatorname{div}(bv) + \int_{\Gamma \setminus \Gamma^-} (b \cdot n)uv + \int_{\Omega} cuv,$$

a standard weak formulation of the problem reads as follows:

Find  $u \in W$  such that

$$a(u, v) = f(v) \quad \text{for all } v \in V.$$

It turns out that for this problem it is useful to work with different ansatz and test spaces.  $\square$

Analogously to this example, consider the general problem: Find  $u \in W$  such that

$$a(u, v) = (f, v) \quad \text{for all } v \in V, \tag{4.6}$$

where  $W$  and  $V$  are Hilbert spaces that are not necessarily identical. The following generalization of the Lax-Milgram lemma goes back to Nečas (1962); its proof is similar to our earlier proof of Lax-Milgram. (see also [EG04])

**Theorem 3.48.** *Let  $W$  and  $V$  be two Hilbert spaces with norms  $\|\cdot\|_W$  and  $\|\cdot\|_V$ . Assume that the bilinear form  $a(\cdot, \cdot)$  on  $W \times V$  has the following properties (with constants  $C$  and  $\gamma > 0$ ):*

$$\begin{aligned} |a(w, v)| &\leq C\|w\|_W\|v\|_V \quad \text{for all } v \in V, w \in W, \\ \sup_{v \in V} \frac{a(w, v)}{\|v\|_V} &\geq \gamma\|w\|_W \quad \text{for all } w \in W, \end{aligned}$$

and

$$\sup_{w \in W} a(w, v) > 0 \quad \text{for all } v \in V.$$

Then (4.6) has for each  $f \in V^*$  a unique solution  $u$  with

$$\|u\|_W \leq \frac{1}{\gamma}\|f\|_*.$$

Babuška (see [8]) formulated the corresponding generalization of Cea's lemma using the discrete condition

$$\sup_{v_h} \frac{a(w_h, v_h)}{\|v_h\|_{V_h}} \geq \gamma_h\|w_h\|_{W_h} \quad \text{for all } w_h \in W_h, \quad (4.7)$$

for some constant  $\gamma_h > 0$ . It is important to note that the discrete condition (4.7) does not in general follow from its continuous counterpart. Nevertheless there are several techniques available to investigate its validity—see Chapter 4.6.

Babuška proved the error estimate

$$\|u - u_h\| \leq (1 + C/\gamma_h) \inf_{v_h \in V_h} \|u - v_h\|. \quad (4.8)$$

Recently it was shown in [124] that one can remove the constant 1 from this estimate.

Finally, as a third extension of  $V$ -ellipticity, we discuss  $V$ -coercivity.

Let  $V \subset H^1(\Omega)$  be a space related to the weak formulation of a problem based on a second-order differential operator. We say that a bilinear form  $a(\cdot, \cdot)$  is  $V$ -coercive if there exist constants  $\beta$  and  $\gamma > 0$  such that

$$a(v, v) + \beta\|v\|_0^2 \geq \gamma\|v\|_1^2 \quad \text{for all } v \in V.$$

In this situation the operator  $A : V \mapsto V^*$  defined by

$$\langle Av, w \rangle := a(v, w),$$

still satisfies the so-called Riesz-Schauder theory. With some further assumptions, one has (see Chapter 8 of [Hac03a]) the following error estimate for the Ritz-Galerkin method:

**Theorem 3.49.** *Assume that the variational equation*

$$a(u, v) = f(v) \quad \text{for all } v \in V,$$

*where the bilinear form is  $V$ -coercive, has a solution  $u$ . If the bilinear form  $a(\cdot, \cdot)$  is moreover continuous and satisfies*

$$\inf \{ \sup \{ |a(u, v)| : v \in V_h, \|v\| = 1 \} : u \in V_h, \|u\| = 1 \} = \gamma_h > 0,$$

*then the Ritz-Galerkin discrete problem has a solution  $u_h$  whose error is bounded by*

$$\|u - u_h\| \leq (1 + C/\gamma_h) \inf_{w \in V_h} \|u - w\|.$$

In [Hac03a] the validity of the inf-sup condition used in this theorem is discussed.

### 3.5 Extensions to Nonlinear Boundary Value Problems

In the previous sections we discussed abstract variational equations that treated only linear boundary value problems. Under certain conditions it is possible, however, to extend the technique used in the proof of the Lax-Milgram lemma—the construction of a suitably chosen contractive mapping—to more general differential operators. In this context monotone operators play an essential role; see [GGZ74, Zei90, ET76]. A different approach to proving the existence of solutions of nonlinear boundary value problems is to combine monotone iteration schemes with compactness arguments. To use this technique one needs assumptions that guarantee the monotonicity of the iteration process and carefully chosen starting points for the iteration; see [LLV85].

We now sketch the basic facts of the theory of monotone operators. This will enable us to apply the Galerkin method to some nonlinear elliptic boundary value problems.

Let  $V$  be a Hilbert space with scalar product  $(\cdot, \cdot)$  and let  $B : V \rightarrow V$  be an operator with the following properties:

i) There exists a constant  $\gamma > 0$  such that

$$(Bu - Bv, u - v) \geq \gamma \|u - v\|^2 \quad \text{for all } u, v \in V.$$

ii) There exists a constant  $M > 0$  such that

$$\|Bu - Bv\| \leq M \|u - v\| \quad \text{for all } u, v \in V.$$

Property (i) is called *strong monotonicity*, and (ii) *Lipschitz continuity* of the operator  $B$ .

Consider the abstract operator equation: find  $u \in V$  with

$$Bu = 0. \tag{5.1}$$

This is equivalent to the nonlinear variational equation

$$(Bu, v) = 0 \quad \text{for all } v \in V. \quad (5.2)$$

The next statement generalizes the Lax-Milgram lemma:

**Lemma 3.50.** *Assume that  $B$  is monotone and Lipschitz continuous. Then equation (5.1) has a unique solution  $u \in V$ . This solution is a fixed point of the auxiliary operator  $T_r : V \rightarrow V$  defined by*

$$T_r v := v - rBv, \quad v \in V,$$

*which is contractive when the parameter  $r$  lies in  $(0, \frac{2\gamma}{M^2})$ .*

**Proof:** As in the proof of the Lax-Milgram lemma we check whether  $T_r$  is contractive:

$$\begin{aligned} \|T_r y - T_r v\|^2 &= \|y - rBy - [v - rBv]\|^2 \\ &= \|y - v\|^2 - 2r(By - Bv, y - v) + r^2\|By - Bv\|^2 \\ &\leq (1 - 2\gamma r + r^2 M^2)\|y - v\|^2 \quad \text{for all } y, v \in V. \end{aligned}$$

Hence  $T_r$  is indeed a contraction mapping for  $r \in (0, \frac{2\gamma}{M^2})$ . Consequently  $T_r$  possesses a unique fixed point  $u \in V$ , i.e.,

$$u = T_r u = u - rBu.$$

That is,  $u$  is a solution of the operator equation (5.1).

Uniqueness of the solution follows immediately from the strong monotonicity property using the same argument as in the proof of the Lax-Milgram lemma. ■

Next we consider operators  $A : V \rightarrow V^*$ . Here  $V^*$  denotes the dual space of  $V$  and  $\langle \cdot, \cdot \rangle$  the dual pairing, i.e.,  $\langle l, v \rangle$  denotes the value of the continuous linear functional  $l \in V^*$  applied to  $v \in V$ . We assume that  $A$  has the following properties:

i) The operator  $A$  is *strongly monotone*, i.e., there exists a constant  $\gamma > 0$  such that

$$\langle Au - Av, u - v \rangle \geq \gamma \|u - v\|^2 \quad \text{for all } u, v \in V.$$

ii) The operator  $A$  is *Lipschitz continuous*, i.e., there exists a constant  $M > 0$  such that

$$\|Au - Av\|_* \leq M \|u - v\| \quad \text{for all } u, v \in V.$$

Then the problem

$$Au = f \quad (5.3)$$

has for each  $f \in V^*$  a unique solution  $u \in V$ .

This statement follows immediately from Lemma 3.50 using the auxiliary operator  $B : V \rightarrow V$  defined by

$$Bv := J(Av - f), \quad v \in V.$$

Here  $J : V^* \rightarrow V$  denotes the Riesz operator that maps each continuous linear functional  $g \in V^*$  to an element  $Jg \in V$  such that

$$\langle g, v \rangle = (Jg, v) \quad \text{for all } v \in V.$$

Problem (5.3) is equivalent to the nonlinear variational equation

$$\langle Au, v \rangle = \langle f, v \rangle \quad \text{for all } v \in V. \quad (5.4)$$

Existence and uniqueness of solutions hold not only for the case of a Hilbert space  $V$ , as in fact it is sufficient that  $V$  be a reflexive Banach space—see [Zei90].

We now discuss two examples of nonlinear elliptic boundary value problems that can be treated with the theory of this section. Note however that some important practical problems cannot be dealt with using monotone and Lipschitz continuous operators; then more sophisticated techniques are necessary. First we present a semi-linear problem and then a special quasi-linear boundary value problem.

*Example 3.51.* Let  $\Omega \subset \mathbb{R}^2$  be a bounded domain with smooth boundary  $\Gamma$ . Consider the weakly nonlinear problem

$$\begin{aligned} -\operatorname{div}(M \operatorname{grad} u) + F(x, u(x)) &= 0 \quad \text{in } \Omega, \\ u|_{\Gamma} &= 0. \end{aligned} \quad (5.5)$$

Here  $M = M(x) = (m_{ij}(x))$  is a matrix-valued function satisfying the estimate

$$\bar{\sigma}\|z\|^2 \geq z^T M(x)z \geq \underline{\sigma}\|z\|^2 \quad \text{for all } x \in \Omega, z \in \mathbb{R}^2, \quad (5.6)$$

for some constants  $\bar{\sigma} \geq \underline{\sigma} > 0$ . Furthermore, let  $F : \bar{\Omega} \times \mathbb{R} \rightarrow \mathbb{R}$  be a continuous function with the properties

$$\left. \begin{aligned} |F(x, s) - F(x, t)| &\leq L|s - t| \\ (F(x, s) - F(x, t))(s - t) &\geq 0 \end{aligned} \right\} \quad \text{for all } x \in \Omega, s, t \in \mathbb{R}, \quad (5.7)$$

where  $L$  is some constant. Choose  $V = H_0^1(\Omega)$ . Define a mapping  $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$  by

$$a(u, v) := \int_{\Omega} [(\nabla u)^T M^T \nabla v + F(x, u(x))v] \, dx, \quad u, v \in V.$$

For fixed  $u \in V$  our assumptions guarantee that  $a(u, \cdot) \in V^*$ . We define the related operator  $A : V \rightarrow V^*$  by  $Au := a(u, \cdot)$  and study its properties.

First we obtain

$$\begin{aligned} & |\langle Au, v \rangle - \langle Ay, v \rangle| \\ &= \left| \int_{\Omega} [(\nabla u - \nabla y)^T M^T \nabla v + (F(x, u(x)) - F(x, y(x)))v(x)] dx \right| \\ &\leq \int_{\Omega} (\bar{\sigma} \|\nabla u - \nabla y\| \|\nabla v\| + L \|u - y\| \|v\|) dx \\ &\leq c \|u - y\| \|v\|. \end{aligned}$$

Hence the operator  $A$  is Lipschitz continuous.

Friedrichs' inequality, (5.6) and (5.7) give the following estimates:

$$\begin{aligned} \langle Au - Av, u - v \rangle &= \int_{\Omega} \nabla(u - v)^T M^T \nabla(u - v) dx \\ &\quad + \int_{\Omega} (F(x, u(x)) - F(x, v(x)))(u(x) - v(x)) dx \\ &\geq \underline{\sigma} \int_{\Omega} \nabla(u - v) \nabla(u - v) dx \\ &\geq \underline{\sigma} \gamma \|u - v\|^2 \quad \text{for all } u, v \in V. \end{aligned}$$

Thus  $A$  is strongly monotone as well and our earlier theory is applicable.  $\square$

The next example sketches the analysis of a quasi-linear boundary value problem. This is more difficult to handle than Example 3.51, so we omit the details which can be found in [Zei90].

*Example 3.52.* Consider the following equation, where a nonlinearity appears in the main part of the differential operator:

$$-\sum_i \frac{\partial}{\partial x_i} \left( \varphi(x, |Du|) \frac{\partial u}{\partial x_i} \right) = f(x) \quad \text{in } \Omega.$$

Assume homogeneous Dirichlet boundary conditions for  $u$ , and that  $\varphi$  is a continuous function satisfying the following conditions:

- (i)  $\varphi(x, t)t - \varphi(x, s)s \geq m(t - s)$  for all  $x \in \Omega$ ,  $t \geq s \geq 0$ ,  $m > 0$ ;
- (ii)  $|\varphi(x, t)t - \varphi(x, s)s| \leq M|t - s|$  for all  $x \in \Omega$ ,  $t, s \geq 0$ ,  $M > 0$ .

If, for instance,  $\varphi(x, t) = g(t)/t$  and  $g$  is differentiable, then both these conditions are satisfied if

$$0 < m \leq g'(t) \leq M.$$

Under these hypotheses one can show that the theory of monotone and Lipschitz continuous operators is applicable, and deduce the existence of weak solutions for this nonlinear boundary value problem. Of course, the conditions (i) and (ii) are fairly restrictive.  $\square$

If one has both strong monotonicity and Lipschitz continuity, then it is not difficult to generalize Cea's lemma:

**Lemma 3.53.** *Let  $A : V \rightarrow V^*$  be a strongly monotone, Lipschitz continuous operator. Let  $f \in V^*$ . If  $V_h \subset V$  is a finite-dimensional subspace of  $V$ , then there exists a unique  $u_h \in V_h$  satisfying the discrete variational equation*

$$\langle Au_h, v_h \rangle = \langle f, v_h \rangle \quad \text{for all } v_h \in V_h. \quad (5.8)$$

Moreover, the error  $u - u_h$  of the Galerkin method satisfies the quasi-optimality estimate

$$\|u - u_h\| \leq \frac{M}{\gamma} \inf_{v_h \in V_h} \|u - v_h\|.$$

**Proof:** The finite dimensionality of  $V_h$  implies that it is a closed subspace of  $V$  and therefore it too is a Hilbert space with the same inner product as  $V$ . Clearly  $A$  is strongly monotone and Lipschitz continuous on  $V_h$ . These properties yield existence and uniqueness of a solution  $u_h \in V_h$  of the discrete problem (5.8).

From (5.4), (5.8),  $V_h \subset V$  and  $u_h \in V_h$  we have

$$\langle Au - Au_h, v_h - u_h \rangle = 0 \quad \text{for all } v_h \in V_h.$$

This identity, strong monotonicity and Lipschitz continuity together imply

$$\begin{aligned} \gamma \|u - u_h\|^2 &\leq \langle Au - Au_h, u - u_h \rangle \\ &= \langle Au - Au_h, u - v_h \rangle \\ &\leq M \|u - u_h\| \|u - v_h\| \quad \text{for all } v_h \in V_h, \end{aligned}$$

and the desired result follows.  $\blacksquare$

Unlike the case of linear boundary value problems, the Galerkin equations (5.8) are now a set of nonlinear equations. With the ansatz

$$u_h(x) = \sum_{j=1}^N s_j \varphi_j(x),$$

the Galerkin equations are equivalent to the nonlinear system

$$\langle A(\sum_{j=1}^n s_j \varphi_j), \varphi_i \rangle = \langle f, \varphi_i \rangle, \quad i = 1, \dots, N.$$

In principle, this nonlinear system can be solved by standard techniques such as Newton's method; see [Sch78, OR70]. But because the number of unknowns is in general large and the conditioning of the problem is bad, one should take advantage of the special structure of the system.

Moreover, one can systematically use information from discretizations on coarser meshes to obtain good starting points for iterative solution of the system on finer meshes. In [4] a variant of Newton's method exploits properties of the discretization on coarse and finer meshes. In [OR70] Newton's method is combined with other iterative methods suited to the discretization of partial differential equations.

Alternatively, one could deal with a nonlinear problem at the continuous level by some successive linearization technique (such as Newton's method) applied in an infinite-dimensional function space. The application of Newton's method does however demand strong regularity assumptions because differentiation is required.

A further linearization technique at the continuous level is the method of *frozen coefficients*. To explain this technique we consider the boundary value problem

$$\begin{aligned} -\operatorname{div}(D(x, u, \nabla u) \operatorname{grad} u) &= f & \text{in } \Omega, \\ u|_T &= 0, \end{aligned} \tag{5.9}$$

with a symmetric positive-definite matrix-valued function  $D(\cdot, \cdot, \cdot)$ . Let  $u^0 \in V = H_0^1(\Omega)$  be a suitably chosen starting point for the iteration. Then a sequence  $\{u^k\} \subset V$  of approximate solutions for (5.9) is generated, where (given  $u^k$ ) the function  $u^{k+1}$  is a solution of the linear problem

$$\int_{\Omega} \nabla u^{k+1} D(x, u^k, \nabla u^k) \nabla v \, dx = \int_{\Omega} f v \, dx \quad \text{for all } v \in V.$$

This technique is also known as the *secant modulus* method or *Kačanov method*. Its convergence properties are examined in [Neč83] and [Zei90].



<http://www.springer.com/978-3-540-71582-5>

Numerical Treatment of Partial Differential Equations

Grossmann, C.; Roos, H.-G.; Stynes, M.

2007, XII, 596 p. 86 illus., Softcover

ISBN: 978-3-540-71582-5