3.2. The general Dirichlet problem

We consider a general linear operator on an open set $\Omega \subset \mathbb{R}^n$

$$L(x, \nabla) = \sum_{|\alpha| \leq k} a_\alpha(x) \nabla^\alpha \qquad x \in \Omega$$

which is elliptic, i.e.

$$L^p(x, i\xi) = i^k \sum_{|\alpha| = k} a_\alpha(x) \xi^\alpha \neq 0 \qquad \forall \xi \in \mathbb{R}^n \backslash \{0\}.$$

It is easy to see, that (for $n \geq 2$) k must be even and with some smoothness assumptions on the coefficients a_α one can show that L can be written in *divergence form*

$$L(x, \nabla)u = \sum_{0 \leq |\sigma|, |\gamma| \leq m} (-1)^{|\sigma|} \nabla^\sigma (a_{\sigma_\gamma}(x) \nabla^\gamma u) \qquad x \in \Omega$$

with suitable coefficients $a_{\sigma\gamma}$ (see Exercise 3.1). The divergence form strongly simplifies a weak formulation of boundary value problems related to L.

As example, let us consider the classical Dirichlet problem, to find $u \in C^{2m}(\Omega) \cap C^m(\bar{\Omega})$ such that

$$\begin{aligned} L(x, \nabla)u &= f \qquad \text{in } \Omega \\ \nabla^\alpha u &= 0 \qquad \text{on } \partial\Omega, \ |\alpha| \leq m - 1 \end{aligned} \tag{3.13}$$

where $f \in C(\Omega)$. To set up a weak formulation, we multiply the equation by a test function $v \in C_0^\infty(\Omega)$ and integrate by parts. This leads to the equation

$$\sum_{0 \leq |\sigma|, |\gamma| \leq m} \int_\Omega a_{\sigma_\gamma} \nabla^\gamma u \nabla^\sigma v \, dx = \int_\Omega f v \, dx.$$

Introducing the bilinear form

$$a(u, v) = \sum_{0 \leq |\sigma|, |\gamma| \leq m} \int_\Omega a_{\sigma_\gamma} \nabla^\gamma u \nabla^\sigma v \, dx \tag{3.14}$$

and the linear form

$$l(v) = \int_\Omega f v \, dx$$

we obtain

Definition 3.14. *Let $\Omega \subset \mathbb{R}^n$ be open. In* (3.14)*, let $a_{\sigma,\gamma} \in \mathbb{L}^\infty(\Omega)$ for all $0 \leq |\sigma|, |\gamma| \leq m$ and $l \in H^{-m}(\Omega)$ (the dual space of $H_0^m(\Omega)$). A function $u \in H_0^m(\Omega)$ is called weak solution of the Dirichlet problem* (3.13) *if*

$$a(u, v) = l(v) \qquad \forall v \in H_0^m(\Omega).$$

Similar to the examples in the previous section, we can generalize the homogeneous case to more general Dirichlet conditions. For example, we can require that $\nabla^\alpha u = \nabla^\alpha g$ on $\partial\Omega$ for all $|\alpha| \leq m - 1$ where $g \in H^m(\Omega)$ is a given function.

3.3. The Lax Milgram Theorem

The weak formulation derived in the previous section can easily be stated completely in functional analytic terms:

Definition 3.15. *Let V be a Hilbert space, $a : V \times V \mapsto \mathbb{R}$ a bilinear form and $l \in V'$ an element of the dual space. The problem to find $u \in V$ such that*

$$a(u,v) = l(v) \qquad \forall v \in V \tag{3.15}$$

is called abstract variational problem.

The name is motivated by the observation that (3.15) is equivalent to an optimization problem, if a is a *symmetric bilinear form.*

Lemma 3.16. *Let a be a symmetric bilinear form on a Hilbert space V which satisfies $a(v,v) \geq 0$ for all $v \in V$. Further, let $l \in V'$ be a bounded linear functional. Then* (3.15) *is equivalent to the variational problem to find $u \in V$ which minimizes the functional*

$$\Theta(u) = \frac{1}{2}a(u,u) - l(u).$$

Proof: If u is a minimizer of the functional Θ, then

$$f(t) := \Theta(u + tv)$$

is minimal at $t = 0$ for any $v \in V$. In particular,

$$\left.\frac{d}{dt}f(t)\right|_{t=0} = 0.$$

To calculate the derivative, we first use the bilinearity of a to write

$$f(t) = \Theta(u) + (a(u,v) - l(v))t + \frac{1}{2}a(v,v)t^2. \tag{3.16}$$

Clearly, the vanishing derivative in $t = 0$ implies

$$a(u,v) - l(v) = 0 \qquad \forall v \in V$$

which shows that u solves (3.15). Conversely, if u is a solution of the abstract variational problem, we find for any $v \in V$ by setting $t = 1$ in (3.16)

$$\Theta(u+v) = \Theta(u) + (a(u,v) - l(v)) + \frac{1}{2}a(v,v).$$

Since $(a(u,v) - l(v)) = 0$ and $a(v,v) \geq 0$ we see that $\Theta(u+v) \geq \Theta(u)$ for all $v \in V$ which shows that u is a minimizer of Θ. ■

For problems arising from physical models, one very often finds that Θ has the meaning of an energy. For example, in electrostatics the equation for the potential ϕ of an electric field $E = \nabla\phi$ which is generated by a charge distribution ρ is just Poisson's equation $-\Delta\phi = \rho$ (for example with boundary conditions $\phi = 0$ if the boundary of the domain is 'grounded'). We already know the bilinear form a corresponding to the Dirichlet Poisson problem. It is given by the $\mathbb{L}^2$ scalar product of the gradients of the functions. Since the gradient of ϕ is the electric field E, the functional Θ has the form

$$\Theta(\phi) = \frac{1}{2}\int_\Omega \|E\|^2\,dx - \int_\Omega \phi\rho\,dx.$$

While the first integral describes the electric energy stored in the field E, the second term is the potential energy of the charge distribution ρ in the potential ϕ so that Θ is indeed an energy functional.

In order to get a general solution theory for abstract variational problems we need as minimal requirements *continuity* and *ellipticity* of the bilinear form a.

Definition 3.17. *Let $a : V \times V \mapsto \mathbb{R}$ be a bilinear form on the Hilbert space V. We say that a is continuous if there exists a constant $C_1 > 0$ such that*

$$|a(u,v)| \le C_1\|u\|_V\|v\|_V \qquad \forall u, v \in V.$$

a is called V–elliptic (or simply elliptic) if, for some $C_2 > 0$

$$a(u,u) \ge C_2\|u\|_V^2 \qquad \forall u \in V.$$

We remark that ellipticity of L is a weaker condition than ellipticity of the corresponding bilinear form a (see Exercise 3.5). Also, $-\Delta$ is an elliptic operator but the corresponding form

$$a(u,v) = \langle \nabla u, \nabla v\rangle_0 \tag{3.17}$$

is in general not $H^1(\Omega)$–elliptic (reflecting the non uniqueness of the Neumann Poisson problem). Obviously, a contains not only information about the differential operator but also about the boundary conditions of the problem (which are hidden in the domain of definition of a). For the bilinear form (3.17) on the Hilbert space $V = H_0^1(\Omega)$ (which appears in the connection with the Dirichlet Poisson problem), continuity follows immediately from the Cauchy-Schwarz inequality in $\mathbb{L}^2$

$$|a(u,v)| \le \|\nabla u\|_{\mathbb{L}^2}\|\nabla v\|_{\mathbb{L}^2} \le \|u\|_{1,2}\|v\|_{1,2}.$$

Typically, the more difficult part is to prove ellipticity. Indeed, for the Poisson problem we needed the Poincaré inequality to show

$$a(u,u) = \langle \nabla u, \nabla u\rangle_0 \ge C_2\|u\|_{1,2}^2 \qquad \forall u \in H_0^1(\Omega).$$

If a is a symmetric bilinear form (as in the Poisson problem) a simple application of the Riesz representation theorem yields a unique solution of the abstract variational problem.

Lemma 3.18. *Let V be a Hilbert space and let $a : V \times V \mapsto \mathbb{R}$ be a symmetric, continuous and elliptic bilinear form. Then* (3.15) *admits a unique solution u which depends continuously on the data (i.e. $\|u\|_V \le C\|l\|_{V'}$).*

Proof: The form a is clearly a scalar product since it is bilinear, symmetric and satisfies $a(u,u) \ge C_2\|u\|_V^2 > 0$ for all $u \ne 0$. Using continuity and ellipticity we find for the associated norm $\|u\|_a := \sqrt{a(u,u)}$

$$\sqrt{C_2}\|u\|_V \leq \|u\|_a \leq \sqrt{C_1}\|u\|_V$$

which shows that $\|\cdot\|_a$ gives rise to the original (complete) topology on V. Altogether, V is Hilbert space also with the scalar product $a(\cdot,\cdot)$. Due to the equivalence of norms, any linear functional $l \in V'$ is also continuous with respect to $\|\cdot\|_a$. Thus Theorem 3.5 can be applied and we obtain a unique solution u of the abstract variational problem which satisfies

$$\|u\|_a = \sup_{\|v\|_a=1} |l(v)|.$$

Using ellipticity twice, we get

$$\sqrt{C_2}\|u\|_V \leq \|u\|_a \leq \|l\|_{V'} \sup_{\|v\|_a=1} \|v\|_V \leq \frac{1}{\sqrt{C_2}}$$

which yields the continuous dependence in the original norm of V. ■

Unfortunately, the Lemma applies only to symmetric bilinear forms so that many elliptic operators are excluded. For example, let us consider a simple, one dimensional situation with a second order operator with constant coefficients

$$Lu = -(a_{11}u')' - (a_{10}u)' + a_{01}u' + a_{00}u.$$

In this case, the bilinear form is

$$a(u,v) = \int_\Omega a_{11}u'v' + a_{10}uv' + a_{01}u'v + a_{00}uv\,dx.$$

Assuming symmetry of a, the relation $0 = a(u,v) - a(v,u)$ implies

$$(a_{10} - a_{01}) \int_\Omega uv' - vu'\,dx = 0 \qquad \forall u,v \in H_0^1(\Omega).$$

Using integration by parts, we get

$$2(a_{10} - a_{01}) \int_\Omega u'v\,dx = 0 \qquad \forall u,v \in H_0^1(\Omega).$$

Finally, by choosing a function $v \geq 0$ in $C_0^\infty(\Omega)$ and a function u with $u' > 0$ on the support of v we can conclude $a_{01} = a_{10}$. Going back to the corresponding operator L

$$Lu = -a_{11}u'' + (a_{01} - a_{10})u' + a_{00}u = -a_{11}u'' + a_{00}u$$

we see that L cannot contain first derivatives if a is a symmetric bilinear form. To incorporate also non symmetric forms in our general treatment we therefore need to extend the Riesz representation theorem. The corresponding result for general bilinear forms is called Lax Milgram Theorem. It plays a central role in the theory of linear elliptic partial differential equations because it provides existence and uniqueness results for weak solutions.

Theorem 3.19. (Lax Milgram)
Let V be a Hilbert space, $l \in V'$ a continuous linear functional and $a : V \times V \mapsto \mathbb{R}$ a continuous, elliptic bilinear form. Then, the abstract variational problem (3.15) admits a unique solution $u \in V$ which satisfies $\|u\|_V \le C\|l\|_{V'}$.

Proof: In the first step, we transform the abstract variational problem into an operator equation of the form $\tilde{A}u = \tilde{l}$. To do this, we fix $u \in V$ and define the functional

$$a_u(v) := a(u, v) \qquad \forall v \in V.$$

Clearly, a_u is linear and we have

$$|a_u(v)| = |a(u, v)| \le C_1\|u\|_V\|v\|_V$$

which implies $\|a_u\|_{V'} \le C_1\|u\|_V$. In particular, $a_u \in V'$ and we can define

$$\tilde{A}u := \mathcal{R}a_u$$

where $\mathcal{R}$ is the Riesz operator from Corollary 3.8. By definition of $\mathcal{R}$, we have

$$a(u, v) = a_u(v) = \langle \mathcal{R}a_u, v\rangle_V = \left\langle \tilde{A}u, v\right\rangle_V$$

which leads to the equivalent reformulation of problem (3.15)

$$a(u, v) = l(v)\ \forall v \in V \quad \Leftrightarrow \quad a_u = l \quad \Leftrightarrow \quad \tilde{A}u = \mathcal{R}l =: \tilde{l}$$

which is the desired operator formulation.
In the second step of the proof we show that $\tilde{A}$ is a linear bounded operator which has a continuous inverse. This implies that $\tilde{A}u = \tilde{l}$ (and thus also the variational problem) has a unique solution $u = \tilde{A}^{-1}\tilde{l}$. Continuity of $\tilde{A}^{-1}$ obviously gives continuous dependence on the data.
Since $\tilde{A}$ is the composition of $\mathcal{R}$ and a_u linearity follows from the observation that

$$\begin{aligned} a_{\alpha_1 u_1 + \alpha_2 u_2}(v) &= a(\alpha_1 u_1 + \alpha_2 u_2, v) \\ &= \alpha_1 a(u_1, v) + \alpha_2 a(u_2, v) = (\alpha_1 a_{u_1} + \alpha_2 a_{u_2})(v). \end{aligned}$$

Boundedness follows from

$$\|\tilde{A}u\|_V = \|\mathcal{R}a_u\|_V = \|a_u\|_{V'} \le C_1\|u\|.$$

Existence of an inverse operator is shown with the Banach fixed point theorem (the idea is taken from [**5**]). We transform $\tilde{A}u = \tilde{l}$ in a fixed point problem $Tu = u$ with

$$Tu = u - \rho(\tilde{A}u - \tilde{l}) \qquad u \in V$$

T is an affine linear map which is a contraction mapping if we choose ρ in a suitable way. Indeed,

$$\begin{aligned}\|Tu_1 - Tu_2\|_V^2 &= \|u_1 - u_2\|_V^2 - 2\rho \langle A(u_1 - u_2), u_1 - u_2\rangle_V + \rho^2 \|A(u_1 - u_2)\|_V^2 \\ &\le (1 + \rho^2 C_1^2)\|u_1 - u_2\|_V^2 - 2\rho a(u_1 - u_2, u_1 - u_2) \\ &\le (1 - 2C_2\rho + \rho^2 C_1^2)\|u_1 - u_2\|_V^2 \\ &= \left(1 - \frac{C_2^2}{C_1^2}\right)\|u_1 - u_2\|_V^2\end{aligned}$$

where the last line follows if we choose $\rho = C_2/C_1^2$. According to the Banach fixed point theorem, T has a unique fixed point u which is the limit of the sequence $u_{n+1} = Tu_n$ (in fact, this is *Richardson iteration with relaxation* – see Appendix A). Due to the construction of T, the fixed point u is a solution of $\tilde{A}u = \tilde{l}$ which shows existence and uniqueness. To get continuity of $\tilde{A}^{-1}$ (or equivalently continuous dependence on the data) we can use again a result from the Banach fixed point theorem

$$\|u\|_V = \|Tu\|_V \le \|Tu - T0\|_V + \|T0\|_V \le \sqrt{1 - \frac{C_2^2}{C_1^2}}\|u\|_V + \rho\|\tilde{l}\|_V.$$

Another possibility is

$$\begin{aligned}\|\tilde{A}^{-1}\tilde{l}\|_V^2 = \frac{1}{C_2}C_2\|u\|_V^2 \le \frac{1}{C_2}a(u,u) = \frac{1}{C_2}\left\langle \tilde{A}u, u\right\rangle_V \\ \le \frac{1}{C_2}\|\tilde{A}u\|_V\|u\|_V = \frac{1}{C_2}\|\tilde{l}\|_V\|\tilde{A}^{-1}\tilde{l}\|_V.\end{aligned}$$

■

In the proof of the Lax Milgram theorem we have transformed the abstract variational problem into an operator equation. This approach turns out to be useful also for other purposes. We slightly reformulate the idea.

Corollary 3.20. *To any continuous bilinear form a on $V \times V$ we can uniquely associate a bounded linear operator $A : V \mapsto V'$ such that*

$$a(u,v) = \langle Au, v\rangle_{V' \times V} \qquad u, v \in V.$$

If a is elliptic, the operator A is a continuous bijection with $\|A^{-1}\| \le 1/C_2$ where C_2 is the ellipticity constant of a.

Proof: Let $\tilde{A}$ be the operator constructed in the proof of the Lax Milgram theorem. We set $A := \mathcal{R}^{-1}\tilde{A}$ so that $Au = \mathcal{R}^{-1}\tilde{A}u = \mathcal{R}^{-1}\mathcal{R}a_u = a_u$. Then

$$\langle Au, v\rangle_{V' \times V} = a_u(v) = a(u,v) \qquad \forall v \in V.$$

The remaining properties of A follow directly from those of $\tilde{A}$ and $\mathcal{R}$. ■

The representation of a in terms of A gives rise to an operator formulation of the abstract variational problem.

Let V be a Hilbert space, $l \in V'$ and let $a : V \times V \mapsto \mathbb{R}$ be a continuous bilinear form. If A denotes the operator associated to a then problem (3.15) *is equivalent to finding $u \in V$ such that*

$$Au = l.$$

Note that if a arises from an elliptic problem like (3.13), the operator A contains information about the differential operator L *and* the boundary conditions (which are hidden in the domain of definition of A).

In the particular case where a arises from the homogeneous Dirichlet problem (Section 3.2), the associated operator A can be considered as a direct extension of the differential operator L. This is due to the fact that $C_0^\infty(\Omega)$ is dense in $V = H_0^m(\Omega)$ and that the action of A on this dense subset is exactly the one of L. More precisely, for $\phi, \psi \in C_0^\infty(\Omega)$ we have by definition

$$\langle A\phi, \psi \rangle_{H^{-m} \times H_0^m} = a(\phi, \psi)$$

and using integration by parts, we find

$$\langle A\phi, \psi \rangle_{H^{-m} \times H_0^m} = \langle L\phi, \psi \rangle_{\mathbb{L}^2} = \left\langle \mathcal{R}_{\mathbb{L}^2}^{-1} L\phi, \psi \right\rangle_{(\mathbb{L}^2)' \times \mathbb{L}^2} \tag{3.18}$$

where $\mathcal{R}_{\mathbb{L}^2}$ is the Riesz operator on $\mathbb{L}^2$. We now want to identify the $\mathbb{L}^2$–functional $\mathcal{R}_{\mathbb{L}^2}^{-1} L\phi$ with an H_0^m–functional (to get an equivalent expression for $A\phi$). This is possible since $H_0^m(\Omega)$ is continuously embedded in $\mathbb{L}^2(\Omega)$.

Definition 3.21. *Let V, U be Banach spaces. We say that V is continuously embedded in U if there exists a bounded linear and injective mapping $I : V \mapsto U$ (a so called inclusion operator). Based on I, we can identify V with the linear subspace $I(V) \subset U$. If there is no ambiguity about I, we also write simply $V \hookrightarrow U$.*

Lemma 3.22. *Let V, U be Hilbert spaces and $V \hookrightarrow U$ based on the inclusion I. If $I(V)$ is dense in U then $U' \hookrightarrow V'$ based on the inclusion I' defined by*

$$\langle I'l, v \rangle_{V' \times V} := \langle l, Iv \rangle_{U' \times U} \qquad \forall l \in U',\ v \in V.$$

Proof: First, we note that I' is well defined. Indeed, since I is linear and bounded

$$\Lambda(v) := \langle l, Iv \rangle_{U' \times U}$$

defines an element of V' for each $l \in U'$. Setting $I'l := \Lambda$ we first check injectivity of the mapping. If $I'l_1 = I'l_2$ then clearly $\langle l_1 - l_2, Iv \rangle_{U' \times U} = 0$ for all $v \in V$. Since $I(V)$ is dense in U this implies $l_1 = l_2$. The linearity of I' is easily checked by using the linearity of the dual product and again the density of $I(V)$ in U. Finally

$$|(I'l)(v)| \le \|l\|_{U'} \|Iv\|_U \le \|I\| \, \|l\|_{U'} \|v\|_V.$$

shows that $\|I'\| \le \|I\|$. ■

In the particular case $V = H_0^m(\Omega)$ and $U = \mathbb{L}^2(\Omega)$ a natural inclusion is the identity mapping on $\mathbb{L}^2(\Omega)$ restricted to $H_0^m(\Omega)$ (since H_0^m is, by definition, a subset of $\mathbb{L}^2$). Moreover, since $C_0^\infty(\Omega) \subset H_0^m(\Omega)$ is dense in $\mathbb{L}^2(\Omega)$ we can apply Lemma 3.22 which shows that $(\mathbb{L}^2)' \hookrightarrow H^{-m}$ based on the adjoint of the natural inclusion of H_0^m into EL^2. Consequently, we can continue equation (3.18) as

$$\langle A\phi, \psi \rangle_{H^{-m} \times H_0^m} = \left\langle \mathcal{R}_{\mathbb{L}^2}^{-1} L\phi, \psi \right\rangle_{H^{-m} \times H_0^m} \qquad \forall \phi, \psi \in C_0^\infty(\Omega).$$

(The inclusion operator corresponding to $(\mathbb{L}^2)' \hookrightarrow H^{-m}$ is typically not mentioned because it is the natural choice.) Since $C_0^\infty(\Omega)$ is dense in $H_0^m(\Omega)$ we get

$$A\phi = \mathcal{R}_{\mathbb{L}^2}^{-1} L\phi \qquad \forall \phi \in C_0^\infty(\Omega)$$

which shows that A is the unique extension of $\mathcal{R}_{\mathbb{L}^2}^{-1} L$ from $C_0^\infty(\Omega)$ to $H_0^m(\Omega)$. We can even suppress the Riesz operator if we agree on the chain of inclusions

$$H_0^m(\Omega) \hookrightarrow \mathbb{L}^2(\Omega) = \mathbb{L}^2(\Omega)' \hookrightarrow H^{-m}(\Omega). \tag{3.19}$$

(This line has to be read in the following way: H_0^m is embedded in $\mathbb{L}^2$ based on the natural inclusion; $\mathbb{L}^2$ is identified with $(\mathbb{L}^2)'$ based on the Riesz isometry $\mathcal{R}_{\mathbb{L}^2}$; $(\mathbb{L}^2)'$ is included in H^{-m} based on the adjoint of the natural inclusion $H_0^m \hookrightarrow \mathbb{L}^2$.) If $I : H_0^m(\Omega) \mapsto H^{-m}(\Omega)$ denotes the inclusion generated by the chain (3.19) (i.e. the composition of natural inclusion, Riesz operator and adjoint of natural inclusion) we get again

$$\langle IL\phi, \psi \rangle_{H^{-m} \times H_0^m} = \langle A\phi, \psi \rangle_{H^{-m} \times H_0^m} \qquad \forall \phi, \psi \in C_0^\infty(\Omega).$$

A is the unique extension of IL which motivates the notation $A = IL$ (or even $A = L$ if it is clear that the inclusion is based on the chain (3.19)). An important feature of the inclusion (3.19) is its compactness. Indeed, $H_0^m(\Omega)$ is a compact subset of $\mathbb{L}^2(\Omega)$ which is equivalent to saying that the natural inclusion operator is compact. Since compositions of compact and bounded operators are compact, we conclude that I is a compact mapping. In particular, if a is an elliptic form, we know that A is invertible (mapping H^{-m} into H_0^m), so that IA^{-1} is a compact operator on $H^{-m}(\Omega)$ (or again, if we agree on I as the natural inclusion we can simply say that A^{-1} is compact). Recalling that $A^{-1}l$ is the solution of the abstract variational problem we conclude that the solution operator corresponding to boundary value problems which are associated to elliptic bilinear forms is a compact operator.

This important result enables us to apply spectral theory of compact operators and thus yields information about eigenvalues and eigensolutions of elliptic boundary value problems.

We conclude with the remark that the inclusion $I : H_0^m(\Omega) \mapsto H^{-m}(\Omega)$ based on (3.19) is not the only possible choice. Indeed, based on the inclusion $\mathcal{R}^{-1} : H_0^m(\Omega) \mapsto H^{-m}(\Omega)$ (the Riesz operator on $H_0^m(\Omega)$) one can even *identify* H_0^m with its dual H^{-m}. Note, however, that this embedding cannot be compact since the inverse of a compact operator cannot be bounded.

Also, the inclusion I is the most natural in connection with eigenvalue problems

$$\begin{aligned} L(x,\nabla)u &= \lambda u \qquad &&\text{in } \Omega \\ \nabla^\alpha u &= 0 &&\text{on } \partial\Omega,\ |\alpha| \le m-1 \end{aligned} \tag{3.20}$$

Indeed, the weak formulation of (3.20) is to find $u \in H_0^m(\Omega)$ such that

$$a(u,v) = \lambda \langle u,v\rangle_{\mathbb{L}^2} \qquad \forall v \in H_0^m(\Omega).$$

Note that the inclusion operator I is just the operator associated to the bilinear form $\langle u,v\rangle_{\mathbb{L}^2}$, i.e.

$$\langle Iu,v\rangle_{H^{-m}\times H_0^m} = \langle u,v\rangle_{\mathbb{L}^2}\,.$$

Consequently, the operator notation of the eigenvalue problem is

$$Au = \lambda I u$$

where A is associated to a. With our considerations above, we can as well write

$$I(Lu - \lambda u) = 0$$

(or even $Lu - \lambda u = 0$ if we suppress the injection I).

3.4. General eigenvalue problems

Up to now, we have tacitly assumed that V is a Hilbert space over $\mathbb{R}$. In the case of spectral analysis, however, it is favorable to work with the field $\mathbb{C}$ of complex numbers. The standard procedure in this case is a complexification of V, where V is embedded in a $\mathbb{C}$–space $\hat{V}$. More precisely, we define

$$\hat{V} := \{(u,v) \,:\, u,v \in V\} = V \times V$$

and use as notation $(u,v) = u + iv$. (Note that the complex unit i is only a symbol which is used to memorize the following definitions of addition, multiplication and the rules for extensions of functionals, mappings and bilinear forms.) We set

$$(u+iv) + (x+iy) := u + x + i(v+y)$$

and define multiplication with $\lambda = \alpha + i\beta \in \mathbb{C}$ by

$$(\alpha + i\beta)(u+iv) := (\alpha u - \beta v) + i(\alpha v + \beta u).$$

The projection on the first component of $V \times V$ is denoted $\Re$ and on the second component $\Im$. It is then natural to identify V with the subspace $\Re(V) = V \times \{0\}$ of *real vectors*. Based on $\langle\cdot,\cdot\rangle_V$ we define a scalar product on $\hat{V}$

$$\langle u+iv, x+iy\rangle_{\hat{V}} := \langle u,x\rangle_V + \langle v,y\rangle_V + i(\langle v,x\rangle_V - \langle u,y\rangle_V).$$

The definition is done in such a way that

$$\langle w,z\rangle_{\hat{V}} = \overline{\langle z,w\rangle_{\hat{V}}} \qquad w,z \in \hat{V}.$$

Moreover, if we apply $\langle\cdot,\cdot\rangle_{\hat{V}}$ to real vectors $w,z \in V$ (i.e. $\Im z = \Im w = 0$) we recover the V product. The same approach is used to extend any other bilinear form a. Note that the extension $\hat{a}$ is a sesquilinear form on V, i.e. it is linear in the first component, additive in the second and $\hat{a}(w, \lambda z) = \bar{\lambda}\hat{a}(w, z)$. All other extensions are done along the same lines. For example, a mapping $A : V \mapsto U$ is extended to $\hat{A} : \hat{V} \mapsto \hat{U}$ by

$$\hat{A}(u + iv) := Au + iAv.$$

(If $\hat{U}$ is already a $\mathbb{C}$ vector space, then $\hat{U} = U$ otherwise $\hat{U}$ is the complexification of U.) In the following, we can always think of V as a subset of a complex vector space. In this sense, spectral theory can be applied to objects defined on V by using their complex extensions. Of course, the results are in general related to the complexification $\hat{V}$ and perhaps have to be translated back into results for the original objects on V. However, since all extensions used in the complexification are very natural it is clear from the context how expressions involving $\mathbb{C}$-objects are to be understood.

With these remarks in mind, we now turn to the eigenvalue problem. It turns out that the underlying bilinear form a need not be elliptic. It suffices if a is V–coercive.

Definition 3.23. *Let V be a Hilbert space which is dense and continuously embedded into another Hilbert space U. A bilinear form $a : V \times V \mapsto \mathbb{R}$ is called V–coercive with respect to U if a is continuous and if there exists $C_V, C_U > 0$ such that*

$$a(v, v) + C_U \|v\|_U^2 \geq C_V \|v\|_V^2 \qquad \forall v \in V.$$

Note that coercivity is closely connected with ellipticity. Indeed, if a is V–elliptic then A is also V–coercive with respect to any U in which V is dense and continuously embedded. Conversely, if a is V–coercive with respect to U we can define a new bilinear form

$$\tilde{a}(u, v) = a(u, v) + C \langle u, v \rangle_U \qquad u, v \in V$$

which is V–elliptic if $C \geq C_U$. Note that the operator $\tilde{A}$ related to $\tilde{a}$ is given by

$$\tilde{A} = A + CI$$

where A is related to a. The operator $I : V \mapsto V'$ is the natural inclusion operator from V into V' related to the chain

$$V \hookrightarrow U = U' \hookrightarrow V'.$$

The first inclusion $V \hookrightarrow U$ of the chain is assumed in the current context. Since V is dense in U, we know from Lemma 3.22 that also $U' \hookrightarrow V'$ which explains the last inclusion of the chain. The identification $U = U'$ is naturally based on the Riesz operator on U. Hence, I is defined by the relation

$$\langle Iu, v \rangle_{V' \times V} = \langle u, v \rangle_U \qquad \forall u, v \in V. \tag{3.21}$$

Typically, the action of inclusion operators is unambiguous so that they need not be mentioned explicitly. In this spirit, we can write (3.21) in the form

$$\langle u, v \rangle_{V' \times V} = \langle u, v \rangle_U \qquad \forall v \in V, u \in U.$$

If the embedding $V \hookrightarrow U$ is compact (i.e. if V–bounded sequences have U–convergent subsequences), we have the following result.

Theorem 3.24. *Let V, U be Hilbert spaces which are arranged in the chain $V \hookrightarrow U = U' \hookrightarrow V'$ where V is assumed to be dense and compactly embedded in U. The inclusion operator from V into V' related to the chain is denoted I. Let $a : V \times V \mapsto \mathbb{R}$ be V–coercive with respect to U and let $A : V \mapsto V'$ be the operator related to a.*

a) *For every $\lambda \in \mathbb{C}$ one of the two alternatives holds*
   i) *$(A - \lambda I)^{-1}$ is a bounded linear operator from V' into V. In particular, $Au - \lambda I u = l$ is uniquely solvable for every $l \in V'$ (i.e. $a(u, v) - \lambda \langle u, v\rangle_U = l(v)$ for all $v \in V$).*
   ii) *The subspace $E(\lambda) = \ker(A - \lambda I)$ has dimension $1 \le \dim E(\lambda) < \infty$. ($E(\lambda)$ is the eigenspace corresponding to the eigenvalue λ). In particular, there exists $0 \neq u \in E(\lambda)$ such that $Au = \lambda I u$ (i.e. $a(u, v) = \lambda \langle u, v\rangle_U$ for all $v \in V$).*
b) *The spectrum $\sigma(A)$ of A consists of at most countably many eigenvalues which cannot have an accumulation point in $\mathbb{C}$.*

Proof: Since a is assumed to be V–coercive with respect to U we have already seen that $\tilde{a}(u, v) = a(u, v) + C_U \langle u, v\rangle_U$ is V–elliptic. Hence, the associated operator $\tilde{A} = A + C_U I$ has a bounded inverse. Using that I is a compact mapping we find that

$$K := (A + C_U I)^{-1} I$$

is a compact operator on V. Applying spectral theory for compact operators [**8**] we conclude that K has a pure point spectrum with zero as only possible accumulation point. Moreover, the eigenspaces are finite dimensional.

In order to use these results for the original mapping A, we consider the relation between eigenvalues of K and those of A. If I_V is the identity mapping on V we have

$$\begin{aligned} K - \mu I_V &= -\mu \left(I_V - \frac{1}{\mu}(A + C_U I)^{-1} I \right) \\ &= -\mu (A + C_U I)^{-1} \left(A + C_U I - \frac{1}{\mu} I \right) \\ &= -\mu (A + C_U I)^{-1} (A - \lambda I) \end{aligned}$$

where

$$\lambda = \frac{1}{\mu} - C_U \qquad \text{respectively} \qquad \mu = \frac{1}{\lambda + C_U}.$$

Given $\lambda \in \mathbb{C}$ and neglecting the case $\lambda = -C_U$ for which we know already that $A - \lambda I$ has a bounded inverse, there are two alternatives: if $\mu = 1/(\lambda + C_U) \notin \sigma(K)$ then K and thus also

$$A - \lambda I = -\frac{1}{\mu}(A + C_U I)(K - \mu I_V) \tag{3.22}$$

has a bounded inverse. On the other hand, if $\mu \in \sigma(K)$ then $\ker(K - \mu I_V) \neq \{0\}$ is a finite dimensional space. Due to (3.22), the kernel of $K - \mu I_V$ is in one to one correspondence with the kernel of $A - \lambda I$ (via the invertible mapping $A + C_U I$). Hence, the eigenspace $E(\lambda) = \ker(A - \lambda I)$ has the same properties as $\ker(K - \mu I_V)$. Finally, since $\mu = 0$ is the only possible accumulation point of $\sigma(K)$, the spectrum of A has no finite accumulation point. ■

A simple consequence of Theorem 3.24 is given by

Corollary 3.25. *On top of the conditions of Theorem 3.24 we assume that $a(u,u) = 0$ for $u \in V$ implies $u = 0$. Then, the variational problem to find $u \in V$ such that for $l \in V'$ the relation $a(u,v) = l(v)$ holds for all $v \in V$ has a unique solution which depends continuously on the data.*

Proof: We will show that under the given assumptions $0 \notin \sigma(A)$ so that A^{-1} is a bounded operator (and hence $u = A^{-1}l$ is the unique solution with $\|u\|_V \leq \|A^{-1}\| \, \|l\|_{V'}$). If $0 \in \sigma(A)$ then we know from Theorem 3.24 that $\dim E(0) \geq 1$. Hence, there exists $w \neq 0$ such that $Aw = 0$. In particular,

$$0 = (Aw)(w) = \langle Aw, w \rangle_{V' \times V} = a(w,w)$$

which leads to the contradiction $w = 0$. ■

3.5. Solving the Dirichlet problem

In order to get existence and uniqueness for solutions of the general Dirichlet problem (together with continuous dependence on the data) we have to check the assumptions of the Lax Milgram Theorem for the particular bilinear form

$$a(u,v) = \sum_{0 \leq |\sigma|, |\gamma| \leq m} \int_\Omega a_{\sigma_\gamma} \nabla^\gamma u \nabla^\sigma v \, dx \tag{3.23}$$

on the Hilbert space $V = H_0^m(\Omega)$. It turns out that ellipticity of the related operator

$$L(x, \nabla)u = \sum_{0 \leq |\sigma|, |\gamma| \leq m} (-1)^{|\sigma|} \nabla^\sigma (a_{\sigma_\gamma}(x) \nabla^\gamma u) \qquad x \in \Omega \tag{3.24}$$

is not sufficient to guarantee $H_0^m(\Omega)$–ellipticity of a (see Exercise 3.5). Recall that L is elliptic if the symbol $L^p(x, i\xi)$ of the principal part does not vanish on $R^n \backslash \{0\}$. For an operator in divergence form, the principal part is given by

$$L^p(x, \nabla) = \sum_{|\sigma| = |\gamma| = m} (-1)^{|\sigma|} a_{\sigma_\gamma}(x) \nabla^\sigma \nabla^\gamma$$

so that

$$L^p(x, i\xi) = \sum_{|\sigma| = |\gamma| = m} a_{\sigma_\gamma}(x) \xi^\sigma \xi^\gamma$$

and L is elliptic if

$$\sum_{|\sigma|=|\gamma|=m} a_{\sigma_\gamma}(x)\xi^\sigma\xi^\gamma > 0 \qquad \forall \xi \in \mathbb{R}^n\backslash\{0\},\, x \in \Omega$$

(we use the convention that the symbol should be positive). To show ellipticity of the associated bilinear form (3.23), we need the following stronger condition.

Definition 3.26. *The general operator L defined in* (3.24) *is called uniformly elliptic on $\Omega \subset \mathbb{R}^n$ if for some $\theta > 0$*

$$\sum_{|\sigma|=|\gamma|=m} a_{\sigma_\gamma}(x)\xi^\sigma\xi^\gamma \geq \theta\|\xi\|^{2m} \qquad \forall \xi \in \mathbb{R}^n.$$

For certain uniformly elliptic operators L we can then show that a is $H_0^m(\Omega)$ elliptic.

Lemma 3.27. *Let $\Omega \subset \mathbb{R}^n$ be open and assume the coefficients $a_{\sigma\gamma}$ of L in* (3.24) *are constants. Moreover, let $a_{00} > 0$ and $a_{\sigma\gamma} = 0$ for $0 < |\sigma|, |\gamma| < m$. If L is uniformly elliptic then the corresponding bilinear form a defined in* (3.23) *is $H_0^m(\Omega)$–elliptic. If Ω is bounded, the result remains true if $a_{00} = 0$.*

Proof: To check ellipticity, we take $u \in C_0^\infty(\Omega)$ and try to estimate $a(u, u)$ from below. By definition

$$a(u,u) = \sum_{|\sigma|,|\gamma|=m} a_{\sigma_\gamma} \int_\Omega \nabla^\gamma u \nabla^\sigma u\, dx + a_{00}\int_\Omega u^2\, dx.$$

Since u is compactly supported in Ω, we can replace the domain of integration by $\mathbb{R}^n$. Using standard properties of the Fourier transform

$$\hat{u}(\xi) = (\mathcal{F}u)(\xi) = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} e^{-ix\cdot\xi} u(x)\, dx$$

we obtain

$$a(u,u) = \sum_{|\sigma|,|\gamma|=m} a_{\sigma_\gamma} \int_{\mathbb{R}^n} (i\xi)^\gamma \hat{u}\overline{(i\xi)^\sigma \hat{u}}\, d\xi + a_{00}\int_{\mathbb{R}^n} |\hat{u}|^2\, d\xi \tag{3.25}$$

where the bar denotes complex conjugation. We have used the property

$$\int_{\mathbb{R}^n} uv\, dx = \int_{\mathbb{R}^n} \hat{u}\overline{\hat{v}}\, d\xi$$

and the fact that $\mathcal{F}$ transforms derivatives into polynomials. (Note that for general operators $L(\nabla)$ with constant coefficients the transform $\mathcal{F}(L(\nabla)u) = L(i\xi)\hat{u}$ leads to the *symbol of L*. Applying $L(\nabla)$ to u thus corresponds to a *multiplication* of $\hat{u}$ with the symbol which is a polynomial in ξ. The principal part $L^p(i\xi)$ involves the highest order terms of this polynomial. For large ξ ('high frequencies'), the principal part obviously dominates the lower order terms and thus dictates the behavior of the differential operator.) We continue with the estimate (3.25) by using the uniform ellipticity of L

$$a(u,u) \geq \int_{\mathbb{R}^n} \left(\theta\|\xi\|^{2m} + a_{00}\right) |\hat{u}|^2 \, d\xi. \tag{3.26}$$

Assuming that $a_{00} > 0$, we easily see that

$$\theta\|\xi\|^{2m} - d_\alpha \xi^{2\alpha} + a_{00} \geq 0$$

for suitable $d_\alpha > 0$. Choosing the smallest constant d_α among $|\alpha| \leq m$, we obtain

$$\sum_{\|\alpha| \leq m} \xi^{2\alpha} \leq \frac{1}{C_2}\left(\theta\|\xi\|^{2m} + a_{00}\right)$$

with $C_2 > 0$. Using this estimate in (3.26) and transforming back yields

$$a(u,u) \geq C_2 \int_{\mathbb{R}^n} \sum_{\|\alpha| \leq m} |\xi^\alpha \hat{u}|^2 \, d\xi = C_2 \int_{\mathbb{R}^n} \sum_{\|\alpha| \leq m} |\nabla^\alpha u|^2 \, dx = C_2\|u\|^2_{m,2}.$$

In the case where Ω is bounded and $a_{00} = 0$ we use $|\xi_i| \leq \|\xi\|$ so that

$$|\xi^\alpha| \leq \|\xi\|^m \qquad \forall |\alpha| = m$$

and thus

$$a(u,u) \geq \theta \int_{\mathbb{R}^n} \sum_{\|\alpha| = m} |\nabla^\alpha u|^2 \, dx = \theta |u|^2_{m,2}$$

with the notation from Theorem 3.10 (Poincaré inequality). Using the result of that theorem, we conclude further

$$a(u,u) \geq C_2 \|u\|^2_{m,2}.$$

Since $C_0^\infty(\Omega)$ is dense in $H_0^m(\Omega)$ the ellipticity estimate carries over to general H_0^m functions due to continuity. ■

In the special case of second order operators we can easily generalize these considerations to x–dependent coefficients.

Lemma 3.28. *If L is a second order operator with the structure as in Lemma 3.27 but with coefficients in $\mathbb{L}^\infty(\Omega)$ (and $a_{00} \geq \mu > 0$ in the general case or $a_{00} \geq 0$ for bounded Ω), the corresponding bilinear form is $H_0^1(\Omega)$–elliptic.*

In this case, Fourier transform is not required because we can apply the ellipticity condition directly to $\xi = \nabla u$ since $\nabla^\sigma u = (\nabla u)^\sigma$ for $|\sigma| = 1$

$$\begin{aligned} a(u,u) &= \sum_{|\sigma|=|\gamma|=1} \int_\Omega a_{\sigma\gamma} (\nabla u)^\sigma (\nabla u)^\gamma \, dx + \int_\Omega a_{00} u^2 \, dx \\ &\geq \theta \int_\Omega \|\nabla u\|^2 dx + \int_\Omega a_{00} u^2 \, dx \end{aligned}$$

If $a_{00}(x) \geq \mu > 0$ then $a(u,u) \geq C_2 \|u\|_{1,2}^2$. Otherwise, the result follows with the Poincaré inequality. ■

If we include first order derivatives in the operator L, ellipticity of a is in general not given. For the special case of first order derivatives with constant coefficients, we find a similar result as in Lemma 3.28 (see Exercise 3.4). For the general second order operator

$$L(x,\nabla)u := -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} a_{ij}(x) \frac{\partial u}{\partial x_j} + \sum_{i=1}^{n} b_i(x) \frac{\partial u}{\partial x_i} + c(x)u, \tag{3.27}$$

with associated bilinear form

$$a(u,v) = \sum_{i,j=1}^{n} \int_\Omega a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i}\, dx + \sum_{i=1}^{n} \int_\Omega b_i(x) \frac{\partial u}{\partial x_i} v\, dx + \int_\Omega cuv\, dx \tag{3.28}$$

we typically get coercivity results with respect to $\mathbb{L}^2(\Omega)$.

Theorem 3.29. *Let Ω be a bounded domain in $\mathbb{R}^n$ and L defined in (3.27) be uniformly elliptic on Ω with $\mathbb{L}^\infty(\Omega)$ coefficients. Then there exists C_3 and $\hat{\lambda} \geq 0$ such that*

$$a(u,u) + \hat{\lambda} \|u\|_{\mathbb{L}^2}^2 \geq C_3 \|u\|_{1,2}^2 \qquad \forall u \in H^1(\Omega)$$

where a is given in (3.28).

Proof: Apart from the estimate of the principal part used in Lemma 3.28, we have

$$\begin{aligned} &+\sum_{i=1}^{n} \int_\Omega b_i(x) \frac{\partial u}{\partial x_i} v\, dx + \int_\Omega cuv\, dx \\ &\qquad\qquad \geq -\max_{i=1}^{n} \|b_i\|_{\mathbb{L}^\infty} \int_\Omega \|\nabla u\|\; |u|\, dx - \|c\|_{\mathbb{L}^\infty} \int_\Omega u^2\, dx. \end{aligned}$$

Applying Cauchy-Schwarz inequality, we get

$$-\int_\Omega \|\nabla u\|\; |u|\, dx \geq -\|\nabla u\|_{\mathbb{L}^2} \|u\|_{\mathbb{L}^2}.$$

The trick

$$-ab = -(\epsilon a)\left(\frac{b}{\epsilon}\right) \geq -\frac{1}{2}\left(\epsilon^2 a^2 + \frac{b^2}{\epsilon^2}\right)$$

then yields

$$-\|\nabla u\|_{\mathbb{L}^2} \|u\|_{\mathbb{L}^2} \geq -\frac{1}{2}\epsilon^2 \|\nabla u\|_{\mathbb{L}^2}^2 - \frac{1}{2\epsilon^2} \|u\|_{\mathbb{L}^2}^2.$$

Choosing ϵ^2 small enough we get

$$a(u,u) \geq \frac{1}{2}\theta \|\nabla u\|_{\mathbb{L}^2}^2 - \tilde{C} \|u\|_{\mathbb{L}^2}^2 \geq C_3 \|u\|_{1,2}^2 - \hat{\lambda} \|u\|_{\mathbb{L}^2}^2.$$

■

According to our considerations in Section 3.4 coercivity enables us to apply spectral theory. In particular, if the spectrum of L (as mapping from $H_0^1(\Omega)$ to $H^{-1}(\Omega)$) does not contain $\lambda = 0$, then the original Dirichlet problem has a unique solution which corresponds continuously on the data. (Even if $0 \in \sigma(L)$, spectral theory yields results about the solution space $\{u \in H_0^m(\Omega) : Lu = f\}$.) In view of these remarks, the following theorem is very important for the case of general uniformly elliptic operators.

Theorem 3.30. **(Gårding's inequality)**
Let $\Omega \subset \mathbb{R}^n$ be a bounded open set and let L in (3.24) be uniformly elliptic with coefficients of the principal part in $C(\bar{\Omega})$ and all other coefficients in $\mathbb{L}^\infty(\Omega)$. Then a given in (3.23) is $H_0^m(\Omega)$–coercive with respect to $\mathbb{L}^2(\Omega)$, i.e.

$$a(u,u) + C_{\mathbb{L}^2}\|u\|_{\mathbb{L}^2}^2 \geq C_{H^m}\|u\|_{H^m}^2 \qquad u \in H_0^m(\Omega)$$

Proof: For the complete proof see the articles quoted in [**5**]. Here, we just consider the case where the coefficients of the principal part are constant. (The full proof reduces the general case to this situation by a partition of unity argument and exploitation of the uniform continuity of the coefficients of the principal part.) In Lemma 3.27 we have already seen that the principal part satisfies

$$a^p(u,u) = \sum_{|\sigma|,|\gamma|=m} a_{\sigma_\gamma} \int_\Omega \nabla^\gamma u \nabla^\sigma u \, dx \geq C_2\|u\|_{m,2}^2.$$

The remaining terms are then estimated by Cauchy-Schwarz inequality and the $\mathbb{L}^\infty$ bounds on the coefficients

$$\begin{aligned} |a(u,u) - a^p(u,u)| \leq & \sum_{\substack{0\leq|\sigma|,|\gamma|\leq m \\ |\sigma|+|\gamma|<2m}} \int_\Omega \left|a_{\sigma_\gamma}\nabla^\gamma u \nabla^\sigma u\right| \, dx \\ & C \sum_{\substack{0\leq|\sigma|,|\gamma|\leq m \\ |\sigma|+|\gamma|<2m}} \|\nabla^\gamma u\|_{\mathbb{L}^2}\|\nabla^\sigma u\|_{\mathbb{L}^2} \\ & \leq C\|u\|_{m,2}\|u\|_{m-1,2}. \end{aligned}$$

Using the ϵ–trick, we can separate the product into a sum of squares with an arbitrarily small contribution to the $\|\cdot\|_{m,2}$ norm

$$|a(u,u) - a^p(u,u)| \leq \frac{\epsilon}{2}\|u\|_{m,2}^2 + C(\epsilon)\|u\|_{m-1,2}^2.$$

Since $X = H^m(\Omega)$ is compactly embedded in $Y = H^{m-1}(\Omega)$ and Y is continuously embedded in $Z = \mathbb{L}^2(\Omega)$, we can find for every $\epsilon > 0$ a constant $D(\epsilon)$ such that

$$\|u\|_Y \leq \epsilon\|u\|_X + D(\epsilon)\|u\|_Z \qquad \forall u \in X. \tag{3.29}$$

(This is *Ehrling's Lemma* which is proved by contradiction: we assume that there is $\epsilon_0 > 0$ for which (3.29) fails. Then, there is a sequence $(x_n)_{n\in\mathbb{N}}$ in X

such that $\|x_n\|_X = 1$ and

$$\|x_n\|_Y > \epsilon_0 + n\|x_n\|_Z. \tag{3.30}$$

Since $X \hookrightarrow Y$, $(x_n)_{n\in\mathbb{N}}$ is bounded in Y so that $(x_n)_{n\in\mathbb{N}}$ converges to zero in Z. A convergent subsequence in Y (which exists since $X \hookrightarrow Y$ is compact) must then also converge to zero which contradicts (3.30).) Using this result, we get

$$\|u\|_{m-1,2}^2 \le \frac{\epsilon}{2}\|u\|_{m,2}^2 + D(\epsilon)\|u\|_{\mathbb{L}^2}^2$$

and thus

$$\begin{aligned} a(u,u) &\ge a^p(u,u) - |a(u,u) - a^p(u,u)| \\ &\qquad\qquad \ge C_2\|u\|_{m,2}^2 - \left(\epsilon\|u\|_{m,2}^2 + E(\epsilon)\|u\|_{\mathbb{L}^2}^2\right). \end{aligned}$$

Hence, for ϵ small enough we get coercivity of a. ∎

3.6. Other boundary conditions

3.6.1. Natural boundary conditions. In Section 3.2 we have seen that the choice of the Hilbert space $V = H_0^m(\Omega)$ is related to homogeneous Dirichlet conditions. Similarly, the unrestricted case $V = H_0^m(\Omega)$ implies certain boundary conditions. Since they are not explicitly given through the definition of V they are called *natural boundary conditions*. For the Helmholtz and Poisson problem we have already seen that the natural conditios are Neumann conditions. In general, the natural boundary conditions are more complicated because they are determined by the coefficients of L so that especially for higher order operators the boundry conditions involve many parameters. For simplicity, we will only consider second order operators in divergence form

$$Lu = -\frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial}{\partial x_j}u\right) - \frac{\partial}{\partial x_i}(a_{i0}u) + a_{0j}\frac{\partial}{\partial x_j} + a_{00}u$$

with corresponding bilinear form

$$a(u,v) = \int_\Omega a_{ij}\frac{\partial v}{\partial x_i}\frac{\partial u}{\partial x_j} + a_{i0}\frac{\partial v}{\partial x_i}u + a_{0j}\frac{\partial u}{\partial x_j}v + a_{00}uv\,dx$$

(Einstein's summation convention is used). If $u \in H^1(\Omega)$ is a solution of the abstract variational problem

$$a(u,v) = l(v) \qquad \forall v \in H^1(\Omega) \tag{3.31}$$

and if u, the coefficients of L, the right hand side and the boundary of the domain have sufficient regularity, we can recover the classical formulation of the problem (3.31). Using $v \in C_0^\infty(\Omega)$ and assuming $l(v) = \langle f, v\rangle_0$ we get after integration by parts

$$\int_\Omega (Lu - f)v\,dx = 0 \qquad \forall v \in C_0^\infty(\Omega)$$

which implies $Lu = f$. If we now take $v \in C^\infty(\bar\Omega)$ in (3.31) integration by parts gives rise to boundary terms

$$0 = a(u,v) - \langle f, v\rangle_0 = \int_\Omega (Lu - f)v\,dx + \int_{\partial\Omega}\left(a_{ij}\frac{\partial u}{\partial x_j} + a_{i0}u\right)n_i v\,d\sigma.$$

Since $Lu - f = 0$ and traces of $C(\bar\Omega)$ functions are dense in $\mathbb{L}^2(\partial\Omega)$ we conclude

$$\left(a_{ij}\frac{\partial u}{\partial x_j} + a_{i0}u\right)n_i = 0. \tag{3.32}$$

Introducing the matrix $A_L = (a_{ij})$ of the second order coefficients of L as well as the vector $a_L = (a_{i0})$ we can rewrite (3.32) in the compact form

$$(A_L\nabla u)\cdot n + (a_L\cdot n)u = 0$$

If $L = -\Delta$ then $a_{ij} = \delta_{ij}$ and $a_L = 0$ so that the natural boundary conditions are the well known Neumann conditions $0 = n\cdot\nabla u = \partial u/\partial n$. In order to get inhomogeneous boundary conditions

$$(A_L\nabla u)\cdot n + (a_L\cdot n)u = g$$

with $g \in \mathbb{L}^2(\partial\Omega)$ it suffices to modify the right hand side of (3.31) by adding an extra linear form

$$v \mapsto \int_{\partial\Omega} gBv\,d\sigma$$

where B is the trace operator which is continuous on $H^1(\Omega)$ (if $\partial\Omega$ is regular enough).

3.6.2. Mixed boundary conditions. If Γ_N is an open subset of $\partial\Omega$ with complement Γ_D we can consider the problem to find a function u such that $Lu = f$ in Ω with Dirichlet conditions on Γ_D and natural conditions on Γ_N. To get a weak formulation we obviously have to generalize the space $H_0^m(\Omega)$ to the case where the zero conditions are only imposed on $\Gamma_D \subset \partial\Omega$. Similar to the definition of H_0^m we introduce the space $H_{\Gamma_D}^m(\Omega)$ as the H^m–closure of all $C^\infty(\bar\Omega)$ functions which vanish in a neighborhood of Γ_D. Then, the variational problem to find $u \in H_{\Gamma_D}^m(\Omega)$ such that $a(u,v) = l(v)$ for all $v \in H_{\Gamma_D}^m(\Omega)$ is the natural weak formulation of the mixed boundary value problem.

3.6.3. General boundary conditions. A linear elliptic operator L of order $2m$ requires in general m boundary conditions to obtain a well posed boundary value problem. The corresponding boundary operators B_j are of the form

$$B_j u = \sum_{|\alpha|\le m_j} b_{j\alpha}B(\nabla^\alpha u) \qquad j = 1,\dots,m$$

where B is the trace operator and $0 \le m_j < 2m$. Of course, $\{B_j : j = 1,\dots,m\}$ must be linearly independent but independency might not be sufficient to obtain well posed problems. A characterization of reasonable boundary conditions is given by the so called *complementing condition for elliptic systems* (see [**14**] for an introduction). For second order operators a single boundary condition is required. Since the order of the boundary operator is at most one, the most general case is given by

$$B_{gen}u := b_{gen} \cdot \nabla u + c_{gen}u.$$

We assume that the vector b_{gen} always has a normal component, i.e.

$$b_{gen}(x) \cdot n(x) \geq \mu > 0 \qquad \forall x \in \partial\Omega \tag{3.33}$$

Note that the natural boundary conditions are of the same form if we set $b_{nat} = {A_L}^T n$ and $c_{nat} = a_L \cdot n$ since

$$n \cdot (A_L \nabla u) + (a_L \cdot n)u = b_{nat} \cdot \nabla u + c_{nat}u.$$

If L is uniformly elliptic, the vector b_{nat} always has a normal component. Indeed,

$$b_{nat} \cdot n = ({A_L}^T n) \cdot n = \sum_{i,j=1}^{n} a_{ij} n_i n_j \geq \theta \|n\|^2 = \theta > 0. \tag{3.34}$$

Starting from the bilinear form a which is related to L and the natural conditions we will apply modifications which leave L unchanged but have some effect on the bilinear form and thus also on the boundary conditions.

The problem can be seperated into three steps. In the first step we transform c_{nat} into the required function c_{gen} without modifying the first order term $b_{nat} \cdot \nabla u$. In a second step, we adjust the first order term without changing the zeroth order part so that both steps are independent of each other. Finally, the results are combined and the condition $b_{gen} \cdot n = b_{nat} \cdot n$ which we require in the second step is relaxed. We remark that the modifications presented here are not unique. However, they are done in such a way that the resulting bilinear form is $H^1(\Omega)$ coercive with respect to $\mathbb{L}^2(\Omega)$ if the same holds for the original form a.

We start with the observaion that due to the product rule

$$L_1 u = -\frac{\partial}{\partial x_i}(\beta_i u) + \beta_i \frac{\partial}{\partial x_i} u + u \frac{\partial}{\partial x_i} \beta_i = 0$$

whenever u and β are smooth enough (we still use Einstein's summation convention over repeated indices). Hence, the modified operator $L + L_1$ acts exactly like L but leads to the different bilinear form $\tilde{a}(u,v) = a(u,v) + a_1(u,v)$ with

$$a_1(u,v) = \int_\Omega \beta_i u \frac{\partial}{\partial x_i} v + \beta_i v \frac{\partial}{\partial x_i} u + uv \frac{\partial}{\partial x_i} \beta_i \, dx.$$

The associated natural boundary conditions are given by

$$\tilde{B}u = b_{nat} \cdot \nabla u + c_{nat}u + (\beta \cdot n)u = 0.$$

Now we use the freedom in β to adjust the zeroth order part. If

$$\beta(x) = c_{gen}(x)n(x) - a_L(x) \qquad \forall x \in \partial\Omega$$

then $\beta \cdot n = c_{gen} - a_L \cdot n = c_{gen} - c_{nat}$ showing that $\tilde{B}u = b_{nat} \cdot \nabla u + c_{gen} u$. Of course, the problem is to find a smooth function β in Ω which has the required values on the boundary. If $\partial\Omega$ and the baoundary data are smooth enough such a function can always be constructed by extension along the normal direction and smooth cut off. Since the bilinear form corresponding to L_1

$$a_1(u,v) = \int_\Omega \beta_i u \frac{\partial}{\partial x_i} v + \beta_i v \frac{\partial}{\partial x_i} u + uv \frac{\partial}{\partial x_i} \beta_i \, dx$$

involves only first order terms we can use the standard trick to estimate

$$\begin{aligned} a_1(u,u) &\geq -2\|\beta\|_{\mathbb{L}^\infty} \|u\|_{\mathbb{L}^2} \|u\|_{H^1} - \|\operatorname{div} \beta\|_{\mathbb{L}^\infty} \|u\|_{\mathbb{L}^2}^2 \\ &\geq -\epsilon \|u\|_{H^1}^2 - C(\epsilon) \|u\|_{\mathbb{L}^2}^2. \end{aligned}$$

Hence, if a is H^1–coercive with respect to $\mathbb{L}^2$ we can choose ϵ so small that $a + a_1$ has the same property.

In the next step we use a similar approach to fix the first order part of the boundary condition. Assuming that $b_{nat} \cdot n = b_{gen} \cdot n$, we have

$$b_{gen} \cdot \nabla + c_{nat} = b_{nat} \cdot \nabla + c_{nat} + d \cdot \nabla$$

where $d = b_{gen} - b_{nat}$ is a tangential vector. The trick is again to find a bilinear form a_2 which has the additional term $d \cdot \nabla$ as natural boundary condition so that together with a the modified conditions are obtained. Since the first order term in the natural boundary conditions has the structure $b_{nat} = {A_L}^T n$ we have to find an operator L_2 such that $L_2 u = 0$ and ${A_{L_2}}^T n = d$. The second condition is easily satisfied by the tensor product $(n \otimes d)_{ij} = n_i d_j$. However, to keep coercivity and to satisfy the condition $L_2 u = 0$ we use the antisymmetric matrix

$$A_{ij}^{(2)} = n_i d_j - n_j d_i.$$

Since $n \perp d$, we have ${A^{(2)}}^T n = -A^{(2)} n = -(n \cdot d) n + (n \cdot n) d = d$. As in the first step the main problem is to extend the matrix valued function $A^{(2)}$ to a function in Ω (keeping antisymmetry). Having achieved that, we define the operator L_2 as

$$L_2 u = -\frac{\partial}{\partial x_i} \left(A_{ij}^{(2)} \frac{\partial}{\partial x_j} u \right) + \left(\frac{\partial}{\partial x_i} A_{ij}^{(2)} \right) \frac{\partial}{\partial x_j} u.$$

Due to the antisymmetry of $A^{(2)}$ we have

$$L_2 u = -A_{ij}^{(2)} \frac{\partial^2}{\partial x_i \partial x_j} u = -\frac{1}{2} \left(A_{ij}^{(2)} + A_{ji}^{(2)} \right) \frac{\partial^2}{\partial x_i \partial x_j} u = 0.$$

Hence, by construction, the modified operator $L + L_2$ acts exactly like L (on smooth functions) and gives rise to the bilinear form $\bar{a}(u,v) = a(u,v) + a_2(u,v)$ with

$$a_2(u,v) = \int_\Omega A_{ij}^{(2)} \frac{\partial v}{\partial x_i} \frac{\partial u}{\partial x_j} + \left(\frac{\partial}{\partial x_i} A_{ij}^{(2)} \right) v \frac{\partial}{\partial x_j} u \, dx.$$

The natural boundary conditions of the combined form $\bar{a}$ is

$$b_{nat} \cdot \nabla + c_{nat} + ({A^{(2)}}^T n) \cdot \nabla = (b_{nat} + d) \cdot \nabla + c_{nat} = b_{gen} \cdot \nabla + c_{nat}.$$

Note that again coercivity is not destroyed since

$$\begin{aligned}
a_2(u,u) &= \int_\Omega A_{ij}^{(2)} \frac{\partial v}{\partial x_i} \frac{\partial u}{\partial x_j} + \left(\frac{\partial}{\partial x_i} A_{ij}^{(2)} \right) v \frac{\partial}{\partial x_j} u \, dx \\
&= \int_\Omega \frac{1}{2} \left(A_{ij}^{(2)} + A_{ji}^{(2)} \right) \frac{\partial v}{\partial x_i} \frac{\partial u}{\partial x_j} + \left(\frac{\partial}{\partial x_i} A_{ij}^{(2)} \right) v \frac{\partial}{\partial x_j} u \, dx \\
&= \int_\Omega \left(\frac{\partial}{\partial x_i} A_{ij}^{(2)} \right) v \frac{\partial}{\partial x_j} u \, dx \\
&\geq -C \|u\|_{\mathbb{L}^2} \|u\|_{H^1} \geq -\epsilon \|u\|_{H^1}^2 - C(\epsilon) \|u\|_{\mathbb{L}^2}^2.
\end{aligned}$$

In the last step we combine the previous results. First, the boundary condition

$$b_{gen} \cdot \nabla u + c_{gen} u = 0$$

is modified by multiplying with $\eta = (b_{nat} \cdot n)/(b_{gen} \cdot n)$. According to (3.34) and assumption (3.33) we have $\eta > \mu/\theta > 0$ so that

$$(\eta b_{gen}) \cdot \nabla u + (\eta c_{gen}) u = 0 \tag{3.35}$$

is an equivalent condition for which we construct a_1 and a_2 as before (observe that $(\eta b_{gen}) \cdot n = b_{nat} \cdot n$). The resulting bilinear form $a + a_1 + a_2$ on the Hilbert space $V = H^1(\Omega)$ yields a weak formulation of the general boundary value problem. As we have seen before, coercivity of a carries over to the new bilinear form. Also, inhomogeneous conditions can be obtained by adding the functional

$$v \mapsto \int_{\partial\Omega} \eta g B v \, d\sigma \qquad v \in H^1(\Omega)$$

on the right hand side of the abstract variational problem.

3.7. The Ritz–Galerkin Method

In the theory of classical solutions it is natural to use approximation procedures like Finite Difference Methods which are based on a point wise evaluation of functions and differential operators. When dealing with weak solutions, this approach cannot be taken over because point values of functions in $H^m(\Omega)$ are in general not well defined. In fact, the formulation of the original problem as abstract variational problem suggests a different strategy to convert the infinite dimensional setup into a finite dimensional one which then allows a numerical treatment. If the abstract variational problem is to find u in a Hilbert space V such that for some $l \in V'$

$$a(u, v) = l(v) \qquad \forall v \in V$$

then a discretization is obtained by replacing V with a finite dimensional subspace V_h. (Here, $h > 0$ is related to the dimension of V_h and will later play the role of a discretization length.) To get a numerical approximation, the idea of the *Ritz–Galerkin Method* is to find $u_h \in V_h$ such that

$$a(u_h, v) = l(v) \qquad \forall v \in V_h \tag{3.36}$$

which is a finite dimensional problem. By introducing a basis $\{\psi_1, \psi_2, ..., \psi_N\}$ of V_h we can reformulate (3.36) as

$$a(u_h, \psi_i) = l(\psi_i), \quad i = 1, \dots, N.$$

Using

$$u_h = \sum_{j=1}^{N} z_j \psi_j$$

we get the system of equations

$$\sum_{j=1}^{N} a(\psi_j, \psi_i) z_j = l(\psi_i), \quad i = 1, \dots, N$$

which we may write in the form

$$Az = b$$

with the matrix $A_{ij} = a(\psi_j, \psi_i)$ and right hand side $b_i = l(\psi_i)$. In engineering science the matrix A is called the *stiffness matrix*. Note that A is positive definite if a is V–elliptic since

$$z^T Az = \sum_{i,j} z_i A_{ij} z_j = a\left(\sum_j z_j \psi_j, \sum_i z_i \psi_i\right) = a(u_h, u_h) \geq C_2 \|u_h\|_V^2$$

In particular, $Az = 0$ implies $u_h = 0$ and thus also $z = 0$ which shows that A is invertible. Consequently, the Ritz–Galerkin solution always exists if a is elliptic. Moreover, if a is symmetric, i.e. $a(u, v) = a(v, u)$ for all $u, v \in V$ then A is a symmetric matrix. The basic estimate on the quality of a numerical approximation u_h based on the sub space V_h is the following lemma.

Lemma 3.31. (Céa Lemma)
Let a be a bilinear, V-elliptic form and let u and u_h be the solutions of the variational problem in V and $V_h \subset V$. Then

$$\|u - u_h\|_V \leq \frac{C_1}{C_2} \inf_{v_h \in V_h} \|u - v_h\|_V$$

where C_1 is the constant of continuity and C_2 the constant of ellipticity of a.

Proof: By definition we have $a(u, v) = l(v)$ for all $v \in V$ and $a(u_h, v) = l(v)$ for all $v \in V_h$. Since $V_h \subset V$ we can subtract the two equations which yields

$$a(u - u_h, v) = 0 \qquad \forall v \in V_h.$$

Now, let $v_h \in V_h$. With $v = v_h - u_h \in V_h$ we have $a(u - u_h, v_h - u_h) = 0$ and

$$\begin{aligned} C_2 \|u - u_h\|_V^2 &\leq a(u - u_h, u - u_h) \\ &= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \\ &\leq C_1 \|u - u_h\|_V \|u - v_h\|_V \end{aligned}$$

Hence, it follows that $C_2 \|u - u_h\|_V \leq C_1 \|u - v_h\|_V$ for all $v_h \in V_h$. ■

The Céa Lemma gives information about the convergence of the Ritz–Galerkin method. Obviously, the distance between approximation u_h and exact solution u vanishes if the distance of u to the space V_h becomes arbitrarily small. (The distance of a point u to a set V_h is defined by $d(u, V_h) := \inf\{\|u - v_h\|_V : v_h \in V_h\}$.) Consequently, convergence can only be obtained if the spaces V_h become larger for $h \to 0$ and eventually exhaust the Hilbert space V.

Corollary 3.32. *Let $V_h \subset V$ $(h > 0)$ be a family of subspaces with the property*

$$d(u, V_h) \xrightarrow[h \to 0]{} 0 \qquad \forall u \in V.$$

If a is a V–elliptic bilinear form and if u is the solution of an abstract variational problem related to a, the Ritz–Galerkin approximation u_h converges

$$\|u - u_h\|_V \xrightarrow[h \to 0]{} 0.$$

Proof: Immediate consequence of Lemma 3.31. ■

The Ritz–Galerkin Method turns into an efficient solution algorithm if V_h has good approximation properties and if the stiffness matrix A remains either small or sparse. The latter property is obtained if the basis is chosen almost a–orthogonal so that each row of A has only a few non zero entries. Apart from that, it is useful if the basis is related to the problem under consideration. More precisely, if V is a function space, the basis functions should be selected in such a way that high resolution is possible where the solution is expected to develop important structures. In the next section, a few exemplary choices of V_h are discussed.