Chapter 3

Abstract Formulation

The first step in a finite element approach is to write an appropriate weak or variational problem. In lieu of deriving existence and uniqueness results for each weak problem we encounter, our strategy is to formulate a general weak problem and prove existence and uniqueness for it. Then, as we encounter specific weak problems, we only need to show that each problem fits into the framework of the general problem and satisfies any conditions required by our analysis of the general problem. We repeat the procedure with the discrete weak problem, but, in addition, derive a general error estimate. The tools introduced in the last chapter easily allow us to formulate a general weak problem; the existence and uniqueness of its solution is established through the Lax-Milgram theorem which is proved with the aid of the Projection and the Riesz Representation theorems from Chapter **??**.

The abstract weak problem which we study is posed on a general Hilbert space, but when we look at specific examples we need to completely specify the particular space. It turns out that the class of Hilbert spaces that are appropriate is Sobolev spaces. Before studying the general problem, we introduce these spaces and the concept of weak derivatives.

Not all weak problems we encounter fit into the framework of the general problem introduced in this chapter. In a later chapter (see Chapter **??**) we consider an obvious generalization to this weak problem and in Chapter **??** we introduce a so-called *mixed weak problem*. Consequently, by the completion of this book, we plan to analyze several general weak problems which can handle a wide variety of linear problems. Nonlinear problems are discussed in Chapter **??**.

When we derived the weak formulation (see (**??**)) to our prototype example in Chapter **??**, we saw that it was equivalent to solving a corresponding minimization problem. Not all variational problems have this corresponding Rayleigh-Ritz formulation. In Section 3.4 we prove a result which gives conditions when the two formulations are equivalent.

3.1 Weak L^2 derivatives and Sobolev spaces

In this section we define the particular class of Hilbert spaces which we use as our spaces of admissible functions; these spaces are called *Sobolev spaces*. We want to generalize the concept of derivative to define what we refer to as a *weak* or *generalized derivative* and do it in such a way that if everything is "smooth enough" then the classical and weak derivatives coincide. The concept of a weak derivative is an extension of the classical concept in which maintains the validity of the integration by parts formula or its analogue in higher dimensions. Our generalization allows functions such as $u(\mathbf{x}) = |\mathbf{x}|$ on $[-1, 1]$ to have a derivative in the weak sense.

We use this weak derivative in our definition of Sobolev spaces, the particular Hilbert spaces we need. We begin this section with some notation which simplifies the exposition, follow with the definition of a weak derivative, and then introduce Sobolev spaces with their associated norms and inner products.

As usual, let Ω be an open, connected subset of $\mathbb{R}^n$ and let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ denote a general point in Ω. The set of all real-valued functions $u(\mathbf{x}) = u(x_1, \cdots, x_n)$ which are defined and continuous on Ω is denoted $C(\Omega)$ and the set of all continuous functions having derivatives of order less than or equal to k continuous in Ω is denoted $C^k(\Omega)$, $k < \infty$. We also need the space C_0^∞ which is the space of infinitely differentiable functions which have compact support. A function ϕ has compact suport if $\phi = 0$ outside a closed and bounded subset of Ω; the support of a function $\phi(x)$ generally refers to the closure of the set of all x for which $\phi(x) \neq 0$.

To simplify the derivative notation we introduce the notation of a *multi-index* α which is defined as an n-tuple of non-negative integers, *i.e.,* $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ where α_i, $i = 1, \ldots, n$ is a non-negative integer. We use the notation

$$|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n \,.$$

In this way we can rewrite the partial differential operator as

$$D^\alpha \equiv \frac{\partial^{\alpha_1+\alpha_2+\cdots+\alpha_n}}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_n^{\alpha_n}} = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_n^{\alpha_n}} \,.$$

For example, in $\mathbb{R}^1$, $\alpha = \alpha_1$ so that D^α denotes the ordinary differential operator; for example, $D^2 = d^2/dx^2$. For $\mathbb{R}^2$, $\alpha = (\alpha_1, \alpha_2)$ and so for $|\alpha| = 1$ we have the first order partial differential operators $D^{(1,0)} = \partial/\partial x_1$ and $D^{(0,1)} = \partial/\partial x_2$. For $|\alpha| = 2$ we have

$$D^{(2,0)} = \frac{\partial^2}{\partial x_1^2} \,, \qquad D^{(0,2)} = \frac{\partial^2}{\partial x_2^2} \,, \quad \text{and} \quad D^{(1,1)} = \frac{\partial^2}{\partial x_1 x_2} \,.$$

Using this notation we can define $C^k(\Omega)$ as

$$C^k(\Omega) = \{u \,:\, D^\alpha u \in C(\Omega),\ |\alpha| \leq k\} \,.$$

3.1.1 Weak derivatives

We now define the concept of the *weak* (or *generalized* or *distributional*) $L^2(\Omega)$ derivative of a function. Let $u \in L^2(\Omega)$; we say that u has a derivative of order α

in the weak L^2-sense if there exists a function $v \in L^2(\Omega)$ such that

$$\int_\Omega u D^\alpha \phi \, d\Omega = (-1)^{|\alpha|} \int_\Omega v\phi \, d\Omega \tag{3.1}$$

holds for all $\phi \in C_0^\infty(\Omega)$.

To help us understand this definition we consider a specific case in $\mathbb{R}^1$ where $\Omega = (0,1)$. Suppose $\phi(x)$ is a continuously differentiable function on Ω which vanishes on the boundary of Ω, *i.e.,* $\phi(0) = \phi(1) = 0$. Let $u \in C^1([0,1])$. Then

$$\int_0^1 u\frac{\partial \phi}{\partial x}\, dx = \phi u\big|_0^1 - \int_0^1 \phi \frac{\partial u}{\partial x}\, dx$$

and thus

$$\int_0^1 u\frac{\partial \phi}{\partial x}\, dx = -\int_0^1 \phi \frac{\partial u}{\partial x}\, dx\,.$$

So the classical derivative $\partial u/\partial x$ can be viewed as a function v satisfying

$$\int_0^1 u\frac{\partial \phi}{\partial x}\, dx = -\int_0^1 \phi v\, dx\,. \tag{3.2}$$

Conversely, if we find a function v satisfying (3.2) then it behaves like the derivative when integrated against functions in $C_0^\infty(\Omega)$. Note that (3.2) is just (3.1) where $|\alpha| = 1$ since Ω is a subset of $\mathbb{R}^1$.

We conclude that the classical derivatives, if they exist and are continuous in the usual sense, coincide with the weak derivatives. However, there are functions which possess a weak L^2-derivative but have no classical derivatives.

Example 3.1 We know that the function $u(x) = |x|$ on $\Omega = (-1,1)$ does not have a classical derivative at $x = 0$; however it does have a generalized L^2-derivative. To see this, let

$$v(x) = \begin{cases} -1 & \text{for} \quad -1 < x \le 0 \\ 1 & \text{for} \quad 0 < x < 1\,. \end{cases}$$

Clearly, $v \in L^2(\Omega)$ and we claim that $v(x)$ is the weak L^2-derivative of $u(x) = |x|$. To show this, we note that if $\phi \in C_0^\infty(-1,1)$ we have

$$\begin{aligned} -\int_{-1}^1 v\phi\, dx &= \int_{-1}^0 \phi\, dx - \int_0^1 \phi\, dx = -\int_{-1}^0 \phi \frac{d}{dx}(-x)\, dx - \int_0^1 \phi \frac{d}{dx}(x)\, dx \\ &= -\big[\phi(-x)\big]_{-1}^0 - \big[\phi x\big]_0^1 + \int_{-1}^0 (-x)\phi'\, dx + \int_0^1 x\phi'\, dx \\ &= \int_{-1}^1 |x|\phi'\, dx = \int_{-1}^1 u\phi'\, dx \end{aligned}$$

and thus (3.1) is satisfied with $|\alpha| = 1$. ∎

There are functions in $L^2(\Omega)$ which do not possess weak or classical derivatives. The reader is referred to [Adams] for a complete exposition of generalized derivatives.

We note that it can be proved that weak L^2-derivatives are unique almost everywhere; that is, unique except on a set of measure zero. For example, in Example 3.1 we could have chosen the generalized L^2-derivative to be

$$w(x) = \begin{cases} -1 & \text{for} \quad -1 < x < 0 \\ 1 & \text{for} \quad 0 \le x < 1 \,. \end{cases}$$

Note that $w(x)$ and $v(x)$ defined in Example 3.1 differ only at the point $x = 0$, *i.e.,* on a set of measure zero.

3.1.2 Sobolev spaces

We are now ready to define the class of Hilbert spaces that we use to pose our weak problems. The *Sobolev space* $H^m(\Omega)$ is the set of functions $u \in L^2(\Omega)$ which possess generalized (weak) L^2-derivatives $D^\alpha u$ which are also in $L^2(\Omega)$ for $0 \le |\alpha| \le m$; *i.e.,*

$$H^m(\Omega) = \{ u \in L^2(\Omega) \ : \ D^\alpha u \in L^2(\Omega) \text{ for } 0 \le |\alpha| \le m \ \} \,. \tag{3.3}$$

Clearly, $H^m(\Omega)$ is a subspace of $L^2(\Omega)$ and $H^0(\Omega) = L^2(\Omega)$. On $H^m(\Omega)$ we define the inner product

$$\begin{aligned} (u, v)_m &= \sum_{|\alpha| \le m} \int_\Omega D^\alpha u D^\alpha v \, d\Omega \\ &= \sum_{|\alpha| \le m} (D^\alpha u, D^\alpha v) \quad \forall \, u, v \in H^m(\Omega) \,, \end{aligned} \tag{3.4}$$

where $(\cdot, \cdot)$ denotes the standard inner product on $L^2(\Omega)$. Using this definition of inner product, we define the norm on $H^m(\Omega)$ as

$$\|u\|_m = (u, u)_m^{1/2} = \left(\sum_{|\alpha| \le m} \|D^\alpha u\|^2 \right)^{1/2} \qquad \forall \, u \in H^m(\Omega) \,, \tag{3.5}$$

where $\|\cdot\|$ denotes the standard norm on $L^2(\Omega)$. Clearly, $\|\cdot\|_0$ is the standard $L^2(\Omega)$ norm so in the sequel we denote the L^2-norm by $\|\cdot\|_0$.

The following result guarantees that $H^m(\Omega)$ is a complete inner product space; for the proof, see [Adams].

Theorem 3.2. *$H^m(\Omega)$, equipped with the inner product and norm defined in (3.4) and (3.5), respectively, is a Hilbert space and thus a Banach space.*

We make extensive use of the space $H^1(\Omega)$; if $\Omega \subset \mathbb{R}^1$ then the norm on $H^1(\Omega)$ is explicitly given by

$$\|u\|_1 = \left(\|u\|_0^2 + \|u'\|^2 \right)^{1/2} \tag{3.6}$$

and if $\Omega \subset \mathbb{R}^2$ then the norm is explicitly given by

$$\|u\|_1 = \left(\|u\|_0^2 + \left\| \frac{\partial u}{\partial x_1} \right\|_0^2 + \left\| \frac{\partial u}{\partial x_2} \right\|_0^2 \right)^{1/2} . \tag{3.7}$$

Note that by construction, for a function $u \in H^m(\Omega)$ we have

$$\|u\|_0 \leq \|u\|_1 \leq \|u\|_2 \cdots \leq \|u\|_m .$$

We also make use of the Sobolev semi-norm on $H^m(\Omega)$ which is denoted by $|\cdot|$ and defined by

$$|u|_m = \left(\sum_{|\alpha|=m} \|D^\alpha u\|_0^2 \right)^{1/2} \qquad \forall\, u \in H^m(\Omega) . \tag{3.8}$$

Thus for $\Omega \subset \mathbb{R}^n$ the H^1 semi-norm is explicitly given by

$$|u|_1 = \left(\sum_{i=1}^{n} \left\| \frac{\partial u}{\partial x_i} \right\|_0^2 \right)^{1/2} \qquad \forall\, u \in H^1(\Omega) . \tag{3.9}$$

Again by definition of the norms, we have that for $u \in H^m(\Omega)$

$$|u|_m \leq \|u\|_m . \tag{3.10}$$

Note that we are using the standard notation for partial derivative and D^α interchangeably; the context should make it clear if we are referring to the classical or weak derivative.

We also make use of the constrained space $H_0^1(\Omega)$ which denotes all functions in $H^1(\Omega)$ which are zero on the boundary; *i.e.,*

$$H_0^1(\Omega) = \{ u \in H^1(\Omega) \ : \ u|_{\partial\Omega} = 0 \} . \tag{3.11}$$

Formally, $H_0^1(\Omega)$ is defined as the completion of $C_0^\infty(\Omega)$ with respect to the norm $\|\cdot\|_1$ and it can be shown that it is a closed subspace of $H^1(\Omega)$ consisting precisely of those functions $u \in H^1(\Omega)$ which almost everywhere satisfy $u = 0$ on the boundary of Ω.

We comment that if $\Omega \subset \mathbb{R}^n$ for $n > 1$, then $H^m(\Omega)$ can contain functions which are not continuous. As an example, if $n = 2$ and Ω is the open unit disk with center at the origin, consider the function $u = (\ln(1/r))^k$ for $k < 1/2$ and $r = (x_1^2 + x_2^2)^{1/2}$. It can be shown that $u \in H^1(\Omega)$ but u is not continuous at the origin. A result known as Sobolev's Theorem (see [Adams]) gives the connection between $H^m(\Omega)$ and $C^m(\Omega)$ for arbitrary m.

We conclude this section with the following result, known as the Poincaré inequality, which is extremely useful in relating the L^2-norm of certain functions in $H^1(\Omega)$ with their corresponding semi-norm. Recall that by the definition of the Sobolev norm, it is always true that $\|u\|_0 \leq \|u\|_1$. However, it is not obvious if the

result holds when we replace the one-norm with the one semi-norm. It turns out that it is true for functions in $H_0^1(\Omega)$ and even for functions which are zero on some portion of their boundary. It is important to realize that this result does not, in general, hold for all functions in $H^1(\Omega)$.

Lemma 3.3. (Poincaré Inequality) *Let $u \in$ hone such that $u = 0$ on some portion of the boundary of Ω. Then there exists a constant C depending on Ω such that*

$$\|u\|_0 \leq C \left(\sum_{i=1}^{n} \left\| \frac{\partial u}{\partial x_i} \right\|_0^2 \right)^{1/2} = C|u|_1 \,. \tag{3.12}$$

Note that the Poincaré inequality, along with (3.10) gives that on $H_0^1(\Omega)$ the H^1-norm and H^1-seminorm are equivalent norms.

3.2 Formulation and analysis of a general weak problem

In this section we use the tools developed in the last chapter to formulate a general weak problem. We state and prove the Lax-Milgram theorem which is central to the theory of the finite element method since it provides us with conditions which guarantee the existence and uniqueness of the solution of our general weak problem.

Let V denote a Hilbert space, let $A(\cdot,\cdot)$ denote a bilinear form on $V \times V$ and let F denote a linear functional on V. The general weak problem we consider is to

$$\begin{cases} \text{seek } u \in V \text{ satisfying} \\ \qquad A(u,v) = F(v) \quad \forall\, v \in V \,. \end{cases} \tag{3.13}$$

Many weak formulations that we encounter can easily be put into the general form of (3.13) with appropriate choices for the Hilbert space, the bilinear form, and the linear functional.

Example 3.4 Consider the simple two-point boundary value problem

$$-u''(x) = \sin \pi x \quad 0 < x < 1 \tag{3.14a}$$

and the boundary conditions

$$u(0) = 0 \tag{3.14b}$$

and

$$u(1) = 0 \,. \tag{3.14c}$$

In choosing the underlying Hilbert space for our weak formulation of (3.14), we must require our solution to be in $L^2(0,1)$ and to possess at least one weak L^2-derivative. In addition, we want to constrain our space so that we only consider functions which satisfy the homogeneous Dirichlet boundary conditions. Thus we

choose $H_0^1(0,1)$ to be the underlying Hilbert space in which we seek a solution $u(x)$. In particular, we seek a $u \in H_0^1(0,1)$ satisfying

$$\int_0^1 u'v'\,dx = \int_0^1 \sin \pi x v\,dx \quad \forall\, v \in H_0^1(0,1)\,. \tag{3.15}$$

Clearly any solution of this two-point boundary value problem is also a solution of (3.15). Now we can easily cast (3.15) into the general form of (3.13) if we let $V = H_0^1(0,1)$,

$$A(u,v) = \int_0^1 u'v'\,dx$$

and

$$F(v) = \int_0^1 \sin \pi x\; v(x)\,dx = (\sin \pi x, v)\,,$$

where $(\cdot,\cdot)$ denotes the $L^2(0,1)$ inner product. Clearly $A(u,v)$ defined in this way is a bilinear form on $H^1(0,1)$ and $F(v)$ is a linear functional on $H^1(0,1)$ and thus on $H_0^1(0,1)$. ∎

If F is a bounded linear functional on the given Hilbert space V and the bilinear form $A(\cdot,\cdot)$ is bounded, or equivalently, continuous on the space V and, in addition, satisfies a property referred to as *coercivity* or equivalently as *V-ellipticity*, then the Lax-Milgram theorem guarantees existence and uniqueness of the solution of (3.13). Moreover, the theorem also provides a bound of the solution of the weak problem in terms of the data. This is analogous to bounds obtained in PDE theory.

Theorem 3.5. (Lax-Milgram Theorem) *Let V be a Hilbert space and let $A(\cdot,\cdot) : V \times V \to \mathbb{R}^1$ be a bilinear form on V which satisfies*

$$|A(u,v)| \le M\,\|u\|\,\|v\| \quad \forall\, u,v \in V \tag{3.16}$$

and

$$A(u,u) \ge m\,\|u\|^2 \quad \forall\, u \in V\,, \tag{3.17}$$

where M and m are positive constants independent of $u, v \in V$. Let $F : V \to \mathbb{R}^1$ be a bounded linear functional on V. Then there exists a unique $u \in V$ satisfying (3.13). Moreover

$$\|u\| \le \frac{1}{m}\,\|F\|\,. \tag{3.18}$$

Proof. In order to prove this result we begin by fixing a $u \in V$ and demonstrating that $Q(v) = A(u,v)$ defines a bounded linear functional on V. We then apply the Riesz representation theorem (Theorem **??**) to obtain a unique element $\hat{u} \in V$ such that

$$Q(v) = A(u,v) = (v,\hat{u}) \quad \forall\, v \in V\,.$$

This allows us to associate to each $u \in V$ a unique $\hat{u} \in V$. If we denote this correspondence by $\hat{u} = \mathcal{A}u$ we have

$$A(u,v) = (v, \mathcal{A}u) \quad \forall\, u,v \in V\,. \tag{3.19}$$

We then demonstrate that $\mathcal{A}$ is a linear operator and that the range of $\mathcal{A}$, denoted $\mathcal{R}(\mathcal{A})$, is a closed subspace of V and finally that $\mathcal{R}(\mathcal{A}) = V$.

Once we have established these facts then we can establish the existence and uniqueness by the following argument. Since F is a bounded linear functional on V then the Riesz representation theorem guarantees the existence of a unique element $\phi \in V$ such that $F(v) = (\phi, v)$ for all $v \in V$. If the $\mathcal{R}(\mathcal{A}) = V$, then there exists a $u \in V$ such that $\mathcal{A}u = \phi$. Hence there exists a $u \in V$ such that

$$F(v) = (\mathcal{A}u, v) = A(u, v) \quad \forall\, v \in V\,.$$

Uniqueness is shown in the standard way of choosing $u_1 \neq u_2$ such that

$$A(u_1, v) = A(u_2, v) = F(v) \quad \forall\, v \in V\,.$$

Then we have that $A(u_1 - u_2, v) = 0$ for all $v \in V$ and choosing $v = u_1 - u_2$, we conclude that $A(u_1 - u_2, u_1 - u_2) = 0$. Using (3.17) we know that $A(u_1 - u_2, u_1 - u_2) \geq m \left\|u_1 - u_2\right\|^2$ which implies the contradiction $0 \geq m \left\|u_1 - u_2\right\|^2$.

We now return to proving the claims necessary to complete the proof of existence. First, we see that $Q(v) \equiv A(u, v)$ is a bounded linear functional on V. Linearity immediately follows from the linearity of $A(\cdot, \cdot)$; the fact that it is bounded follows from (3.16); *i.e.,*

$$|Q(v)| = |A(u, v)| \leq C \left\|u\right\| \left\|v\right\|$$

and thus $\|Q\| \leq C \left\|u\right\| < \infty$. It is now required to show that the operator $\mathcal{A}$ is linear. Given $\phi, \psi \in V$

$$\begin{aligned}(v, \mathcal{A}(\alpha\phi + \beta\psi)) &= A(\alpha\phi + \beta\psi, v) = \alpha A(\phi, v) + \beta A(\psi, v) \\ &= \alpha\Phi(v) + \beta\Psi(v) \quad \forall\, v \in V\,.\end{aligned}$$

Using the same argument as we did for $Q(v)$, we see $\Phi(v)$ and $\Psi(v)$ are bounded linear functionals on V and so we can apply the Riesz representation theorem and the definition of $\mathcal{A}$ to write

$$\Phi(v) = (v, \hat{\phi}) = (v, \mathcal{A}\phi)$$

and similarly for $\Psi(v)$. Combining these results we obtain

$$(v, \mathcal{A}(\alpha\phi + \beta\psi)) = \alpha\,(v, \mathcal{A}\phi) + \beta\,(v, \mathcal{A}\psi) \quad \forall\, v \in V$$

and hence $\mathcal{A}(\alpha\phi + \beta\psi) = \alpha\mathcal{A}(\phi) + \beta\mathcal{A}(\psi)$; *i.e.,* linearity is proved. It remains to show that $\mathcal{R}(\mathcal{A})$ is a closed subspace of V and, in fact, $\mathcal{R}(\mathcal{A}) = V$. The fact that $\mathcal{R}(\mathcal{A})$ is a subspace is obvious from its definition; to show that it is closed we choose a sequence $\{\hat{\phi}_n\} \in \mathcal{R}(\mathcal{A})$ which converges to $\hat{\phi} \in V$ and demonstrate that $\hat{\phi} \in \mathcal{R}(\mathcal{A})$. Since $\hat{\phi}_n \in \mathcal{R}(\mathcal{A})$ we can write $\hat{\phi}_n = \mathcal{A}\phi_n$ for $\phi_n \in V$; we want to demonstrate that $\{\phi_n\}$ is a Cauchy sequence in V. Now by the definition of $\mathcal{A}$, $(v, \mathcal{A}\phi_n) = A(\phi_n, v)$ for all $v \in V$ and thus $A(\phi_n - \phi_m, v) = (v, \mathcal{A}(\phi_n - \phi_m))$ for all $v \in V$. Choosing $v = \phi_n - \phi_m$ and using (3.17) we have that

$$m \left\|\phi_n - \phi_m\right\|^2 \leq a(\phi_n - \phi_m, \phi_n - \phi_m) = (\phi_n - \phi_m, \mathcal{A}(\phi_n - \phi_m))\,.$$

Using the Cauchy-Schwartz inequality and the linearity of A we have $\|\phi_n - \phi_m\| \leq \frac{1}{m}\|\mathcal{A}\phi_n - \mathcal{A}\phi_m)\| = \frac{1}{N}\|\hat{\phi}_n - \hat{\phi}_m\|$; we thus conclude that $\{\phi_n\}$ is a Cauchy sequence in V. Since V is complete, there exists a $\phi \in V$ such that $\phi_n \to \phi$. If we now show that $\hat{\phi} = \mathcal{A}\phi$ we have demonstrated that the limit of the sequence $\{\hat{\phi}_n\}$ is in $\mathcal{R}(\mathcal{A})$ and thus $\mathcal{R}(\mathcal{A})$ is closed. To see this we note that by using the linearity of $A(\cdot,\cdot)$ and (3.16) we have

$$|A(\phi_n, v) - A(\phi, v)| \leq M \|\phi_n - \phi\| \|v\| \quad \forall\, v \in V\,.$$

Thus $A(\phi_n, v) \to A(\phi, v)$ as $n \to \infty$ for all $v \in V$. In terms of an inner product, this yields $(v, \mathcal{A}\phi_n) \to (v, \mathcal{A}\phi)$ as $n \to \infty$. But $(\mathcal{A}\phi_n, v) = (\hat{\phi}_n, v) \to (\hat{\phi}, v)$. So $(\mathcal{A}\phi_n, v) \to (\hat{\phi}, v)$ and $(\mathcal{A}\phi_n, v) \to (v, \mathcal{A}\phi)$; thus $\hat{\phi} = \mathcal{A}\phi$ and the $\mathcal{R}(\mathcal{A})$ is closed. To show that $\mathcal{R}(\mathcal{A}) = V$ we assume that $\mathcal{R}(\mathcal{A}) \subset V$; *i.e.,* there exists a $z \in \mathcal{R}(\mathcal{A})^{\perp}$. This implies $(z, \hat{v}) = 0$ for all $\hat{v} \in \mathcal{R}(\mathcal{A})$; or equivalently for all $v \in V$, $(z, \mathcal{A}v) = 0$. In particular, if we set $v = z$ we have $A(z, z) = (z, \mathcal{A}z) = 0$, but from (3.17)) $A(z, z) \geq N \|z\|^2$ implying that $z = 0$, a contradiction.

To conclude the proof we must demonstrate (3.18)). Since $A(u, u) = F(u)$ we have that

$$m \|u\|^2 \leq |A(u, u)| = |F(u)|$$

from which we have for $u \neq 0$

$$\|u\| \leq \frac{1}{m} \frac{|F(u)|}{\|u\|}\,.$$

Therefore

$$\|u\| \leq \sup_{u \neq 0} \frac{1}{m} \frac{|F(u)|}{\|u\|} = \frac{1}{m} \|F\|\,.$$

■

3.3 Galerkin approximations

In the previous section we defined a general weak problem, (3.13), which is posed on an infinite-dimensional Hilbert space V. We then stated and proved the Lax-Milgram theorem which gave conditions guaranteeing existence and uniqueness of its solution. Since in finite elements, our objective is to approximate the solution of this weak problem, we want to state a general discrete weak problem, give conditions which guarantee existence and uniqueness of its solution, and finally to bound the error between the solution of (3.13) and the discrete solution.

We begin by letting $\{S^h\}$, $0 < h < 1$, be a family of finite dimensional subspaces of the Hilbert space V. Then the discrete problem corresponding to (3.13) for a fixed h is to

$$\left\{ \begin{array}{l} \text{seek } u^h \in S^h \text{ satisfying} \\ \qquad A(u^h, v^h) = F(v^h) \quad \forall\, v^h \in S^h\,. \end{array} \right. \tag{3.20}$$

If the conditions of the Lax-Milgram theorem hold over the whole space V, then clearly they hold over any subspace S^h. Consequently, existence and uniqueness of (3.20) is automatically guaranteed by the Lax-Milgram theorem. The following result, known as Galerkin's or Cea's theorem, provides us with an error estimate for $\|u - u^h\|$ where $u \in V$ satisfies (3.13), $u^h \in S^h \subset V$ satisfies 3.20, and $\|\cdot\|$ denotes the norm on V. Simply stated, this result says that the error in the solution to the weak problem and its Galerkin approximation is less than or equal to a constant (which is ≥ 1) times the best approximation to the solution of (3.13) in S^h.

Lemma 3.6. (Galerkin's or Cea's Lemma) *Let $A(\cdot,\cdot)$ be a bilinear form on V satisfying (3.16) and (3.17), and let $F(\cdot)$ be a bounded linear functional on V. Let u be the unique solution of*

$$A(u, v) = F(v) \quad \forall\, v \in V$$

guaranteed by the Lax-Milgram theorem. Let $\{S^h\}$, $0 < h < 1$, be a family of finite dimensional subspaces of V. Then for every h there exists a unique $u^h \in S^h$ such that

$$A(u^h, v^h) = F(v^h) \quad \forall\, v^h \in S^h$$

and moreover,

$$\|u - u^h\| \leq \frac{M}{m} \inf_{\chi^h \in S^h} \|u - \chi^h\| \,, \tag{3.21}$$

where M, m are the constants appearing in the Lax-Milgram theorem and $\|\cdot\|$ denotes the norm on V.

Proof. As indicated in the discussion preceding the theorem, the existence and uniqueness of (3.20) is guaranteed by the Lax-Milgram theorem. In order to prove our error estimate, we begin by establishing the so-called Galerkin *orthogonality condition*. We note that (3.13) holds for all $v \in V$ so, in particular, it holds for all $v^h \in S^h \subset V$; *i.e.,*

$$A(u, v^h) = F(v^h) \quad \forall\, v^h \in S^h \,.$$

Subtracting this expression from (3.20) we have that

$$A(u - u^h, v^h) = 0 \quad \forall\, v^h \in S^h \tag{3.22}$$

which says that the error $u - u^h$ is orthogonal to S^h. Using the coercivity property of $A(\cdot,\cdot)$ given in (3.17) we have

$$m \|u - u^h\|^2 \leq A(u - u^h, u - u^h) \,; \tag{3.23}$$

adding and subtracting an arbitrary element $\chi^h \in S^h$ and using the linearity of $A(\cdot,\cdot)$ gives

$$A(u-u^h, u-u^h) = A(u-u^h, u-\chi^h+\chi^h-u^h) = A(u-u^h, u-\chi^h)+A(u-u^h, \chi^h-u^h) \,.$$

Now the orthogonality condition (3.22) tells us that the last term is zero since $\chi^h - u^h \in S^h$. Combining this result with (3.23) and using the bound on $A(\cdot,\cdot)$ given in (3.16) we have

$$m\left\|u-u^h\right\|^2 \le A(u-u^h, u-\chi^h) \le M\left\|u-\chi^h\right\|\left\|u-u^h\right\| \quad \forall\, \chi^h \in S^h$$

and thus

$$\left\|u-u^h\right\| \le \frac{M}{m}\left\|u-\chi^h\right\| \quad \forall\, \chi^h \in S^h\,.$$

Taking the infimum over all $\chi^h \in S^h$ provides the final result. ∎

As an immediate corollary to this result we have that if the family of subspaces S^h has the property that the norm of u minus its best approximation in S^h approaches zero as $h \to 0$ then we have convergence of u^h to u as $h \to 0$.

Corollary 3.7. *If $\{S^h\}$, $0 < h < 1$, is a family of subspaces of V which satisfy*

$$\lim_{h\to 0} \inf_{\chi^h \in S^h} \left\|u-\chi^h\right\| = 0 \tag{3.24}$$

then $\left\|u-u^h\right\| \to 0$ as $h \to 0$.

It is important to note that if w^h is any element of S^h then

$$\inf_{\chi^h \in S^h} \left\|u-\chi^h\right\|_1 \le \left\|u-w^h\right\|_1\,.$$

This is particularly useful when we want to bound the error $\left\|u-u^h\right\|_1$ in terms of powers of h. From the study of approximation theory, we know that bounds are not readily available for the best approximation but bounds are easy to obtain for particular elements of S^h such as the S^h-interpolant of u. Thus if we can bound the error in u and its S^h-interpolant in terms of powers of h, then we have a useful bound for $\left\|u-u^h\right\|_1$. We return to this when we consider particular examples in the next chapter.

The discrete weak problem (3.20) results in a linear algebraic system of equations once a basis is chosen for the n-dimensional space S^h. In particular, let $\{\phi_i(x)\}$, $i = 1,\ldots,n$ be a basis for S^h. Then $u^h \in S^h$ can be written as a linear combination of these basis vectors, *i.e.,*

$$u^h = \sum_{j=1}^{n} \xi_j \phi_j(x)$$

and thus (3.20) becomes

$$a\Big(\sum_{j=1}^{n} \xi_j \phi_j(x), v^h\Big) = F(v) \quad \forall\, v^h \in S^h\,. \tag{3.25}$$

Now testing (3.25) against each $v^h \in S^h$ is equivalent to testing it against each element in the basis for S^h so that we have

$$\sum_{j=1}^{n} \xi_j a\big(\phi_j(x), \phi_i(x)\big) = F\big(\phi_i(x)\big) \quad i = 1, 2, \ldots, n$$

or in matrix form $Ac = b$ where A is an $n \times n$ matrix, $c, b \in \mathbb{R}^n$ with

$$A_{ij} = A(\phi_j, \phi_i)\,, \qquad c_i = \xi_i \quad \text{and} \quad b_i = F(\phi_i)\,. \tag{3.26}$$

Properties of the bilinear form $A(\cdot, \cdot)$ are inherited by the matrix A. From numerical linear algebra, we know that a symmetric, positive definite matrix is easily solved by Cholesky factorization or by an iterative method. Consequently, it is worthwhile to note the conditions on $A(\cdot, \cdot)$ which guarantee that the resulting matrix is symmetric, positive definite.

Lemma 3.8. *Let $A(\cdot, \cdot)$ be a symmetric bilinear form defined on $V \times V$. If $A(\cdot, \cdot)$ satisfies the coercivity condition (3.17), then the matrix defined by (3.26) is symmetric and positive definite.*

Proof. See exercises. ∎

Of course we have not discussed choices of the finite dimensional subspaces S^h; we address some simple choices in the next chapter when we consider examples and Chapter **??** are devoted entirely to the study of finite element spaces. However, it is important to keep in mind that of all possible choices for S^h, finite element methods usually employ continuous piecewise polynomial spaces.

We have seen that if our bilinear form is symmetric and coercive, then the resulting matrix is symmetric, positive definite. However, since the size of our linear system can be quite large, especially in two and three dimensions, we would also like to have a sparse, banded matrix. The choice of basis for S^h governs this sparsity. In particular, we choose basis functions which have *compact support*, *i.e.,* are zero outside of a compact set. So, for example, in one dimension we choose basis functions which are nonzero on as few intervals as possible.

Example 3.9 Returning to (3.15), the variational formulation in Example 3.1, we see that the corresponding discrete weak problem is to seek $u^h \in S^h \subset H_0^1(0,1)$ satisfying

$$\int_0^1 \frac{\partial u^h}{\partial x} \frac{\partial v^h}{\partial x}\, dx = \int_0^1 \sin \pi x\, v^h\, dx \quad \forall\, v^h \in S^h$$

and Galerkin's lemma provides us with an error bound using the norm on $H^1(0,1)$. In particular we have that

$$\big\|u - u^h\big\|_1 \leq \inf_{\chi^h \in S^h} \big\|u - \chi^h\big\|_1$$

where

$$\big\|u - u^h\big\|_1 = \left(\int_0^1 (u - u^h)^2\, dx + \int_0^1 (\frac{du}{dx} - \frac{du^h}{dx})^2\, dx \right)^{1/2}.$$

■

3.4 The Rayleigh-Ritz problem

Recall from linear algebra that determining an $x \in \mathbb{R}^n$ satisfying the linear system $Ax = b$, where A is an $n \times n$ *symmetric, positive definite* matrix, $b \in \mathbb{R}^n$, is equivalent to solving the minimization problem

$$\min_{y \in \mathbb{R}^n} \left(\frac{1}{2} y^T A y - y^T b \right) . \tag{3.27}$$

See exercises. Although we rarely solve a linear system as a minimization problem, the equivalence between the two problems is often useful. In this section we want to show that an analogous relationship exists between the solution of the weak problem (3.13) and an appropriate minimization problem; this minimization problem is often called the Ritz problem or the Rayleigh-Ritz problem.

Consider the minimization problem

$$\min_{v \in V} \mathcal{J}(v) \tag{3.28}$$

where $\mathcal{J} : V \to \mathbb{R}$ is the functional defined by

$$\mathcal{J}(v) = \frac{1}{2} A(v, v) - F(v) \quad \forall\, v \in V . \tag{3.29}$$

It turns out that if $A(\cdot, \cdot)$ satisfies the hypotheses of the Lax-Milgram theorem and is *symmetric* then solving the minimization problem (3.28) is equivalent to solving the weak problem (3.13). Consequently, once we discretize a symmetric problem, we have the choice of solving it as a system of linear algebraic equations or as a minimization problem. The following result demonstrates the equivalence of the two problems.

Theorem 3.10. *Let $A(\cdot, \cdot)$ be a symmetric bilinear form satisfying the hypotheses of the Lax-Milgram Theorem. Then the problem of finding a u satisfying the weak problem (3.13) and finding a solution to the minimization problem (3.28) are equivalent.*

Proof. First assume that $u \in V$ satisfies the weak problem (3.13) and let $w \in V$ be artibrary. Then using the definition (3.29) of $\mathcal{J}$ and the linearity of $A(\cdot, \cdot)$ and $F\cdot$), we obtain

$$\begin{aligned} \mathcal{J}(u + w) &= \frac{1}{2} A(u + w, u + w) - F(u + w) \\ &= \frac{1}{2} A(u, u) + \frac{1}{2} \big(A(w, u) + A(u, w) \big) + A(w, w) - F(u) - F(w) \\ &= \mathcal{J}(u) + A(u, w) - F(w) + A(w, w) , \end{aligned}$$

where in the last step we have used the symmetry of $A(\cdot, \cdot)$ and the definition of $\mathcal{J}$. Since $w \in V$ and u satisfies (3.13), $A(u, w) - F(w) = 0$. Also since $A(\cdot, \cdot)$ is coercive,

$A(w,w) > 0$ for $w \neq 0$. Therefore $\mathcal{J}(u+w) > \mathcal{J}(u)$ and thus u is a minimizer of (3.28).

Now assume that u minimizes $\mathcal{J}(v)$ for all $v \in V$. Then for any scalar σ and $v \in V$, $u + \sigma v \in V$ and so $\mathcal{J}(u+\sigma v) \geq \mathcal{J}(u)$. Then the function $g(\sigma) = \mathcal{J}(u+\sigma v)$ has a minimum at $\sigma = 0$. From calculus, we know that

$$\frac{dg}{d\sigma}\Big|_{\sigma=0} = 0\,.$$

Since

$$\begin{aligned}
\frac{dg}{d\sigma} &= \frac{d}{d\sigma}\left(\frac{1}{2}A(u+\sigma v, u+\sigma v) - F(u+\sigma v)\right) \\
&= \frac{d}{d\sigma}\left(\frac{1}{2}A(u,u) + \sigma A(u,v) + \frac{1}{2}\sigma^2 A(v,v) - F(u) - \sigma F(v)\right) \\
&= A(u,v) + \sigma A(v,v) - F(v)
\end{aligned}$$

where we have used the properties of $A(\cdot,\cdot)$ and the inner product. Evaluating this derivative at $\sigma = 0$, we arrive at $A(u,v) - F(v) = 0$ for all $v \in V$, *i.e.,* if u minimizes (3.28) then u satisfies (3.13). ∎

Exercises

3.1. Let P^h be the projection operator $P^h : V \to S^h$. Demonstrate that

$$\left\|u - u^h\right\| \leq \frac{M}{m} \left\|u - P^h u\right\| , \tag{3.30}$$

where M, m are the constants appearing in the Lax Milgram Theorem 3.5.

3.2. Prove Lemma 3.8.

3.3. Show that on $H_0^1(\Omega)$ the H^1-norm and the H^1-seminorm are equivalent norms.

3.4. Show that determining an $x \in \mathbb{R}^n$ satisfying the linear system $Ax = b$, where A is an $n \times n$ *symmetric, positive definite* matrix, $b \in \mathbb{R}^n$, is equivalent to solving the minimization problem (3.27).

3.5. Give an example of a weak formulation for a linear two-point boundary value problem on $[0, 1]$ which is not equivalent to a Rayleigh-Ritz minimization problem. Explain your reasoning.