Chapter 2

Ritz and Galerkin methods for elliptic problems

In Section 1. we have reformulated the Dirichlet problem to seek weak solutions and we showed its well-posedness. The problem being infinite dimensional, it is not computable.

Question: Can we construct an approximation to Problem (1.1) which is also well-posed?

2.1 Approximate problem

In the previous section we showed how a classical PDE problem such as Problem (1.1) can be reformulated as a weak problem. The abstract problem for this class of PDE reads then:

$$\left| \begin{array}{l} \text{Find } u \in V, \text{ such that:} \\ \\ a(u,v) = L(v) \quad , \forall\, v \in V \end{array} \right. \tag{2.1}$$

with $a(\,\cdot\,,\,\cdot\,)$ a coercive continuous bilinear form on $V \times V$ and $L(\,\cdot\,)$ a continuous linear form on V.

Since in the case of the Poisson problem the bilinear form is continuous, coercive and symmetric, the well-posedness follows directly from Riesz–Fréchet representation Theorem. If the bilinear form is still coercive but not symmetric then we will see that the well-posedness is proven by the Lax–Milgram Theorem.

But for the moment, let us focus on the symmetric case: we want now to construct an approximate solution u_n to the Problem (2.1) then prove that the solution to the obtained approximate problem exists and is unique.

2.2 Ritz method for symmetric bilinear forms

2.2.1 Variational formulation and minimization problem

Ritz's idea is to replace the solution space V (which is infinite dimensional) by a finite dimensional subspace $V_n \subset V$, $\dim(V_n) = n$.

Problem (2.2) is the approximate weak problem by Ritz's method:

$$\left|\begin{array}{l} \text{Find } u_n \in V_n,\ V_n \subset V, \text{ such that:} \\ a(u_n, v_n) = L(v_n) \quad ,\forall\ v_n \in V_n \end{array}\right. \tag{2.2}$$

with $a(\,\cdot\,,\,\cdot\,)$ a coercive symmetric continuous bilinear form on $V \times V$ and $L(\,\cdot\,)$ a continuous linear form on V.

Provided that the bilinear form is symmetric, Problem (2.3) is the equivalent approximate variational problem under minimization form:

$$\left|\begin{array}{l} \text{Find } u_n \in V_n,\ V_n \subset V, \text{ such that:} \\ J(u_n) \leq J(v_n) \quad ,\forall\ v_n \in V_n \\ \text{with } J(v_n) = \dfrac{1}{2}a(v_n, v_n) - L(v_n) \end{array}\right. \tag{2.3}$$

Proposition 2.2.1. *Equivalence of weak and variational formulations Problem 2.2 and 2.3 are equivalent.*

Before moving to the well-posedness of the approximate variational problem some definitions are introduced to caracterize the solution of mimimization problems, then the equivalence of formulations for the Poisson problem with homogeneous Dirichlet boundary conditions in one dimension of space is given as example.

Definition 2.2.2 (Directional derivative)**.** Let V be a Hilbert space, for any $u \in V$ the relation:

$$J'(u; w) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\big(J(u + \varepsilon w) - J(u)\big) \tag{2.4}$$

defines $J'(\cdot;\cdot) : V \times V \to \mathbb{R}$ derivative of the functional J at u in the direction w.

Definition 2.2.3 (Fréchet derivative)**.** Let V be a Hilbert space, J is Fréchet-derivable at u if:

$$J(u + v) = J(u) + L_u(v) + \varepsilon(v)\|v\|_V \tag{2.5}$$

with L_u a continuous linear form on V and $\varepsilon(v) \to 0$ as $v \to 0$.

Proposition 2.2.4 (Optimality conditions)**.** *Let V be a Hilbert space and J a twice Fréchet-derivable functional, $u_0 \in V$ is solution to*

$$\inf_{v \in V} J(v) \tag{2.6}$$

if the following conditions are satisfied:

1. $J'(u_0) = 0$ *(Euler condition).*

2. $\langle\, J''(u)w\,,\, w\,\rangle \geq 0$ *(Legendre condition).*

Both conditions can be interpreted in terms of the simpler case of real functions: the first one requires that the first derivative cancels so that u_0 is an extremum, while the second condition is a convexity argument. Moreover, a sufficient condition is given by $\langle\, J''(\tilde{u})w\,,\, w\,\rangle \geq 0$ for any $\tilde{u}$ in a neighbourhood of u_0 (strong Legendre condition). The coercivity of the bilinear form $a(\cdot,\cdot)$ is an even stronger condition equivalent to: $\exists \alpha > 0$ such that $\langle\, J''(\tilde{u})w\,,\, w\,\rangle \geq \alpha(w,w)$.

Example 2.2.5. Equivalence of weak and variational formulations for the Dirichlet problem posed on $\Omega = (0,1)$. Let us derive the expression of $J'(u;w)$ defined by (2.5) given $\varepsilon > 0$ and $w \in V$.

First let us verify that if u solves the minimization problem then it solves the corresponding weak problem.

$$\begin{aligned} J(u+\varepsilon w) &= \frac{1}{2}\int_\Omega \left[(u+\varepsilon w)'\right]^2 \mathrm{d}x - \int_\Omega f(u+\varepsilon w)\,\mathrm{d}x \\ &= \frac{1}{2}\int_\Omega \left[(u')^2 + 2\varepsilon u'w' + \varepsilon^2 (v')^2\right]\mathrm{d}x - \int_\Omega fu\,\mathrm{d}x - \varepsilon\int_\Omega fw\,\mathrm{d}x \\ &= J(u) + \varepsilon\left[\int_\Omega u'w'\,\mathrm{d}x - \int_\Omega fw\,\mathrm{d}x\right] + \frac{1}{2}\,\varepsilon^2\int_\Omega (w')^2\,\mathrm{d}x \end{aligned}$$

Writing the derivative gives,

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\big(J(u+\varepsilon w) - J(u)\big) = \lim_{\varepsilon\to 0}\left[\int_\Omega u'w'\,\mathrm{d}x - \int_\Omega fw\,\mathrm{d}x + \frac{1}{2}\,\varepsilon|w|_{\mathrm{H}^1_0}\right]$$

so that the Euler condition holds if for any $w \in V = \mathrm{H}^1_0(\Omega)$

$$J'(u;w) = \int_\Omega u'w'\,\mathrm{d}x - \int_\Omega fw\,\mathrm{d}x = 0$$

In this case the functional J is Fréchet-derivable as L_u is linear.

Secondly, the other way around considering that the weak formulation holds for the test function $\varepsilon w \in V$ then in the relation

$$J(u+\varepsilon w) = J(u) + \varepsilon\left[\int_\Omega u'w'\,\mathrm{d}x - \int_\Omega fw\,\mathrm{d}x\right] + \frac{1}{2}\,\varepsilon^2\int_\Omega (w')^2\,\mathrm{d}x$$

the second term of the right-hand side cancels, and the third term is non-negative, then

$$J(u+\varepsilon w) \geq J(u)$$

so that u is solution to the minimization problem.

The same result holds for the continuous problem in V and the approximation in V_n since only requirement is to work in a Hilbert space. Actually the following result for the Dirichlet problem is due to Stampacchia which characterizes the solution to the weak problem in term of minimization.

Theorem 2.2.6 (Stampacchia)**.** *Let $a(\cdot,\cdot)$ be a bilinear coercive continuous form on* H *a Hilbert space, and K be a convex closed non-empty subset of H. Given $\phi \in H'$, $\exists! u \in K$ such that*

$$a(\, u\, ,\, v-u\,) \geq \langle\, \phi\, ,\, v-u\,\rangle_{H',H}, \qquad \forall\, v \in K$$

and if a is symmetric then

$$u = \underset{v\in K}{\operatorname{argmin}} \left\{ \frac{1}{2} a(\, v\, ,\, v\,) - \langle\, \phi\, ,\, v\,\rangle_{H',H} \right\}$$

The solution can be seen as satisfying a minimization of energy, also called Dirichlet principle.

2.2.2 Well-posedness

Theorem 2.2.7 (Well-posedness)**.** *Let V be a Hilbert space and V_n a finite dimensional subspace of V, $\dim(V_n) = n$, Problem* (2.2) *admits a unique solution u_n.*

Proof. Given that the weak formulation differs only by introducting finite dimensional subspaces the proof could conclude directly with the Lax–Milgram Theorem. Instead we show that there exists a unique solution to the equivalent minimisation problem (2.3) by explicitly constructing an approximation $u_n \in V_n$ decomposed uniquely on a basis $(\varphi_1, \cdots, \varphi_n)$ of V_n:

$$u_n = \sum_{j=1}^{n} u_j\, \varphi_j$$

In practice this basis is not any basis but the one constructed to define the approximation space V_n: to one chosen approximation space will correspond one carefully constructed basis. In so doing, the constructive approach paves the way to the Finite Element Method and is thus chosen as a prequel to establishing the Galerkin method.

Writing the minimisation functional for u_n reads:

$$\begin{aligned} J(u_n) &= \frac{1}{2}\, a(u_n, u_n) - L(u_n) \\ &= \frac{1}{2}\, a(\sum_{j=1}^{n} u_j\varphi_j, \sum_{i=1}^{n} u_i\varphi_i) - L(\sum_{j=1}^{n} u_i\varphi_i) \\ &= \frac{1}{2} \sum_{i=1}^{n}\sum_{j=1}^{n} a(u_j\varphi_j, u_i\varphi_i) - \sum_{j=1}^{n} L(u_i\varphi_i) \\ &= \frac{1}{2} \sum_{i=1}^{n}\sum_{j=1}^{n} u_j u_i a(\varphi_j, \varphi_i) - \sum_{j=1}^{n} u_i L(\varphi_i) \end{aligned}$$

Collecting the entries by index i, the functional can be rewritten under algebraic form:

$$\mathrm{J}(\mathbf{u}) = \frac{1}{2}\mathbf{u}^{T}\mathrm{A}\mathbf{u} - \mathbf{u}^{T}\mathbf{b}$$

where $\mathbf{u}$ is the vector of algebraic unknowns also called *degrees of freedom*

$$\mathbf{u}^T = (u_1, \ldots, u_n)$$

and A, $\mathbf{b}$ are respectively the stiffness matrix and the load vector:

$$\mathrm{A}_{ij} = a(\varphi_j, \varphi_i), \mathbf{b}_i = L(\varphi_i)$$

Proposition 2.2.8 (Convexity of a quadratic form)**.**

$$\mathrm{J}(\mathbf{u}) = \mathbf{u}^T \mathrm{K}\mathbf{u} - \mathbf{u}^T \mathrm{G} + \mathrm{F}$$

is a strictly convex quadratic functional iff K *symmetric positive definite non-singular.*

As a consequence to Proposition 2.2.8 J is a strictly convex quadratic form, then there exists a unique $\mathbf{u} \in \mathbb{R}^n \ : \ \mathrm{J}(\mathbf{u}) \leq \mathrm{J}(\mathbf{v}), \forall\, \mathbf{v} \in \mathbb{R}^n$, which in turns proves the existence and uniqueness of $u_n \in V_n$.

The minimum is achieved with $\mathbf{u}$ satisfying $\mathrm{A}\mathbf{u} = \mathbf{b}$ which corresponds to the Euler condition $J'(u_n) = 0$ □

The general setting for Galerkin methods will be to construct approximate solutions of the form:

$$u_n = \sum_{j=1}^{n} u_j \varphi_j \tag{2.8}$$

where $\{u_j\}_{1 \leq j \leq n}$ is a family of real numbers and $\mathcal{B} = (\varphi_1, \ldots, \varphi_n)$ a basis of V_n. Since V_n is finite dimensional, there exist a unique decomposition (2.8) on the basis. This basis can be chosen in a way that seems natural so that in practice we will construct a unique basis for a given type of space V_n and which will define the approximation properties (the basis itself is not unique but we need to choose one that possesses good properties).

2.2.3 Convergence

The question in this section is: considering a sequence of discrete solutions $(u_n)_{n \in \mathbb{N}}$, with each u_n belonging to V_n, can we prove that $u_n \to u$ in V as $n \to \infty$? The ingredients are similar to the Lax principle: stability and consistency implies convergence.

Lemma 2.2.9 (Estimate in the energy norm)**.** *Let V be a Hilbert space and V_n a finite dimensional subspace of V. We denote by $u \in V$, $u_n \in V_n$ respectively the solution to Problem* (2.1) *and the solution to approximate Problem* (2.2)*. Let us define the energy norm $\| \cdot \|_a = a(\cdot, \cdot)^{1/2}$, then the following inequality holds:*

$$\|u - u_n\|_a \leq \|u - v_n\|_a \quad , \forall\, v_n \in V_n$$

Proof. Using the coercivity and the continuity of the bilinear form, we have:

$$\alpha\|u\|_V^2 \leq \|u\|_a^2 \leq M\|u\|_V^2$$

then $\|u\|_a$ is norm equivalent to $\|u\|_V$, thus $(V, \|\cdot\|_a)$ is a Hilbert space.

$$a(u - \mathrm{P}_{V_n} u, v_n) = 0 \quad , \forall\, v_n \in V_n$$

by definition of P_{V_n} as the orthogonal projection of u onto V_n with respect to the scalar product defined by the bilinear form a.

$$\|u - u_n\|_a^2 = a(u - u_n, u - v_n) + a(u - u_n, v_n - u_n) \quad , \forall\, v_n \in V_n$$

Since the second term of the right-hand side cancels due to the consistency of the approximation, we deduce $u_n = \mathrm{P}_{V_n} u$, then u_n minimizes the distance from u to V_n:

$$\|u - u_n\|_a^2 \leq \|u - v_n\|_a^2 \quad , \forall\, v_n \in V_n$$

which means that the error estimate is *optimal* in the energy norm. □

Lemma 2.2.10 (Céa's Lemma)**.** *Let V be a Hilbert space and V_n a finite dimensional subspace of V. we denote by $u \in V$, $u_n \in V_n$ respectively the solution to Problem* (2.1) *and the solution to approximate Problem* (2.2) *, then the following inequality holds:*

$$\|u - u_n\|_V \leq \sqrt{\frac{M}{\alpha}}\|u - v_n\|_V \quad , \forall\, v_n \in V_n$$

with $M > 0$ the continuity constant and $\alpha > 0$ the coercivity constant.

Proof. Using the coercivity and continuity of the bilinear form, we bound the left-hand side of the estimate (2.2.9) from below and its right-hand side from above:

$$\alpha\|u - u_n\|_V^2 \leq M\|u - v_n\|_V^2 \quad \forall\, v_n \in V_n$$

Consequently:

$$\|u - u_n\|_V \leq \sqrt{\frac{M}{\alpha}}\|u - v_n\|_V \quad , \forall\, v_n \in V_n$$

□

Lemma (2.2.10) gives a control on the discretisation error $e_n = u - u_n$ which is *quasi-optimal* in the V-norm (*i.e.* bound multiplied by a constant).

Lemma 2.2.11 (Stability)**.** *Any solution $u_n \in V_n$ to Problem* (2.2) *satisfies:*

$$\|u_n\|_V \leq \frac{\|L\|_{V'}}{\alpha}$$

Proof. Direct using the coercivity and the dual norm. □

2.2.4 Method

Algorithm 2.2.12 (Ritz's method)**.** *The following procedure applies:*

1. *Chose an approximation space* V_n
2. *Construct a basis* $\mathcal{B} = (\varphi_1, \ldots, \varphi_n)$
3. *Assemble stiffness matrix* A *and load vector* $\mathbf{b}$
4. *Solve* $\mathrm{A}\mathbf{u} = \mathbf{b}$ *as a minimisation problem*

2.3 Galerkin method

2.3.1 Formulation

We use a similar approach as for Ritz's method, except that the abstract problem does not require the symmetry of the bilinear form. Therefore we cannot endow V with a norm defined from the scalar product based on $a(\,\cdot\,,\,\cdot\,)$.

Problem (2.9) is the approximate weak problem by Galerkin's method:

$$\left|\begin{array}{l} \text{Find } u_n \in V_n,\ V_n \subset V, \text{ such that:} \\ a(u_n, v_n) = L(v_n) \quad ,\forall\ v_n \in V_n \end{array}\right. \tag{2.9}$$

with $a(\,\cdot\,,\,\cdot\,)$ a coercive continuous bilinear form on $V \times V$ and $L(\,\cdot\,)$ a continuous linear form on V.

2.3.2 Convergence

The following property is merely a consequence of the consistency, as the continuous solution u is solution to the discrete problem (*i.e.* the bilinear form is the "same"), but it is quite useful to derive error estimates. Consequently, whenever needed we will refer to the following proposition:

Proposition 2.3.1 (Galerkin orthogonality)**.** *Let* $u \in V$, $u_n \in V_n$ *respectively the solution to Problem* (2.1) *and the solution to approximate Problem* (2.9)*, then:*

$$a(\, u - u_n, v_n \,) = 0 \quad ,\ \forall\ v_n \in V_n$$

Proof. Direct consequence of the consistency of the method. □

Lemma 2.3.2 (Consistency)**.** *Let* V *be a Hilbert space and* V_n *a finite dimensional subspace of* V*. we denote by* $u \in V$, $u_n \in V_n$ *respectively the solution to Problem* (2.1) *and the solution to approximate Problem* (2.9)*, then the following inequality holds:*

$$\|u - u_n\|_V \leq \frac{M}{\alpha}\|u - v_n\|_V \quad ,\ \forall\ v_n \in V_n$$

with $M > 0$ *the continuity constant and* $\alpha > 0$ *the coercivity constant.*

Proof. Using the coercivity:

$$\begin{aligned}
\alpha\|u-u_n\|_V^2 &\leq a(u-u_n, u-u_n) \\
&\leq a(u-u_n, u-v_n) + \underbrace{a(u-u_n, v_n-u_n)}_{0} \\
&\leq a(u-u_n, u-v_n) \\
&\leq M\|u-u_n\|_V\|u-v_n\|_V \\
\|u-u_n\|_V &\leq \frac{M}{\alpha}\|u-v_n\|_V
\end{aligned}$$

□

The only difference with the symmetric case is that the constant is squared due to the loss of the symmetry.

2.3.3 Method

Algorithm 2.3.3 (Galerkin's method)**.** *The following procedure applies:*

1. *Chose an approximation space* V_n
2. *Construct a basis* $\mathcal{B} = (\varphi_1, \ldots, \varphi_n)$
3. *Assemble stiffness matrix* A *and load vector* $\mathbf{b}$
4. *Solve* $\mathrm{A}\mathbf{u} = \mathbf{b}$

2.4 Exercises

Exercise 2.4.1. Given an abstract weak problem posed in a Hilber space V:

$$\left|\begin{array}{l} \text{Find } u \in V,\ V, \text{ such that:} \\ a(u,v) = L(v) \quad ,\forall\ v \in V \end{array}\right.$$

and a minimization problem

$$\left|\begin{array}{l} \text{Find } u \in V,\ V, \text{ such that:} \\ J(u) \leq J(v) \quad ,\forall\ v \in V \\ \text{with } J(v) = \dfrac{1}{2}a(v,v) - L(v) \end{array}\right.$$

1. Show the equivalence of the formulation when a is bilinear s.p.d and L linear.
2. Show that if $V = \mathbb{R}^n$ the minimization problem can be recast into a stricly convex quadratic form $J(u) = \frac{1}{2}u^T Au - u^T b$ and the unique solution satisfies $Au = b$.

Exercise 2.4.2. Let us consider the Poisson problem posed on the domain $\Omega = (0,1)$:

$$-u''(x) = f(x), \qquad \forall\ x \in \Omega \tag{2.10a}$$

with $f \in \mathrm{L}^2(\Omega)$, and satisfying the boundary condition on $\partial\Omega$

$$u(x) = 0, \qquad \forall\ x \in \partial\Omega \tag{2.10b}$$

1. For $f \equiv 1$ give a solution to Problem (2.10).
2. Find the weak formulation (WF) of Problem (2.10) and specify the function spaces.
3. Is this problem well-posed?
4. Justify that it is possible to reformulate this problem into a minimization problem?
5. Derive the minimisation functional $J(u)$.
6. Let $w_1 = a_1 \sin(\pi x)$. Find the value of the amplitude a_1 minimizes $J(w_1)$. How does a_1 compare with the maximum of the exact solution u?
7. Show that $J(w_1) > J(u)$ and interpret.
8. Let $\phi_i = \sin(2i-1)\pi x)$, $i \in \mathbb{N}$. Verify that these function are infinitely differentiable and satisfy $\phi_i(0) = \phi_i(1) = 0$. Compute coefficients
$$a_{ij} = \int_0^1 \phi_i'(x)\phi_j'(x)\,\mathrm{d}x\ , \qquad b_i = \int_0^1 \phi_i(x)\,\mathrm{d}x$$
9. Given a finite dimensional space $V_n = span\{\phi_i\}_{1\leq i\leq n}$, express the linear system obtained by the Galerkin method and give the solution.