

A Finite Element Primer ^{*}

David J. Silvester
School of Mathematics, University of Manchester
d.silvester@manchester.ac.uk.

Version 1.0, updated 3 January 2007

Contents

1	A Model Diffusion Problem	1
	x.1 Domain	1
	x.2 Continuous Function	2
	x.3 Normed Vector Space	2
	x.4 Square Integrable Function	3
	x.5 Inner Product Space	4
	x.6 Cauchy-Schwarz Inequality	5
	x.7 Sobolev Space	5
	x.8 Weak Derivative	11
2	Galerkin Approximation	12
	x.9 Cauchy Sequence	16
	x.10 Complete Space	16
3	Finite Element Galerkin Approximation	18

^{*}This is a summary of finite element theory for a diffusion problem in one dimension. It provides the mathematical foundation for Chapter 1 of our reference book *Finite Elements and Fast Iterative Solvers with Applications in Incompressible Fluid Dynamics*, see <http://www.oup.co.uk/isbn/0-19-852868-X>

1. A Model Diffusion Problem

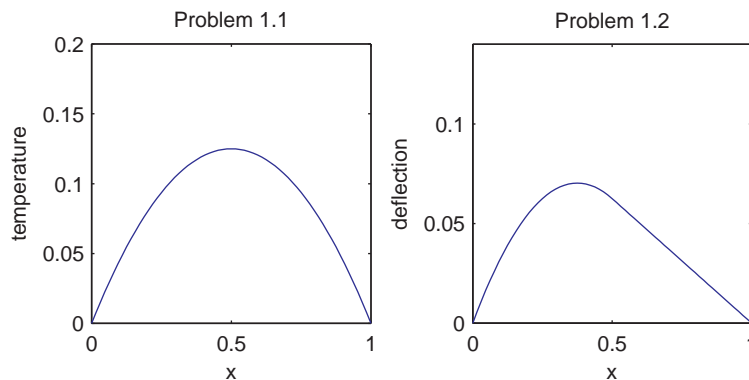
The problem we consider herein is a two point boundary value problem. A formal statement is: given a real function $f \in C^0(0, 1)$ (see Definition x.2 below), we seek a function $u \in C^2(0, 1) \cap C^0[0, 1]$ (see below) satisfying

$$\left. \begin{aligned} -\frac{d^2u}{dx^2} &= f \quad \text{for } 0 < x < 1 \\ u(0) &= 0; \quad u(1) = 0. \end{aligned} \right\} \quad (D)$$

The term $-\frac{d^2u}{dx^2}$ represents “diffusion”, and f is called the “source” term. A sufficiently smooth function u satisfying (D) is called a “strong” (or “classical”) solution.

Problem 1.1 ($f = 1$) This is a model for the temperature in a wire with the ends kept in ice. There is a current flowing in the wire which generates heat. Solving (D) gives the parabolic “hump”

$$u(x) = \frac{1}{2}(x - x^2).$$



Some basic definitions will need to be added if our statement of (D) is to make sense.

Definition x.1 (Domain)

A **domain** is a bounded open set; for example, $\Omega = (0, 1)$, which identifies where a differential equation is defined. The “closure” of the set, denoted $\overline{\Omega}$, includes all the points on the boundary of the domain; for example, $\overline{\Omega} = [0, 1]$.

Definition x.2 (Continuous function)

A real **function** f is mapping which assigns a unique real number to every point in a domain: $f : \Omega \rightarrow \mathbb{R}$.

- $C^0(\Omega)$ is the set of all continuous functions defined on Ω .
- $C^k(\Omega)$ is the set of all continuous functions whose k th derivatives are also continuous over Ω .
- $C^0(\overline{\Omega})$ is the set of all functions $u \in C^0(\Omega)$ such that u can be extended to a continuous function on $\overline{\Omega}$.

The standard way of categorizing spaces of functions is to use the notion of a “norm”. This is made explicit in the following definition.

Definition x.3 (Normed vector space)

A **normed vector space** V , has (or more formally is “equipped with”) a mapping $\|\cdot\| : V \rightarrow \mathbb{R}$ which satisfies four axioms:

- ① $\|u\| \geq 0 \quad \forall u \in V;$ (where \forall means “for all”)
- ② $\|u\| = 0 \iff u = 0;$ (where \iff means “if and only if”)
- ③ $\|\alpha u\| = |\alpha| \|u\|, \quad \forall \alpha \in \mathbb{R} \text{ and } \forall u \in V;$
- ④ $\|u + v\| \leq \|u\| + \|v\| \quad \forall u, v \in V.$

Note that, if the second axiom is relaxed to the weaker condition $\|u\| = 0 \Leftarrow u = 0$ then V is only equipped with a **semi-norm**. A normed vector space that is “complete” (see Definition x.10 below) is called a **Banach Space**.

Two examples of normed vector spaces are given below.

Example x.3.1 Suppose that $V = \mathbb{R}^2$, that is, all vectors $\mathbf{u} = \begin{bmatrix} u_x \\ u_y \end{bmatrix}$.

Valid norms are

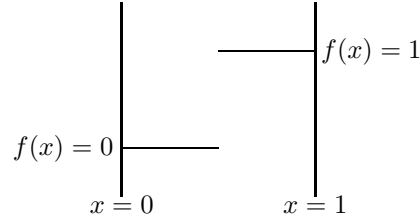
$$\begin{aligned} \|\mathbf{u}\|_1 &= |u_x| + |u_y|; & \ell_1 \text{ norm} \\ \|\mathbf{u}\|_2 &= (u_x^2 + u_y^2)^{1/2}; & \ell_2 \text{ norm} \quad \heartsuit \\ \|\mathbf{u}\|_\infty &= \max\{|u_x|, |u_y|\}. & \ell_\infty \text{ norm} \end{aligned}$$

Example x.3.2 Suppose that $V = C^0(\Omega)$. A valid norm is

$$\|u\| = \max_{x \in \overline{\Omega}} |u(x)|. \quad L_\infty \text{ norm.} \quad \heartsuit$$

Returning to (D) , the source function $f(x)$ may well be “rough”, $f \notin C^0(\Omega)$. An example is given below.

Problem 1.2 ($f(x) = H(1/2)$; where $H(x)$ is the “unit step” function)



This is a model for the deflection of a simply supported elastic beam subject to a discontinuous load. Solving the differential equation over the two intervals and imposing continuity of the solution and the first derivative at the interface point $x = 1/2$ gives the “generalized” solution shown in the figure on page 1:

$$u(x) = \begin{cases} -\frac{x^2}{2} + \frac{3}{8}x & 0 \leq x < \frac{1}{2} \\ -\frac{x}{8} + \frac{1}{8} & \frac{1}{2} \leq x \leq 1. \end{cases}$$

Solving (D) means finding two functions:

$$\begin{aligned} \text{first, } v \text{ such that } \frac{dv}{dx} &= -f, \\ \text{second, } u \text{ such that } \frac{du}{dx} &= v. \end{aligned}$$

We will see that an appropriate starting point for constructing function spaces for v (and hence u) is the space of square integrable functions.

Definition x.4 (Square integrable function)

$L_2(\Omega)$ is the vector space of square integrable functions defined on Ω :

$$u \in L_2(\Omega) \text{ if and only if } \int_{\Omega} u^2 < \infty.$$

Functions that are not continuous in $[0, 1]$ may still be square integrable. We give two examples below.

Example x.4.1 Consider $f = x^{-1/4}$.

$$\int_0^1 f^2 dx = \int_0^1 x^{-1/2} dx = 2$$

hence $\int_0^1 f^2 < \infty$ so that $f \in L_2(\Omega)$. \square

Example x.4.2 Consider $f = \begin{cases} 0 & 0 \leq x < \frac{1}{2} \\ 1 & \frac{1}{2} \leq x \leq 1 \end{cases}$.

$$\begin{aligned} \int_0^1 f^2 dx &= \int_0^{1/2} f^2 dx + \int_{1/2}^1 f^2 dx \\ &= \underbrace{\int_0^{1/2} 0 dx}_0 + \underbrace{\int_{1/2}^1 dx}_{1/2} \end{aligned}$$

hence $\int_0^1 f^2 < \infty$ so that $f \in L_2(\Omega)$. \square

$L_2(\Omega)$ is a Banach space.

Example x.3.3 Suppose that $V = L_2(\Omega)$ with $\Omega = (0, 1)$. A valid norm is

$$\|u\| = \left(\int_0^1 u^2 dx \right)^{1/2}. \quad L_2 \text{ norm} \quad \heartsuit$$

$L_2(\Omega)$ is pretty special—it is also equipped with an inner product.

Definition x.5 (Inner product space)

An **inner product** space V , has a mapping $(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$ which satisfies four axioms:

- ❶ $(u, w) = (w, u) \quad \forall u, w \in V;$
- ❷ $(u, u) \geq 0 \quad \forall u \in V;$
- ❸ $(u, u) = 0 \iff u = 0;$
- ❹ $(\alpha u + \beta v, w) = \alpha(u, w) + \beta(v, w) \quad \forall \alpha, \beta \in \mathbb{R}; \quad \forall u, v, w \in V.$

Example x.5.1 Suppose that $V = \mathbb{R}^2$. A valid inner product is given by

$$(\mathbf{u}, \mathbf{w}) = u_x w_x + u_y w_y = \mathbf{u} \cdot \mathbf{w} \quad \heartsuit$$

Example x.5.2 Suppose that $V = L_2(\Omega)$, with $\Omega = (0, 1)$.

A valid inner product is given by

$$(u, w) = \int_0^1 uw. \quad \heartsuit$$

A complete inner product space like $L_2(\Omega)$ is called a **Hilbert Space**.

Note that an inner product space is also a normed space. There is a “natural” (or “energy”) norm

$$\|u\| = (u, u)^{\frac{1}{2}}.$$

Inner products and norms are related by the Cauchy-Schwarz inequality.

Definition x.6 (Cauchy-Schwarz inequality)

$$|(u, v)| \leq \|u\| \|v\| \quad \forall u, v \in V. \quad (C-S)$$

Example x.6.1 Suppose that $V = \mathbb{R}^2$. We have the discrete version of C–S:

$$\mathbf{u} \cdot \mathbf{w} \leq |\mathbf{u} \cdot \mathbf{w}| \leq (u_x^2 + u_y^2)^{1/2} (w_x^2 + w_y^2)^{1/2}. \quad \heartsuit$$

Example x.6.2 Suppose that $V = L_2(\Omega)$, with $\Omega = (0, 1)$. We have

$$\int_0^1 uw \leq \left| \int_0^1 uw \right| \leq \left(\int_0^1 u^2 \right)^{1/2} \left(\int_0^1 w^2 \right)^{1/2}. \quad \heartsuit$$

Returning to Problem 1.2, we now address the question of where (in which function space) do we look for the generalized solution u when the function f is square integrable but *not* continuous? The answer to this is “in a Sobolev space”.

Definition x.7 (Sobolev space)

For a positive index k , the Sobolev space $H^k(0, 1)$ is the set of functions $v : (0, 1) \rightarrow \mathbb{R}$ such that v and all derivatives up to and including k are square integrable:

$$u \in H^k(0, 1) \iff \int_0^1 u^2 < \infty, \int_0^1 \left(\frac{du}{dx} \right)^2 < \infty, \dots, \int_0^1 \left(\frac{d^k u}{dx^k} \right)^2 < \infty.$$

Note that $H^k(0, 1)$ defines a Hilbert space with inner product

$$(u, w)_k = \int_0^1 uw + \int_0^1 \left(\frac{du}{dx} \right) \left(\frac{dw}{dx} \right) + \dots + \int_0^1 \left(\frac{d^k u}{dx^k} \right) \left(\frac{d^k w}{dx^k} \right)$$

and norm

$$\|u\|_k = \left(\int_0^1 u^2 + \int_0^1 \left(\frac{du}{dx} \right)^2 + \dots + \int_0^1 \left(\frac{d^k u}{dx^k} \right)^2 \right)^{1/2}.$$

This gives a model for the deflection of a simply supported elastic beam subject to a point load. The solution is called a “fundamental solution” or a Green’s function. In this case, since $v \in H^1(0, 1) \Rightarrow v \in C^0(0, 1)$, we have that $\int_0^1 f v = v(1/2) < \infty$, so that (M) is well defined. (Note that this statement is only true for domains in \mathbb{R}^1 —in higher dimensions the Dirac delta function is not admissible as load data.)

Returning to (M), we can compute u by solving the following “variational formulation” : given $f \in L_2(0, 1)$ find $u \in H_0^1(0, 1)$ such that

$$\int_0^1 \frac{du}{dx} \frac{dv}{dx} = \int_0^1 f v \quad \forall v \in H_0^1(0, 1). \quad (V)$$

A solution to (V) (or, equivalently a solution to (M)) is called a “weak” solution. The relationship between (D), (M) and (V) is explored in the following three theorems.

Theorem 1.1 ((D) \Rightarrow (V))

If u solves (D) then it solves (V).

Proof. Let u satisfy (D). Since continuous functions are square integrable then $u \in L_2(0, 1)$ and $\frac{du}{dx} \in L_2(0, 1)$. Furthermore since $u(0) = 0 = u(1)$ from the statement of (D), we have that $u \in H_0^1(0, 1)$.

To show (V), let $v \in H_0^1(0, 1)$, multiply (D) by v and integrate over Ω :

$$-\int_0^1 \frac{d^2u}{dx^2} v = \int_0^1 f v.$$

Using integrating by parts gives

$$-\int_0^1 \frac{d^2u}{dx^2} v = \int_0^1 \frac{du}{dx} \frac{dv}{dx} - \left[\frac{du}{dx} v \right]_0^1,$$

where

$$\left[\frac{du}{dx} v \right]_0^1 = \frac{du}{dx}(1)v(1) - \frac{du}{dx}(0)v(0),$$

and since $v \in H_0^1(0, 1)$ we have $v(0) = v(1) = 0$ so that the boundary term is zero. Thus we have that u satisfies

$$\int_0^1 \frac{du}{dx} \frac{dv}{dx} = \int_0^1 f v \quad \forall v \in H_0^1(0, 1)$$

as required. \square

The above proof shows us how to “construct” a weak formulation from a classical formulation. We now use the properties of an inner product in Definition x.5 to show that (V) has a unique solution.

Theorem 1.2 *A solution to (V) is unique.*

Proof. Let $V = H_0^1(0,1)$ and assume that there are two weak solutions $u_1(x) \in V$, $u_2(x) \in V$ such that

$$\begin{aligned} \left(\frac{du_1}{dx}, \frac{dv}{dx} \right) &= (f, v) \quad \forall v \in V, & (a, b) &= \int_0^1 ab; \\ \left(\frac{du_2}{dx}, \frac{dv}{dx} \right) &= (f, v) \quad \forall v \in V. \end{aligned}$$

Subtracting

$$\left(\frac{du_1}{dx}, \frac{dv}{dx} \right) - \left(\frac{du_2}{dx}, \frac{dv}{dx} \right) = 0 \quad \forall v \in V,$$

and then using ❹ gives

$$\left(\frac{du_1}{dx} - \frac{du_2}{dx}, \frac{dv}{dx} \right) = 0 \quad \forall v \in V.$$

We now define $w = u_1 - u_2$, so that

$$\left(\frac{dw}{dx}, \frac{dv}{dx} \right) \quad \forall v \in V. \quad (\ddagger)$$

Our aim is to show that $w = 0$ in $(0,1)$ (so that $u_1 = u_2$.) To do this we note that $w \in V$ (by the definition of a vector space) and set $v = w$ in (\ddagger) . Using ❺ then gives

$$\left(\frac{dw}{dx}, \frac{dw}{dx} \right) = 0 \Rightarrow \frac{dw}{dx} = 0. \quad (*)$$

From which we might deduce that w is constant. Finally, if we use the fact that functions in V are continuous over $[0,1]$, and are zero at the end points, we can see that $w = 0$ as required. \square

The “hole” in the above argument is that two square integrable functions which are identical in $[0,1]$ except at a finite set of points are **equivalent** to each other. (They cannot be distinguished from each other in the sense of taking their L_2 norm.) Thus, a more precise statement of $(*)$ is that $dw/dx = 0$ “almost everywhere”. Thus a more rigorous way of establishing uniqueness is to use the famous Poincaré–Friedrich inequality.

Lemma 1.3 (Poincaré–Friedrich)

If $w \in H_0^1(0,1)$ then

$$\int_0^1 w^2 \leq \int_0^1 \left(\frac{dw}{dx} \right)^2. \quad (P-F)$$

Proof. See below. \square

Thus, starting from (*) and using P - F gives

$$\underbrace{\int_0^1 w^2}_{\geq 0} \leq \int_0^1 \left(\frac{dw}{dx} \right)^2 = 0$$

and we deduce that $w = 0$ almost everywhere in $(0, 1)$, so that there is a unique solution to (V) in the L_2 sense. \square

Proof. (of P - F)

Suppose $w \in H_0^1(0, 1)$, then

$$w(x) = w(0) + \int_0^x \frac{dw}{dx}(\xi) d\xi.$$

Thus, since $w(0) = 0$ we have

$$\begin{aligned} w^2 &= \left| \int_0^x \frac{dw}{dx} \right|^2 \\ &\leq \left(\int_0^x 1^2 \right) \left(\int_0^x \left(\frac{dw}{dx} \right)^2 \right) && \text{using } C\text{-}S \\ &\leq \underbrace{\left(\int_0^1 1^2 \right)}_{=1} \left(\int_0^1 \left(\frac{dw}{dx} \right)^2 \right) && \text{because } x \leq 1. \end{aligned}$$

Hence $w^2(x) \leq \int_0^1 \left(\frac{dw}{dx} \right)^2$, and integrating over $(0, 1)$ gives

$$\int_0^1 w^2 \leq \int_0^1 \underbrace{\left\{ \int_0^1 \left(\frac{dw}{dx} \right)^2 \right\}}_{\in \mathbb{R}^+} dx = \left\{ \int_0^1 \left(\frac{dw}{dx} \right)^2 \right\} \underbrace{\int_0^1 dx}_{=1}$$

as required. \square

Theorem 1.4 $((V) \Leftrightarrow (M))$

If u solves (V) then u solves (M) and vice versa.

Proof.

(I) $(V) \Rightarrow (M)$

Let $u \in H_0^1(0, 1)$ be the solution of (V) , that is,

$$\left(\frac{du}{dx}, \frac{dv}{dx} \right) = (f, v) \quad \forall v \in H_0^1(0, 1).$$

Suppose $v \in H_0^1(0, 1)$, and define $w = v - u \in H_0^1(0, 1)$, then using the symmetry **1** and linearity **4** of the inner product gives

$$\begin{aligned}
F(v) &= F(u + w) \\
&= \frac{1}{2} \left(\frac{d}{dx}(u + w), \frac{d}{dx}(u + w) \right) - (f, u + w) \\
&= \frac{1}{2} \left(\frac{du}{dx}, \frac{du}{dx} \right) - (f, u) + \frac{1}{2} \left(\frac{dw}{dx}, \frac{dw}{dx} \right) \\
&\quad + \frac{1}{2} \underbrace{\left(\frac{dw}{dx}, \frac{du}{dx} \right) - (f, w)}_{\frac{1}{2} \left(\frac{dw}{dx}, \frac{dw}{dx} \right)} + \frac{1}{2} \left(\frac{dw}{dx}, \frac{dw}{dx} \right) \\
&= \frac{1}{2} \left(\frac{du}{dx}, \frac{du}{dx} \right) - (f, u) + \frac{1}{2} \left(\frac{dw}{dx}, \frac{dw}{dx} \right) + \underbrace{\left(\frac{du}{dx}, \frac{dw}{dx} \right) - (f, w)}_{=0} \\
&= F(u) + \frac{1}{2} \left(\frac{dw}{dx}, \frac{dw}{dx} \right).
\end{aligned}$$

Finally, using **2**, we have that $F(v) \geq F(u)$ as required. \square

(II) $(V) \Leftarrow (M)$.

Let $u \in H_0^1(0, 1)$ be the solution of (M) , that is

$$F(u) \leq F(v) \quad \forall v \in H_0^1(0, 1).$$

Thus, given a function $v \in H_0^1(0, 1)$ and $\varepsilon \in \mathbb{R}$, $u + \varepsilon v \in H_0^1(0, 1)$, so that $F(u) \leq F(u + \varepsilon v)$. We now define $g(\varepsilon) := F(u + \varepsilon v)$. This function is minimised when $\varepsilon = 0$, so that we have that $\left. \frac{dg}{d\varepsilon} \right|_{\varepsilon=0} = 0$. Now,

$$\begin{aligned}
g(\varepsilon) &= \frac{1}{2} \left(\frac{d}{dx}(u + \varepsilon v), \frac{d}{dx}(u + \varepsilon v) \right) - (f, u + \varepsilon v) \\
&= \frac{1}{2} \varepsilon^2 \left(\frac{dv}{dx}, \frac{dv}{dx} \right) + \varepsilon \left(\frac{du}{dx}, \frac{dv}{dx} \right) - \varepsilon (f, v) \\
&\quad + \frac{1}{2} \left(\frac{du}{dx}, \frac{du}{dx} \right) - (f, u),
\end{aligned}$$

$$\text{and so, } \frac{dg}{d\varepsilon} = \varepsilon \left(\frac{dv}{dx}, \frac{dv}{dx} \right) + \left(\frac{du}{dx}, \frac{dv}{dx} \right) - (f, v).$$

Finally, setting $\left. \frac{dg}{d\varepsilon} \right|_{\varepsilon=0} = 0$ we see that u solves (V) .

\square

In summary, we have that $(D) \Rightarrow (V) \Leftrightarrow (M)$. The converse implication $(D) \Leftarrow (V)$ is not true unless $u \in H_0^1(0,1)$ is smooth enough to ensure that $u \in C^2(0,1)$. In this special case, we have that $(D) \Leftrightarrow (V)$.

Returning to (M) and the space $H_0^1(0,1)$, a very important property of $L_2(\Omega)$ is the concept of a “weak derivative”.

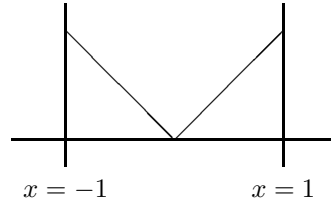
Definition x.8 (Weak derivative)

$u \in L_2(\Omega)$ possesses a weak derivative $\partial u \in L_2(\Omega)$ satisfying

$$(\phi, \partial u) = - \left(\frac{d\phi}{dx}, u \right) \quad \forall \phi \in C_0^\infty(\Omega), \quad (W-D)$$

where $C_0^\infty(\Omega)$ is the space of infinitely differentiable functions which are zero outside Ω .

Example x.8.1 Consider the function $u = |x|$ with $\Omega = (-1,1)$.



This function is not differentiable in the classical sense. However, starting from the right hand side of $W-D$, and integrating by parts gives

$$\begin{aligned} - \left(\frac{d\phi}{dx}, u \right) &= - \int_{-1}^1 |x| \frac{d\phi}{dx} \\ &= - \int_{-1}^0 (-x) \frac{d\phi}{dx} - \int_0^1 x \frac{d\phi}{dx} \\ &= \int_{-1}^0 \frac{d}{dx} (-x) \phi + \int_0^1 \frac{d}{dx} (x) \phi \\ &\quad - \underbrace{\left[(-x)\phi \right]_{-1}^0}_{0\phi(0) - 1\phi(-1)} - \underbrace{\left[(x)\phi \right]_0^1}_{1\phi(1) - 0\phi(0)} \\ &= \int_{-1}^0 \phi \left\{ \frac{d}{dx} (-x) \right\} + \int_0^1 \phi \left\{ \frac{d}{dx} (x) \right\} = (\phi, \partial u). \end{aligned}$$

Thus the weak derivative is a step function,

$$\partial u = \begin{cases} -1, & -1 < x < 0 \\ 1, & 0 < x < 1 \end{cases}.$$

Note that the value of ∂u is not defined at the origin. \square

Example x.8.2 Consider the step function $u(x) = H(1/2)$. Constructing the weak derivative using W - D gives

$$-\left(\frac{d\phi}{dx}, u\right) = -\underbrace{\int_0^{\frac{1}{2}} (0) \frac{d\phi}{dx}}_{=0} - \int_{\frac{1}{2}}^1 (1) \frac{d\phi}{dx}.$$

Integrating by parts then gives

$$-\left(\frac{d\phi}{dx}, u\right) = \underbrace{\int_{\frac{1}{2}}^1 \phi \frac{d}{dx}(1)}_0 - \underbrace{[1\phi]_{\frac{1}{2}}^1}_{\phi(1) - \phi(1/2)}.$$

Finally, since $\phi(1) = 0$, we conclude that

$$-\left(\frac{d\phi}{dx}, u\right) = \phi\left(\frac{1}{2}\right) = (\phi, \partial u).$$

Thus, if we relax the requirement that $\partial u \in L_2(\Omega)$, we see that the weak derivative of a step function is a delta function. \square

2. Galerkin Approximation

We now introduce a finite dimensional subspace $V_k \subset H_0^1(0, 1)$. This is associated with a set of basis functions

$$V_k = \text{span} \{\phi_1(x), \phi_2(x), \dots, \phi_k(x)\}$$

so that every element of V_k , say u_k , can be uniquely written as

$$u_k = \sum_{j=1}^k \alpha_j \phi_j, \quad \alpha_j \in \mathbb{R}.$$

To compute the Galerkin approximation, we pose the variational problem (V) over V_k . That is, we seek $u_k \in V_k$ such that

$$\left(\frac{du_k}{dx}, \frac{dv_k}{dx}\right) = (f, v_k) \quad \forall v_k \in V_k. \quad (V_h)$$

Equivalently, since $\{\phi_i\}_{i=1}^k$ are a basis set, we have that

$$\begin{aligned} \left(\frac{du_k}{dx}, \frac{d\phi_i}{dx} \right) &= (f, \phi_i), \quad i = 1, 2, \dots, k \\ \left(\frac{d}{dx} \left(\sum_{j=1}^k \alpha_j \phi_j \right), \frac{d\phi_i}{dx} \right) &= (f, \phi_i), \\ \sum_{j=1}^k \alpha_j \left(\frac{d\phi_j}{dx}, \frac{d\phi_i}{dx} \right) &= (f, \phi_i). \end{aligned}$$

This can be written in matrix form as

$$\mathbf{Ax} = \mathbf{f} \quad (V'_h)$$

$$\begin{aligned} \text{with } A_{ij} &= \left(\frac{d\phi_j}{dx}, \frac{d\phi_i}{dx} \right), \quad i, j = 1, \dots, k; \\ x_j &= \alpha_j, \quad j = 1, \dots, k; \\ \text{and } f_i &= (f, \phi_i), \quad i = 1, \dots, k. \end{aligned}$$

(V'_h) is called the Galerkin system, A is called the “stiffness matrix”, \mathbf{f} is the “load vector” and $u_k = \sum_{j=1}^k \alpha_j \phi_j$ is the “Galerkin solution”.

Theorem 2.1 *The stiffness matrix is symmetric and positive definite.*

Proof. Symmetry follows from **1**. To establish positive definiteness we consider the quadratic form and use **4**:

$$\begin{aligned} \mathbf{x}^T \mathbf{Ax} &= \sum_{j=1}^k \sum_{i=1}^k \alpha_j A_{ji} \alpha_i \\ &= \sum_{j=1}^k \sum_{i=1}^k \alpha_j \left(\frac{d\phi_j}{dx}, \frac{d\phi_i}{dx} \right) \alpha_i \\ &= \left(\sum_{j=1}^k \alpha_j \frac{d\phi_j}{dx}, \sum_{i=1}^k \alpha_i \frac{d\phi_i}{dx} \right) \\ &= \left(\frac{du_k}{dx}, \frac{du_k}{dx} \right). \end{aligned}$$

Thus from **2** we see that A is at least semi-definite. Definiteness follows from the fact that $\mathbf{x}^T \mathbf{Ax} = 0$ if and only if $du_k/dx = 0$. But since $u_k \in H_0^1(0, 1)$ then $du_k/dx = 0$ implies that $u_k = 0$. Finally, since $\{\phi_i\}_{i=1}^k$ are a basis set, we have that $u_k = 0$ implies that $\mathbf{x} = \mathbf{0}$. \square

Theorem 2.1 implies that A is nonsingular. This means that the solution \mathbf{x} (and hence u_k) exists and is unique.

An alternative approach, the so-called Rayleigh–Ritz method, is obtained by posing the minimization problem (M) over the finite dimensional subspace V_k . That is, we seek $u_k \in V_k$ such that

$$F(u_k) \leq F(v_k) \quad \forall v_k \in V_k. \quad (M_h)$$

Doing this leads to the matrix system (V'_h) so that the Ritz solution and the Galerkin solution are one and the same.

The beauty of Galerkin's method is the “best approximation” property.

Theorem 2.2 (Best approximation)

If u is the solution of (V) and u_k is the Galerkin solution, then

$$\left\| \frac{d}{dx}(u - u_k) \right\| \leq \left\| \frac{d}{dx}(u - v_k) \right\| \quad \forall v_k \in V_k. \quad (B-A)$$

Proof. The functions u_k and u satisfy the following

$$\begin{aligned} u_k \in V_k; \quad \left(\frac{du_k}{dx}, \frac{dv_k}{dx} \right) &= (f, v_k) \quad \forall v_k \in V_k \\ u \in V; \quad \left(\frac{du}{dx}, \frac{dv}{dx} \right) &= (f, v) \quad \forall v \in V. \end{aligned}$$

But since $V_k \subset V$ we have that

$$\left(\frac{du}{dx}, \frac{dv_k}{dx} \right) = (f, v_k) \quad \forall v_k \in V_k.$$

Subtracting equations and using \bullet gives

$$\begin{aligned} \left(\frac{du}{dx}, \frac{dv_k}{dx} \right) - \left(\frac{du_k}{dx}, \frac{dv_k}{dx} \right) &= 0 \quad \forall v_k \in V_k \\ \left(\frac{du}{dx} - \frac{du_k}{dx}, \frac{dv_k}{dx} \right) &= 0 \\ \left(\frac{d}{dx}(u - u_k), \frac{dv_k}{dx} \right) &= 0 \quad \forall v_k \in V_k \quad (G-O) \end{aligned}$$

This means that the error $u - u_k$ is “orthogonal” to the subspace V_k —a property known as **Galerkin orthogonality**. To establish the best approximation property we start with the left hand side of $B-A$ and use Galerkin

orthogonality as follows:

$$\begin{aligned}
 \left\| \frac{d}{dx}(u - u_k) \right\|^2 &= \left(\frac{d}{dx}(u - u_k), \frac{d}{dx}(u - u_k) \right) \\
 &= \left(\frac{d}{dx}(u - u_k), \frac{du}{dx} \right) - \underbrace{\left(\frac{d}{dx}(u - u_k), \frac{du_k}{dx} \right)}_{=0 \quad G-O} \quad u_k \in V_k \\
 &= \left(\frac{d}{dx}(u - u_k), \frac{du}{dx} \right) - \underbrace{\left(\frac{d}{dx}(u - u_k), \frac{dv_k}{dx} \right)}_{=0 \quad G-O} \quad v_k \in V_k \\
 &= \left(\frac{d}{dx}(u - u_k), \frac{d}{dx}(u - v_k) \right) \\
 &\leq \left\| \frac{d}{dx}(u - u_k) \right\| \left\| \frac{d}{dx}(u - v_k) \right\|. \quad \text{using } C-S
 \end{aligned}$$

Hence, dividing by $\left\| \frac{d}{dx}(u - u_k) \right\| > 0$ ¹

$$\left\| \frac{d}{dx}(u - u_k) \right\| \leq \left\| \frac{d}{dx}(u - v_k) \right\| \quad \forall v_k \in V_k,$$

as required. \square

An important observation here is that we have a natural norm to measure errors—which is inherited from the minimization problem (M).

Example x.3.4 Suppose $V = H_0^1(0, 1)$. A valid norm is

$$\|v\|_E = \left(\frac{dv}{dx}, \frac{dv}{dx} \right)^{1/2} = \left\| \frac{dv}{dx} \right\|.$$

This is called the **energy norm**. \heartsuit

A technical issue that arises here is that the best approximation property does not automatically imply that the Galerkin method converges in the sense that

$$\|u - u_k\|_E \rightarrow 0 \quad \text{as } k \rightarrow \infty.$$

For convergence, we really need to introduce the concept of a “complete” space that was postponed earlier.

Definition x.9 (Cauchy sequence)

A sequence $(v^{(k)}) \in V$ is called a **Cauchy sequence** in a normed space V if for any $\varepsilon > 0$, there exists a positive integer $k_0(\varepsilon)$ such that

$$\|v^{(\ell)} - v^{(m)}\|_V < \varepsilon \quad \forall \ell, m \geq k_0.$$

¹If $\left\| \frac{d}{dx}(u - u_k) \right\| = 0$ then $B-A$ holds trivially.

A Cauchy sequence is convergent, so the only issue is whether the limit of the sequence is in the “correct space”. This motivates the following definition.

Definition x.10 (Complete space)

A normed space V is **complete** if it contains the limits of all Cauchy sequences in V . That is, if $(v^{(k)})$ is a Cauchy sequence in V , then there exists $\xi \in V$ such that

$$\lim_{k \rightarrow \infty} \|v^{(k)} - \xi\|_V = 0.$$

We write this as $\lim_{k \rightarrow \infty} v^{(k)} = \xi$.

Example x.10.1 The space $H_0^1(0, 1)$ is complete with respect to the energy norm $\|\cdot\|_E$. The upshot is that the Galerkin approximation is guaranteed to converge to the weak solution in the limit $k \rightarrow \infty$. \heartsuit

We now introduce a simple-minded Galerkin approximation based on “global” polynomials. That is, we choose

$$V_k = \text{span} \{1, x, x^2, \dots, x^{k-1}\}.$$

To ensure that $V_k \subset V$, the function $u_k = \sum_j \alpha_j x^{j-1}$ must satisfy two conditions:

(I) $u_k \in H^1(0, 1)$;

(II) $u_k(0) = 0 = u_k(1)$.

The first condition is no problem, $u_k \in C^\infty(0, 1)$! To satisfy the second condition we need to modify the basis set to the following,

$$V_k^* = \text{span} \{x(x-1), x^2(x-1), x^3(x-1), \dots, x^k(x-1)\}.$$

Problem 2.1 ($f = 1$)

Consider

$$V_2^* = \text{span} \left\{ \underbrace{x(x-1)}_{\phi_1}, \underbrace{x^2(x-1)}_{\phi_2} \right\}$$

Then constructing the matrix system (V'_h) , we have that

$$\begin{aligned} A_{11} &= \int_0^1 \left(\frac{d\phi_1}{dx} \right)^2 = \int_0^1 \left(\frac{d}{dx}(x^2 - x) \right)^2 = \frac{1}{3} \\ A_{12} &= \int_0^1 \frac{d\phi_2}{dx} \frac{d\phi_1}{dx} = \int_0^1 \frac{d}{dx}(x^2 - x) \frac{d}{dx}(x^3 - x^2) = \frac{1}{6} = A_{21} \\ A_{22} &= \int_0^1 \left(\frac{d\phi_2}{dx} \right)^2 = \int_0^1 \left(\frac{d}{dx}(x^3 - x^2) \right)^2 = \frac{2}{15} \\ f_1 &= \int_0^1 \phi_1 = \int_0^1 x^2 - x = -\frac{1}{6} \\ f_2 &= \int_0^1 \phi_2 = \int_0^1 x^3 - x^2 = -\frac{1}{12}. \end{aligned}$$

This gives

$$\begin{pmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{2}{15} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} -\frac{1}{6} \\ -\frac{1}{12} \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} \\ 0 \end{pmatrix}.$$

So the Galerkin solution is

$$u_2(x) = -\frac{1}{2}\phi_1 + 0\phi_2 = \frac{1}{2}(x - x^2) = u(x).$$

The fact that the Galerkin approximation agrees with the exact solution is to be expected given the best approximation property and noting that $u \in V_2^*$.

Problem 2.2 ($f(x) = H(1/2)$; where $H(x)$ is the “unit step” function)

Consider V_2^* as above. In this case,

$$\begin{aligned} f_1 &= \int_{\frac{1}{2}}^1 \phi_1 = -\frac{1}{12} \\ f_2 &= \int_{\frac{1}{2}}^1 \phi_2 = \int_0^1 x^3 - x^2 = -\frac{5}{192}. \end{aligned}$$

This gives the Galerkin system

$$\begin{pmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{2}{15} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} -\frac{1}{12} \\ -\frac{5}{192} \end{pmatrix}.$$

Note that the Galerkin solution is not exact in this case.

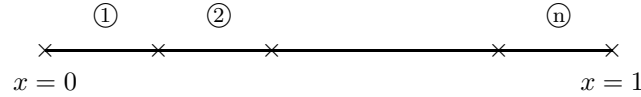
The big problem with global polynomial approximation is that the Galerkin matrix becomes increasingly ill conditioned as k is increased. Computationally, it behaves like a Hilbert matrix and so reliable computation for $k > 10$ is not possible. For this reason, **piecewise polynomial** basis functions are used in practice instead of global polynomial functions.

3. Finite Element Galerkin Approximation

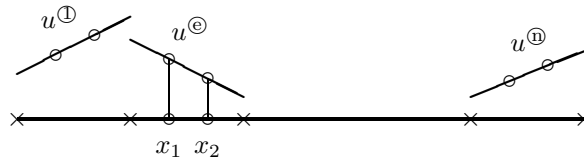
A piecewise polynomial approximation space can be constructed in four steps.

Step (i) Subdivision of $\bar{\Omega}$ into “elements”.

For $\bar{\Omega} = [0, 1]$ the elements are intervals as illustrated below.



Step (ii) Piecewise approximation of u using a low-order polynomial (e.g. linear):



In general, a linear function $u^{\textcircled{e}}(x)$ is defined by its values at two distinct points $x_1 \neq x_2$

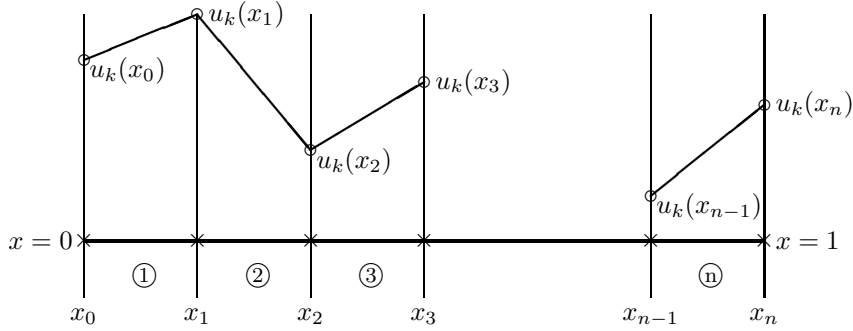
$$u^{\textcircled{e}}(x) = \underbrace{\frac{(x - x_2)}{(x_1 - x_2)}}_{\ell_1^{\textcircled{e}}(x)} u^{\textcircled{e}}(x_1) + \underbrace{\frac{(x - x_1)}{(x_2 - x_1)}}_{\ell_2^{\textcircled{e}}(x)} u^{\textcircled{e}}(x_2).$$

$\ell_i(x)$ are called “nodal” basis functions and satisfy the interpolation conditions

$$\ell_1^{\textcircled{e}}(x) = \begin{cases} \text{linear over } \textcircled{e} \\ 1 & \text{if } x = x_1 \\ 0 & \text{if } x = x_2 \end{cases} ; \quad \ell_2^{\textcircled{e}}(x) = \begin{cases} \text{linear over } \textcircled{e} \\ 1 & \text{if } x = x_2 \\ 0 & \text{if } x = x_1 \end{cases} .$$

Step (iii) Satisfaction of the smoothness requirement (I).

This is done by carefully positioning the nodes, so that x_1 and x_2 are at the end points of the interval.

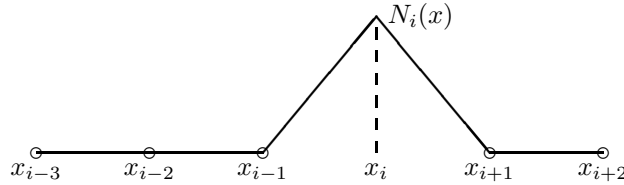


Concatenating the element functions $u^\ominus(x)$ gives a global function $u_k(x)$

$$u_k(x) = \bigwedge_{\ominus=1}^n u^\ominus(x).$$

Thus $u_k(x)$ is defined by $k = n - 1$ internal values, and the two boundary values $u_k(x_0) = u_k(0)$ and $u_k(x_n) = u_k(1)$. We can then define $N_i(x)$, the so called “global” basis function, so that

$$N_i(x) = \begin{cases} \text{linear over } [0, 1] \\ 1 \text{ if } x = x_i \\ 0 \text{ if } x = x_j \text{ (} j \neq i \text{)} \end{cases},$$



and write

$$u_k(x) = \alpha_0 N_0(x) + \alpha_1 N_1(x) + \dots + \alpha_n N_n(x).$$

Note that $u_k(x_i) = \alpha_i$, so that the “unknowns” are the function values at the nodes.

Step (iv) Satisfaction of the essential boundary condition requirement (II). This is easy—we simply remove the basis functions $N_0(x)$ and $N_n(x)$ from the basis set. The modified Galerkin approximation is

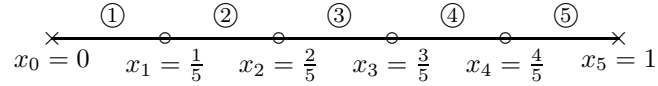
$$u_k^*(x) = \alpha_1 N_1(x) + \dots + \alpha_{n-1} N_{n-1}(x),$$

and is associated with the approximation space

$$V_{n-1}^* = \text{span} \{N_1(x), \dots, N_{n-1}(x)\}.$$

Problem 3.1 ($f(x) = 1$)

Consider five equal length elements



so that

$$V_6 = \text{span} \{N_i(x)\}_{i=0}^5.$$

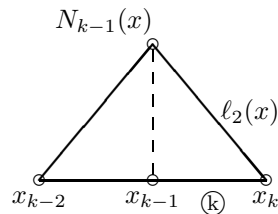
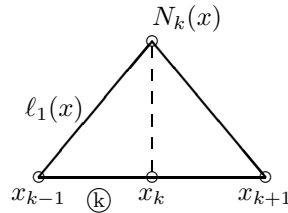
For computational convenience the Galerkin system coefficients

$$A_{ij} = \int_0^1 \frac{dN_j}{dx} \frac{dN_i}{dx}, \quad b_i = \int_0^1 N_i,$$

may be computed element-by-element. That is,

$$A_{ij} = \sum_{\mathbb{K}=1}^5 \underbrace{\int_{x_{k-1}}^{x_k} \frac{dN_j}{dx} \frac{dN_i}{dx}}_{\int_{x_{k-1}}^{x_k} \frac{d\ell_s}{dx} \frac{d\ell_t}{dx}}, \quad b_i = \sum_{\mathbb{K}=1}^5 \underbrace{\int_{x_{k-1}}^{x_k} N_i}_{\int_{x_{k-1}}^{x_k} \ell_t},$$

where s and t are local indices referring to the associated element basis functions illustrated below.



In particular, in element \textcircled{k} there are two nodal basis functions

$$\ell_1(x) = \frac{(x - x_{k-1})}{(x_k - x_{k-1})}, \quad \ell_2(x) = \frac{(x - x_k)}{(x_{k-1} - x_k)},$$

so that

$$\frac{d\ell_1}{dx} = \frac{1}{h}, \quad \frac{d\ell_2}{dx} = -\frac{1}{h},$$

with $h = x_k - x_{k-1} = 1/5$. This generates a 2×2 “element contribution” matrix

$$A^{\textcircled{k}} = \begin{bmatrix} \int_{x_{k-1}}^{x_k} \left(\frac{d\ell_1}{dx}\right)^2 & \int_{x_{k-1}}^{x_k} \left(\frac{d\ell_2}{dx}\right)\left(\frac{d\ell_1}{dx}\right) \\ \int_{x_{k-1}}^{x_k} \left(\frac{d\ell_1}{dx}\right)\left(\frac{d\ell_2}{dx}\right) & \int_{x_{k-1}}^{x_k} \left(\frac{d\ell_2}{dx}\right)^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{h} & -\frac{1}{h} \\ -\frac{1}{h} & \frac{1}{h} \end{bmatrix}$$

and a (2×1) “element contribution” vector

$$b^{\textcircled{k}} = \begin{bmatrix} \int_{x_{k-1}}^{x_k} \ell_1 \\ \int_{x_{k-1}}^{x_k} \ell_2 \end{bmatrix} = \begin{bmatrix} \frac{h}{2} \\ \frac{h}{2} \end{bmatrix}.$$

Thus, the “assembled ” Galerkin system is

$$\begin{bmatrix} \frac{1}{h} & -\frac{1}{h} & 0 & 0 & 0 & 0 \\ -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & 0 & 0 & 0 \\ 0 & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & 0 & 0 \\ 0 & 0 & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & 0 \\ 0 & 0 & 0 & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} \\ 0 & 0 & 0 & 0 & -\frac{1}{h} & \frac{1}{h} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{bmatrix} = \begin{bmatrix} \frac{h}{2} \\ h \\ h \\ h \\ h \\ \frac{h}{2} \end{bmatrix}.$$

$A \qquad x \qquad b$

Note that this system is singular (and inconsistent!). All the columns sum to zero, so that $A\mathbf{1} = \mathbf{0}$. This problem arises because we have not yet imposed the essential boundary conditions. To do this we simply need to remove $N_0(x)$ and $N_5(x)$ from the basis set, that is, delete the first and last row and column from the system. Doing this gives the nonsingular reduced system

$$\begin{bmatrix} \frac{2}{h} & -\frac{1}{h} & 0 & 0 \\ -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & 0 \\ 0 & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} \\ 0 & 0 & -\frac{1}{h} & \frac{2}{h} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} = \begin{bmatrix} h \\ h \\ h \\ h \end{bmatrix}.$$

$A^* \qquad x^* \qquad b^*$

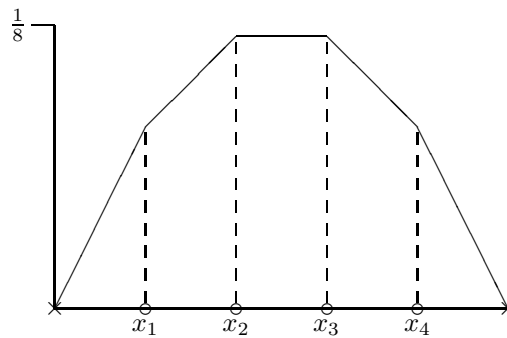
Setting $h = 1/5$, and solving gives

$$\alpha_1 = \alpha_4 = 0.08; \quad \alpha_2 = \alpha_3 = 0.12;$$

so the Galerkin finite element solution is

$$u_4^*(x) = 0.08N_1(x) + 0.12N_2(x) + 0.12N_3(x) + 0.08N_4(x).$$

It is illustrated below. Note that $u_4^*(x_i) = u(x_i)$ so the finite element solution is exact at the nodes! (It is not exact at any point between the nodes though.)



Note also that the generic equation

$$-\frac{1}{h} \xrightarrow{x_{i-1}} \frac{2}{h} \xrightarrow{x_i} -\frac{1}{h} \xrightarrow{x_{i+1}} = h$$

corresponds to a centered finite difference approximation to $-\frac{d^2 u}{dx^2} = 1$.

To complete the discussion we would like to show that the finite element solution converges to the weak solution in the limit $h \rightarrow 0$.

Theorem 3.1 (*Convergence in the energy norm*)

If u is the solution of (V) and u_k is the finite element solution based on linear approximation then

$$\|u - u_k\|_E \leq h \|f\|$$

where h is the length of the longest element in the subdivision (which does not have to be uniform.)

Proof. We now formally introduce the linear interpolant, $u^* \in V_k$ of the exact solution, so that

$$u(x_i) = u^*(x_i) \quad i = 0, 1, 2, \dots, n.$$

Note that we cannot assume that $u_k = u^*$ in general. Introducing $e(x) = u(x) - u^*(x)$, we see that $e \in V_k$ and that

$$e(x_i) = 0, \quad i = 0, 1, 2, \dots, n.$$

We can now bound the element interpolation error using standard tools from approximation theory

$$\begin{aligned} \int_{x_{i-1}}^{x_i} \left(\frac{de}{dx} \right)^2 &\leq (x_i - x_{i-1})^2 \int_{x_{i-1}}^{x_i} \left(\frac{d^2 e}{dx^2} \right)^2 \\ &= (x_i - x_{i-1})^2 \int_{x_{i-1}}^{x_i} \left(\frac{d^2 u}{dx^2} \right)^2 \\ &\leq h^2 \int_{x_{i-1}}^{x_i} \left(\frac{d^2 u}{dx^2} \right)^2 \end{aligned}$$

where $h = \max_i |x_i - x_{i-1}|$. Summing over the intervals then gives the estimate

$$\int_0^1 \left(\frac{de}{dx} \right)^2 \leq h^2 \int_0^1 \left(\frac{d^2 u}{dx^2} \right)^2.$$

Finally, using B - A

$$\left\| \frac{d}{dx}(u - u_k) \right\| \leq \left\| \frac{d}{dx}(u - u^*) \right\| \leq h \left\| \frac{d^2 u}{dx^2} \right\| = h \|f\| < \infty.$$

Thus $\lim_{h \rightarrow 0} u_k = u$ in the energy norm. \square

To get an error estimate in L_2 we use a very clever “duality argument”.

Theorem 3.2 (*Aubin-Nitsche*)

If u is the solution of (V) and u_k is the finite element solution based on linear approximation then

$$\|u - u_k\| \leq h^2 \|f\|.$$

Proof. Let w be the solution of the dual problem

$$-\frac{d^2 w}{dx^2} = u - u_k \quad x \in (0, 1); \quad w(0) = 0 = w(1).$$

Then we have

$$\begin{aligned} \|u - u_k\|^2 &= (u - u_k, u - u_k) \\ &= (u - u_k, -\frac{d^2 w}{dx^2}) \\ &= \left(\frac{d}{dx}(u - u_k), \frac{dw}{dx} \right) \quad (\text{since } w(0) = w(1) = 0) \\ &= \left(\frac{d}{dx}(u - u_k), \frac{dw}{dx} \right) - \left(\frac{d}{dx}(u - u_k), \frac{dw^*}{dx} \right), \quad w^* \in V_k, \quad G-O \end{aligned}$$

where w^* is the interpolant of w in V_k . Hence

$$\begin{aligned} \|u - u_k\|^2 &= \left(\frac{d}{dx}(u - u_k), \frac{d}{dx}(w - w^*) \right) \\ &\leq \left\| \frac{d}{dx}(u - u_k) \right\| \underbrace{\left\| \frac{d}{dx}(w - w^*) \right\|}_{C-S} \\ &\leq h \left\| \frac{d^2 w}{dx^2} \right\| = h \|u - u_k\|. \end{aligned}$$

Hence, assuming that $\|u - u_k\| > 0$, we have that

$$\|u - u_k\| \leq h \left\| \frac{d}{dx}(u - u_k) \right\| \leq h^2 \|f\|,$$

as required. \square

A more complete discussion of these issues can be found in Chapters 11 and 14 of the following reference book.

- ENDRE SÜLI & DAVID MAYERS, *An Introduction to Numerical Analysis*, Cambridge University Press, 2003.