3.20 Let's revisit Example 3.17 and set $b(x) = c(x) = 0$ and $k(x) \geq k_0 > 0$, for all x, so the bilinear form is

$$a(u,v) = \int_0^L k(x)u'(x)v'(x)dx.$$

In this case, the proof of coercivity in Example 3.17 breaks down, since, following the argument, $c_0 = 0$ and hence the coercivity constant would be zero as well. In fact, the bilinear form is no longer coercive in $\mathscr{W} = H^1([0,L])$. For example, a constant function u different than zero is in $\mathscr{W}$, but $a(u,u) = 0$.

However, what matters for Problem 1.3 is coercivity on $\mathscr{V}$, not on $\mathscr{W}$; specifically, on

$$\mathscr{V} = \{v \in \mathscr{W} \mid w(0) = 0\}.$$

Coercivity on $\mathscr{V}$ follows because the boundary condition at $x = 0$ enables the of use Poincaré's inequality. To wit, for $u \in \mathscr{V}$,

$$\begin{aligned} a(u,u) &= \int_0^L k(x)u'(x)^2\,dx \\ &\geq k_0 \int_0^L u'(x)^2\,dx \\ &= k_0 |u|_{1,2}^2 \\ &\geq \frac{k_0}{2}|u|_{1,2}^2 + \frac{1}{C_p^2}\frac{k_0}{2}\|u\|_{0,2}^2 \qquad \text{Poincaré's inequality, (3.30)} \\ &\geq \frac{k_0}{2}\min\{1, C_p^{-2}\}\|u\|_{1,2}^2. \end{aligned}$$

3.21 We revisit now Example 3.15, which involved a particular case of the the diffusion problem in two-dimension, Problem 2.2. Example 3.15 showed that the bilinear form and the linear functional are continuous if $\mathscr{W} = H^1(\Omega)$. However, a constant function $u \neq 0$ is in $\mathscr{W}$, and since its gradient is zero, $a(u,u) = 0$. This shows that the bilinear form is not coercive on $\mathscr{W}$.

Instead, consider the test space in Problem 2.2,

$$\mathscr{V} = \{v \in \mathscr{W} \mid v(x) = 0 \text{ for all } x \in \partial\Omega_D\},$$

where the length of $\partial\Omega_D$ is assumed to be positive (Hypothesis **H1** in Problem 2.1). In this case, we can apply Poincaré's inequality, and obtain coercivity on $\mathscr{V}$. We will also use that K is positive definite with a smallest eigenvalue greater or equal than $\kappa_0 > 0$ (Hypothesis **H2** in Problem 2.1), to get

$$(K\nabla u)\cdot\nabla u = \sum_{i,j=1}^{2} K_{ij}\frac{\partial u}{\partial x_j}\frac{\partial u}{\partial u_i} \geq \kappa_0\|\nabla u\|^2. \tag{3.31}$$

We can now show the coercivity on $\mathcal{V}$: For $u \in \mathcal{V}$,

$$\begin{aligned} a(u,u) &= \int_\Omega (K\nabla u)\cdot \nabla u \, dx \\ &\geq \kappa_0 \int_\Omega \|\nabla u\|^2 \, d\Omega && \text{from (3.31)} \\ &= \frac{\kappa_0}{2}|u|_{1,2}^2 + \frac{\kappa_0}{2}|u|_{1,2}^2 && \text{split to use Poincaré next} \\ &\geq \frac{\kappa_0}{2}|\nabla u|_{1,2}^2 + \frac{\kappa_0}{2}\frac{1}{C_p^2}\|u\|_{0,2}^2 && \text{Poincaré's inequality, (3.30)} \\ &\geq \frac{\kappa_0}{2}\min\{1, C_p^{-2}\}\|u\|_{1,2}^2. \end{aligned}$$

Notice how the two hypotheses in Problem 2.1 play a crucial role in establishing the coercivity of the bilinear form in $\mathcal{V}$, and hence in the uniqueness of solutions of the weak form.

Invertibility of the stiffness matrix. We now establish that if $a(\cdot,\cdot)$ is coercive on $\mathcal{V}_h$, then the stiffness matrix K is invertible, and hence the solution to Galerkin Method exists and is unique.

To see this, remember that the solution $u_h \in \mathcal{S}_h$ of Galerkin Method satisfies that

$$a(u_h, v_h) = \ell(v_h) \qquad \forall v_h \in \mathcal{V}_h \Longleftrightarrow KU = F,$$

where K is the stiffness matrix, F is the load vector, and U is the column vector of components of u_h in the chosen basis. The matrix K maps vectors in $\mathbb{R}^m$ to vectors in $\mathbb{R}^m$, where m is the dimension of $\mathcal{W}_h$. To prove that K is invertible, we prove that the null space of K (or the kernel of K) contains only the zero vector. From elementary linear algebra results (the Rank-Nullity theorem, e.g. [1, 6, 7]), this implies that the image of K is $\mathbb{R}^m$, and hence that K is invertible and that the solution U, and hence u_h, exists and is unique.

The null space of K is the set $\mathcal{N}(K) = \{U \in \mathbb{R}^m \mid KU = 0\}$. Consider $U \in \mathcal{N}(K)$; we will prove that $U = 0$, and hence that $\mathcal{N}(K) = \{0\}$, as sought. Because $K_{ab} = \delta_{ab}$ if $a \in \eta_g$, we can conclude that $U_a = 0$ if $a \in \eta_g$. This implies that U defines a function $u_h \in \mathcal{V}_h$, since $\overline{u}_h = \sum_{a\in\eta_g} U_a N_a + v_h = v_h$, where $\{N_a\}$ are the basis functions for $\mathcal{W}_h$, and $v_h \in \mathcal{V}_h$. Additionally, since $KU = 0$, the function u_h satisfies that

$$a(u_h, v_h) = 0 \qquad \forall v_h \in \mathcal{V}_h,$$

and, in particular, that $0 = a(u_h, u_h)$. From the coercivity of $a(\cdot,\cdot)$ on $\mathcal{V}_h$,

$$0 \leq c_{\mathcal{V}_h}\|u_h\|_{\mathcal{V}_h}^2 \leq a(u_h, u_h) = 0,$$

from where it follows that $u_h = 0$, and hence $U = 0$.

Energy Norm. In addition to guaranteeing uniqueness, a coercive bilinear form has another interesting feature: it defines a natural norm for the problem, the energy norm.

Definition 3.6 (Energy norm)**.** *Let* $(V, \|\cdot\|_V)$ *be a normed space, and let* $a\colon V \times V \to \mathbb{R}$ *be a coercive bilinear form on* V*. The* **energy norm** *of* $a(\cdot,\cdot)$ *is defined for any* $u \in V$ *as*

$$\|v\|_a = \sqrt{a(u,u)}. \tag{3.32}$$

It can be verified that $\|\cdot\|_a$ indeed satisfies Definition 3.1, but we shall skip the proof. The denomination of energy norm originates in the interpretation of $a(u,u)/2$ as the elastic energy (or strain energy) in structural mechanics problems, as we shall have the opportunity to see when we discuss linear elasticity.

3.2.4 Interpolation Errors

We now turn our attention to the last ingredient in our path to show convergence: We need to show that any potential solution of our problem can be approximated to any degree of accuracy by an appropriately selected finite element space. Under mild conditions on the mesh and the smoothness of the function, a finite element space will have this property. Before stating a general result, let us take a look at a simple example.

For the forthcoming discussion we will slightly modify the notation for norms to include the domain over which the norm is computed. Precisely, for a domain $\Omega \subset \mathbb{R}^n$ and any function $f\colon \Omega \to \mathbb{R}$ that takes values over it we will denote

$$\|f\|_{0,2,\Omega} = \left[\int_\Omega f(x)^2\, dx\right]^{1/2}, \qquad |f|_{1,2,\Omega} = \left[\int_\Omega \|\nabla f(x)\|^2\, dx\right]^{1/2}. \tag{3.33}$$

Example 3.22 Consider a one-dimensional finite element mesh of P^1-elements over a domain $\Omega = [0,L]$, and a smooth function $u\colon [0,L] \to \mathbb{R}$. The mesh has n_{el} equal elements with mesh size $h = L/n_{\text{el}}$ and domain $\Omega_e = [x_{e-1}, x_e]$ for $e = 1, \ldots, n_{\text{el}}$, and hence $m = n_{\text{el}} + 1$ nodes at $x_a = (a-1)h$ for $a \in \{1, \ldots, m\}$. Let $\{N_1, \ldots, N_m\}$ be the hat functions over this mesh. We define the interpolant $\mathscr{I} u$ of u as

$$\mathscr{I} u = \sum_{a=1}^{m} u(x_a) N_a. \tag{3.34}$$

We would like to estimate an upper bound for the errors $\|u - \mathscr{I} u\|_{0,2}$ and $|u - \mathscr{I} u|_{1,2}$ over $[0,L]$. To this end, we will split the error as a sum of errors over each element, namely,

$$\|u - u_h\|_{0,2,\Omega}^2 = \int_0^L (u - u_h)^2\, dx = \sum_{e=1}^{n_{\text{el}}} \int_{x_e}^{x_{e+1}} (u - u_h)^2\, dx = \sum_{e=1}^{n_{\text{el}}} \|u - u_h\|_{0,2,\Omega_e}^2,$$

and similarly,

$$|u-u_h|^2_{1,2,\Omega}=\sum_{e=1}^{n_{\mathrm{el}}}|u-u_h|^2_{1,2,\Omega_e}.$$

Errors over Ω can then be computed by first estimating errors over each element.

Let's see then how the error over an element can be obtained. Over element e the error function between u and $\mathscr{I}u$,

$$g(x)=u(x)-\mathscr{I}u(x)=u(x)-\left[u(x_e)N_1^e(x)+u(x_{e+1})N_2^e(x)\right],$$

satisfies that $g(x_e)=g(x_{e+1})=0$, its derivative is

$$g'(x)=u'(x)-\mathscr{I}u'(x)=u'(x)-\frac{u(x_{e+1})-u(x_e)}{h},$$

and its second derivative coincides with that of u, namely,

$$g''(x)=u''(x). \tag{3.35}$$

Since u is smooth, by Rolle's theorem there exists $\xi\in(x_e,x_{e+1})$ such that

$$g'(\xi)=0.$$

Performing a Taylor expansion of g and g' at ξ we obtain that for $x\in\Omega_e$

$$\begin{aligned}g'(x)&=g''(\zeta_x)(x-\xi)\\ g(x)&=g(\xi)+\frac{1}{2}g''(\zeta_x)(x-\xi)^2\end{aligned}$$

for some ζ_x between x and ξ that depends on x, and that may not be the same in the two identities above. An estimate for the H^1-seminorm of the error follows, namely,

$$\begin{aligned}|u-\mathscr{I}u|^2_{1,2,\Omega_e}&=\|g'(x)\|^2_{0,2,\Omega_e}\\ &=\int_{x_e}^{x_{e+1}}g''(\zeta_x)^2(x-\xi)^2\,dx\\ &\le\max_{x\in\Omega_e}|g''(x)|^2\int_{x_e}^{x_{e+1}}h^2\,dx\\ &=\max_{x\in\Omega_e}|u''(x)|^2h^3 &&\text{from (3.35).}\end{aligned}$$

A similar process leads to the estimate for the L^2-norm. To this end, we will need that for $a,b\in\mathbb{R}$

$$0\le(a-b)^2\Longrightarrow 2ab\le a^2+b^2, \tag{3.36}$$

and that $g(a)=0$ implies that

$$0=g(\xi)+\frac{1}{2}g''(\zeta_a)(a-\xi)^2\Longrightarrow 4|g(\xi)|^2\le\max_{x\in\Omega_e}|g''(x)|^2h^4. \tag{3.37}$$

Then,

$$\begin{aligned}
\|u-\mathscr{I}u\|^2_{0,2,\Omega_e} &= \|g(x)\|^2_{0,2,\Omega_e} \\
&= \int_{x_e}^{x_{e+1}} \left[g(\xi) + \frac{1}{2} g''(\zeta_x)(x-\xi)^2 \right]^2 \, dx \\
&\leq \int_{x_e}^{x_{e+1}} 2g(\xi)^2 + \frac{1}{2} g''(\zeta_x)^2 (x-\xi)^4 \, dx && \text{from (3.36)} \\
&\leq \frac{1}{2} \max_{x\in\Omega_e} |g''(x)|^2 \int_{x_e}^{x_{e+1}} h^4 \, dx + \frac{1}{2} \max_{x\in\Omega_e} |g''(x)|^2 \int_{x_e}^{x_{e+1}} h^4 \, dx && \text{from (3.37)} \\
&\leq \max_{x\in\Omega_e} |u''(x)|^2 h^5 && \text{from (3.35).}
\end{aligned}$$

Now that we have the two errors over the elements, we can add up over all elements to obtain errors over the entire domain, using that $n_{\text{el}} = L/h$:

$$\begin{aligned}
\|u-\mathscr{I}u\|^2_{0,2,\Omega} &= \sum_{e=1}^{n_{\text{el}}} \|u-\mathscr{I}u\|^2_{0,2,\Omega^e} \\
&\leq \sum_{e=1}^{n_{\text{el}}} \max_{x\in\Omega_e} |u''(x)|^2 h^5 \\
&\leq \max_{x\in\Omega} |u''(x)|^2 \sum_{e=1}^{n_{\text{el}}} h^5 \\
&\leq \max_{x\in\Omega} |u''(x)|^2 h^5 n_{\text{el}} \\
&= \max_{x\in\Omega} |u''(x)|^2 L h^4. \\
|u-\mathscr{I}u|^2_{1,2,\Omega} &= \sum_{e=1}^{n_{\text{el}}} |u-\mathscr{I}u|^2_{1,2,\Omega^e} \\
&\leq \max_{x\in\Omega} |u''(x)|^2 L h^2,
\end{aligned}$$

where the second estimate follows in a similar way to the first one.

An interesting aspect of this last derivation is that the local errors scale like h^{r+1}, but after adding over $\sim 1/h$ elements, the order of the error is reduced by one because $h^{r+1}/h = h^r$. Hence, convergence would not be guaranteed if a local error scaled like h only.

We hence derived the following interpolation error estimates

$$\|u-\mathscr{I}u\|_{0,2,\Omega} \leq \sqrt{L} \max_{x\in\Omega} |u''(x)| h^2 \tag{3.38a}$$

$$|u-\mathscr{I}u|_{1,2,\Omega} \leq \sqrt{L} \max_{x\in\Omega} |u''(x)| h. \tag{3.38b}$$

We conclude by making the following observations:

- The interpolation error for the function decreases more rapidly than that for the derivative as $h \searrow 0$.
- A function u whose second derivative is identically zero will be interpolated exactly. These are all affine functions over Ω.

> - The magnitude of the interpolation error does not depend on how large the derivative of the function is, but rather on how large the second derivative is.

We are now positioned to introduce the two main results of this section: the local and global interpolation inequalities. The former is adapted from [2, Thm. 3.1.4].

Theorem 3.5 (Local interpolation inequality)**.** *Let* $(\Omega_e, \mathcal{N}_e)$ *be a finite element,* $k \geq 0$ *and* $m \geq 0$ *be integers, and assume that* $\mathbb{P}^k(\Omega_e) \subset \mathscr{P}^e = span(\mathcal{N}_e)$*. Let* $\mathscr{I}_e \colon H^{k+1}(\Omega_e) \to \mathscr{P}^e$ *be a linear map that satisfies*

1. *(Continuity) There exists* $M > 0$ *such that for all* $u \in H^{k+1}(\Omega_e)$*,*

$$\|\mathscr{I}_e u\|_{m,2,\Omega_e} \leq M\|u\|_{k+1,2,\Omega_e}. \tag{3.39}$$

2. *If* $u \in \mathbb{P}^k(\Omega_e)$*, then* $\mathscr{I}_e u = u$*.*

Then, there exists a constant $C_{\mathscr{I}}(\Omega_e)$ *such that for all* $u \in H^{k+1}(\Omega_e)$*,*

$$|u - \mathscr{I}_e u|_{m,2,\Omega_e} \leq C_{\mathscr{I}}(\Omega_e) h_e^{k+1-m} |u|_{k+1,2,\Omega_e}, \tag{3.40}$$

where

$$h_e = \max_{x,y \in \Omega^e} \|x - y\| \tag{3.41}$$

is the **diameter** *of* Ω_e*, or element size. The constant* $C_{\mathscr{I}}$ *depends only on the shape of* Ω_e*, but is independent of* h_e*. The map* $\mathscr{I}_e$ *is called the* **local interpolation operator**.

Let's illustrate the theorem with a few examples.

> **Examples:**
>
> 3.23 Let e be a P^1-element with domain $\Omega_e = [a, b]$. In this case $\mathscr{P}^e$ contains all polynomials of degree 0 and 1, so we can set $k = 0, 1$ in Thm. 3.5. Let's consider $k = 1$ for now, and we will comment on $k = 0$ later. We can define the local interpolation operator
>
> $$\mathscr{I}_e u(x) = u(x_a) N_1^e(x) + u(x_b) N_2^e(x) \tag{3.42}$$
>
> for any $u \in H^2(\Omega_e)$ $(k + 1 = 2)$. It is possible to show that this interpolation operator satisfies the continuity condition for any $m \geq 0$. However, if $m \neq 0, 1$, $|\mathscr{I}_e u|_{m,2,\Omega_e} = 0$, since $I_e u$ is a first-order polynomial, and hence its second derivative is identically zero. Therefore, $\|\mathscr{I}_e u\|_{m,2,\Omega_e} = \|\mathscr{I}_e u\|_{1,2,\Omega_e}$ for any $m \geq 1$.

It also follows from the discussion in Example 1.52 that if $u \in \mathbb{P}^1(\Omega_e)$, then $\mathscr{I}_e u = u$, so that the second condition in Thm. 3.5 is satisfied. We can then conclude that for all $u \in H^2(\Omega_e)$,

$$\|u - \mathscr{I}_e u\|_{0,2,\Omega_e} \leq C_{\mathscr{I}}(\Omega_e) h_e^2 |u|_{2,2,\Omega_e} \tag{3.43a}$$

$$|u - \mathscr{I}_e u|_{1,2,\Omega_e} \leq C_{\mathscr{I}}(\Omega_e) h_e |u|_{2,2,\Omega_e}, \tag{3.43b}$$

where we have used that the exponent of h_e is $k+1-m = 2-m$ for $m = 0, 1$. For $m \geq 2$, $|u - \mathscr{I}_e u|_{m,2,\Omega_e} = |u|_{m,2,\Omega_e}$, so the local interpolation operator does not approximate derivatives of order larger than 1. This is also reflected in the exponent of h_e, $k+1-m = 2-m \leq 0$ for $m \geq 2$. These local interpolation inequalities in the L^2-norm and the H^1-seminorm show the same scaling with h_e than those obtained in Example 3.22 (provided that one power of h_e is used to mimic the integral of the second derivative over Ω_e).

When should we choose $k = 0$? Only when the function u that we are trying to interpolate is not smooth enough, $u \in H^1(\Omega_e)$ but $u \notin H^2(\Omega_e)$, such as the function $(x-a)\log(x-a)$. In this case, we can only get a meaningful estimate for the L^2-norm,

$$\|u - \mathscr{I}_e u\|_{0,2,\Omega_e} \leq C_{\mathscr{I}}(\Omega_e) h_e |u|_{1,2,\Omega_e},$$

since the exponent of h_e for $m = 1$ would be $k+1-m = 0+1-1 = 0$, so the value of the derivative is not guaranteed to be approximated despite the fact that we still have affine interpolation of u. In this case, the use of P_1 elements may only provide a better approximation than a P^0 one, but it would not provide a faster convergence rate.

3.24 Next, we obtain an interpolation error estimate with a larger order of convergence. To this end, let e be a P_l-element, $l \geq 1$, with domain $\Omega_e = [a, b]$, as defined in Example 1.52. To satisfy the condition $\mathbb{P}^k(\Omega_e) \subset \mathscr{P}^e$ in Thm. 3.5, we need $0 \leq k \leq l$. Given $u \in H^{k+1}(\Omega_e)$, it is possible to show (not here) that the local interpolation operator defined as

$$I_e u(x) = u(x_1^e) N_1^e(x) + \ldots + u(x_{k+1}^e) N_{k+1}^e(x) \tag{3.44}$$

satisfies the continuity condition in Thm. 3.5 for any $m \geq 0$. The second conditions in the theorem, that $\mathscr{I}_e u = u$ for $u \in \mathbb{P}^k(\Omega_e)$ is also satisfied, since in Example 1.52 we saw that if $u \in \mathbb{P}^l(\Omega_e)$, then $\mathscr{I}_e u = u$. We can then conclude that for all $u \in H^{k+1}(\Omega_e)$,

$$\begin{gathered}
\|u - \mathscr{I}_e u\|_{0,2,\Omega_e} \leq C_{\mathscr{I}}(\Omega_e) h_e^{k+1} |u|_{k+1,2,\Omega_e} \\
|u - \mathscr{I}_e u|_{1,2,\Omega_e} \leq C_{\mathscr{I}}(\Omega_e) h_e^{k} |u|_{k+1,2,\Omega_e}, \\
\vdots \\
|u - \mathscr{I}_e u|_{k,2,\Omega_e} \leq C_{\mathscr{I}}(\Omega_e) h_e |u|_{k+1,2,\Omega_e}.
\end{gathered}$$

Therefore, the largest order of convergence we can obtain with a P_l-element is for functions $u \in H^{l+1}(\Omega_e)$ ($k = l$), in which case the error in the L^2-norm decreases with order $l+1$. For example, quadratic elements P_2 can achieve third order of convergence in the L^2-norm for a function that is at least in $H^3(\Omega_e)$. The convergence rate for functions that are less smooth will drop, as discussed in Example 3.23.

3.25 Consider next a P_1-element in $\mathbb{R}^2$, so Ω_e is a triangle. In this case $\mathbb{P}^k(\Omega_e) \subset \mathscr{P}^e$ for $k = 0, 1$. We can define the local interpolation operator as

$$\mathscr{I} u(\mathbf{x}) = u(\mathbf{X}^1) N_1^e(\mathbf{x}) + u(\mathbf{X}^2) N_2^e(\mathbf{x}) + u(\mathbf{X}^3) N_3^e(\mathbf{x}) \tag{3.45}$$

for any $\mathbf{x} \in \Omega_e$, where $\mathbf{X}^a \in \mathbb{R}^2$ for $a = 1, 2, 3$ are the position of the three nodes of the element. This interpolation operator can be shown to be continuous for $u \in H^2(\Omega_e)$ ($k = 1$) for any $m \geq 0$. It is also true that if $u \in \mathbb{P}^1(\Omega_e)$, then $\mathscr{I}_e u = u$, since the values at the three nodes uniquely determine an affine polynomial. From Theorem 3.5 then, the exists $C_{\mathscr{I}}$ such that for all $u \in H^2(\Omega_e)$

$$\|u - \mathscr{I} u\|_{0,2,\Omega_e} \leq C_{\mathscr{I}}(\Omega_e) h_e^2 |u|_{2,2,\Omega_e} \tag{3.46a}$$

$$|u - \mathscr{I} u|_{1,2,\Omega_e} \leq C_{\mathscr{I}}(\Omega_e) h_e |u|_{2,2,\Omega_e}. \tag{3.46b}$$

The continuity condition (3.39) is similar to the boundedness condition we request for a linear form, and it is in fact a necessary and sufficient condition for the local interpolation operator to be continuous. We shall skip the details here.

You may also be wondering why Thm. 3.5 states results in terms of seminorms, and not of norms. This is because by having the convergence in each seminorm, we know the convergence behavior of the approximation to each derivative of the function. The convergence rate in a norm will be equal to the convergence rate of the highest derivative involved in the norm. Precisely, the convergence rate in the $\|\cdot\|_{m,2}$ will be that of $|\cdot|_{m,2}$. This is because as $h \searrow 0$, the contribution to the error that decreases with the smallest power of h will be significantly larger than the rest, and will dominate the convergence behavior of the entire norm, just as illustrated in Fig. 3.8.

Theorem 3.6 (Global interpolation inequality). *Let $\Omega \subset \mathbb{R}^n$ be a domain such that Ω is meshed by a collection of elements $\{(\Omega_e, \mathscr{N}_e)\}_e$. Assume that each one of the elements has a local interpolation operator $\mathscr{I}_e$, and that the conditions of Theorem 3.5 are satisfied for integers $k \geq 0$ and $m \geq 0$. Given a function $u \colon \Omega \to \mathbb{R}$, the* **global interpolation operator** *$\mathscr{I}$ builds a function $\mathscr{I} u$ according to*

$$\mathscr{I} u(x) = \mathscr{I}_e u(x) \qquad \forall x \in \mathring{\Omega}_e. \tag{3.47}$$

Then, $\mathscr{I} u$ is well defined for any $u \in H^{k+1}(\Omega)$, and there exists a constant C that depends only on the mesh such that for all $u \in H^{k+1}(\Omega)$,

$$\sum_e |u - \mathscr{I} u|_{m,2,\Omega_e} \leq C h^{k+1-m} |u|_{k+1,2,\Omega}, \tag{3.48}$$

where $C = \max_e C_{\mathscr{I}}(\Omega_e)$ *and* $h = \max_e h_e$. *If, additionally,* $\mathscr{I} u \in H^{k+1}(\Omega)$, *then*

$$|u - \mathscr{I} u|_{m,2,\Omega} \le C h^{k+1-m} |u|_{k+1,2,\Omega}. \tag{3.49}$$

Proof. The proof is straightforward. First, the global inteporlation operator is well-defined for $u \in H^{k+1}(\Omega)$, since the restriction of u to Ω_e is in $H^{k+1}(\Omega_e)$ for each e, the domain of $\mathscr{I}_e$. Additionally, $\mathscr{I}_e u$ satisfies (3.40). Therefore,

$$\begin{aligned}
\sum_e |u - \mathscr{I}_e u|_{m,2,\Omega_e} &\le \sum_e C_{\mathscr{I}}(\Omega_e) h_e^{k+1-m} |u|_{m,2,\Omega_e} && \text{from (3.40)} \\
&\le \left(\max_e C_{\mathscr{I}}(\Omega_e)\right) \left(\max_e h_e\right)^{k+1-m} \sum_e |u|_{m,2,\Omega_e} \\
&= C h^{k+1-m} |u|_{m,2,\Omega}.
\end{aligned}$$

If $\mathscr{I} u \in H^{k+1}(\Omega)$, then

$$|u - \mathscr{I} u|_{m,2,\Omega} = \sum_e |u - \mathscr{I} u|_{m,2,\Omega_e}.$$

□

The global interpolation inequality introduces the constant $C = \max_e C_{\mathscr{I}}(\Omega_e)$ that depends only on the mesh. We will have an opportunity to discuss more about this constant later. For the moment, it is important to know that the "quality" of the elements in the mesh enters the error through it. For example, we will see that $C_{\mathscr{I}}(\Omega_e)$ grows unbounded as the smallest angle of a triangular element decreases to zero. This imposes restrictions on the construction of a mesh; in this case, triangular elements in a mesh should have a guaranteed minimum angle. Alternatively, if we considered meshes in which the smallest angle in the mesh decreased to zero as h did, the uncontrolled error in the interpolation inequality could be reflected in an uncontrolled error in the finite element solution, potentially decreasing the convergence rate or, even worse, leading to lack of convergence.

The presence of small angles in a mesh does not necessarily leads to large errors, and in some circumstances, it is desired. For example, in the presence of sharp directional gradients in a solution, such as in a boundary layer in a fluid or a shear band in a solid, it is convenient and computationally efficient to have elements that are elongated along the direction of the smallest partial derivative. Because elements are elongated, their smallest angle will be very small enough to better capture changes of the solution in the direction of the largest partial derivative, but not arbitrarily small.

As a sidenote, notice that in principle $\mathscr{I} u$ does not need to be a function in $H^{k+1}(\Omega)$, and in general it will not be except for low values of m, like $0, 1, 2$, since the continuity of higher-order derivatives is typically computationally expensive to attain for most but a selected class of carefully crafted elements and meshes. Locally, however, the approximation would still hold.

Examples:

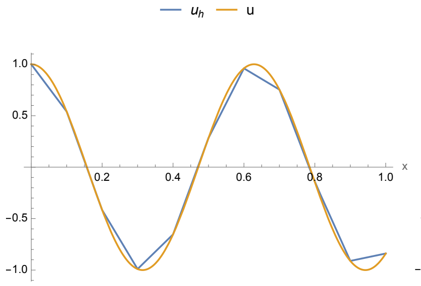
3.26 Consider a mesh of $n_{\text{el}} \in \mathbb{N}$ P^1-elements over a domain $\Omega = [0, L]$, and let The local interpolation operator for this element was defined in Example 3.23, (3.42). The global interpolation operator was already introduced in Example 3.22, (3.34). It is simple to see that they satisfy (3.47), simply because the only two basis functions that are non-zero in element e are N_e and N_{e+1}. Then, from Thm. 3.6, for all $u \in H^2(\Omega)$

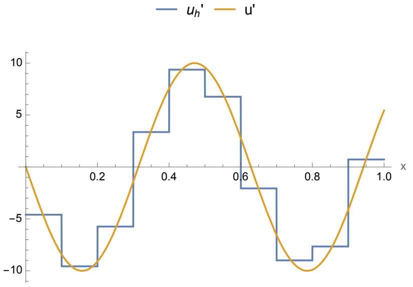
$$\|u - \mathscr{I} u\|_{0,2,\Omega} \le Ch^2 |u|_{2,2,\Omega} \tag{3.50a}$$

$$|u - \mathscr{I} u|_{1,2,\Omega} \le Ch |u|_{2,2,\Omega}, \tag{3.50b}$$

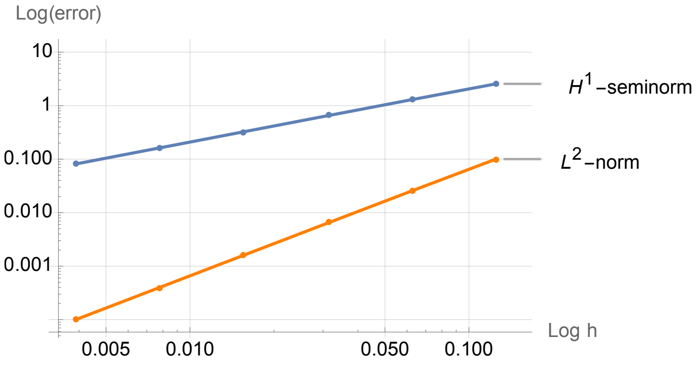
for a constant C independent of u. We cannot state anything about the approximation of derivatives of order 2 or higher, since the second derivative of the local interpolation operator (which is equal to zero) does not approximate u''.

For a concrete example, let $L = 1$, $h = 1/n_{\text{el}}$, and $u(x) = \cos(10x) \in H^2([0, L])$. The interpolant and its first derivative are plotted below for $n_{\text{el}} = 11$.





The L^2-norm and H^1-seminorm of the error $u - \mathscr{I} u$ is plotted next, in a log-log scale. Why a log-log plot? Just as in Fig. 3.8, because if we compute the log on both sides of a relation of the form error$\sim Ch^r$, we get that $\log(\text{error}) \sim \log C + r \log h$, so log(error) decreases linearly with $\log h$, and the slope of the line is equal to the order r.



Notice that the L^2-error decreases 2 units in the vertical axis per unit in the horizontal axis, a slope equal to 2, in agreement with the expected order of convergence in (3.50). Similarly, the H^1-seminorm decreases 1 unit in the vertical axis per unit in the horizontal one, indicating a slope equal to 1, in agreement with the first order of convergence in (3.50). The square of the H^1-norm is the sum of the squares of the two curves, and when plotted, it essentially overlaps with the curve of the H^1-seminorm, so it was not plotted. A quick inspection reveals that at the largest value of h, the L^2-norm of the error is already more than 10 times smaller than the H^1-seminorm, so when the two squared values are added, the error on the value of the function contributes less than 1% to the sum. The H^1-error is dominated by the error in the derivative.

3.27 Consider next a conformal triangulation of a domain Ω in $\mathbb{R}^2$, and the space of continuous P^1 finite elements over it. Similar to Example 3.26 the global interpolant $\mathscr{I} u$ of a function $u \colon \Omega \to \mathbb{R}$ is defined as

$$\mathscr{I} u(x) = \sum_{A=1}^{m} N_a(\mathbf{x}) u(\mathbf{X}^a), \tag{3.51}$$

given that $\mathscr{I} u(x) = \mathscr{I}_e u(x)$ for $x \in \mathring{\Omega}_e$ and any e, where $\mathscr{I}_e$ was defined in Example 3.25. The equality is obtained because the only non-zero basis functions in Ω_e are the ones that are equal to 1 at each one of the vertices of Ω_e. The global interpolant is continuous in Ω, and hence it can be proved that $\mathscr{I} u \in H^1(\Omega)$.

Then, for $u \in H^2(\Omega)$, we conclude that

$$\|u - \mathscr{I} u\|_{0,2,\Omega} \leq C h^2 |u|_{2,2,\Omega} \tag{3.52a}$$

$$|u - \mathscr{I} u|_{1,2,\Omega} \leq C h |u|_{2,2,\Omega}. \tag{3.52b}$$

If $u \notin H^2(\Omega)$, interpolation error estimates are more complex and their discussion is beyond the scope of this notes.

3.2.5 Convergence

We are now ready to state the main convergence result for finite element approximations, Céa's lemma, which establishes the convergence of finite element approximations in the space in which the bilinear form and linear functional are continuous. For linear second-order problems this is typically a subspace of H^1.

In the next four lemmas we use the definitions introduced in Problems 3.1 and 3.2.

Lemma 3.2 (Galerkin orthogonality)**.** *The error $u - u_h$ satisfies that*

$$a(u - u_h, v_h) = 0 \tag{3.53}$$

for any $v_h \in \mathcal{V}_h$.

Proof. Since $\mathcal{V}_h \subset \mathcal{V}$, the exact and finite element solutions satisfy that for any $v_h \in \mathcal{V}_h$,

$$a(u, v_h) = \ell(v_h) \tag{3.54}$$

$$a(u_h, v_h) = \ell(v_h). \tag{3.55}$$

Subtracting the two equations, we conclude that

$$a(u - u_h, v_h) = 0. \tag{3.56}$$

□

A version of the next lemma was first introduced by Jean Céa, and it provides the basic convergence results for finite element methods for elliptic problems.

Lemma 3.3 (Céa's Lemma)**.** *Let* $\|\cdot\|$ *be a norm on* $\mathcal{W}$*. Assume that* $a(\cdot,\cdot)$ *is continuous on* $\mathcal{W}$*,*

$$|a(w, v)| \le M\|w\|\|v\| \quad \textit{for some } M > 0 \textit{ and all } w, v \in \mathcal{W}, \tag{3.57}$$

and coercive on $\mathcal{V}_h$*,*

$$c_h\|v_h\|^2 \le a(v_h, v_h) \quad \textit{for some } c_h \textit{ and all } v_h \in \mathcal{V}_h. \tag{3.58}$$

Then,

$$\|u - u_h\| \le \left(1 + \frac{M}{c_h}\right)\|u - w_h\| \quad \textit{for all } w_h \in \mathcal{S}_h. \tag{3.59}$$

Proof. For $w_h \in \mathcal{S}_h$ we have that

$$\begin{aligned} c_h\|u_h - w_h\|^2 &\le a(u_h - w_h, u_h - w_h) && \text{coercivity, (3.58)} \\ &= \underbrace{a(u_h - u, u_h - w_h)}_{=0} + a(u - w_h, u_h - w_h) && \text{add and subtract } u \\ &= a(u - w_h, u_h - w_h) && \text{Galerkin orthogonality, (3.56)} \\ &\le M\|u - w_h\|\|u_h - w_h\| && \text{continuity, (3.57).} \end{aligned}$$

Notice that it is possible to use Galerkin orthogonality above because $w_h - u_h \in \mathcal{V}_h$. We can then conclude that

$$\|u_h - w_h\| \le \frac{M}{c_h}\|u - w_h\|.$$

Application of the triangle inequality leads to (3.59), namely,

$$\begin{aligned} \|u - u_h\| &\le \|u - w_h\| + \|w_h - u_h\| \\ &\le \|u - w_h\| + \frac{M}{c_h}\|u - w_h\| = \left(1 + \frac{M}{c_h}\right)\|u - w_h\|. \end{aligned}$$

□

Let's comment on the significance of this result. Notice that since (3.59) in Céa's lemma holds for any $w_h \in \mathscr{S}_h$, it holds in particular for the best approximation we can find for u in $\mathscr{S}_h$, namely

$$\|u - u_h\| \leq \left(1 + \frac{M}{c_h}\right) \min_{w_h \in \mathscr{S}_h} \|u - w_h\|. \tag{3.60}$$

This means that the finite element solution is at most $(1 + M/c_h)$ worse than the best function to approximate u in $\mathscr{S}_h$. The key observation here is that since we do not know what u is, we cannot generally find the best approximation, but we can find u_h *without* knowing u.

The right hand side (3.59) has two quantities that depend on h, c_h and $\mathscr{S}_h$. Most often, the bilinear form $a(\cdot,\cdot)$ is coercive in $\mathscr{V} \ni \mathscr{V}_h$ (c.f. Examples 3.20 and 3.21), so there exists a coercivity constant that is independent of h. Specifically, the constant $c_\mathscr{V}$ in $\mathscr{V}$ is necessarily smaller or equal than c_h for any h, and hence we have that

$$a(v_h, v_h) \geq c_h \|v_h\|_2 \geq c_\mathscr{V} \|v_h\|^2.$$

In this case (3.59) becomes

$$\|u - u_h\| \leq \left(1 + \frac{M}{c_\mathscr{V}}\right) \|u - w_h\| \qquad \forall w_h \in \mathscr{S}_h, \tag{3.61}$$

in which the only dependence on h is in $\mathscr{S}_h$.

To approximate u with u_h in this case, it is sufficient to have the ability to construct spaces of functions that, by appropriately selecting h, can approximate *any* function in $\mathscr{S}$ to any desired accuracy. In other words, if for each h we can choose $w_h \in \mathscr{V}_h$ such that $w_h \to u$ as $h \to 0$, then it would follow from Céa's lemma that $u_h \to u$ as well. This is where the interpolation error estimates in Finite Element spaces discussed in §3.2.4 play a crucial role. Let's put all the ingredients together next.

Mimicking the results on interpolation (Thm. 3.6), assume that there exists a global interpolation operator $\mathscr{I} : \mathscr{S} \to \mathscr{S}_h$ and a constant $C > 0$ independent of h such that for all $u \in \mathscr{S}$,

$$\|u - \mathscr{I}u\| \leq Ch^r |u|$$

for some seminorm $|\cdot|$ of u. Then, we can set $w_h = \mathscr{I}u$ in 3.61, and conclude that

$$\|u - u_h\| \leq \left(1 + \frac{M}{c_\mathscr{V}}\right) Ch^r |u|, \tag{3.62}$$

the desired convergence result.

Let's look at a specific example next.

Examples:

3.28 Let's obtain error estimates for the finite element approximation to the solution of the diffusion problem in $\mathbb{R}^2$, Problem 2.2 with W_h being a continuous P^1 finite element space over a conformal triangulation. For simplicity again, we consider the case in with $K = k\mathbb{I}_{2\times 2}$ with $k > 0$ constant in the domain. In Example 3.15, we have already seen that if we set $\mathscr{W} = H^1(\Omega)$, then the bilinear form 2.17 and linear functional 2.18 of the problem are continuous. In particular, we found that

$$|a(u,v)| \leq k\|u\|_{1,2,\Omega}\|v\|_{1,2,\Omega}.$$

Additionally, if $\partial\Omega_D$ has positive length, we found in Example 3.21 that a is also coercive in the space

$$\mathscr{V} = \{u \in \mathscr{W}) \mid u = 0 \text{ on } \partial\Omega_D\},$$

with coercivity constant $c_{\mathscr{V}} = \frac{k}{2}\min\{1, C_p^{-2}\}$. In particular, the coercivity on $\mathscr{V}_h = \mathscr{V} \cap W_h$ is satisfied as well.

Therefore, taking $(\mathscr{W}, \|\cdot\|) = (H^1(\Omega), \|\cdot\|_{1,2,\Omega})$, the conditions of Céa's lemma are satisfied, and we can immediately conclude that

$$\|u - u_h\|_{1,2,\Omega} \leq \left(1 + 2\max\{1, C_p^2\}\right)\|u - w_h\|_{1,2,\Omega} \qquad \forall w_h \in \mathscr{S}_h. \tag{3.63}$$

where we used that

$$\left(1 + \frac{M}{c_{\mathscr{V}}}\right) = \left(1 + \frac{2k}{\min\{1, C_p^{-2}\}k}\right) = 1 + 2\max\{1, C_p^2\}.$$

It remains to find $w_h \in \mathscr{S}_h$ that is close enough to $u \in H^1(\Omega)$. With this goal, two important observations should be made.

First, notice that we need an estimate of the H^1-norm of $u - w_h$. So, to take advantage of the global interpolation inequality in Thm. 3.6, we will need to require the exact solution u to be in at least $H^2(\Omega)$, for otherwise Thm. 3.6 does not provide an estimate for the H^1-seminorm $|u - \mathscr{I}u|_{1,2,\Omega}$.

Second, (3.63) is valid for $w_h \in \mathscr{S}_h$, so to set $w_h = \mathscr{I}u$ we need $\mathscr{I}u \in \mathscr{S}_h$. For this problem, functions in $\mathscr{S}$ and hence $\mathscr{S}_h$ satisfy that they are equal to g on $\partial\Omega_D$, (2.14). This imposes a nontrivial constraint on $\mathscr{I}u$, since g may not be equal to *any* continuous P^1 function on $\partial\Omega_D$.

For example, let $\Omega = [-1,1] \times [-1,1]$, $\partial\Omega_D = \{1\} \times [-1,1]$, or the right edge of the square, and denote the Cartesian coordinates by (x_1, x_2). The function g should be defined for $x_2 \in [-1,1]$. If $g(x_2) = ax_2 + b$, for $a, b \in \mathbb{R}$, then $\mathscr{I}u(1, x_2) = g(x_2)$ on $\partial\Omega_D$, since all affine functions over $\partial\Omega_D$ are in $\mathscr{S}_h$. However, if $g(x_2) = \cos(x_2)$, $\mathscr{I}u(x_2) \neq g(x_2)$ for most values of x_2, regardless of what mesh we choose. In this case, $\mathscr{I}u \notin \mathscr{S}_h$. Even worse, in this case $\mathscr{S}_h = \emptyset$.

The situation is not as dire as it seems, and we are going to discuss how we proceed in this case when we talk about *variational crimes*.

For the moment, *we will assume that g is such that* $\mathscr{I} u = g$ *on* $\partial\Omega_D$, so that $\mathscr{I} u \in \mathscr{S}_h$.

Summarizing, assuming that $u \in H^2(\Omega)$ and that $\mathscr{I} u \in \mathscr{S}_h$, we can set $w_h = \mathscr{I} u$ in (3.63), and conclude that

$$\|u - u_h\|_{1,2,\Omega} \leq \hat{C} h |u|_{2,2,\Omega}, \tag{3.64}$$

where $\hat{C} = \left(1 + 2\max\{1, C_p^2\}\right)\sqrt{2}C$, and where we used that from Thm. 3.6,

$$\begin{aligned}\|u - \mathscr{I} u\|^2_{1,2,\Omega} &= \|u - \mathscr{I} u\|^2_{0,2,\Omega} + |u - \mathscr{I} u|^2_{1,2,\Omega} \\ &\leq C^2 h^4 |u|^2_{2,2,\Omega} + C^2 h^2 |u|^2_{2,2,\Omega} \leq 2C^2 |u|^2_{2,2,\Omega} h^2,\end{aligned}$$

where the last inequality follows from requesting $h < 1$, without loss of generality.

Best approximation property. In the case in which $a(\cdot,\cdot)$ is symmetric, Galerkin orthogonality has a nice and intuitive interpretation.

Lemma 3.4 (Best approximation property)**.** *If* $a(\cdot,\cdot)$ *is symmetric and coercive on* $\mathscr{V}_h$ *as in* (3.58)*, then*

$$a(u_h - u, u_h - u) < a(w_h - u, w_h - u) \qquad \textit{for any } w_h \in \mathscr{S}_h, w_h \neq u_h. \tag{3.65}$$

This can be stated in terms of the energy norm as

$$\|u - u_h\|_a \leq \|u - w_h\|_a \qquad \textit{for any } w_h \in \mathscr{S}_h, w_h \neq u_h. \tag{3.66}$$

This lemma says that the error of the finite element solution u_h is the smallest among all functions in the trial space $\mathscr{S}_h$ when measured in the energy norm! This is called the **best approximation property** of Galerkin method.

Proof. This follows easily from Galerkin orthogonality. For a more compact notation, set $e_u = u_h - u$, $e_w = w_h - u$, and $\Delta u_h = u_h - w_h$. Then,

$$\begin{aligned}a(e_w, e_w) &= a(w_h - u, w_h - u) \\ &= a(w_h - u_h + u_h - u, w_h - u_h + u_h - u) \\ &= a(\Delta u_h + e_u, \Delta u_h + e_u) \\ &= a(\Delta u_h, \Delta u_h) + a(\Delta u_h, e_u) + a(e_u, \Delta u_h) + a(e_u, e_u) \\ &= a(\Delta u_h, \Delta u_h) + 2a(u - u_h, w_h - u_h) + a(e_u, e_u) && \text{symmetry of } a(\cdot,\cdot) \\ &= a(\Delta u_h, \Delta u_h) + a(e_u, e_u) && \text{Galerkin orthogonality} \\ &> a(e_u, e_u) && \text{coercivity on } \mathscr{V}_h.\end{aligned}$$

In the next to last line we used that $w_h - u_h \in \mathscr{V}_h$, to conclude that $a(u - u_h, w_h - u_h) = 0$. In the last line we used that $a(\Delta u_h, \Delta u_h) \geq c_h \|\Delta u_h\|^2 > 0$. □

Galerkin orthogonality and the best approximation property have the intuitive interpretation we are used to in finite dimensional spaces, as illustrated in Fig. 3.9. The closest point from the exact solution to a subspace $\mathscr{S}_h$ is the orthogonal projection to it, so that the error $u - u_h$ is orthogonal to $\mathscr{S}_h$.

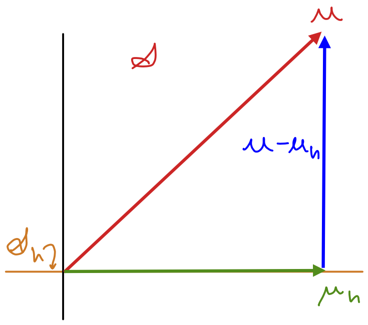


Figure 3.9