Advanced Finite Element methods

Ivan Graham, University of Bath

This is the final version of the printed notes for 2009, apart from any typos which may be pointed out to me.

Please let me know of any queries or errors which you spot.

For the exam, students should study both what is written on the whiteboard and what is given in these printed notes following the guidance on the conduct of the oral exam which was issued previously.

Last updated : April 20, 2009

1 Background on Sobolev Spaces

In this section I briefly review background material in Sobolev spaces. Most of this will be familiar if you have taken the graduate unit on this topic (although I will also quote some results for Lipschitz domains which were not dealt with there). Suitable references for futher detail are [GRB,BS,EG,Ha,McL] from the reference list.

Throughout we assume that Ω is a domain (i.e. an open set) in $\mathbb{R}^n$. In most of the course Ω will be bounded and n will be $1, 2$ or 3. With $\overline{\Omega}$ denoting the closure of Ω (i.e. all points obtained as limits of sequences in Ω) we define the boundary of Ω as

$$\partial\Omega = \overline{\Omega} \setminus \Omega \ .$$

We assume readers are familiar with the space $L^p(\Omega)$ for $1 \leq p \leq \infty$. For $p = 2$, this is a Hilbert Space (see Appendix for functional analysis background material), the inner product is

$$(f, g)_{L^2(\Omega)} = \int_\Omega f(x)g(x)dx \ ,$$

and the norm is

$$\|f\|_{L^2(\Omega)} \ = \ (f, f)^{1/2}_{L^2(\Omega)} \ .$$

(This is the form of the inner product for real-valued functions: all our functions will be real-valued unless otherwise stated.)

For $x, y \in \mathbb{R}^n$, we define the usual dot product and Euclidean norm: $x.y \ = \ \sum_i x_i y_i$, $|x| \ = \ \sqrt{x.x}$.

Multiindices A multi-index is an ordered list of n non-negative integers $\alpha = (\alpha_1, ..., \alpha_n)$ with each $\alpha_i \in \mathbb{N} \cup \{0\}$. The order of α is $|\alpha| = \alpha_1 + ... + \alpha_n$. Given α we have associated *polynomial functions*

$$x^\alpha := x_1^{\alpha_1} x_2^{\alpha_2} \ \ldots \ x_n^{\alpha_n}$$

and *partial differential operators*

$$(D^\alpha v) = \left(\frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \frac{\partial^{\alpha_2}}{\partial x_2^{\alpha_2}} \ \ldots \ \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}} \right) v.$$

For $f : \Omega \to \mathbb{R}$, we define the support of f as the closed set:

$$\texttt{supp}\, f = \overline{\{x \in \Omega : f(x) \neq 0\}}.$$

If this is a bounded subset of the open domain Ω then it is compact and we say u has compact support in Ω. Note that the constant function with value 1 on Ω does not have compact support in Ω, since its support is actually $\overline{\Omega}$, which is larger than Ω.

For any non-negative k, $C^k(\Omega)$ denotes the space of functions which are $k-$ times continuously differentiable on Ω (without any condition on behaviour at the boundary),

whereas $C^k(\overline{\Omega})$ denotes the space of functions which are $k-$ times continuously differentiable on Ω, and each derivatve has a continuous extension to the boundary and hence then all derivatives are bounded. When $k = 0$, we write $C(\Omega)$ and $C(\overline{\Omega})$. Also $C^\infty(\Omega)$ and $C^\infty(\overline{\Omega})$ denote the analogous spaces of functions with infinitely many continuous derivatives. Finally we set

$$C_0^\infty(\Omega) = \{f \in C^\infty(\Omega) : \texttt{supp}\, f \text{ is compact in } \Omega\}$$

In many references (e.g. [GRB]) $C_0^\infty(\Omega)$ is denoted $\mathcal{D}(\Omega)$, "the space of test functions". The following result is standard, e.g. [BS,GRB]:

Lemma 1.1. *$C_0^\infty(\Omega)$ is dense in $L^p(\Omega)$ for all $1 \le p < \infty$.*

This means that every function in $L^p(\Omega)$ can be arbitrarily closely approximated by a function from $C_0^\infty(\Omega)$ (with the error measured in the $L^p(\Omega)$ norm).

Example 1.2. *(i) Any polynomial is in $C^\infty(\Omega)$ for any Ω. As long as Ω is bounded, then any polynomial is also in $C^\infty(\overline{\Omega})$.*

(ii) Suppose Ω is the triangle with vertices $(0,0)$ $(1,0)$ and $(0,1)$. Then the function $f(x) := x_1x_2(1 - x_1 - x_2)$ is in $C^\infty(\Omega)$ and $C^\infty(\overline{\Omega})$. However, even though f vanishes on the boundary of Ω, f is not in $C_0^\infty(\Omega)$, since its support is all of $\overline{\Omega}$.

Exercise 1.3. *Define the function φ on all of $\mathbb{R}^n$ by:*

$$\varphi(x) = \begin{cases} e^{\frac{1}{|x|^2-1}}, & |x| < 1 \\ 0, & |x| \ge 1 \end{cases}$$

Prove that $\varphi(x) \in C_0^\infty(\Omega)$ for any Ω containing $\overline{B}(0,1) = \{x \in \mathbb{R}^n : |x| \le 1\}$.

Hint: *For $|x| < 1$, write $\varphi(x) = e^{-t}$, with $t = \dfrac{1}{1-|x|^2}$,. Then*

$$\frac{\partial \varphi}{\partial x_j} = e^{-t}\frac{1}{(1-|x|^2)^2}.(-2x_j) = -2x_j e^{-t}t^2 \qquad (\textit{and} \quad e^{-t}t^2 \to 0 \quad \textit{as} \quad t \to \infty).$$

Prove by induction that for all multiindices α, there $\exists$ a polynomial P_α such that $(D^\alpha\varphi)(x) = P_\alpha(x)e^{-t}t^{2|\alpha|}$, $|x| < 1$. Also, $(D^\alpha\varphi)(x) = 0$, $|x| \ge 1$. Hence deduce the result.

Up to now all derivatives have been defined in the classical way. We need to work with more general *weak derivatives*. To do this, following the standard texts, we introduce the space:

$$L^1_{loc}(\Omega) = \{f : \Omega \to \mathbb{R} : \ f|_K \in L^1(K) \quad \text{for all compact subsets} \quad K \subset \Omega\}.$$

Functions in L_{loc} are "locally integrable" but may behave badly at the boundary $\partial\Omega$. The notion of *weak derivative* is defined through the integration by parts formula:

Definition 1.4. *Let α be any multi-index. Then a function $f \in L^1_{loc}(\Omega)$ has a weak derivative $D^\alpha f$ provided there exists $g \in L^1_{loc}(\Omega)$ such that*

$$\int_\Omega g\varphi = (-1)^{|\alpha|} \int_\Omega f D^\alpha \varphi \qquad \text{for all} \quad \varphi \in C_0^\infty(\Omega) \ .$$

If this formula holds then we write $D^\alpha f = g$.

Example 1.5. *As a very simple example suppose $f : (-1,1) \to \mathbb{R}$ is defined by*

$$f(x) \ = \ \begin{cases} 1+x & x \in (-1,0) \\ 1-x & x \in (0,1) \ . \end{cases}$$

The function g defined by

$$g(x) \ = \ \begin{cases} 1 & x \in (-1,0) \\ -1 & x \in (0,1) \end{cases}$$

is certainly in $L^1_{loc}(-1,1)$. Moreover, (with prime denoting the classical derivative), we have $g = f'$ on $(-1,0)$ and on $(0,1)$ (but not on $(-1,1)$). Hence, for $\varphi \in C_0^\infty(-1,1)$,

$$\begin{aligned} \int_{-1}^1 g\varphi \ &= \ \int_{-1}^0 (1+x)'\varphi(x)\ dx \ + \ \int_0^1 (1-x)'\varphi(x)\ dx \\ &= \ \Big[(1+x)\varphi(x)\Big]_{-1}^0 \ - \ \int_{-1}^0 (1+x)\varphi'(x)\ dx \ + \ \Big[(1-x)\varphi(x)\Big]_0^1 \ - \ \int_0^1 (1-x)\varphi'(x)\ dx \\ &= \ -\int_{-1}^1 f\varphi' \ . \end{aligned}$$

(The first and third terms vanish since $\varphi \in C_0^\infty(-1,1)$.)

Hence $g = f'$ $(\ = \ Df)$ in the sense of weak derivatives.

In the above example we have been able to construct the weak derivative of f by stitching together its classical derivatives on each of $(-1,0)$ and $(0,1)$. (Even though f is not differentiable in the classical sense at 0.) It is dangerous to think that this is a way of computing the weak derivative in all cases as the following exercise shows.

Exercise 1.6. *Let g be defined as in the previous example. Show that if the distributional derivative Dg exists then*

$$\int_{-1}^1 (Dg)(x)\varphi(x)dx \ = \ -2\varphi(0) \qquad \text{for all } \varphi \in C_0^\infty(-1,1). \qquad (*)$$

Hence show that $Dg \notin L^1_{loc}(-1,1)$. (Hint: use the fact that $C_0^\infty(-1,1)$ is dense in $L^1(-1,1)$, from Lemma 1.1) . (Note that (*) implies that $-(1/2)Dg$ is the classical delta function on $(-1,1)$.)

Sobolev Spaces

Definition 1.7.

$$H^k(\Omega) = \{f \in L^2(\Omega) : \ \|f\|_{H^k(\Omega)} < \infty\}$$

where the norm is defined by

$$\|f\|_{H^k(\Omega)} = \left\{ \sum_{0 \le |\alpha| \le k} \|D^\alpha f\|^2_{L^2(\Omega)} \right\}^{\frac{1}{2}} .$$

This is a Hilbert space with inner product

$$(f,g)_{H^k(\Omega)} = \sum_{0 \le |\alpha| \le k} (D^\alpha f, D^\alpha g)_{L^2(\Omega)}$$

and is the special case $p = 2$ of the more general $W^{k,p}(\Omega)$ spaces found in many references (e.g. [BS, GRB]) .

We shall often be working with $H^1(\Omega)$. Note that

$$(f,g)_{H^1(\Omega)} \ = \ (f,g)_{L^2(\Omega)} \ + \ \sum_{j=1}^{n} \left(\frac{\partial f}{\partial x_j}, \frac{\partial g}{\partial x_j} \right)_{L^2(\Omega)}$$

and so

$$\|f\|_{H^1(\Omega)} \ = \ \left\{ \|f\|^2_{L^2(\Omega)} \ + \ |f|^2_{H^1(\Omega)} \right\}^{1/2} ,$$

where the H^1 seminorm is defined by

$$|f|_{H^1(\Omega)} \ = \ \left\{ \sum_{j=1}^{n} \left\| \frac{\partial f}{\partial x_j} \right\|^2_{L^2(\Omega)} \right\}^{1/2} \ = \ \left\{ \int_\Omega |\nabla f|^2 \right\}^{1/2} ,$$

where ∇f denotes the gradient of f and $|x| := \{x_1^2 + x_2^2 + \ldots + x_n^2\}^{1/2}$ for $x \in \mathbb{R}^n$.

Definition 1.8. *$H_0^k(\Omega)$ is defined to be the closure of $C_0^\infty(\Omega)$ with respect to $\| \, . \, \|_{H^k(\Omega)}$.*

Loosely speaking $H_0^k(\Omega)$ consists of all functions in $H^k(\Omega)$ which vanish on the boundary of Ω. However as we shall see below, functions in $H^k(\Omega)$ do not necessarily have point values, and so this loose notion is not quite correct.

One of the most important results for this course is:

Theorem 1.9 (The Poincaré or Poincaré-Friedrichs Inequality). *If Ω is a bounded domain then there exists a constant C (which depends on Ω) such that*

$$\|u\|_{H^1(\Omega)} \ \le \ C|u|_{H^1(\Omega)} \qquad \textit{for all } u \in H_0^1(\Omega).$$

A proof can be found in [GRB, Theorem 2.2], [BS,5.3.3], and [EG, Lemma B61].

Remark 1.10.

(i) By definition, $C_0^\infty(\Omega)$ is a dense subset of $H_0^k(\Omega)$.

(ii) A deeper fact is that $C^\infty(\Omega) \cap H^k(\Omega)$ is dense in $H^k(\Omega)$. (This is the famous Meyers-Serrin Theorem - e.g. [GRB, Theorem 3.8]).

Exercise 1.11. *Remind yourself of what is means for two norms on a Hilbert space to be equivalent and show that Theorem 1.9 implies that $\|\,.\,\|_{H^k(\Omega)}$ and $|\,.\,|_{H^k(\Omega)}$ are equivalent norms on $H_0^k(\Omega)$.*

Theorem 1.12 (Trace Theorem, version 1)**.** *(e.g. [GRB, Theorem 3.12]) Let Ω be bounded with a C^1 boundary, $\partial\Omega$. Then there exists a bounded linear operator $\mathtt{tr} : H^1(\Omega) \to L^2(\partial\Omega)$, with the property that $\mathtt{tr}(v)$ and $v|_{\partial\Omega}$ coincide on $\partial\Omega$ for $v \in C(\overline{\Omega}) \cap H^1(\Omega)$.*

Remark 1.13. *By definition of a bounded linear operator, the trace theorem implies that*

$$\|\mathtt{tr}(v)\|_{L^2(\partial\Omega)} \leq C\|v\|_{H_1(\Omega)} \quad \text{for all } v \in H^1(\Omega)\ ,$$

with C independent of v. Because $\mathtt{tr}(v)$ coincides with v on $\partial\Omega$ for all sufficiently smooth v, this is often written, more flippantly, as

$$\|v\|_{L^2(\partial\Omega)} \leq C\|v\|_{H^1(\Omega)}\ , \tag{1.1}$$

although we have to bear in mind that v may not have point values on all of $\partial\Omega$.

We shall give a more sophisticated version of this theorem below, which is valid for a more general class of boundaries, which we now define.

First recall that for any domain $\mathcal{D} \subseteq \mathbb{R}^n$, a function $f : \mathcal{D} \to \mathbb{R}$ is called Lipschitz continuous if $f \in C(\mathcal{D})$ and

$$\sup_{x,y\in\mathcal{D},x\neq y} \frac{|f(x) - f(y)|}{|x - y|} < \infty\ .$$

Definition 1.14. *A bounded domain $\Omega \subset \mathbb{R}^n$ is called Lipschitz (more precisely, has a Lipschitz boundary, $\partial\Omega$) if for all $x \in \partial\Omega$, there is an orthonormal coordinate system and a rectangle $\mathcal{B}(x)$ in this coordinate system such that $x \in \mathcal{B}(x)$ and such that $\partial\Omega \cap \mathcal{B}(x)$ is the graph of a Lipschitz function of $n-1$ variables, and $\Omega \cap \mathcal{B}(x)$ lies below this graph.*

Expressing this more explicitly, the definition implies that for each $x \in \partial\Omega$, we can choose an orthonormal coordinate system $\xi = (\xi_1, ..., \xi_n)$ and a vector $\mathbf{a} \in \mathbb{R}^n_+$ such that the rectangle

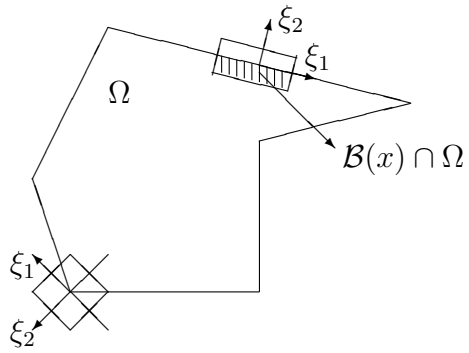
$$\mathcal{B}(x) := \{\xi \in \mathbb{R}^n : -a_j < \xi_j < a_j,\ \forall j = 1, ..., n\}$$

contains x. Moreover there exists a Lipschitz continuous function φ defined on the $n-1$ dimensional rectangle $\mathcal{B}(x)' := \{\xi' \in \mathbb{R}^{n-1} : -a_j < \xi'_j < a_j,\ j = 1, ..., n-1\}$ such that

$$\begin{aligned} \Omega \cap \mathcal{B}(x) &= \{\xi \in \mathcal{B}(x) : \xi_n < \varphi(\xi'),\ \xi' \in \mathcal{B}(x)'\} \\ \text{and} \quad \partial\Omega \cap \mathcal{B}(x) &= \{\xi \in \mathcal{B}(x) : \xi_n = \varphi(\xi'),\ \xi' \in \mathcal{B}(x)'\}\ . \end{aligned}$$

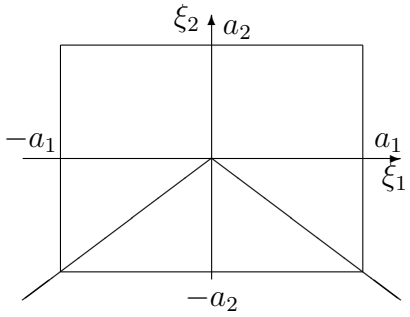
Remark 1.15. *Since the definition of Lipschitz domain is given for the case when Ω is bounded, the boundary $\partial\Omega$ can be covered by a finite set of rectanges, inside each of which $\partial\Omega$ is the graph of a Lipschitz function.*

Exercise 1.16. *A polygon in $\mathbb{R}^2$, is Lipschitz. It does not have to be convex.*



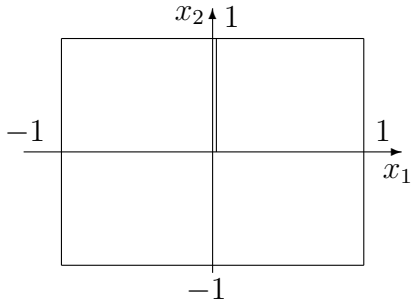
When x is not a corner point, choose ξ_1 in direction of the tangent, and ξ_2 in the direction of the normal, with the correct choice of orientation (as shown). Then $\varphi(\xi_1) = 0$ is the required Lipschitz function.

When x is a corner choose an axis system to get the "local picture":

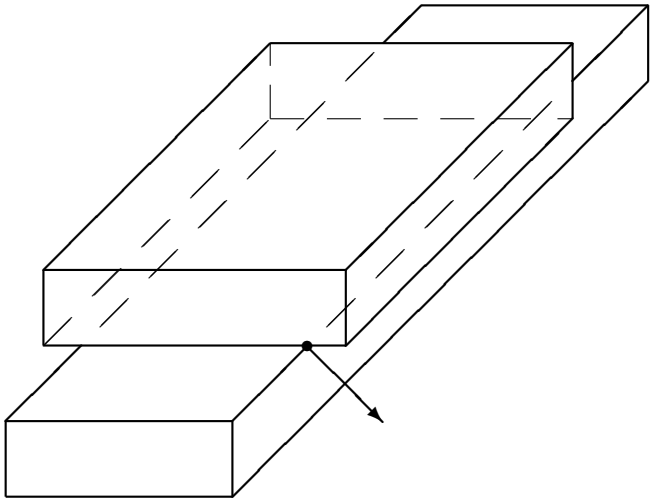


Show that the resulting function φ is Lipschitz.

Example 1.17. *The slit domain $\Omega = (-1,1) \times (-1,1) \setminus \{(0,y) : y \in (0,1)\}$ is not Lipschitz.*



Example 1.18. *The following is an example of a non-Lipschitz domain in* $\mathbb{R}^3$*, showing that not all polyhedra in* $\mathbb{R}^3$ *are Lipschitz.*



It can be shown that a Lipschitz continuous function has an L^∞ gradient. (This is Rademacher's theorem (see [McL, p.96] or [BS, p. 39] and allows an easy definition of the integral over the boundary of a Lipschitz domain as the following example shows.)

Example 1.19. *In the simplest case when the whole of* $\partial\Omega$ *is the graph of a Lipschitz function* φ*, every* $x \in \partial\Omega$ *may be written* $x = (x', \varphi(x'))$ *for* $x' \in \mathbb{R}^{n-1}$*. Then the tangent plane to* $\partial\Omega$ *at* $(x', \varphi(x'))$ *is spanned by the vectors* $(\partial/\partial x_i')\{(x', \varphi(x'))^T\} = (\mathbf{e}_i^T, \partial\phi/\partial x_i'(x'))^T$*, for* $i = 1, 2, \ldots, n-1$*, where* $\mathbf{e}_i$ *denotes the* i*th standard basis vector in* $\mathbf{R}^{n-1}$*. Hence for any suitable* f*, we define the integral by*

$$\int_{\partial\Omega} f(x)\ dS(x) = \int_{\mathbb{R}^{n-1}} f(x', \varphi(x'))\ \sqrt{1 + |\nabla\varphi(x')|^2}\ dx'\ .$$

and the outward unit normal is

$$\boldsymbol{\nu}(x') = \frac{(-\nabla\varphi(x'), 1)^T}{\sqrt{1 + |\nabla\varphi(x')|^2}}. \tag{1.2}$$

When Ω *is a general Lipschitz domain, the normal and integral may be computed locally on each piece of* $\partial\Omega$ *which is represented as the graph of a Lipschitz function.*

Through this mechanism, function spaces such as $\|\,.\,\|_{L^2(\partial\Omega)}$ make sense also for Lipschitz domains.

Theorem 1.20 (Divergence Theorem)**.** *Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain and consider $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$. Suppose each component F_k of $\mathbf{F}$ is in $C_0^\infty(\mathbb{R}^n)$. Then*

$$\int_\Omega \nabla.\mathbf{F}\ dx = \int_{\partial\Omega} \mathbf{F}.\boldsymbol{\nu}\ dS\ .$$

Proof. We give this proof to illustrate the arguments needed to deal with Lipschitz domains. It is taken from [McL, Theorem 3.34].

Suppose first that $\partial\Omega$ is the graph of a Lipschitz function on $\mathbb{R}^{n-1}$, as in Example 1.19. For $k = 1, \ldots, n-1$, consider the function

$$u_k(x') \ = \ \int_{-\infty}^{\varphi(x')} F_k(x', x_n)dx_n\ , \quad x' \in \mathbb{R}^{n-1}\ .$$

Then

$$\frac{\partial u_k}{\partial x_k}(x') \ = \ F_k(x', \varphi(x'))\frac{\partial \varphi}{\partial x_k}(x') \ + \ \int_{-\infty}^{\varphi(x')} \frac{\partial F_k}{\partial x_k}(x', x_n)dx_n\ .$$

(By Rademacher's theorem, this holds for almost all $x' \in \mathbb{R}^{n-1}$.)

Using the fact that u_k has compact support, we have

$$\int_{\mathbb{R}^{n-1}} \frac{\partial u_k}{\partial x_k}(x')dx' \ = \ 0\ .$$

and so

$$\underbrace{\int_{\mathbb{R}^{n-1}} \int_{-\infty}^{\varphi(x')} \frac{\partial F_k}{\partial x_k}(x', x_n)dx_n dx'}_{\int_\Omega \frac{\partial F_k}{\partial x_k}(x)dx\ .} \ = \ -\int_{\mathbb{R}^{n-1}} F_k(x', \varphi(x'))\frac{\partial \varphi(x')}{\partial x_k}(x')dx'\ , \tag{1.3}$$

Also

$$\int_\Omega \frac{\partial F_n}{\partial x_n}(x)dx \ = \ \int_{\mathbb{R}^{n-1}} \int_{-\infty}^{\varphi(x')} \frac{\partial F_n}{\partial x_n}(x', x_n)dx_n dx' \ = \ \int_{\mathbb{R}^{n-1}} F_n(x', \varphi(x'))dx'\ . \tag{1.4}$$

Combining (1.3) and (1.4), we obtain

$$\int_\Omega \nabla.\mathbf{F}(x)\ dx \ = \ \int_{\mathbb{R}^{n-1}} \mathbf{F}(x', \varphi(x')).\left(\frac{(-\nabla\varphi(x'), 1)}{\{1 + |\nabla\varphi|^2\}^{1/2}}\right)\{1 + |\nabla\varphi|^2\}^{1/2}dx' \ = \ \int_{\partial\Omega} \mathbf{F}.\boldsymbol{\nu}\ dS$$

(see Example 1.19).

This argument can then be extended to general Lipschitz domains by localising to pieces of the boundary, each of which is the graph of a Lipschitz function and then glueing together the estimates on each piece. We do not give the technical details here, but the argument is not difficult - see [McL, Thm 3.34] □

Remark 1.21. *By using the density and extension arguments, the divergence theorem can be extended also to the case when each component of* $\mathbf{F}$ *is in* $H^1(\Omega)$*. (This requires also the trace theorem, Theorem 1.25 below.)*

Corollary 1.22. *If* $v, w \in H^1(\Omega)$ *then for each* $i = 1, \ldots, n$*,*

$$\int_\Omega \frac{\partial w}{\partial x_i} v \; dx \;=\; -\int_\Omega w \frac{\partial v}{\partial x_i} \; dx \;+\; \int_{\partial\Omega} w v \nu_i \; dS.$$

Proof. Put $F := wv\mathbf{e}_i$ into Theorem 1.20. □

Corollary 1.23. *Suppose* $A(x)$ *is an* $n \times n$ *matrix with* $A_{ij} \in C^1(\overline{\Omega})$ *and suppose* $u \in H^2(\Omega)$*,* $v \in H^1(\Omega)$*. Then*

$$\int_\Omega [\nabla.(A\nabla u)] v \; dx \;=\; -\int_\Omega (A\nabla u).(\nabla v) \; dx \;+\; \int_{\partial\Omega} [(A\nabla u).\nu] \, v \; dS.$$

Proof. Apply Corollary 1.22 with $w = (A\nabla u)_i$ and sum over $i = 1, \ldots, n$. □

As the first example of a boundary-value problem we consider:

Example 1.24.

$$\left.\begin{array}{ll} -\nabla.A\nabla u + \mathbf{b}.\nabla u + b_0 u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{array}\right\} (D)$$

with $A = (A_{ij}) \in C^1(\overline{\Omega})$*,* $b_i \in C(\overline{\Omega})$*,* $i = 0, ..., n$*,* $f \in C(\overline{\Omega})$*.*

If u *solves (D) then multiply by* $v \in H_0^1(\Omega)$ *and integrating, applying Corollary 1.23, we get*

$$\underbrace{\int_\Omega \{(A\nabla u).\nabla v + (\mathbf{b}.\nabla u)v + b_0 uv\}}_{a(u,v)} = \underbrace{\int_\Omega fv}_{F(v)} \; .$$

The weak form of D is:

$$\textit{Find} \quad u \in H_0^1(\Omega) \quad \textit{such that} \quad a(u,v) = F(v) \quad \textit{for all} \quad v \in H_0^1(\Omega) \; .$$

Note that $A_{i,j}, b_i \in L^\infty(\Omega)$*,* $i, j = 0, ..., n$ *and* $f \in L^2(\Omega)$ *would be sufficient for the weak form to be well-defined.*

Before proceeding we state a generalisation of the trace theorem which is valid for all Lipschitz domains.

Theorem 1.25 (Trace Theorem, Version 2)**.** *If* Ω *is Lipschitz then there exists a constant* $C > 0$ *such that*

$$\|u\|_{L^2(\partial\Omega)} \leq C \|u\|_{L^2(\Omega)}^{\frac{1}{2}} \|u\|_{H^1(\Omega)}^{\frac{1}{2}} \qquad \forall u \in H^1(\Omega) \; ,$$

where the boundary values of u *are to be understood in the sense of the image of a trace operator, as described above.*

Proof. This is stated without proof in [BS, Theorem 1.6.6]. A proof can be constructed by following for example the arguments in §1.5 of Grisvard's book [GR].

Recall the definition of a Lipschitz domain above and see Remark 1.15 which follows it. This implies that $\partial\Omega$ can be covered by an overlapping set of open $n-$ dimensional rectanges $\mathcal{B}(x^k)$, for some points $x^k \in \partial\Omega$, $k = 1, \ldots, s$. In each $\mathcal{B}(x^k)$, $\partial\Omega \cap \mathcal{B}(x_j)$ is the graph of a Lipschitz function with outward unit normal denoted $\boldsymbol{\nu}^k \in L^\infty$. With the notation as in Definition 1.14, for each $k = 1, \ldots, s$, let $\boldsymbol{\xi}^k$ denote the local coordinate system in $\mathcal{B}(x^k)$ and choose unit vectors $\boldsymbol{\mu}^k \in \mathbb{R}^n$ in the direction ξ_n^k. Then recalling (1.2), we have

$$\boldsymbol{\mu}^k \cdot \boldsymbol{\nu}^k(\xi') \;=\; \frac{1}{\sqrt{1 + |\nabla\phi(\xi')|^2}} \;,$$

for almost all $\xi' \in \mathcal{B}(x^k)'$. Now since ϕ is Lipschitz in $\mathcal{B}(x^k)'$, we have, for almost all ξ', $|\nabla\phi^k(\xi')| \;\leq\; L^k$ where L^k is the Lipschitz constant for ϕ on $\mathcal{B}(x^k)'$. Thus

$$\boldsymbol{\mu}^k \cdot \boldsymbol{\nu}^k(\xi') \;\geq\; \frac{1}{\sqrt{1 + (L^k)^2}} \quad , \quad \text{almost all} \quad \xi' \in \mathcal{B}(x^k)'.$$

Now we choose a partition of unity $\left\{\theta^k : k = 1, \ldots, s\right\}$. (This is a set of functions which are $\mathcal{C}^\infty$, and take values between 0 and 1, such that supp $\theta^k \;\subset\; \mathcal{B}(x^k)$ and $\sum_{k=1}^s \theta_k = 1$ on $\partial\Omega$.) The existence of such P.O.U.'s is standard in analysis (see, e.g. [McL, p.83]). Then define, for $x \in \partial\Omega$,

$$\boldsymbol{\mu}(x) \;=\; \sum_{k=1}^{s} \theta^k(x)\boldsymbol{\mu}^k \;.$$

This is a $\mathcal{C}^\infty$ function defined on $\partial\Omega$ with values in $\mathbb{R}^n$. Then any $x \in \partial\Omega \cap \mathcal{B}(x^k)$ may be parametrised by $\xi' \in \mathcal{B}(x^k)'$, and we have

$$\begin{aligned}
\boldsymbol{\mu} \cdot \boldsymbol{\nu}(x) \quad &= \quad \sum_k (\boldsymbol{\mu}^k \cdot \boldsymbol{\nu}^k(\xi'))\; \theta^k(x) \;\geq \min_k \left(\frac{1}{\sqrt{1 + (L^k)^2}}\right) \sum_k \theta^k(x) \\
&= \quad \min_k \left(\frac{1}{\sqrt{1 + (L^k)^2}}\right) =: \delta > 0 \;.
\end{aligned}$$

Hence we have found a $\mathcal{C}^\infty$ vector-valued function which satisfies $\boldsymbol{\mu}\cdot\boldsymbol{\nu} \;\geq\; \delta > 0$ almost everywhere on $\partial\Omega$.

Then for $u \in C^\infty(\overline{\Omega})$, we have, by the Divergence theorem, Cauchy-Schwarz and the

Poincaré inequality,

$$\begin{aligned}
\int_{\partial\Omega} u^2 &\leq \frac{1}{\delta}\int_{\partial\Omega} u^2 \boldsymbol{\mu}\cdot\boldsymbol{\nu} \;=\; \frac{1}{\delta}\int_{\Omega} \nabla\cdot(u^2\boldsymbol{\mu}) \\
&= \frac{1}{\delta}\int_{\Omega} \left(u^2\nabla\cdot\boldsymbol{\mu} + 2u\nabla u\cdot\boldsymbol{\mu}\right) \\
&\leq C\int_{\Omega} \left(u^2 + |u||\nabla u|\right) \\
&\leq C\left(\|u\|^2_{L^2(\Omega)} \;+\; \|u\|_{L^2(\Omega)}|u|_{H^1(\Omega)}\right) \\
&\leq C\left(\|u\|_{L^2(\Omega)}\|u\|_{H^1(\Omega)}\right)\;,
\end{aligned}$$

where C denotes a "generic constant" whose value may change from line to line but is always independent of u.

□

Remark 1.26. *Note that our trace theorem Version 2, implies as a consequence the trace theorem, Version 1.*

Note also that $H^1_0(\Omega) = \{u \in H^1(\Omega) : \mathtt{tr}(u) \;=\; 0 \text{ in } L^2(\partial\Omega)\}$.

2 Abstract Theory of Variational Problems

In this chapter we shall study the abstract problem:

Find $u \in V$ such that

$$a(u,v) \;=\; F(v)\;, \quad \text{for all } v \in V\;, \tag{2.1}$$

where V is a real Hilbert space (or a closed subspace of a bigger Hilbert space H with norm $\|\cdot\|$ and inner product $(\cdot,\cdot)$) and $a : V \times V \mapsto \mathbb{R}$ is a bilinear form (i.e. linear in each argument).

We shall also study the approximation of (2.1) in a finite dimensional space (the finite element method is an example of this).

We shall make use of the dual space V' of all bounded linear functionals on V, with norm $\|\cdot\|_{V'}$ and inner product $(\cdot,\cdot)_{V'}$. A key theorem is the Riesz Representation theorem (Theorem A4).

Definition 2.1. *The bilinear form a is called* bounded *if there is a constant C such that*

$$|a(u,v)| \;\le\; C\|u\|\|v\|, \text{ for all } u,v \in V.$$

Definition 2.2. *The bilinear form a is called* coercive *if there is a constant $\varepsilon > 0$ such that*

$$a(v,v) \;\ge\; \varepsilon\|v\|^2, \text{ for all } v \in V\;, \quad v \neq 0.$$

Example 2.3. *(i) Put $A = I$, $\mathbf{b} = 0$ and $b = 1$ in Example 1.24. Then $a(u,v) = (u,v)_{H^1(\Omega)}$ so boundedness and coercivity in $H^1(\Omega)$ follow immediately.*

(ii) Put $A = I$, $\mathbf{b} = 0$ and $b = 0$ in Example 1.24. (Poisson's equation). Then $a(u,v) = \int_\Omega \nabla u.\nabla v$. Boundedness in $H^1(\Omega)$ follows from Cauchy-Schwarz. This bilinear form is not coercive on $H^1(\Omega)$ because with $v = 1$, we have $a(v,v) = 0 < \varepsilon\|v\|^2_{H^1(\Omega)}$ for any $\varepsilon > 0$. However it is coercive on $H^1_0(\Omega)$ because of the Poincaré inequality (Theorem 1.9.)

The Lax-Milgram Lemma provides the theory of (2.1) under the assumption that a is bounded and coercive. In these notes we prove a more general version where coercivity is replaced with the following conditions.

Definition 2.4. *The bilinear form a is said to satisfy the* Babuška conditions *if there exists $\varepsilon > 0$ such that*

$$\inf_{0\neq u\in V} \sup_{0\neq v\in V} \frac{a(u,v)}{\|u\|\|v\|} \;\ge\; \varepsilon \tag{Ba1}$$

$$\text{and} \quad \sup_{u\neq 0} \frac{a(u,v)}{\|u\|} \;> 0, \text{ for all } v \in V\;, v \neq 0. \tag{Ba2}$$

Exercise 2.5. *Show that if a is coercive then it satisfies the Babuška conditions. Show also that the condition (Ba2) yields the following statement :*

$$a(u,v) = 0 \quad \textit{for all } \; u \in V \quad \textit{implies} \quad v = 0 \; .$$

Theorem 2.6 (Generalised Lax-Milgram Lemma)**.** *Let a be a bounded bilinear form satisfying the Babuška conditions. Then $\forall F \in V'$ the problem* (2.1) *has a unique solution u satisfying*

$$\|u\| \leq \frac{1}{\varepsilon}\|F\|_{V'} \; .$$

Proof. Define a mapping $\mathcal{A} : V \to V'$ by

$$(\mathcal{A}u)(v) \; := \; a(u,v) \; .$$

The problem (2.1) may be recast as the problem of finding $u \in V$ satisfying the equation

$$\mathcal{A}u = F \quad \text{in} \quad V'. \tag{2.2}$$

It is clear that $\mathcal{A}$ is a linear operator from V to V', and, because of the boundedness of a,

$$\|\mathcal{A}u\|_{V'} \; = \; \sup_{0 \neq v \in V} \frac{|\mathcal{A}u(v)|}{\|v\|} \; = \; \sup_{0 \neq v \in V} \frac{|a(u,v)|}{\|v\|} \; \leq \; C\|u\| \; .$$

So $\mathcal{A}$ is a bounded linear operator and $\|\mathcal{A}\| \leq C$.

To complete the proof we shall use Theorem A2 to show that $\mathcal{A}^{-1} : V' \to V$ exists and its norm is bounded by ε^{-1}. This requires that we show $\mathcal{A}$ is bijective.

Note first that, using condition (Ba1),

$$\|\mathcal{A}u\| \; = \; \sup_{0 \neq v \in V} \left(\frac{|a(u,v)|}{\|u\|\|v\|} \right) \|u\| \; \geq \; \epsilon\|u\| \; ,$$

which implies $\|\mathcal{A}\| \; \geq \; \epsilon$.

Then injectivity of $\mathcal{A}$ follows, since if $\mathcal{A}u_1 = \mathcal{A}u_2$ for some $u_1, u_2 \in V$, then

$$0 \; = \; \|\mathcal{A}u_1 - \mathcal{A}u_2\|_{V'} \; = \; \|\mathcal{A}(u_1 - u_2)\|_{V'} \; \geq \; \varepsilon\|u_1 - u_2\|_V$$

which implies $u_1 = u_2$.

To show A is surjective, consider the set

$$W = \{\mathcal{A}u : u \in V\} \; \subseteq \; V' \; .$$

We want to show $W = V'$. First we claim W is closed in V'. To see why, consider a sequence $\{\mathcal{A}u_n\}$ in W with $\mathcal{A}u_n \to w_* \in V'$. Then, for all n, n',

$$\|\mathcal{A}u_n - \mathcal{A}u_{n'}\|_{V'} \;=\; \|\mathcal{A}(u_n - u_{n'})\|_{V'} \;\geq\; \epsilon\|u_n - u_{n'}\|_V$$

Hence, since the sequence $\{\mathcal{A}u_n\}$ is Cauchy in V', so must $\{u_n\}$ be Cauchy in V and so by completeness, $u_n \to u_*$, for some $u_* \in V$. Now,

$$\|\mathcal{A}u_n - \mathcal{A}u_*\|_{V'} \;=\; \|\mathcal{A}(u_n - u_*)\|_{V'} \;\leq\; C\|u_n - u^*\| \;\to 0.$$

So $\mathcal{A}u_n \to \mathcal{A}u_* \in W$ and hence W is closed. Appealing to the Hilbert space projection theorem, Theorem A3, we have

$$V' \;=\; W \oplus W^{\perp}\,.$$

Now, if $\mathcal{A}$ is not surjective then $W \neq V'$ and there exists $0 \neq g \in W^{\perp}$, i.e. $(\mathcal{A}u, g)_{V'} = 0$, for all $u \in V$. Using the Reisz representation theorem (Theorem A4), we have $(\tau(\mathcal{A}u), \tau g) = 0$ and thus $\mathcal{A}u(\tau g) = 0$, and hence $a(u, \tau g) = 0$. Since this holds for all $u \in V$ and since $\tau g \neq 0$, this contradicts (Ba2) (see Exercise 2.5). Thus $\mathcal{A}$ is surjective and by Theorem A.2, $\mathcal{A}^{-1}$ exists and is bounded. This means that (2.2) has a unique solution and

$$\varepsilon\|u\|_H \;\leq\; \|\mathcal{A}u\|_{V'} \;=\; \|F\|_{V'}$$

and dividing by ε completes the proof. □

Exercise 2.7. *Show that condition (Ba1) is equivalent to:*

$$\inf_{0\neq u\in V} \sup_{0\neq v\in V} \frac{|a(u,v)|}{\|u\|\|v\|} \;\geq\; \varepsilon \tag{Ba1'}$$

Show that the condition (Ba2) is equivalent to the condition:

$$\textit{and} \quad \sup_{u\neq 0} \frac{|a(u,v)|}{\|u\|} \;> 0, \;\textit{for all } v \in V \;, v \neq 0. \tag{Ba2'}$$

Hint: Note that (Ba1) implies the statement: "For all $u \in V$, $u \neq 0$, there exists a sequence $v_m \in V\backslash\{0\}$ such that

$$\frac{a(u, v_m)}{\|u\|\|v_m\|} \;\geq\; \varepsilon - \frac{1}{m}\,''\,.$$

Abstract Finite Element Method

Now we consider approximation of (2.1) in a finite dimensional subspace V_h of V, $\dim(V_h) = N$. That is we seek to solve the problem:

$$\text{Find } u_h \in V_h \text{ such that } a(u_h, v_h) = F(v_h) \quad \forall v_h \in V_h. \tag{2.3}$$

u_h is often called the Galerkin solution to (2.1).

Let $\{\varphi_i : i = 1, ..., N\}$ be a basis for V_h, then (2.3) is equivalent to the linear system

$$K\mathbf{U} = \mathbf{f}, \tag{2.4}$$

where $K_{ij} = a(\varphi_j, \varphi_i)$, $\quad u_h = \sum_{j=1}^{N} U_j \varphi_j \quad$ and $f_i = F(\varphi_i)$.

Theorem 2.8. *Suppose a is bounded on V and also satisfies the "discrete* $\inf - \sup$ *condition",*

$$\inf_{0 \neq u_h \in V_h} \sup_{0 \neq v_h \in V_h} \frac{a(u_h, v_h)}{\|u_h\| \|v_h\|} \;\geq\; \varepsilon_h \;>\; 0. \tag{Ba1h}$$

Then for all $F \in V_h'$ problem (2.3) has a unique solution u_h and $\|u_h\| \;\leq\; \dfrac{1}{\varepsilon_h} \|F\|_{V_h'}$.

Proof. We only have to verify (Ba2h) (i.e. the condition that (Ba2) holds in V_h) and the result follows from Theorem 2.6. Note that by Exercise 2.7, this is equivalent to verification of (Ba2') in V_h, i.e. we have to show

$$\text{and} \quad \sup_{0 \neq u_h \in V_h} \frac{|a(u_h, v_h)|}{\|u_h\|} > 0, \text{ for all } v_h \in V_h \,, v_h \neq 0. \tag{Ba2h'}$$

Consider the "stiffness matrix" K. If K is singular then $K\mathbf{U} = \mathbf{0}$ for some vector $\mathbf{U} \neq 0$ and then $u_h := \sum_{j=1}^{N} U_j \varphi_j \;\in\; V_h \backslash \{0\}$ satisfies $a(u_h, \varphi_i) = 0$ for all $i = 1, ..., N$ and so $\sup_{0 \neq v_h \in V_h} \dfrac{a(u_h, v_h)}{\|v_h\|_V} = 0$, which contradicts (Ba1h). So, K must be non-singular and K^T (the transpose of K) must also be non-singular.

Now suppose (Ba2h') does not hold then there exists $v_h \in V_h \backslash \{0\}$ such that

$$|a(u_h, v_h)| = 0 \quad \text{for all} \quad u_h \in V_h \,.$$

With V_i satisfying $v_h = \sum_i V_i \phi_i$, we then have

$$0 \;=\; a(\varphi_j, v_h) \;=\; (K^T \mathbf{V})_j \quad \text{for all } j = 1, \ldots, N \,,$$

which contradicts the non-singularity of K^T . Hence (Ba2h') (and (Ba2h)) must hold and the result follows.

□

The next result is the fundamental error estimate for finite element methods.

Theorem 2.9 ("Céa's Lemma")**.** *Let u be a solution of* (2.1)*. Then under the assumptions of Theorem 2.8,*

$$\|u - u_h\| \ \leq \ \left(1 + \frac{C}{\varepsilon_h}\right) \inf_{w_h \in V_h} \|u - w_h\| \qquad (\text{"quasi-optimality"})$$

where C is the boundedness constant of a and ε_h is the constant appearing in (Ba1h).

Proof. Since V_h is a subspace of V, using (2.1) on V_h and (2.3) we have

$$a(u - u_h, v_h) = 0 \qquad \forall v_h \in V_h \qquad \text{"Galerkin Orthogonality"}. \tag{2.5}$$

Thus for all $w_h,\ v_h \in V_h$,

$$a(u - w_h, v_h) = a(u - u_h, v_h) + a(u_h - w_h, v_h) = a(u_h - w_h, v_h). \tag{2.6}$$

So, by (Ba2h), if $u_h - w_h \neq 0$, then

$$\begin{aligned} \varepsilon_h \|u_h - w_h\| \ &\leq \ \sup_{0 \neq v_h \in V_h} \left\{ \frac{a(u_h - w_h, v_h)}{\|u_h - w_h\|_V \|v_h\|_V} \right\} \|u_h - w_h\|_V \\ &= \ \sup_{0 \neq v_h \in V_h} \left\{ \frac{a(u - w_h, v_h)}{\|v_h\|_V} \right\} \quad \text{by} \quad (2.6) \\ &\leq \ C\|u - w_n\|_V \qquad \text{using boundedness of } a\ . \end{aligned}$$

Hence

$$\|u - u_h\| \ \leq \ \|u - w_n\| + \|u_h - w_h\| \ \leq \ \left(1 + \frac{C}{\varepsilon_h}\right) \|u - w_h\| \quad \forall w_h \in V_h.$$

Now take inf on RHS to get the result. □

Exercise 2.10. *With the same notation as in Theorem 2.9. Suppose a is bounded,* **symmetric** *and coercive on V. Prove the sharper error estimate:*

$$\|u - u_h\| \leq \left(\frac{C}{\alpha}\right)^{1/2} \inf_{w_h \in V_h} \|u - w_h\| \ ,$$

where C is the boundedness constant and α is the coercivity constant. [Hint: show that the bilinear form induced by a is an inner product on V.]

Application: to bilinear form appearing in Example 1.24

The remainder of this chapter is devoted to study of this application.

Here a is given by

$$a(u,v) = \int_\Omega (A\nabla u).\nabla v + (\mathbf{b}.\nabla u)v + b_0 uv \tag{2.7}$$

We assume that each component of $A, \mathbf{b}$ and b_0 is in $L^\infty(\Omega)$. Moreover we assume that A is uniformly posiive definite, which means that A is symmetric and there exists $\alpha > 0$ such that

$$\boldsymbol{\xi}.(A(\mathbf{x})\boldsymbol{\xi}) \;\geq\; \alpha\, \boldsymbol{\xi}.\boldsymbol{\xi} \qquad \forall \boldsymbol{\xi} \in \mathbb{R}^n, \quad \mathbf{x} \text{ a.e. in } \Omega. \tag{2.8}$$

A special case would be $A(x) = \alpha_1(x)I$ where $\alpha_1(x) \geq \alpha > 0$, x a.e. in Ω. More generally (2.8) says that $A(x)$ should be symmetric and all eigenvalues of $A(x)$ should be bounded below by $\alpha > 0$, for all $x \in \Omega$.

Lemma 2.11. *Under the above assumptions,*

(i) a is bounded on $H^1(\Omega)$.

(ii) When $\mathbf{b} = 0$ and $b_0 \geq 0$ then

$$a(v,v) \geq \alpha |v|^2_{H^1(\Omega)} \tag{2.9}$$

and in particular, a is coercive on $H^1_0(\Omega)$ because of Remark 1.11 .

(iii) In general, there exists a scalar $K \geq 0$ such that

$$a(v,v) + K(v,v)_{L^2(\Omega)} \;\geq\; \frac{\alpha}{2}\|v\|^2_{H^1(\Omega)} \qquad \forall v \in H^1(\Omega). \tag{2.10}$$

This is called Gårding's inequality. In general K depends on $\mathbf{b}, b_0$ and α.

Exercise 2.12. *Prove parts (i) and (ii) of the Lemma. They are simple applications of Cauchy-Schwarz and of the assumed property* (2.8)*.*

Proof. Proof of (iii):

First note that by Cauchy-Schwarz in $\mathbb{R}^n$,

$$\left|\int_\Omega (\mathbf{b}.\nabla v)v\right| \leq \int_\Omega |\mathbf{b}||\nabla v||v| \;\leq \underbrace{\sqrt{n}\max_i \|b_i\|_{L^\infty(\Omega)}}_{=:B} \int_\Omega |\nabla v||v| \;.$$

So

$$\left|\int_\Omega (\mathbf{b}.\nabla v)v\right| \leq B|v|_{H^1(\Omega)}\|v\|_{L^2(\Omega)} \quad \text{by Cauchy-Schwarz in } L^2(\Omega).$$

Combine with result (ii) above to get

$$\begin{aligned} a(v,v) + K\|v\|^2_{L^2(\Omega)} &\geq \int_\Omega (A\nabla v).\nabla v - \left|\int_\Omega (b\nabla v)v\right| + \int_\Omega (b_0 + K)|v|^2 \\ &\geq \alpha|v|^2_{H^1(\Omega)} - B|v|_{H'(\Omega)}\|v\|_{L^2(\Omega)} + (\beta + K)\|v\|^2_{L^2(\Omega)}, \end{aligned}$$

where $\beta = \operatorname{ess\,inf}(b_0) \geq -\|b_0\|_{L^\infty(\Omega)}$.

To complete the proof, note that for all $p, q \in \mathbb{R}$, $\quad pq \leq \frac{1}{2}(p^2 + q^2)$. Hence, for all $\delta > 0$,

$$pq \;=\; (\sqrt{\delta}p)\left(\frac{q}{\sqrt{\delta}}\right) \;\leq\; \frac{1}{2}\left(\delta p^2 + \frac{q^2}{\delta}\right) \;. \tag{2.11}$$

(This is known as the Arithmetic-Geometric Mean Inequality.) Using this,

$$\underbrace{B|v|_{H^1(\Omega)}}_{p}\underbrace{\|v\|_{L^2(\Omega)}}_{q} \;\leq\; \frac{\delta B^2 |v|^2_{H^1(\Omega)}}{2} + \frac{1}{2\delta}\|v\|^2_{L^2(\Omega)} \;,$$

and so,

$$a(v,v) + K\|v\|^2_{L^2(\Omega)} \;\geq\; \left(\alpha - \frac{\delta B^2}{2}\right)|v|^2_{H^1(\Omega)} + \left(\beta + K - \frac{1}{2\delta}\right)\|v\|^2_{L^2(\Omega)}.$$

Now if $B \neq 0$, choose $\delta \;=\; \dfrac{\alpha}{B^2}$ and $K \;\geq\; \left(\dfrac{1}{2\delta} - \beta + \dfrac{\alpha}{2}\right)$ to obtain the result. If $B = 0$, just choose any $K \geq \frac{\alpha}{2} - \beta$. □

Remark 2.13. *In the special case of the Helmholtz equation, $A = I$, (so $\alpha = 1$), $\mathbf{b} = \mathbf{0}$ (so $B = 0$) and $b_0 = -k^2 = \beta$. So in Lemma 2.11, K must grow with $O(k^2)$ in order to ensure Gårding's inequality. (k was the frequency of the time oscillation in the original wave equation - discussed in lecture 1).*

We finish this section by considering the finite element approximation of the problem

$$\text{Find} \quad u \in H^1_0(\Omega) \quad \text{such that} \quad a(u,v) = (f,v)_{L^2(\Omega)} \quad \text{for all} \quad v \in H^1_0(\Omega) \;, \tag{2.12}$$

where a is given by (2.7). The finite element approximation is obtained by choosing a finite dimensional subspace V_h of $H^1_0(\Omega)$ and the seeking the solution of the problem:

$$\text{Find} \quad u_h \in V_h \quad \text{such that} \quad a(u_h, v_h) = (f, v_h)_{L^2(\Omega)} \quad \text{for all} \quad v_h \in V_h \;. \tag{2.13}$$

Theorem 2.14. *Suppose the following assumptions are satisfied.*

(A1) Well-posedness: *For all $f \in L^2(\Omega)$ each of the problems:*

$$\begin{array}{llllll} \textit{find} & u \in H^1_0(\Omega) & \textit{such that} & a(u,v) = (f,v)_{L^2(\Omega)} & \textit{for all} & v \in H^1_0(\Omega) \\ \textit{find} & \tilde{u} \in H^1_0(\Omega) & \textit{such that} & a(v,\tilde{u}) = (f,v)_{L^2(\Omega)} & \textit{for all} & v \in H^1_0(\Omega) \quad \textit{“adjoint problem”} \end{array}$$

have unique solutions $u, \tilde{u} \in H^1_0(\Omega)$.

(A2) Regularity *The solutions u and $\tilde{u}$ in* **(A2)** *satisfy:*

$$\max\{\|u\|_{H^2(\Omega)} \;,\; \|\tilde{u}\|_{H^2(\Omega)}\} \;\leq\; C_R\|f\|_{L^2(\Omega)} \;.$$

(A3) Approximation *The space V_h provides convergent approximations to any sufficiently smooth $v \in H_0^1(\Omega)$, in the following sense:*

$$\inf_{v_h \in V_h} \|v - v_h\|_{H^1(\Omega)} \;\leq\; C_A h |v|_{H^2(\Omega)} \quad \textit{for all} \quad v \in H_0^1(\Omega) \cap H^2(\Omega).$$

Then, for h sufficiently small, the problem (2.13) *has a unique solution u_h and*

$$\|u - u_h\|_{H^1(\Omega)} \;\leq\; \left(1 + \frac{C}{\epsilon_h}\right) C_A C_R h \|f\|_{L^2(\Omega)} \;, \tag{2.14}$$

where C is the boundedness constant for a and ϵ_h is given below. The threshold of smallness for h is also given below.

Remark 2.15. *Sufficient conditions for* **(A1)** *are in Theorem 2.6. Sufficient conditions for* **(A2)** *and* **(A3)** *will be investigated in later lectures.*

Proof. Provided we can verify the discrete inf-sup condition (Ba1h) of Theorem 2.8, then from Céa's lemma (Theorem 2.9), we can deduce

$$\|u - u_h\|_{H^1(\Omega)} \;\leq\; \left(1 + \frac{C}{\epsilon_h}\right) \inf_{w_h \in V_h} \|u - w_h\|_{H^1(\Omega)}$$

where C is the boundedness constant of a, and (2.14) follows from the approximation and regularity assumptions.

To verify (Ba1h), let K, α be as in (2.10) and choose any $u_h \in V_h \backslash \{0\}$. Then by **(A1)**, there exists $w \in H_0^1(\Omega)$ which solves the adjoint problem with the special right-hand side:

$$a(v, w) \;=\; K(u_h, v)_{L^2(\Omega)} \quad \text{for all} \quad v \in H_0^1(\Omega) \tag{2.15}$$

Then **(A2)** implies that

$$\|w\|_{H^2(\Omega)} \;\leq\; C_R K \|u_h\|_{L^2(\Omega)} \;. \tag{2.16}$$

Now choose $v_h = u_h + w_h \in V_h$ (where w_h is to be chosen below) and note that

$$\begin{aligned} a(u_h, v_h) &= a(u_h, u_h + w) + a(u_h, w - w_h) \\ &= a(u_h, u_h) + K\|u_h\|^2_{L^2(\Omega)} + a(u_h, w - w_h) \\ &\geq \frac{\alpha}{2}\|u_h\|^2_{H^1(\Omega)} - C\|u_h\|_{H^1(\Omega)}\|w - w_h\|_{H^1(\Omega)} \;. \end{aligned} \tag{2.17}$$

Now by **(A3)** and (2.16), there exists $w_h \in V_h$ such that

$$\|w - w_h\|_{H^1(\Omega)} \;\leq\; C_A h |w|_{H^2(\Omega)} \;\leq\; C_A C_R K h \|u_h\|_{L^2(\Omega)} \;. \tag{2.18}$$

So

$$a(u_h, v_h) \;\geq\; \left(\frac{\alpha}{2} - C C_A C_R K h\right) \|u_h\|^2_{H^1(\Omega)} \;.$$

Also

$$\begin{aligned} \|v_h\|_{H^1(\Omega)} &\leq \|u_h\|_{H^1(\Omega)} + \|w\|_{H^1(\Omega)} + \|w - w_h\|_{H^1(\Omega)} \\ &\leq (1 + C_R K + C_A C_R K h)\|u_h\|_{H^1(\Omega)} \ . \end{aligned} \tag{2.19}$$

Moreover, using boundedness of a and the Gårding inequality,

$$C\|u_h\|_{H^1(\Omega)}\|v_h\|_{H^1(\Omega)} \ \geq \ a(u_h, v_h) \ = \ a(u_h, u_h) + (K - \mathcal{O}(h))\|u_h\|^2_{L^2(\Omega)} \ \geq \ (\frac{\alpha}{2} - \mathcal{O}(h))\|u_h\|^2_{H^1(\Omega)} \, ,$$

for h sufficiently small.

Since $u_h \neq 0$, this implies $v_h \neq 0$ and, moreover, by (2.17), (2.18) and (2.19) we have

$$\sup_{v_h \in V_h \setminus \{0\}} \frac{a(u_h, v_h)}{\|u_h\|_{H^1(\Omega)}\|v_h\|_{H^1(\Omega)}} \ \geq \ \underbrace{\frac{(\frac{\alpha}{2} - CC_A C_R K h)}{(1 + KC_R + C_A C_R K h)}}_{=:\varepsilon_h} \, ,$$

which implies (Ba1h) provided h is below a certain threshold, given by

$$h \ < \ \frac{\alpha}{2CC_A C_R K} \ .$$

□

Remark 2.16. *In the case of the Helmholtz equation because K grows with $O(k^2)$, h must decrease with at least $O(k^{-2})$ to guarantee the existence of the finite element solution.*

3 Construction of finite element spaces

Definition 3.1. *A finite element is a triple $(K, \mathcal{P}_K, \mathcal{N}_K)$, where*

(i) $K \subset \mathbb{R}^n$ *is the closure of a bounded Lipshitz domain.*

(ii) $\mathcal{P}_K$ *is a finite dimensional space of functions:* $K \to \mathbb{R}$.

(iii) $\mathcal{N}_K$ *is a basis for* $\mathcal{P}_K'$, *the dual space of* $\mathcal{P}_K$.

Later we need the elements of $\mathcal{N}_K$ to be defined on a bigger space which contains $\mathcal{P}_K$.

Exercise 3.2.

1. *Show* $\dim \mathcal{P}_K' = \dim \mathcal{P}_K$.
2. *If* $\dim \mathcal{P}_K = d$ *and* $\mathcal{N}_K = \{N_1, ..., N_d\} \subset \mathcal{P}_K'$, *then the statements*

   (i) $\mathcal{N}_K$ *is a basis for* $\mathcal{P}_K'$

   and

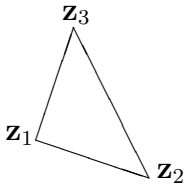
   $$(ii) \qquad \text{for all} \quad p \in \mathcal{P}_K, \quad N_i(p) = 0 \text{ for all } \quad i = 1, ..., d \quad \text{implies} \quad p = 0 \tag{3.1}$$

   are equivalent. If (3.1) holds we say "$\mathcal{N}_K$ *determines* $\mathcal{P}_K$*".*
3. *If* $\{N_1, ..., N_d\}$ *is a basis for* $\mathcal{P}_K'$ *then* $\exists$ *a basis* $\{p_1, ..., p_d\}$ *of* $\mathcal{P}_K$ *such that* $N_i(p_j) = \delta_{ij}$. *This is called the* Nodal basis.

Example 3.3. *In* $\mathbb{R}^2$.
$K =$ *triangle with vertices* $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$ - *"nodes".*
$\mathcal{P}_K = \mathbb{P}_1(2) \equiv$ *polynomials of degree* 1 *in* 2 *variables* x_1 *and* x_2 *on* K.
$\mathcal{N}_K = \{N_1, N_2, N_3\}$ *where* $N_i(p) = p(\mathbf{z}_i)$, $i = 1, 2, 3$.

$\mathcal{N}_K$ *determines* $\mathcal{P}_K$ *since if* $p(\mathbf{x}) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$ *then* $p(\mathbf{z}_i) = 0$ *for each* $i = 1, 2, 3$ *implies*

$$\underbrace{\begin{bmatrix} 1 & \mathbf{z}_1^T \\ 1 & \mathbf{z}_2^T \\ 1 & \mathbf{z}_3^T \end{bmatrix}}_{=:M} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$



then

$$\det(M) = \det\left(\begin{bmatrix} \mathbf{z}_2^T - \mathbf{z}_1^T \\ \mathbf{z}_3^T - \mathbf{z}_1^T \end{bmatrix}\right) = 2 \times \textit{area of } K$$

.

$$\textit{So} \qquad (\alpha_0, \alpha_1, \alpha_2) = \mathbf{0}^T \quad \textit{and} \quad p \equiv 0.$$

In this special case, we denote the nodal basis for $\mathcal{P}_K$ *by* λ_j, $j = 1, 2, 3$ *with* $\lambda_j(\mathbf{z}_i) = \delta_{ij}$.

Note that $\lambda_1 + \lambda_2 + \lambda_3 - 1 = 0$ *(since this holds at each of the three nodes and a plane is uniquely determined by its values at three non-collinear points). Hence* $\sum_{j=1}^{3} \lambda_j = 1$. *The* λ_j *are sometimes called "barycentric coordinates" since (again by uniqueness of linear interpolation at 3 points), any* x *may be written as the weighted sum:*

$$\mathbf{x} = \sum_{j=1}^{3} \lambda_j(\mathbf{x}) \mathbf{z}_j. \tag{3.2}$$

Example 3.4. *Higher order triangle.*
$K =$ *triangle in* $\mathbb{R}^2$.
$\mathcal{P}_K = \mathbb{P}_k(2) \equiv$ *polynomials of total degree* $\leq k$ *in 2 variables, a space of dimension* $\frac{(k+1)(k+2)}{2}$ *(e.g.* $\mathbb{P}_2(2) = \mathtt{span}\{1, x_1, x_2, x_1^2, x_1x_2, x_2^2\}$.*)*

Define $(k+1)(k+2)/2$ *functionals* $\mathcal{N}_K$ *as follows:*

For $k \geq 1$ *we always use the three functionals:*

$$p \mapsto p(\mathbf{z}_i), \quad i = 1, 2, 3. \tag{3.3}$$

introduced above.

For $k \geq 2$, *we add* $3 \times (k-1)$ *additional functionals defined by:*

$$p \mapsto \frac{1}{\mathit{length}(e)} \int_e pq \qquad \textit{for all } q \in \{\textit{a basis for } \mathbb{P}_{k-2}(1)\} \quad \textit{for each edge } e \textit{ of } K. \tag{3.4}$$

For $k \geq 3$, *we add in addition the* $(k-2)(k-1)/2$ *functionals:*

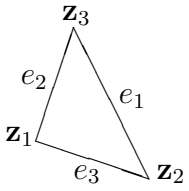
$$p \mapsto \frac{1}{\mathit{area}(K)} \int_K pq \qquad \textit{for all } q \in \{\textit{a basis for } \mathbb{P}_{k-3}(2)\}\ . \tag{3.5}$$

Note that the number of functionals is 6 *when* $k = 2$ *and is* $3 + 3(2) + 1 = 10$ *when* $k = 3$. *In general there are* $(k+1)(k+2)/2$ *functionals.*

To show $\mathcal{N}_K$ *determines* $\mathcal{P}_K$, *let* $p \in \mathcal{P}_K$ *and suppose all of (3.3)-(3.5) vanish. For any edge* e *of* K, *restrict* p *to* e *to get a polynomial of degree* k *in one variable* (s). *Then,* $\forall q \in \mathbb{P}_{k-1}(1)$, *integrate by parts to get*

$$\int_e \frac{dp}{ds} q \underset{(3.3)\ \textit{vanish}}{=} -\int_e p \underbrace{\frac{dq}{ds}}_{\in \mathbb{P}_{k-2}(1)} \underset{(3.4)\ \textit{vanish}}{=} 0.$$

Now put $q = \dfrac{dp}{ds}$ *to get* $\int_e \left(\frac{dp}{ds}\right)^2 = 0$ *and so* $\frac{dp}{ds} = 0$ *on* e. *Integrating on* e *and using the vanishing of (3.3) again, implies* $p = 0$ *on* e. *This holds for all edges* e, *so* $p = 0$ *on* ∂K.



Recalling the barycentric coordinates $\lambda_i(x)$. *Since* λ_i *vanishes on* e_i, *when* $k \geq 3$, *polynomial factorization implies* $p = (\lambda_1\lambda_2\lambda_3)r$ *for some* $r \in \mathbb{P}_{k-3}(2)$. *Also, since* (3.5) *vanish,*

$$0 = \int_K pr = \int_K \underbrace{(\lambda_1\lambda_2\lambda_3)}_{>0\,on\,interior}\underbrace{r^2}_{\geq 0} \quad \Rightarrow \quad r = 0 \quad \Rightarrow \quad p = 0.$$

Exercise 3.5. *Complete the proof of example 3.4 in the case* $k = 2$.

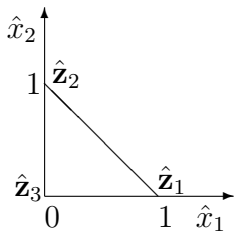
Given an element $(K, \mathcal{P}_K, \mathcal{N}_K)$, let $\mathcal{N}_K = \{N_1, ..., N_d\}$ and choose $\{p_i : i = 1, ..., d\}$ to be the nodal basis for $\mathcal{P}_K$ then for all v in the domain of N_i, $\forall i$, we define the (local) interpolant

$$I_K v = \sum_{i=1}^{d} (N_i v) p_i \in \mathcal{P}_K. \tag{3.6}$$

This interpolates v in the sense that

$$N_j(I_K v) = N_j(v) \qquad j = 1, ..., d. \tag{3.7}$$

Example 3.6. *(Special case of Example 3.3 with* $K = \hat{K} =$ *"unit triangle"),*

$\widehat{K} =$ 

The nodal basis is

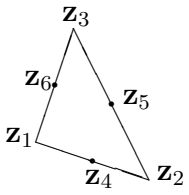
$$\hat{\lambda}_1(\hat{\mathbf{x}}) = \hat{x}_1$$
$$\hat{\lambda}_2(\hat{\mathbf{x}}) = \hat{x}_2$$
$$\hat{\lambda}_3(\hat{\mathbf{x}}) = 1 - \hat{x}_1 - \hat{x}_2$$

(Note $\hat{\lambda}_i(\mathbf{z}_j) = \delta_{ij}$.)

The interpolant is

$$(\hat{I}_{\hat{K}} v) = \sum_{i=1}^{3} v(\hat{\mathbf{z}}_i)\hat{\lambda}_i$$

Exercise 3.7. *Take any triangle with nodes $\mathbf{z_i}$, $i = 1, 2, 3$. Let $\mathbf{z}_4$, $\mathbf{z}_5$, $\mathbf{z}_6$ be any points on interiors of edges as shown.*



$$\mathcal{P}_K = \mathbb{P}_2(2)$$
$$\mathcal{N}_K = \{N_i(p) = p(\mathbf{z}_i) : i = 1, 2, ..., d\}$$

Show that $\mathcal{N}_K$ determines $\mathcal{P}_K$.

Exercise 3.8. *Let K be a tetrahedron in $\mathbb{R}^3$ with vertices $\mathbf{z}_i$, $i = 1, 2, 3, 4$ and define $\mathcal{P}_K = \mathbb{P}_1(3)$, and*

$$\mathcal{N}_K = \{p(\mathbf{z}_i), i = 1, 2, 3, 4\}.$$

Show that $\mathcal{N}_K$ determines $\mathcal{P}_K$.

Exercise 3.9. *(***An arbitrarily high order tetrahedral element** *)*

$$\begin{array}{rcll} K & = & \text{a tetrahedron with vertices } \mathbf{z}_i \ , \ i = 1, 2, 3, 4 \ . & \\ \mathcal{P}_K & = & \mathbb{P}_k(3) \text{ for any } k. & \\ \mathcal{N}_K & : & p(\mathbf{z}_i) \ , \ i = 1, 2, 3, 4 & \textbf{vertex freedoms (F1)} \\ & & \text{length}(e)^{-1} \int_e pq \quad \text{for } q \in \{\text{a basis for } \mathbb{P}_{k-2}(1) & \textbf{edge freedoms (F2)} \\ & & \text{area}(f)^{-1} \int_f pq \quad \text{for } q \in \text{a basis for } \mathbb{P}_{k-3}(2) & \textbf{face freedoms (F3)} \\ & & \text{volume}(K)^{-1} \int_K pq \quad \text{for } q \in \text{a basis for } \mathbb{P}_{k-4}(3) & \textbf{volume freedoms (F4)} \end{array}$$

The total number of freedoms is

$$\begin{aligned} 4 + 6(k-1) + 4\frac{1}{2}(k-2)(k-1) + \frac{1}{6}(k-3)(k-2)(k-1) \\ = \frac{1}{6}(k+3)(k+2)(k+1) = \dim(\mathbb{P}_k(3)) \ . \end{aligned}$$

Show that $\mathcal{N}_K$ determines $\mathcal{P}_K$. [Hint: start by showing that p vanishes on each face using the same sort of argument as in Example 3.4.]

Construction of finite element spaces on a mesh

Now we define finite element spaces on a domain Ω by knitting together finite elements defined on each element of a mesh.

Definition 3.10. *A mesh on $\Omega \subset \mathbb{R}^n$ $(n = 2, 3)$ is a subdivision $\mathcal{T}$ of Ω into closed triangles $(n = 2)$, respectively tetrahedra $(n = 3)$ or simplices (general n), with the properties:*

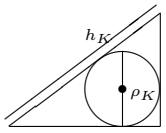
1. *$\overline{\Omega} = \bigcup\{K : K \in \mathcal{T}\}$ and the elements $K \in \mathcal{T}$ have pairwise disjoint interiors.*
2. *If $K, K' \in \mathcal{T}$ and $K \neq K'$, then $K \cap K'$ is either empty or a lower dimensional sub-simplex of of* **both elements** *K, K'. e.g.*

not allowed *OK*

Note that these assumptions presuppose that Ω is polyhedral. Curved boundaries can also be accommodated - we will discuss this if time permits. The regions K are also themselves called elements. Let

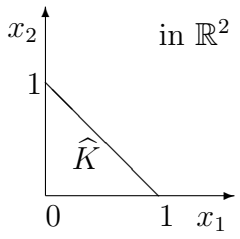
$$h_K \;=\; \max\{|\mathbf{x}-\mathbf{y}| : \mathbf{x}, \mathbf{y} \in \overline{K}\} \quad \text{and} \quad h = \max_{K\in\mathcal{T}} h_K \;. \tag{3.8}$$

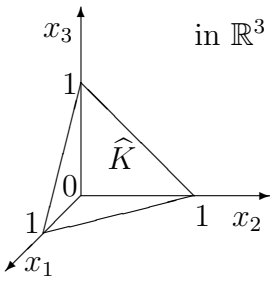
We usually write $\mathcal{T}_h$ for $\mathcal{T}$ and consider a sequence of meshes with $h \to 0$. Also, let $\rho_K =$ diameter of largest ball in $\mathbb{R}^n$ contained inside K. $\rho_K \leq h_K$.



We shall now define an element $(K, \mathcal{P}_K, \mathcal{N}_K)$ for each $K \in \mathcal{T}_h$ by "lifting" from a standard element $\widehat{K}$ as follows.

Let $\widehat{K}$ be the unit simplex in $\mathbb{R}^n$ (equivalently $\widehat{K}$ is the the closed compact hull of the standard unit basis vectors in $\mathbb{R}^n$, $n = 2, 3$).





Then we can define an affine map

$$F_K(\hat{\mathbf{x}}) = A_K\hat{\mathbf{x}} + \mathbf{b}_K \quad \hat{\mathbf{x}} \in \widehat{K} \tag{3.9}$$

with constant A_K and $\mathbf{b}_K$ which maps $\widehat{K}$ to K and maps edges to edges and nodes to nodes. e.g.

$$(n = 3) \qquad F_K(\hat{\mathbf{x}}) = \underbrace{[\mathbf{a}_2\ \mathbf{a}_3\ \mathbf{a}_4]}_{A_K}\hat{\mathbf{x}} + \underbrace{\mathbf{z}_1}_{\mathbf{b}_K},$$

where $\mathbf{a}_i = \mathbf{z}_i - \mathbf{z}_1$ and $\mathbf{z}_i$ are the nodes of K.

Note that $F_K(\hat{\mathbf{z}}_i) = \mathbf{z}_i$ for $i = 1, 2, 3, 4$ where $\hat{\mathbf{z}}_1 = (0, 0, 0)^T, \hat{\mathbf{z}}_2 = (1, 0, 0)^T, \hat{\mathbf{z}}_3 = (0, 1, 0)^T, \hat{\mathbf{z}}_4 = (0, 0, 1)^T$ are the nodes of $\hat{K}$.

Now suppose we have a finite element $(\widehat{K}, \widehat{\mathcal{P}}, \widehat{\mathcal{N}})$ defined on $\widehat{K}$ (sometimes called the "parent element") and write:

$$\begin{aligned} \widehat{\mathcal{P}} &= \text{span}\{\hat{p}_i : i = 1, ..., d\} \\ \widehat{\mathcal{N}} &= \text{span}\{\widehat{N}_i : i = 1, ..., d\}. \end{aligned}$$

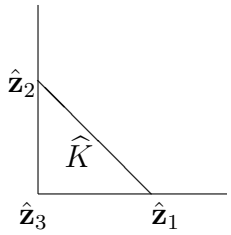
Then, for each $K \in \mathcal{T}_h$, we define an element $(K, \mathcal{P}_K, \mathcal{N}_K)$ on K by defining $\mathcal{P}_K$ to be the span of the functions

$$p_{K,i}(\mathbf{x}) = \hat{p}_i(F_K^{-1}(\mathbf{x})), \quad \mathbf{x} \in K \tag{3.10}$$

and $\mathcal{N}_K$ is the span of the functionals

$$N_{K,i}(p) = \widehat{N}_i(p \circ F_K) \quad \forall p \in \mathcal{P}_K. \tag{3.11}$$

Example 3.11.



with $\hat{\mathbf{z}}_1 = (0,0), \quad \hat{\mathbf{z}}_2 = (1,0), \quad \hat{\mathbf{z}}_1 = (0,1)$. *Then the affine map is*

$$F_K(\hat{\mathbf{x}}) \;=\; [\mathbf{a}_2\ \mathbf{a}_3]\,\hat{\mathbf{x}} \;+\; \hat{\mathbf{z}}_1 \;, \quad \textit{where} \quad \mathbf{a}_i = \mathbf{z}_i - \mathbf{z}_1 \;, \quad i = 2,3 \;.$$

Suppose we take the linear element with nodal interpolation on $\hat{K}$ *(as in Example 3.3):*

$$\widehat{\mathcal{P}} = \mathbb{P}_1(2)$$
$$\widehat{\mathcal{N}} = \{\hat{p}(\hat{\mathbf{z}}_i) : i = 1,2,3\}$$

Then, for any $K \in \mathcal{T}_h$,

$\mathcal{P}_K$ *defined by* (3.10) *is* $\mathbb{P}_1(2)$ *(since composition of two linear functions is linear) and* $\mathcal{N}_K = \{p(\mathbf{z}_i) : i = 1,2,3\}$, *since, by* (3.11), $N_{K,i}(p) = (p \circ F_K)(\hat{\mathbf{z}}_i) = p(\hat{\mathbf{z}}_i)$.

which is the same as the direct definition (Example 3.3).

Exercise 3.12. *Let* $\hat{p}_i$ *be the nodal basis for* $\widehat{\mathcal{P}}$. *Then show that in the construction above,*

- $N_{K,i}$ *are a basis for* $\mathcal{N}_K$
- $p_{K,i}$ *are the nodal basis for* $\mathcal{P}_K$.

Also, show that if $\widehat{I}$ *and* I_K *are the corresponding interpolation operators then*

$$I_K v = (\widehat{I}\hat{v}) \circ F_K^{-1}, \tag{3.12}$$

where $\hat{v} = v \circ F_K$ *and this holds for all* $\hat{v}$ *in the domain of the* $\widehat{N}_i$. *By using* I_K *on each* K *we can define a global interpolation operator* I_h *for functions on* Ω *by*

$$(I_h v)|_K = I_K v \quad \textit{for all} \quad K \in \mathcal{T}_h.$$

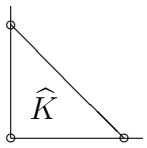
(Note, $I_h v$ *may not be well-defined on the interface between 2 neighboring elements.)*

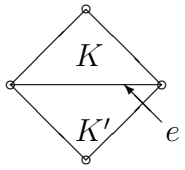
Definition 3.13. *The parent element* $(\widehat{K}, \widehat{\mathcal{P}}, \widehat{\mathcal{N}})$ *is called* H^m *conforming if*

$$I_h v \in H^m(\Omega) \quad \textit{for } v \in C(\overline{\Omega}).$$

Example 3.14.

a) With $(\widehat{K}, \widehat{\mathcal{P}}, \widehat{\mathcal{N}})$ *as in Example 3.11. Since two points uniquely determine a linear function on the edge* e, $I_h v$ *is continuous across* e *and* $I_h v \in H^1(K \cup K')$. *(see Exercise 3.15):*





$\circ$ = *point of interpolation*

b) With $\widehat{K}$ denoting the unit simplex in $\mathbb{R}^2$ choose $\widehat{\mathcal{P}} = \mathbb{P}_2(2)$, and choose $\widehat{\mathcal{N}}$ to be point evaluations at 3 nodal points and 3 interior points of sides as in Exercise 3.7. Then this is H^1 conforming if $\hat{\mathbf{z}}_4, \hat{\mathbf{z}}_5, \hat{\mathbf{z}}_6$ are midpoints of edges.

Exercise 3.15. *Suppose that $\Omega \subset \mathbb{R}^n$ is a Lipschitz domain and that Ω has a decomposition into two non-overlapping Lipschitz subdomains Ω_1, Ω_2, with interface $\Sigma := \overline{\Omega}_1 \cap \overline{\Omega}_2$.*

Suppose that $p_i \in H^1(\Omega_i) \quad i = 1, 2$ and define $p \in L^2(\Omega)$ by

$$p = \begin{cases} p_1 & \text{on } \Omega_1 \\ p_2 & \text{on } \Omega_2 \ . \end{cases}$$

Show that if the traces of p_1 and p_2 coincide on Σ, then $p \in H^1(\Omega)$.

Convergence of approximations determined by interpolation

We now study the convergence of approximations provided by the finite element interpolation operator I_h as $h \to 0$. (Recall that $h = \max_K h_K$, and h_K is the diameter of K.

Throughout, to make statements simpler to write down, if $A(h), B(h)$ are mesh dependent quantities, we write $A(h) \lesssim B(h)$ if $\frac{A(h)}{B(h)}$ is bounded independently of h.

Throughout we assume that the functionals $\widehat{\mathcal{N}}$ in the parent finite element are defined on $C(\widehat{K})$. (This is the case in all our examples above.)

Definition 3.16. *The mesh sequence $\mathcal{T}_h$ is called regular if*

$$h_K \lesssim \rho_K \ , \quad \text{for all} \quad K \in \mathcal{T}_h \quad \text{as} \quad h \to 0$$

Our main aim in the rest of this chapter is:

Theorem 3.17. *Assume the mesh sequence $\mathcal{T}_h$ is regular. Then if $m \geq 2$ and*

$$\mathbb{P}_{m-1}(n) \subseteq \widehat{\mathcal{P}}, \tag{3.13}$$

for $i = 0, \ldots, m$, we have for each $K \in \mathcal{T}_h$,

$$\|v - I_K v\|_{H^i(K)} \ \lesssim \ h_K^{m-i} |v|_{H^m(K)} \ , \quad \text{for all} \quad v \in H^m(K) \ . \tag{3.14}$$

Moreover if the element is H^i conforming with $0 \leq i \leq m$, then

$$\|v - I_h v\|_{H^i(\Omega)} \ \lesssim \ h^{m-i} |v|_{H^m(\Omega)} \ . \tag{3.15}$$

For example the linear element in Example 3.11 satisfies these estimates with $m = 2$ and $i = 0, 1$. The choice $i = 1$ gives yields assumption **(A3)** of Theorem 2.14.

To prove Theorem 3.17, we need some lemmas.

Lemma 3.18.

$$\forall K \in \mathcal{T}_h, \quad |A_K| \lesssim h_K \qquad \text{and} \qquad |A_K^{-1}| \lesssim \rho_K^{-1},$$

where A_K is as in (3.9) (i.e. $F_K(\hat{\mathbf{x}}) = A_K\hat{\mathbf{x}} + \mathbf{b}_K$) and $|\,.\,|$ denotes the matrix norm induced by the Euclidean norm $|\,.\,|$ on vectors.

Proof. Choose and $\boldsymbol{\xi} \in \mathbb{R}^n$ $(n = 2, 3)$. Then, by definition of $\rho_{\widehat{K}}$, there exists $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2 \in \widehat{K}$ with

$$\hat{\mathbf{x}}_1 - \hat{\mathbf{x}}_2 = \rho_{\widehat{K}} \frac{\boldsymbol{\xi}}{|\boldsymbol{\xi}|} .$$

Then

$$\rho_{\widehat{K}} \frac{|A_K\boldsymbol{\xi}|}{|\boldsymbol{\xi}|} = |A_K\hat{\mathbf{x}}_1 - A_K\hat{\mathbf{x}}_2| = |\underbrace{F_K(\hat{\mathbf{x}}_1)}_{\in K} - \underbrace{F_K(\hat{\mathbf{x}}_2)}_{\in K}| \leq h_K.$$

So

$$\frac{|A\boldsymbol{\xi}|}{|\boldsymbol{\xi}|} \leq \frac{h_K}{\rho_{\widehat{K}}} \quad \text{which implies} \qquad |A_K| \leq \frac{h_K}{\rho_{\widehat{K}}} \lesssim h_K.$$

(Note that since $\hat{K}$ is fixed $\rho_{\hat{K}}$ is a fixed constant independent of h.)

For the second estimate, note that $F_K^{-1}(\mathbf{y}) = A_K^{-1}\mathbf{y} - A_K^{-1}\mathbf{b}_K$ is just another affine map from K to $\widehat{K}$. So apply the first part to get

$$|A_K^{-1}| \leq \frac{h_{\widehat{K}}}{\rho_K} \lesssim \rho_K^{-1}.$$

Note that $h_{\widehat{K}}$ is also independent of h. □

The next lemma describes how the $H^i(K)$ semi-norms are related to $H^i(\widehat{K})$ semi-norms.

Lemma 3.19. *Let $\hat{v} \in H^i(\widehat{K})$, $i \geq 0$ and set $v = \hat{v} \circ F_K^{-1}$. Then $v \in H^i(K)$ and*

(i) $|v|_{H^i(K)} \lesssim |A_K^{-1}|^i (\det A_K)^{\frac{1}{2}} |\hat{v}|_{H^i(\widehat{K})}$;

(ii) $|\hat{v}|_{H^i(\widehat{K})} \lesssim |A_K|^i (\det A_K)^{-\frac{1}{2}} |v|_{H^i(K)}$.

Proof. (part (ii) for i=1) Since $\hat{v}(\hat{\mathbf{x}}) = v(F_K(\hat{\mathbf{x}}))$ we have

$$\begin{aligned}\frac{\partial \hat{v}}{\partial \hat{x}_p}(\hat{\mathbf{x}}) &= \sum_{i=1}^{n} \left(\frac{\partial v}{\partial x_i}\right)(F_K(\hat{\mathbf{x}}))\frac{\partial}{\partial \hat{x}_p}((F_K(\hat{\mathbf{x}}))_i) \\ &= (\nabla v)(F_K(\hat{\mathbf{x}})).(p^{th} \text{ column of } A_K) \\ &\leq |(\nabla v)(F_K(\hat{\mathbf{x}}))| \left\{\sum_{i=1}^{n}(A_K)_{ip}^2\right\}^{\frac{1}{2}}. \qquad \text{(Cauchy-Schwarz)}\end{aligned}$$

So (gradient with respect to "hat" variables)

$$|\widehat{\nabla}\hat{v}(\hat{\mathbf{x}})| \leq |(\nabla v)(F_K(\hat{\mathbf{x}}))|\|A_K\|_F,$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix (root sum of squares of all entries). So, by equivalence of norms of matrices,

$$|(\widehat{\nabla}\hat{v})(\hat{\mathbf{x}})| \lesssim |(\nabla v)(F_K(\hat{\mathbf{x}}))||A_K|.$$

Thus

$$\begin{aligned}|\hat{v}|_{H'(\widehat{K})} &\lesssim \left\{\int_{\widehat{K}} |(\nabla v)(F_K(\hat{\mathbf{x}}))|^2 d\hat{x}\right\}^{\frac{1}{2}} |A_K| \\ &= \left\{\int_K |(\nabla v)(x)|^2 dx\right\}^{\frac{1}{2}} (\det A_K)^{-\frac{1}{2}}|A_K|.\end{aligned}$$

(Note: $\int_K g(x)dx = \int_{\widehat{K}} g(F_K(\hat{\mathbf{x}}))(\det A_K)d\hat{x}$.) For $i > 1$ we have to apply the chain rule recursively (see [Ci: Theorem 3.1.2]). □

Exercise 3.20. *Prove part (i) of Lemma 3.19 for $i = 1$.*

Lemma 3.21. *Let $\Omega \subset \mathbb{R}^n$ be any bounded connected Lipschitz domain. Then there exists a constant $C > 0$ which may depend on Ω such that, for all $m \geq 1$*

$$\inf_{p\in\mathbb{P}_{m-1}(n)} \|v - p\|_{H^m(\Omega)} \leq C|v|_{H^m(\Omega)}, \quad \textit{for all} \quad v \in H^m(\Omega).$$

(This is [Ci: Theorem 3.1.1] Its proof requires the Hahn-Banach theorem and an embedding theorem for Sobolev spaces.)

Proof. Let $d = \dim(\mathbb{P}_{m-1}(n))$. Let $\{M_i : i = 1, ..., d\}$ be a basis for $(\mathbb{P}_{m-1}(n))'$ and let $\{q_i : i = 1, ..., d\}$ be the corresponding nodal basis for $\mathbb{P}_{m-1}(n)$ (see Exercise 3.2). Using the Hahn-Banach theorem, each M_i can be extended without increase of norm to a functional $M_i \in H^m(\Omega)'$. We shall prove that there exists a constant $C > 0$ such that

$$\|v\|_{H^m(\Omega)} \leq C\left\{|v|_{H^m(\Omega)} + \sum_{i=1}^{d}|M_i(v)|\right\}, \quad \text{for all} \quad v \in H^m(\Omega). \tag{3.16}$$

The lemma then follows since if $v \in H^m(\Omega)$, then define $q \in \mathbb{P}_{m-1}(n)$ by

$$q = \sum_{j=1}^{d} M_j(v) q_j \ ,$$

and observe that

$$M_i(v-q) \ = \ M_i(v) - M_i(v) = 0 \ .$$

Hence (3.16) implies

$$\|v-q\|_{H^m(\Omega)} \ \leq \ C|v-q|_{H^m(\Omega)} \ = \ C|v|_{H^m(\Omega)} \ ,$$

where the last step uses the fact that $D^\alpha q = 0$ for all $|\alpha| = m$ when $q \in \mathbb{P}_{m-1}(n)$.

To obtain (3.16) suppose for a contradiction that $\exists \{v^l\}_{l=1}^{\infty}$ such that

$$\lim_{l\to\infty} \left\{ |v^l|_{H^m(\Omega)} + \sum_{i=1}^{d} |M_i(v^l)| \right\} \to 0, \tag{3.17}$$

but

$$\|v^l\|_{H^m(\Omega)} \ = \ 1. \tag{3.18}$$

Then since $H^m(\Omega)$ is compactly embedded in $H^{m-1}(\Omega)$ (This holds for Lipschitz domains - [Ci, p.114]), there exists a convergent subsequence (again denoted $\{v^l\}$) such that

$$v^l \to v^* \in H^{m-1}(\Omega).$$

Also, by (3.17),

$$|v^l|_{H^m(\Omega)} \to 0 \quad \text{as} \quad l \to \infty \ . \tag{3.19}$$

So, v^l is Cauchy in $H^m(\Omega)$, so v^l converges in $H^m(\Omega)$ and hence $v^l \to v^*$ in $H^m(\Omega)$ also.

Also, (3.19) implies $|v^*|_{H^m(\Omega)} = 0$ and so $D^\alpha v^* = 0$ on Ω for all $|\alpha| = m$. By the connectedness of Ω, $v^* \in \mathbb{P}_{m-1}(n)$ (This is the "Fundamental theorem of calculus" see e.g. [GRB, Theorem 3.4], [EG, Lemma B.29].)

Also, the continuity of M_i on $H^m(\Omega)$ implies

$$M_i(v^*) = \lim_{l\to\infty} M_i(v_l) = 0 \quad \text{by (3.17).}$$

This holds for all $i = 1, \ldots, d$ and $v^* \in \mathbb{P}_{m-1}(n)$ so by Exercise 3.2, $v^* = 0$. This contradicts (3.18). □

Corollary 3.22. The Friedrichs inequality *Let Ω be a bounded Lipschitz domain. Then there exists a positive constant $C = C(\Omega)$, such that*

$$\|v - \overline{v}\|_{H^1(\Omega)} \ \leq \ C|v|_{H^1(\Omega)} \ , \quad \textit{for all} \quad v \in H^1(\Omega) \ ,$$

where

$$\overline{v} = \frac{1}{|\Omega|} \int_\Omega v(x)dx \qquad \textit{and} \qquad |\Omega| \ = \ \int_\Omega dx \ .$$

Proof. The result follows from Lemma 3.21 with $m = 1$, since the average $\overline{v}$ is the orthogonal projection of v onto the constant functions with respect to the $L^2(\Omega)$ norm. □

Proof of Theorem 3.17. Since $m \geq 2$, for any $v \in H^m(K)$, the Sobolev Embedding Theorem (e.g. [GRB, Thm 4.2], [Ci, p.114]) implies that $v \in C(\overline{K})$ and so $I_K v$ is well-defined. Defining $\hat{v} = v \circ F_K$ on $\widehat{K}$, then, by Exercise 3.12,

$$(v - I_K v)(F_K(\hat{x})) = (\hat{v} - \widehat{I}\hat{v})(\hat{x}). \tag{3.20}$$

We shall first obtain an estimate on the right-hand side and then obtain the required estimate on the left-hand side by a "scaling argument".

First observe that

$$\begin{aligned}
\|\widehat{I}\hat{v}\|_{H^m(\widehat{K})} &= \left\| \sum_{i=1}^{d} (\widehat{N}_i \hat{v}) \hat{p}_i \right\|_{H^m(\widehat{K})} \leq \sum_{i=1}^{d} |\widehat{N}_i \hat{v}| \|\hat{p}_i\|_{H^m(\widehat{K})} \\
&\leq \underbrace{\left[\sum_{i=1}^{d} \|\widehat{N}_i\|_{C(\widehat{K})'} \|\hat{p}_i\|_{H^m(\widehat{K})} \right]}_{\text{independent of } v \text{ and independent of } \mathcal{T}_h} \|\hat{v}\|_{C(\widehat{K})} \quad \text{(by definition of bounded linear functional)} \\
&\leq C \|\hat{v}\|_{H^m(\widehat{K})} \quad \text{(by Sobolev Embedding Theorem),}
\end{aligned} \tag{3.21}$$

where C is generic constant independent of $\hat{v}$.

Since by assumption (3.13), $\mathbb{P}_{m-1}(n) \subseteq \widehat{\mathcal{P}}$, we know $\widehat{I}\hat{p} = \hat{p}$ when $\hat{p} \in \mathbb{P}_{m-1}(n)$ (see Exercise 3.12). So,

$$\begin{aligned}
\|\hat{v} - \widehat{I}\hat{v}\|_{H^m(\widehat{K})} &= \|(\hat{v} - \hat{p}) - \widehat{I}(\hat{v} - \hat{p})\|_{H^m(\widehat{K})} \\
&\leq \|\hat{v} - \hat{p}\|_{H^m(\widehat{K})} + \|\widehat{I}(\hat{v} - \hat{p})\|_{H^m(\widehat{K})} \\
&\leq C \|\hat{v} - \hat{p}\|_{H^m(\widehat{K})} \quad \text{for all} \quad \hat{p} \in \mathbb{P}_{m-1}(n) \quad \text{by (3.21)} \\
&\leq C |\hat{v}|_{H^m(\widehat{K})} \quad \text{by Lemma 3.21.}
\end{aligned} \tag{3.22}$$

Hence, remembering that $\lesssim$ means there is a hidden constant independent of $\mathcal{T}_h$,

$$\begin{aligned}
|v - I_K v|_{H^i(K)} &\lesssim |A_K^{-1}|^i (\det A_K)^{\frac{1}{2}} |\hat{v} - \widehat{I}\hat{v}|_{H^i(\widehat{K})} \quad \text{by Lemma 3.19 and (3.20)} \\
&\lesssim \rho_K^{-i} (\det A_K)^{\frac{1}{2}} |\hat{v} - \widehat{I}\hat{v}|_{H^i(\widehat{K})} \quad \text{by Lemma 3.18} \\
&\lesssim \rho_K^{-i} (\det A_K)^{\frac{1}{2}} |\hat{v}|_{H^m(\widehat{K})} \quad \text{by (3.22)} \\
&\lesssim \rho_K^{-i} h_K^m |v|_{H^m(\widehat{K})} \quad \text{by Lemmas 3.18 and 3.19.}
\end{aligned}$$

Now use the mesh regularity to get the result. □

Remark 3.23. *If we remove the assumption of mesh regularity we have the estimates*

$$|v - I_K v|_{H^i(K)} \lesssim \rho_K^{-i} h_K^m |v|_{H^m(K)} ,$$

and

$$|v - I_K v|_{H^i(\Omega)} \lesssim \left\{ \sum_{K \in \mathcal{T}_h} \rho_K^{-2i} h_K^{2m} |v|^2_{H^m(K)} \right\}^{1/2} .$$

We finish this section by now quantifying the rate of convergence of finite element approximation of the following abstract problem posed on some subspace V of $H^1(\Omega)$:

$$\text{Find} \quad u \in V \quad \text{such that} \quad a(u,v) = (f,v)_{L^2(\Omega)} \quad \text{for all} \quad v \in V, \tag{3.23}$$

where $f \in L_2(\Omega)$.

We shall approximate this on the subspace

$$V_h := I_h(V \cap C(\overline{\Omega})) \tag{3.24}$$

and we assume that

$$V_h \subseteq V \qquad \text{the finite element space is "conforming".} \tag{3.25}$$

The finite element approximation is

$$\text{Find} \quad u_h \in V_h \quad \text{such that} \quad a(u_h, v_h) = (f, v_h)_{L^2(\Omega)} \quad \text{for all} \quad v_h \in V_h \ . \tag{3.26}$$

Theorem 3.24. *Suppose a is bounded on V and the well-posedness and regularity assumptions* **(A1)** *and* **(A2)** *of Theorem 2.14 are satisfied in V. Suppose the discrete inf-sup condition of Theorem 2.8 is satisfied on V with $\epsilon_h \geq \epsilon$ for all $h \geq h_0$ and suppose $\mathcal{T}_h$ is a regular sequence of meshes. Finally suppose $\mathbb{P}_1(n) \subseteq \widehat{\mathcal{P}}$.*

Then for all $h \geq h_0$, we have the finite element error estimate:

$$\|u - u_h\|_{L^2(\Omega)} + h\|u - u_h\|_{H^1(\Omega)} \lesssim h^2 \|f\|_{L^2(\Omega)} \ .$$

Remark 3.25. *The assumption of conformity* (3.25) *may place a restriction on the mesh and the choice of elements. For example when $V = H^1_0(\Omega)$, we require that the interpolant of a continous function is continuous and that the interpolant of a function with zero trace has zero trace. This is true for example when when any of the the elements in Examples 3.3, 3.4, 3.7, 3.8 and 3.9 are used and when $\partial\Omega$ consists entirely of edges (or faces) of elements.*

Proof. By Theorem 2.9, and Theorem 3.17,

$$\begin{aligned} \|u - u_h\|_{H^1(\Omega)} &\lesssim \inf_{w_h \in V_h} \|u - w_h\|_{H^1(\Omega)} \\ &\leq \|u - I_h u\|_{H^1(\Omega)} \lesssim h\|u\|_{H^2(\Omega)} \lesssim h\|f\|_{L^2(\Omega)} \ . \end{aligned} \tag{3.27}$$

(The last step follows from well-posedness and regularity for (3.23).)

To obtain the estimate for the error in $L^2(\Omega)$, we apply a "duality argument". Letting $e_h = u - u_h$, and letting $v \in V$ be the solution to the "auxiliary dual problem":

$$a(\phi, v) = (e_h, \phi)_{L^2(\Omega)} \ , \quad \text{for all} \quad \phi \in V \ . \tag{3.28}$$

Then by well-posedness and regularity (for the dual problem) this has a unique solution and

$$\|e_h\|^2_{L^2(\Omega)} = (e_h, e_h)_{L^2(\Omega)} = a(e_h, v) \ .$$

Now recalling Galerkin orthogonality (2.5) we have $a(e_h, v) = a(e_h, v - I_h v)$. So, using boundedness of a and then (3.27), we have

$$\|e_h\|^2_{L^2(\Omega)} \ \lesssim \ \|e_h\|_{H^1(\Omega)} \|v - I_h v\|_{H^1(\Omega)} \ \lesssim \ h \|f\|_{L^2(\Omega)} \|v - I_h v\|_{H^1(\Omega)} \ . \tag{3.29}$$

Finally Theorem 3.17 and again well-posedness and regularity for the dual problem (3.28) we have

$$\|v - I_h v\|_{H^1(\Omega)} \ \lesssim \ h \|v\|_{H^2(\Omega)} \ \lesssim \ h \|e_h\|_{L^2(\Omega)}$$

and substitution of this into (3.29) yields the result.

□

Remark 3.26. *Note that higher rates of convergence may be obtained by assuming higher regularity and allowing $\widehat{\mathcal{P}}$ to contain higher order polynomials $\mathbb{P}_{m-1}(n)$. If we fix the mesh and let $m \to \infty$ we get the so-called "p-version of the FEM" which is closely related to spectral methods. Letting $m \to \infty$ on some elements and $h \to 0$ on others is the "hp− version" FEM, which has the potential to converge exponentially with respect to the number of degrees of freedom.*

All the assumptions of Theorem 3.24 except the regularity assumption have now been investigated in the context of the convection-diffusion reaction problem Example 1.24. We shall explore regularity in the following chapter. To make the material manageable we shall restrict this chapter to the duffusion problem only.

4 Detailed study of some examples

4.1 The Diffusion Equation

We study

$$-\nabla.A\nabla u \;=\; f \quad \text{in} \quad \Omega \tag{4.1}$$

where $A \in (L^\infty(\Omega))^{n\times n}$ is uniformly positive definite on Ω (see (2.8)) and $f \in L^2(\Omega)$.

Example 4.1. Homogeneous Dirichlet Problem

Solve (4.1) *subject to*

$$u \;=\; 0 \quad \textit{on} \quad \partial\Omega \tag{4.2}$$

with $V = H_0^1(\Omega)$. *The weak form is* (3.23) *with*

$$a(u,v) \;=\; \int_\Omega (A\nabla u).\nabla v \ .$$

Here a *is symmetric, bounded with constant* C *and coercive with constant* ε *on* V *(see Exercise 2.12), so* a *is an inner product on* V. *Choose* V_h *as in* (3.24), (3.25). *The finite element method is as in* (3.26).

Exercise 4.2. *Show that* (3.26) *is equivalent to an* $N \times N$ *system of linear equations with a symmetric positive definite stiffness matrix, and deduce that it has a unique solution for all* h, *i.e. there is no mesh threshold for existence of a finite element solution in this case. The error estimate for this problem is as given in Theorem 3.24, assuming sufficient regularity for the solution. (See also Exercise 2.10 for a sharper estimate in the* H^1 *seminorm.)*

Example 4.3. Inhomogeneous Dirichlet Problem *Solve* (4.1) *subject to*

$$u \;=\; g \quad \textit{on} \quad \partial\Omega \ . \tag{4.3}$$

If g *is sufficiently smooth there exists an extension* $\tilde{g} \in H^1(\Omega) \cap C(\Omega)$ *of* g *with trace* $\tau(\tilde{g})$ *coinciding with* g. *Equivalently, "the trace operatopr has an inverse". A sufficient condition would be* $g \in H^1(\partial\Omega)$ *(see, e.g. [McL: Theorem 3.37]). Then, for any* $v \in V := H_0^1(\Omega)$, *we have, by Corollary 1.23,*

$$a(u,v) \;=\; \int_\Omega A\nabla u.\nabla v = (f,v)_{L^2(\Omega)} \ . \tag{4.4}$$

Seeking u *such that* $u - \tilde{g} \in V$ *we have*

$$a(u-g,v) = (f,v)_{L^2(\Omega)} - a(g,v) \tag{4.5}$$

Since a *is bounded and coercive on* V *and* $v \mapsto (f,v)_{L^2(\Omega)} - a(g,v)$ *is in* V', *Theorem 2.6 implies that* (4.5) *and hence* (4.4) *has a unique solution* u *with the trace of* u *coinciding with* g *on* $\partial\Omega$.

For the finite element solution of this problem we take $V_h \subseteq V$ as in (3.24), (3.25) *and seek u_h such that $u_h - I_h\tilde{g} \in V_h$ and*

$$a(u_h, v_h) \;=\; (f, v_h)_{L^2(\Omega)} \;, \quad \textit{for all} \quad v_h \in V_h \;. \tag{4.6}$$

Then, as in Examples 4.1, 4.2, there exists a unique solution u_h for all h and u_h still satisfies the Galerkin orthogonality property:

$$a(u - u_h, v_h) = 0 \quad \text{for all} \quad v_h \in V_h$$

. But note that now $u - u_h \notin V$ in general. However by the uniform positive definiteness of A,

$$\begin{aligned}
\alpha|u - u_h|^2_{H^1(\Omega)} &\leq a(u - u_h, u - u_h) \\
&= a(u - u_h, u - \tilde{g} - \underbrace{(u_h - I_h\tilde{g})}_{\in V_h} + \tilde{g} - I_h\tilde{g}) \\
&= a(u - u_h, u - \tilde{g} - v_h + \tilde{g} - I_h\tilde{g}) \;, \quad \text{for all} \quad v_h \in V_h \\
&\leq C|u - u_h|_{H^1(\Omega)} \left(|u - \tilde{g} - v_h|_{H^1(\Omega)} \;+\; |\tilde{g} - I_h\tilde{g}|_{H^1(\Omega)}\right)
\end{aligned}$$

Hence

$$|u - u_h|^2_{H^1(\Omega)} \;\leq\; \left(\frac{C}{\alpha}\right) \left(|u - \tilde{g} - v_h|_{H^1(\Omega)} \;+\; |\tilde{g} - I_h\tilde{g}|_{H^1(\Omega)}\right) \;.$$

Now we can prove rates of convergence using Theorem 3.17 and assumption of sufficient regularity for u, $\tilde{g}$.

Example 4.4. *Consider the implementation of Example 4.3 in the case of piecewise linear elements (see Example 3.3). Let*

$$\{x_i : i \in \mathcal{N}(\Omega)\}$$

denote the mesh nodes in the interior of Ω and let

$$\left\{x_i : i \in \mathcal{N}(\overline{\Omega})\right\}$$

denote the mesh nodes in $\overline{\Omega} = \Omega \cup \partial\Omega$. Then $V_h = \text{span}\{\phi_i : i \in \mathcal{N}(\Omega)\}$ where ϕ_i is the linear "hat function" at node x_i, i.e. ϕ_i is linear on each element, has the value 1 at node x_i and value 0 at all other nodes, so that the support of ϕ_i is the union of all elements containing node x_i. Set $\overline{V}_h = \text{span}\{\phi_i : i \in \mathcal{N}(\overline{\Omega})\}$. Then (4.6) *holds if and only if $u_h \in \overline{V}_h$ and $u_h(x_i) = g(x_i)$ for all $i \in \mathcal{N}(\overline{\Omega})\backslash\mathcal{N}(\Omega)$, which is equivalent to*

$$u_h \;=\; \sum_{j \in \mathcal{N}(\Omega)} u_h(x_j)\phi_j \;+\; \sum_{j \in \mathcal{N}(\overline{\Omega})\backslash\mathcal{N}(\Omega)} g(x_j)\phi_j \;.$$

Substitution into (4.6) *yields a system for $\{u(x_i) : i \in \mathcal{N}(\Omega)\}$ with stiffness matrix $A_i = a(\phi_j, \phi_i)$, for $i, j \in \mathcal{N}(\Omega)$ (c.f.* (2.4)*).*

Example 4.5. Homogeneous Neumann Problem.

Solve (4.1) *subject to*

$$(A\nabla u).\boldsymbol{\nu} \ = 0 \quad on \quad \partial\Omega \ .$$

Multiplying (4.1) *by* $v \in H^1(\Omega)$*, and using Corollary 1.23 shows that* $u \in H^1(\Omega)$ *solves*

$$a(u,v) = (f,v)_{L^2(\Omega)} \ , \quad for\ all \quad v \in H^1(\Omega) \ . \tag{4.7}$$

The solution u *is not unique , since* $u \equiv 1$ *solves the homogeneous version of equation* (4.7)*.*

Introduce

$$\widehat{V} \ = \left\{ v \in H^1(\Omega) : \int_\Omega v = 0 \right\}$$

and consider the related problem: Seek $u \in \widehat{V}$ *such that*

$$a(u,v) = (f,v)_{L^2(\Omega)} \ , \quad for\ all \quad v \in \widehat{V} \ . \tag{4.8}$$

Note that a *is bounded on* V *and hence on* $\widehat{V}$*. Also we have for all* $v \in \widehat{V}$*,*

$$a(v,v) \ = \ \int_\Omega A\nabla V.\nabla v \ \geq \ \alpha |v|^2_{H^1(\Omega)} \ \geq \ \alpha \frac{1}{C} \|v\|^2_{H^1(\Omega)}$$

with C *as in Corollary 3.22. So* a *is coercive on* $\widehat{V}$*. By Theorem 2.6,* (4.8) *has a unique solution* $u \in \widehat{V}$*. For the linear finite element approximation of this problem, we should work in the space:*

$$\widehat{\overline{V}}_h := \left\{ v_h \in \overline{V}_h : \int_\Omega v_h = 0 \right\} \ \subseteq \ \widehat{V} \ , \tag{4.9}$$

where $\overline{V}_h$ *is as in Example 4.4. We then seek* $u_h \in \widehat{\overline{V}}_h$ *such that*

$$a(u_h, v_h) = (f, v_h)_{L^2(\Omega)} \quad for\ all \quad v_h \in \widehat{\overline{V}}_h \ . \tag{4.10}$$

By Theorems 2.8 and 2.9, there is a unique solution u_h *and*

$$\|u - u_h\|_{H^1(\Omega)} \ \lesssim \ \inf_{w_h \in \widehat{\overline{V}}_h} \|u - w_h\|_{H^1(\Omega)} \ . \tag{4.11}$$

Now if we choose $w_h = I_h u - \int_\Omega I_h u \in \widehat{\overline{V}}_h$*, then*

$$\begin{aligned} \|u - w_h\|_{H^1(\Omega)} \quad &= \quad \|u - I_h u - \int_\Omega (u - I_h u)\|_{H^1(\Omega)} \\ &\lesssim \ \|u - I_h u\|_{H^1(\Omega)} + \int_\Omega |u - I_h u| \ \lesssim \ \|u - I_h u\|_{H^1(\Omega)} \ . \end{aligned} \tag{4.12}$$

Combining (4.11) *and* (4.12) *with Theorem 3.17 yields a rate of convergence estimate for* $\|u - u_h\|_{H^1(\Omega)}$.

To implement this method appears not to be so nice, since a basis for the space (4.9) *is not so obvious. (Note that hat functions cannot be in this space since they are everywhere non-negative.) Instead we formulate the problem as an extended system: Seek* $(u_h, \lambda) \in \overline{V}_h \times \mathbb{R}$ *such that*

$$\left.\begin{array}{rcl} a(u_h, v_h) + \lambda \int_\Omega v_h & = & (f, v_h)_{L^2(\Omega)} \quad \textit{for all} \quad v_h \in \overline{V}_h \\ \int_\Omega u_h & = & 0 \end{array}\right\} \tag{4.13}$$

This is an $(N+1) \times (N+1)$ *system (where* $N = \#\mathcal{N}(\overline{\Omega})$*). Also* (4.13) *has a unique solution since if* (w_h, γ) *solves the homogeneous version of* (4.13) *(i.e. with* $f = 0$*), then setting* $v_h = \phi_i$ *and summing over all* $i \in \mathcal{N}(\overline{\Omega})$*, shows that*

$$0 \;=\; a(w_h, \underbrace{\sum_{i=1}^{N} \phi_i}_{=1}) = -\gamma \int_\Omega \sum_{i=1}^{N} \phi_i = \gamma |\Omega| \ .$$

This implies that $\gamma = 0$*, so* $a(w_h, w_h) = 0$ *and so, by coercivity,* $w_h = 0$*. So* (4.13) *has a unique solution and the solution satisfies* (4.10)*.*

Example 4.6. Mixed Boundary Value Problems

Solve (4.1) *subject to*

$$\begin{array}{rcl} u & = & g_D \quad \textit{on} \quad \partial\Omega_D \\ (A\nabla u).\boldsymbol{\nu} & = & g_N \quad \textit{on} \quad \partial\Omega_N \end{array}$$

where $\partial\Omega_D$ *and* $\partial\Omega_N$ *partition* $\partial\Omega$*. Suppose* g_D *has an extension* $\tilde{g}$ *in*$H^1(\Omega)$*. Then with*

$$V = \left\{ v \in H^1(\Omega) : \tau(v) = 0 \quad \textit{on} \quad \partial\Omega_D \right\} \ ,$$

we seek u *such that* $u - g_D \in V$ *and Corollary 1.23 implies*

$$a(u, v) \;=\; \int_\Omega \nabla u \cdot \nabla v \;=\; (f, v)_{L^2(\Omega)} + (g_N, v)_{L^2(\partial\Omega_N)} \quad \textit{for all} \quad v \in V \ .$$

The theory proceeds as before but requires the following a generalisation of the Poincaré inequality: If $\gamma = \text{measure}(\Gamma_D) > 0$ *then there exists* $C > 0$ *(which depends on* γ*) such that*

$$\|v\|_{H^1(\Omega)} \;\le\; C|v|_{H^1(\Omega)} \ , \quad \textit{for all} \quad v \in V \ .$$

This can be found in [McL] for example.

4.2 Regularity of solutions for the diffusion equation

Here we consider again the equation:

$$-\nabla.A\nabla u \;=\; f \tag{4.14}$$

in its weak form subject to some boundary conditions. In this subsection we explore conditions under which the regularity assumption (A2) of Theorem 2.14 holds. This is done by a sequence of examples.

Example 4.7. *Consider*

$$-\Delta u \;=\; f \quad on \quad \Omega = [0, 2\pi]^2$$

where f and u are both $2\pi-$periodic in each variable. The solution u can be written in Fourier series as

$$u(x) \;=\; \left(\frac{1}{2\pi}\right)^2 \sum_{j,k\in\mathbb{Z}} \widehat{u}_{jk} \exp(i(jx_1{+}kx_2)) \;, \quad \textit{where } u_{j,k} \in \mathbb{C} \textit{ are the Fourier coefficients of } u \;,$$

$$\widehat{u}_{j,k} \;=\; \left(\frac{1}{2\pi}\right)^2 \int_0^{2\pi}\int_0^{2\pi} u(x_1, x_2) \exp((-i(jx_1 + kx_2))dx_1dx_2 \;.$$

Then by Parseval's equality,

$$\|u\|^2_{L^2(\Omega)} \;=\; \left(\frac{1}{2\pi}\right)^2 \sum_{j,k\in\mathbb{Z}} |\widehat{u}_{jk}|^2 \;. \tag{4.15}$$

By direct calculation,

$$\Delta u = -\left(\frac{1}{2\pi}\right)^2 \sum_{j,k\in\mathbb{Z}} \widehat{u}_{jk}(j^2 + k^2) \exp(i(jx_1 + kx_2))$$

and so

$$\|f\|^2_{L^2(\Omega)} \;=\; \|-\Delta u\|^2_{L^2(\Omega)} \;=\; \left(\frac{1}{2\pi}\right)^2 \sum_{j,k\in\mathbb{Z}} (j^2 + k^2)^2 |\widehat{u}_{jk}|^2 \;=\; \|f\|^2_{L^2(\Omega)} \;. \tag{4.16}$$

Thus

$$\left\|\frac{\partial^2 u}{\partial x_1^2}\right\|^2_{L^2(\Omega)} \;=\; \left(\frac{1}{2\pi}\right)^2 \sum_{j,k\in\mathbb{Z}} j^4 |u_{jk}|^2 \;\le\; \|f\|^2_{L^2(\Omega)}$$

and similarly,

$$\left\|\frac{\partial^2 u}{\partial x_2^2}\right\|^2_{L^2(\Omega)} \;\le\; \|f\|^2_{L^2(\Omega)} \;.$$

Also,

$$\left\|\frac{\partial^2 u}{\partial x_1 \partial x_2}\right\|^2_{L^2(\Omega)} \;=\; \left(\frac{1}{2\pi}\right)^2 \sum_{j,k\in\mathbb{Z}} (jk)^2 |u_{jk}|^2 \;\le\; \left(\frac{1}{2\pi}\right)^2 \sum_{j,k\in\mathbb{Z}} \left(\frac{1}{2}(j^2 + k^2)\right)^2 |u_{jk}|^2 \;\le\; \frac{1}{4}\|f\|^2_{L^2(\Omega)} \;.$$

So overall we have established the regularity estimate

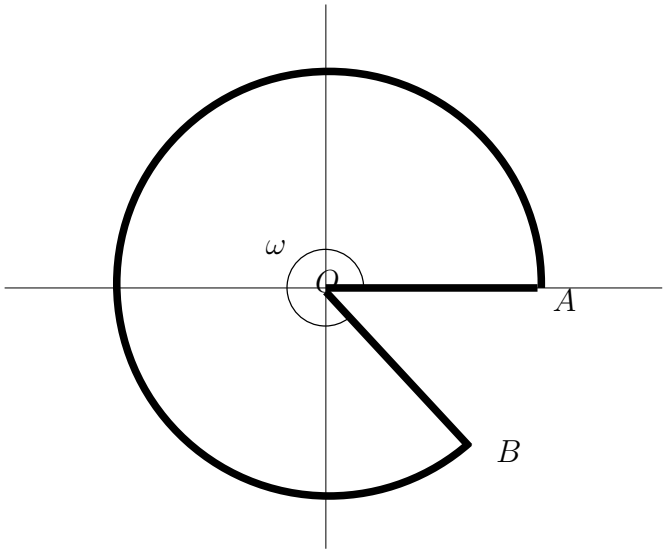
$$\|u\|_{H^2(\Omega)} \;\le C\; \|f\|_{L^2(\Omega)} \;. \tag{4.17}$$

A general result which includes this example is found in many references.

Theorem 4.8. *If f is sufficiently smooth and $\partial\Omega$ is sufficiently smooth or (in 2D) convex and A is sufficiently smooth then* (4.17) *holds for the Dirichlet or Neumann boundary value problem for* (4.14)*. For references to a proof, see [BrSc:p.139].*

For the purposes of this result, "sufficiently smooth" means "two continuous derivatives". This regulartity theorem may fail if $\partial\Omega$ is not smooth and non-convex or if A is non-smooth.

Example 4.9. *With the domain Ω pictured, a segment of the unit circle, with interior angle at the origin equal to ω. Let $\pi < \omega < 2\pi$, i.e. we have a "reentrant corner and $\beta := \pi/\omega \in (1/2, 1)$. We shall show that there exists a solution of $-\Delta u = f$ on Ω, satisfying homogeneous Dirichlet boundary conditions with $f \in L^2(\Omega)$, but $u \notin H^2(\Omega)$. To show this, first consider the function $v(r,\theta) := r^\beta \sin\beta\theta$.*



Since v is the imaginary part of z^β, and this is an analytic function in Ω (note $(0,0) \notin \Omega$), the Cauchy-Riemann equations imply that $\Delta v = 0$ in Ω. Also v vanishes on OA and on OB. Now, defining $u(r,\theta) = (1-r^2)v(r,\theta)$, we note that $u = 0$ on $\partial\Omega$ and a calculation (exercise below) shows that

$$-\Delta u \;=\; 4(1+\beta)v \;=:\; f \in L^2(\Omega)\ . \tag{4.18}$$

Also,

$$\frac{\partial^2 u}{\partial r^2} = \cos^2\theta \frac{\partial^2 u}{\partial x_1^2} + \sin^2\theta \frac{\partial^2 u}{\partial x_2^2}\ . \tag{4.19}$$

Now suppose $u \in H^2(\Omega)$. This implies, from (4.19) *that*

$$\int_0^\omega \int_0^1 \left|\frac{\partial^2 u}{\partial r^2}\right|^2 r dr d\theta \;<\; \theta$$

and this implies

$$\int_0^1 (\beta(\beta-1))^2 (r^{\beta-2})^2 r dr d\theta \;<\; \infty\ ,$$

which is impossible unless $2\beta - 3 > -1$, i.e. $\beta > 1$. So the Dirichlet problem for (4.18) is an example of a boundary-value problem in which the regularity assumption fails.

In this case a more spohisticated regularity result of the form

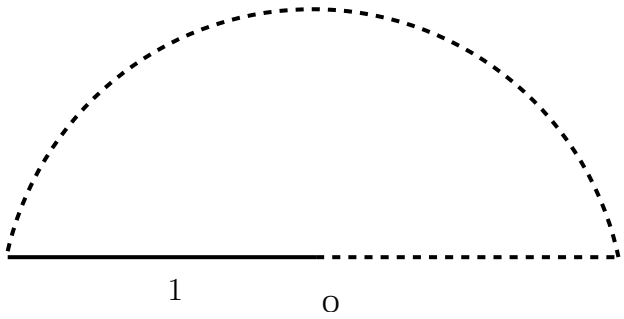
$$\|u\|_{H^{1+s}(\Omega)} \;\leq\; C\|f\|_{l^2(\Omega)} \tag{4.20}$$

does hold for $s < \beta$, but this requires the introduction of fractional order Sobolev spaces, which we have avoided so far. (See, e.g. [Ha, p.122] or [EG, p.484] for the definition of $H^s(\Omega)$.) In this case the result in Theorem 3.24 changes to

$$\|u - u_h\|_{L^2(\Omega)} + h^s\|u - u_h\|_{H^1(\Omega)} \;\lesssim\; h^{2s}\|f\|_{L^2(\Omega)} \;.$$

Note that as w increases to 2π, β decreases to $1/2$ and so s decreases to $1/2$ and the limiting case of the slit domain (see Example 1.17) has the worst regularity.

Example 4.10. *This example shows that if mixed boundary conditions are present we cannot in general expect the regularity assumption (A2) to hold either.*



In this example the domain is a semicircle with unit radius. Let

$$\Gamma = \{(x,0) : 0 \leq x \leq 1\} \cup \{(1,\theta) : \theta \in [0,\pi]\}$$

(the dashed part of the boundary in the figure) and consider the problem

$$\begin{aligned} -\Delta u &= f \quad in \quad \Omega && (4.21)\\ \text{subject to} \quad u &= 0 \quad on \quad \Gamma && (4.22)\\ \text{and} \quad \partial u/\partial\nu &= 0 \quad on \quad \partial\Omega\backslash\Gamma \;. && (4.23) \end{aligned}$$

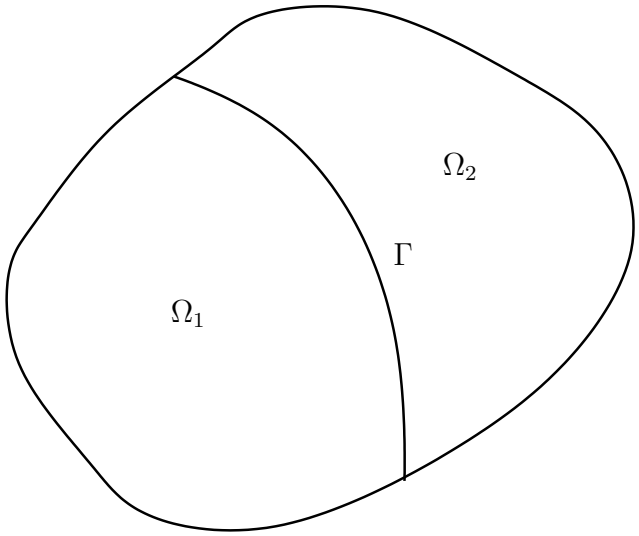
If we consider

$$u(r,\theta) \;=\; (1-r^2)r^{1/2}\sin(\theta/2) \;=\; (r^{1/2} - r^{5/2})\sin(\theta/2) \;,$$

then, as in Example 4.9, $-\Delta u \in L^2(\Omega)$, u satisfies (4.22), (4.23) . In this case (4.20) holds for all $s < 1/2$.

Finally we give an example showing that discontinuous coefficients can also lead to a loss of regularity.

Example 4.11. *Consider the homogeneous Dirichlet problem for* (4.14) *with* $A = aI$ *and* Ω *partitioned into two Lipschitz subdomains as pictured with* $a = a_i$ *on each* Ω_i, $i = 1, 2$. *Supposing* $a_1 \neq a_2$.



Then for $f \in L^2(\Omega)$ u *satsifies, as before,*

$$\int_\Omega a\nabla u.\nabla v \;=\; \int_\Omega fv \;, \quad v \in H_0^1(\Omega) \;. \tag{4.24}$$

With $u_i = u|_{\Omega_i}$ *and* $f_i = f|_{\Omega_i}$, *for* $i = 1, 2$, *since* $C_0^\infty(\Omega_i) \subset H_0^1(\Omega)$, *it follows that*

$$\int_{\Omega_i} a_i\nabla u_i.\nabla\varphi \;=\; \int_{\Omega_i} f_i\varphi \;, \quad \varphi \in C_0^\infty(\Omega_i) \;.$$

Integrating by parts using Corollary 1.23 and recalling that a_i *is constant, implies that*

$$\int_{\Omega_i} (a_i\Delta u_i + f_i)\varphi \;=\; 0$$

for all $\varphi \in C_0^\infty(\Omega_i)$, *which implies that*

$$-a_i\Delta u_i \;=\; f_i \quad \textit{on} \quad \Omega_i \tag{4.25}$$

(in the sense of weak derivatives).

Now consider any $\varphi \in C_0^\infty(\Omega)$. *Then* (4.24) *and integration by parts implies*

$$\begin{aligned}\int_\Omega f\varphi \;&=\; \int_{\Omega_1} a_1\nabla u_1.\nabla\varphi \;+\; \int_{\Omega_2} a_2\nabla u_2.\nabla\varphi \\ &=\; \int_\Gamma (a_1\frac{\partial u_1}{\partial\nu} \;+\; a_2\frac{\partial u_2}{\partial\nu})\varphi \;-\; \int_{\Omega_1} a_1\Delta u_1\varphi \;-\; \int_{\Omega_2} a_2\Delta u_2\varphi\end{aligned}$$

for all $\varphi \in C_0^\infty(\Omega)$, *so by* (4.25),

$$\int_\Gamma (a_1\frac{\partial u_1}{\partial\nu} \;+\; a_2\frac{\partial u_2}{\partial\nu})\varphi \;=\; 0 \;.$$

This holds for arbitrary $\varphi \in C_0^\infty(\Omega)$ *and so*

$$a_1 \frac{\partial u_1}{\partial \nu} \; + \; a_2 \frac{\partial u_2}{\partial \nu} \; = \; 0 \quad \textit{almost everywhere on} \quad \Gamma. \tag{4.26}$$

If now we assume $u \in H^2(\Omega)$ *then* $g := -\Delta u \in L^2(\Omega)$ *and repeating the above argument with* a_1 *and* a_2 *replaced by* 1 *implies*

$$\frac{\partial u_1}{\partial \nu} \; + \; \frac{\partial u_2}{\partial \nu} \; = \; 0 \quad \textit{on} \quad \Gamma,$$

which contradicts (4.26)*, so the regularity assumption fails. In fact in this case* $u \in H^{3/2-\epsilon}(\Omega)$*, for all* $\epsilon > 0$*.*

4.3 An example of a system: Linear elasticity

This is the first example of a system of PDEs and the only one we shall study in this course. Further examples which could be studied include the Stokes and Navier-Stokes equations and Maxwell's equations.

Let $\Omega \subset \mathbb{R}^3$ be an elastic body in its unstressed state. Under stress it undergoes a deformation

$$\boldsymbol{\phi} : \overline{\Omega} \to \mathbb{R}^3 \; .$$

We write $\boldsymbol{\phi} = I + \mathbf{u}$, where I is the 3×3 identity matrix and $\mathbf{u} : \overline{\Omega} \to \mathbb{R}^3$ is the displacement vector which is "small" in the linear theory.

For scalar f and vector $\underset{\sim}{g}$, we introduce the notation

$$\text{grad} f = \nabla f \quad \text{and} \quad \text{div} \underset{\sim}{g} = \nabla . \underset{\sim}{g} \; .$$

Tensor quantities can be identified with 3×3 matrices. Given a vector $\underset{\sim}{u}$ we introduce its gradient, which is a tensor quantity:

$$\underset{\approx}{\text{grad}}(\underset{\sim}{u}) \; = \; \left(\frac{\partial u_i}{\partial x_j} \right)_{i,j=1}^{3} \; .$$

Also if $\underset{\approx}{\tau}$ is a tensor, then its divergence is a vector defined by

$$\underset{\sim}{\text{div}} \left(\underset{\approx}{\tau} \right) \; = \; \sum_{j=1}^{3} \begin{bmatrix} \partial \tau_{1,j} / \partial x_j \\ \partial \tau_{2,j} / \partial x_j \\ \partial \tau_{3,j} / \partial x_j \end{bmatrix} \; .$$

The dot product of two tensors is defined by

$$\underset{\approx}{\tau} : \underset{\approx}{\sigma} = \sum_{i,j} \tau_{i,j} \sigma_{i,j} \; .$$

Now introduce the *stress tensor*

$$\underset{\sim}{\sigma}(\underset{\sim}{u}) = 2\mu\underset{\sim}{\varepsilon}(\underset{\sim}{u}) + \lambda \mathrm{tr}(\underset{\sim}{\varepsilon}(\underset{\sim}{u}))I$$

where I is the 3×3 identity matrix and tr denotes the trace of a matrix, λ and μ are the *Lamé constants*, and $\underset{\sim}{\varepsilon}(\underset{\sim}{u})$ is defined by

$$\underset{\sim}{\varepsilon}(\underset{\sim}{u}) \;=\; \frac{1}{2}\left(\underset{\approx}{\mathrm{grad}}(\underset{\sim}{u}) \;+\; (\underset{\approx}{\mathrm{grad}}(\underset{\sim}{u}))^T\right).$$

Then the displacement $\underset{\sim}{u}$ satisfies the Lamé equation:

$$-\underset{\sim}{\mathrm{div}}\left(\underset{\sim}{\sigma}\left(\underset{\sim}{u}\right)\right) \;=\; \underset{\sim}{f} \quad \text{on} \quad \Omega \tag{4.27}$$

subject to

$$\underset{\sim}{u} = \underset{\sim}{g} \quad \text{on} \quad \partial\Omega_D \tag{4.28}$$

and

$$\underset{\sim}{\sigma}(\underset{\sim}{u})\underset{\sim}{\nu} = \underset{\sim}{t} \quad \text{on} \quad \partial\Omega_N\ , \tag{4.29}$$

with $\underset{\sim}{f}, \underset{\sim}{g}$ and $\underset{\sim}{t}$ given (respectively these are the *body force*, *boundary displacement* and *boundary traction*) and $\underset{\sim}{\nu}$ denoting the outward normal vector from Ω on the boundary $\partial\Omega$.

To get the weak form of (4.27), let $\underset{\sim}{v} \in H^1(\Omega)^3$ with $\underset{\sim}{v} = \underset{\sim}{0}$ on $\partial\Omega_D$. Then, abbreviating $\underset{\sim}{\sigma}(\underset{\sim}{u})$ by $\underset{\sim}{\sigma}$ we have

$$\begin{aligned}
-\int_\Omega \underset{\sim}{\mathrm{div}}(\underset{\sim}{\sigma}).\underset{\sim}{v} \;&=\; -\sum_i \sum_j \int_\Omega \frac{\partial \sigma_{i,j}}{\partial x_j} v_i \\
&=\; \sum_i \sum_j \left\{ \int_\Omega \sigma_{i,j} \frac{\partial v_i}{\partial x_j} - \int_{\partial\Omega} \sigma_{i,j} v_i \nu_j \right\}\ , \quad \text{(by Corollary (1.22))} \\
&=\; \int_\Omega \underset{\sim}{\sigma} : \underset{\approx}{\mathrm{grad}}(\underset{\sim}{v}) - \int_{\partial\Omega} \underbrace{\left(\underset{\sim}{\sigma}\underset{\sim}{\nu}\right)}_{=\underset{\sim}{t}} .\underset{\sim}{v}\ .
\end{aligned}$$

So (4.27) implies that $\underset{\sim}{u}$ satisfies the weak form:

$$\int_\Omega \underset{\sim}{\sigma}(\underset{\sim}{u}) : \underset{\approx}{\mathrm{grad}}(\underset{\sim}{v}) \;=\; (\underset{\sim}{f}, \underset{\sim}{v})_{L^2(\Omega)^3} \;+\; (\underset{\sim}{t}, \underset{\sim}{v})_{L^2(\Omega_N)^3} \tag{4.30}$$

where

$$(\underset{\sim}{f}, \underset{\sim}{g})_{L^2(\Omega)^3} \;:=\; \sum_i \int_\Omega f_i g_i\ .$$

The bilinear form on the right-hand side of (4.30) is not symmetric. However, note that

$$\underset{\sim}{\underset{\sim}{\varepsilon}}(\underline{u}) : \underset{\sim}{\underset{\sim}{\mathrm{grad}}}(\underline{v}) \quad = \quad \frac{1}{2}\sum_{i,j}\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right)\frac{\partial v_i}{\partial x_j}$$

Since the double sum on the right-hand side is over all i and j, the result is unchanged if we interchange i and j in the summand, i.e.

$$\underset{\sim}{\underset{\sim}{\varepsilon}}(\underline{u}) : \underset{\sim}{\underset{\sim}{\mathrm{grad}}}(\underline{v}) \quad = \quad \frac{1}{2}\sum_{i,j}\left(\frac{\partial u_j}{\partial x_i}+\frac{\partial u_i}{\partial x_j}\right)\frac{\partial v_j}{\partial x_i}$$

Summing these last two identities, and then dividing by 2, we obtain

$$\underset{\sim}{\underset{\sim}{\varepsilon}}(\underline{u}) : \underset{\sim}{\underset{\sim}{\mathrm{grad}}}(\underline{v}) \;=\; \underset{\sim}{\underset{\sim}{\varepsilon}}(\underline{u}) : \underset{\sim}{\underset{\sim}{\varepsilon}}(\underline{v}) \;. \tag{4.31}$$

Also it is easy to see by direct calculation that

$$(\mathrm{tr}(\underset{\sim}{\underset{\sim}{\varepsilon}}(\underline{u}))I) : \underset{\sim}{\underset{\sim}{\mathrm{grad}}}(\underline{v}) \;=\; \mathrm{div}(\underline{u})\;\mathrm{div}(\underline{v}) \;. \tag{4.32}$$

Thus, combining (4.30) with (4.31) and (4.32) we have the symmetric form:

$$\int_\Omega \left[2\mu\;\underset{\sim}{\underset{\sim}{\varepsilon}}(\underline{u}) : \underset{\sim}{\underset{\sim}{\varepsilon}}(\underline{v}) \;+\; \lambda\;\mathrm{div}(\underline{u})\;\mathrm{div}(\underline{v})\right] \;=\; (\underset{\sim}{f},\underline{v})_{L^2(\Omega)} \;+\; (\underline{t},\underline{v})_{L^2(\partial\Omega_N)} \;, \tag{4.33}$$

for all $\underline{v} \in H^1(\Omega)^3 \quad$ with $\quad \underline{v} = \underline{0} \quad$ on $\quad \partial\Omega_D$.

Writing (4.33) in the abstract form:

$$\underline{a}(\underline{u},\underline{v}) \;=\; \underline{F}(\underline{v}),$$

this is the weak form of (4.27) - (4.29). It is easy to see that $\underline{a}$ is bounded on $H^1(\Omega)^3$. Note that the norm on $H^1(\Omega)^3$ is induced by the inner product

$$(\underline{u},\underline{v})_{H^1(\Omega)^3} \;:=\; \int_\Omega \underline{u}.\underline{v} \;+\; \int_\Omega \underset{\sim}{\underset{\sim}{\mathrm{grad}}}(\underline{u}) : \underset{\sim}{\underset{\sim}{\mathrm{grad}}}(\underline{v}) \;.$$

Moreover it turns out that $\underline{a}$ is coercive on the space $V = \left\{\underline{v} \in H^1(\Omega)^3 \;: \underline{v} = \underline{0} \quad \text{on} \quad \partial\Omega_D\right\}$, this is given in the following lemma.

Lemma 4.12. *(***Korn's inequality***) If $|\partial\Omega_D| > 0$ then there exists a constant $C' > 0$ such that*

$$\int_\Omega \underset{\sim}{\underset{\sim}{\varepsilon}}(\underline{v}) : \underset{\sim}{\underset{\sim}{\varepsilon}}(\underline{v}) \;\geq\; C'\|\underline{v}\|_{H^1(\Omega)^3} \;. \tag{4.34}$$

Proof. We give the proof only for the special case when $\partial\Omega_N = \emptyset$, the general case is found for example in [BS]. The proof makes use of the following identity which should be checked as an exercise. (Just calculate each side by brute force!)

$$2\underset{\sim}{\underset{\sim}{\varepsilon}}(\underset{\sim}{u}) : \underset{\sim}{\underset{\sim}{\varepsilon}}(\underset{\sim}{u}) \;=\; \mathrm{div}\left([\underset{\sim}{v}\ \mathrm{grad}]\underset{\sim}{v} - (\mathrm{div}\underset{\sim}{v})\underset{\sim}{v}\right) \;+\; \underset{\sim}{\underset{\sim}{\mathrm{grad}}}(\underset{\sim}{v}) : \underset{\sim}{\underset{\sim}{\mathrm{grad}}}(\underset{\sim}{v}) + (\mathrm{div}(\underset{\sim}{v}))^2\ , \qquad (4.35)$$

where "$[\underset{\sim}{v}\ \mathrm{grad}]$" means the operator $\sum_i v_i \partial/\partial x_i$. Using (4.35) and employing the divergence theorem, and the fact that $\underset{\sim}{v} = \underset{\sim}{0}$ on $\partial\Omega = \partial\Omega_D$, we obtain

$$\underset{\sim}{a}(\underset{\sim}{v}, \underset{\sim}{v}) \;\geq\; \mu \int_\Omega \underset{\sim}{\underset{\sim}{\mathrm{grad}}}(\underset{\sim}{v}) : \underset{\sim}{\underset{\sim}{\mathrm{grad}}}(\underset{\sim}{v}) \;=\; \mu\ |\underset{\sim}{v}|^2_{H^1(\Omega)^3} \;\geq\; C' \|\underset{\sim}{v}\|_{H^1(\Omega)^3}$$

with the last step due to the Poincaré inequality, and this proves the result.

□

So now the linear elasticity problem is formulated as a bounded and coercive bilinear form on a subspace of $H^1(\Omega)^3$. All the theory developed earlier in the course can be applied directly and any of the families of finite elements which are conforming for $H^1(\Omega)$ can be applied in a vectorised way to discretise this problem.

5 Adaptivity and Conditioning

These notes are a longer version of the material I gave in the last 2 lectures of the course, where the argument was very condensed. In particular the result in Exercise 5.7 was proved as part of the lectures in 2009.

In this chapter we study issues related to the iterative solutions of finite element approximations of

$$\left.\begin{array}{rcl} -\nabla.\mathcal{A}\nabla u &=& f \quad \text{in} \quad \Omega \\ u &=& 0 \quad \text{on} \quad \partial\Omega \end{array}\right\} \tag{5.1}$$

Throughout we assume that the coefficient matrix $\mathcal{A}$ is $(L^\infty(\Omega))^{n\times n}$ and is uniformly positive definite, so that the induced bilinear form $a(\cdot,\cdot)$ is bounded and coercive.

We will concentrate on linear finite elements on simplices in $\Omega \subset \mathbb{R}^n,\ n = 2$ or 3.

Let $\mathcal{T}_h$ be a sequence of regular meshes on Ω consisting of simplices with $h \to 0$. Let $\{\phi_i : i \in \mathcal{N}(\Omega)\}$ be the nodal basis, with ϕ_i linear on each element $K \in \mathcal{T}_h$,

$$\phi_i(x_j) \ = \ \delta_{ij}, \quad i,j \in \mathcal{N}(\Omega).$$

Then the finite element approximation of (5.1) leads to

$$A\mathbf{U} \ = \ \mathbf{f} \ , \tag{5.2}$$

where

$$A_{ij} \ = \ a(\phi_j, \phi_i) \ . \tag{5.3}$$

A is symmetric positive definite and

$$\begin{array}{rcl} \kappa(A) &=& \|A\|_2\|A^{-1}\|_2 \\ &=& \dfrac{\lambda_{max}(A)}{\lambda_{min}(A)} \ . \end{array} \tag{5.4}$$

The condition number $\kappa(A)$ determines the rate of convergence of many iterative methods, e.g. conjugate gradients, for solving (5.2).

If $\mathbf{U}^{(k)}$ is the k-th conjugate gradient iterate for (5.2), then

$$\frac{\|\mathbf{U}-\mathbf{U}^{(k)}\|_A}{\|\mathbf{U}-\mathbf{U}^{(0)}\|_A} \ \le \ 2\left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^k \tag{5.5}$$

where $\|\mathbf{x}\|_A^2 = \mathbf{x}^T A\mathbf{x}$ (see [ToWi: p.403]).

Hence if

$$k \log\left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}}\right) \;\leq\; \log\left(\frac{\epsilon}{2}\right)$$

then the RHS of (5.5) is less than ϵ.

So if

$$k\left(-\frac{1}{\sqrt{\kappa(A)}}\right) \;\leq\; \log\left(\frac{\epsilon}{2}\right)$$

then the relative error on the LHS of (5.5) is $\leq$ than ϵ. Equivalently

$$k \;\geq\; \sqrt{\kappa(A)}\log\left(\frac{2}{\epsilon}\right).$$

So the rate of convergence (measured in terms of a lower bound for the number of iterations needed) is determined by $\kappa(A)$.

We will now analyse $\kappa(A)$. Since A is symmetric positive definite we have

Lemma 5.1.

$$\begin{aligned}\lambda_{max}(A) &= \max_{\mathbf{V}\neq 0}\frac{\mathbf{V}^T A\mathbf{V}}{\mathbf{V}^T\mathbf{V}}\\ \lambda_{min}(A) &= \min_{\mathbf{V}\neq 0}\frac{\mathbf{V}^T A\mathbf{V}}{\mathbf{V}^T\mathbf{V}}\end{aligned}$$

Proof. See [ToWi, Lemma C1] □

Exercise 5.2. *Read [BS, §9.2]. There it is explained how to obtain error estimates of the form*

$$|u-u_h|_{H^1(\Omega)} \;\lesssim\; \left\{\sum_e \mathcal{E}_e(u_h)^2\right\}^{1/2}, \tag{5.6}$$

where $\mathcal{E}_e(u_h)$ is an "a posteriori" error estimated associated with each e denoting an interior face (in 3D) or edge (in 2D) of the mesh and the sum is over all such e. The main point is that $\mathcal{E}_e(u_h)$ can be computed only from the numerical solution u_h on the current mesh, and so (5.6) *is an error estimate in terms of computable right hand side. (In general of course there is a hidden constant on the right hand side, but it can be proved that this constant does not depend on u_h, or u or on the mesh – more details are in [BS])*

Adaptive methods *then proceed as follows:*

- *Compute u_h on an initial (coarse) mesh;*
- *Compute $\mathcal{E}_e(u_h)$ for each interior face (3D) or edge (2D);*

- *Refine the elements touching e for all those e for which $\mathcal{E}_e(u_h)$ is (relatively) high. Refinement in 2D can be obtained for example by subdividing a triangle into four congruent triangles by joining the mid-points of edges - see e.g. [BS, fig. 9.2] for some pictures.*

- *Solve again on the refined mesh; repeat until the error estimate is sufficiently small.*

It turns out to be useful to consider scaling the nodal basis ϕ_i. (This has only an effect if local mesh refinement has been done and has no impact if we are dealing with uniform meshes.)

Let us recall that $\mathcal{T}_h$ is regular and therefore the elements cannot become long and thin. Elements near each other must be of comparable size (but elements far away from each other might be of different sizes - this is called *local mesh refinement*). For any pair of elements K, K' with $x_i \in \overline{K}, \overline{K'}$ the regularity implies

$$h_K \ \sim \ h_{K'}$$

(Recall the definition of $\lesssim$ and $\gtrsim$ in chapter 3; $\sim$ means $\lesssim$ and $\gtrsim$.)

So, by choosing

$$h_i \ = \ \max_{K: x_i \in \overline{K}} h_K \ ,$$

we have

$$h_i \ \sim \ h_K \quad \forall K \quad \text{s.t.} \quad x_i \in \overline{K} \ . \tag{5.7}$$

For each x_i there is a hat function ϕ_i and now an h_i which we use to obtain the scaled basis

$$\psi_i \ = \ h_i^{\frac{2-n}{2}} \phi_i \tag{5.8}$$

Although the mesh is regular, it could be locally refined and the scaling for $n = 3$ gives ψ_i a relatively large gradient in regions where the elements are relatively small. If the mesh is (nearly) uniform, the scaling (5.8) leads to a (nearly) uniform scaling of all elements of the stiffness matrix and so this has no effect on the condition number.

So we now approximate the problem (5.1) using the scaled basis ψ_i in which case the stiffness matrix now becomes:

$$A_{ij} \ = \ a(\psi_j, \psi_i)$$

Let

$$v_h \ = \ \sum_{i \in \mathcal{N}(\Omega)} V_i \psi_i$$

where $v_h \in V_h = \text{span}\{\psi_i\} = \text{span}\{\phi_i\}$.

Note

$$\begin{aligned} v_h(x_j) &= \sum_{i\in\mathcal{N}(\Omega)} V_i h_i^{\frac{2-n}{2}} \phi_i(x_j) \\ &= h_j^{\frac{2-n}{2}} V_j \end{aligned} \tag{5.9}$$

To analyse $\kappa(A)$ we need two lemmas:

Lemma 5.3. *For all $K \in \mathcal{T}_h$*

$$\sum_{x_i\in\overline{K}} V_i^2 \;\sim\; h_K^{n-2}\|v_h\|^2_{L^\infty(K)}$$

Proof. Since v_h is linear on K

$$\begin{aligned} h_K^{n-2}\|v_h\|^2_{L^\infty(K)} &= h_K^{n-2}\max_{x_i\in\overline{K}}(v_h(x_i))^2 \\ &= h_K^{n-2}\max_{x_i\in\overline{K}}(h_i^{2-n}V_i^2) \\ &\sim (\max_{x_i\in\overline{K}}|V_i|)^2 \\ &\sim \sum_{x_i\in\overline{K}}|V_i|^2 \qquad \text{(by norm equivalence on } \mathbb{R}^{d+1}) \end{aligned}$$

which completes the proof. □

Lemma 5.4 (Inverse estimates)**.** *For all $v_h \in V_h$*

(i)

$$\|v_h\|_{H^1(K)} \;\lesssim\; h_K^{-1}\|v_h\|_{L^2(K)} \quad , \quad \forall K \in \mathcal{T}_h \ .$$

(ii) For $p \geq q$

$$\|v_h\|_{L^p(K)} \;\sim\; h_K^{n\left(\frac{1}{p}-\frac{1}{q}\right)}\|v_h\|_{L^q(K)} \ , \quad \forall K \in \mathcal{T}_h \ .$$

Remark. Of course estimates like $\|v\|_{H^1(K)} \leq C\|v\|_{L^2(K)}$ cannot hold for general $v \in H^1(K)$. The point in the lemma above is that it can hold when v lies in a finite dimensional space (here the linear functions on K) but then with a constant C which blows up as $h_K \to \infty$. Note that the sort of estimates in Lemma 5.4 are not restricted to linear elements, any polynomial elements have the same property (but the hidden constant depends on the degree of the elements).

Proof. Part (i):
With F_K denoting the affine map from $\widehat{K} \to K$ in (3.9) we set

$$\hat{v} \;=\; v \circ F_K \quad \text{on} \quad \widehat{K} \ .$$

Then, using Lemmas 3.18 and 3.19 , and recalling that the meshes were assumed regular,

$$\begin{aligned} |v_h|_{H^1(K)} &\lesssim |A_K^{-1}| \, (\det A_K)^{\frac{1}{2}} \, |\hat{v}_h|_{H^1(\widehat{K})} \\ &\lesssim h_K^{-1} \, (\det A_K)^{\frac{1}{2}} \, |\hat{v}_h|_{H^1(\widehat{K})} \end{aligned} \tag{5.10}$$

Now note that by the equivalence of norms on the finite dimensional space of linear $\hat{p}$ on $\widehat{K}$,

$$\|\hat{p}\|_{H^1(\widehat{K})} \lesssim \|\hat{p}\|_{L^2(\widehat{K})}$$

So

$$|\hat{v}_h|_{H^1(\widehat{K})} \leq \|\hat{v}_h\|_{H^1(\widehat{K})} \lesssim \|\hat{v}_h\|_{L^2(\widehat{K})}$$

Hence (5.10) implies

$$\begin{aligned} |v_h|_{H^1(K)} &\lesssim h_K^{-1} (\det A_K)^{\frac{1}{2}} \|\hat{v}_h\|_{L^2(\widehat{K})} \\ &\lesssim h_K^{-1} \|v_h\|_{L^2(K)} \end{aligned}$$

This implies the first result (i).

Part (ii)

$$\begin{aligned} \|v_h\|_{L^p(K)}^p &= \int_K |v_h|^p \\ &= \int_{\widehat{K}} (\det A_K) |\hat{v}_h|^p \\ &= (\det A_K) \int_{\widehat{K}} |\hat{v}_h|^p \\ &= (\det A_K) \|\hat{v}_h\|_{L^p(\widehat{K})}^p \end{aligned}$$

So

$$\|v_h\|_{L^p(K)} = (\det A_K)^{\frac{1}{p}} \|\hat{v}_h\|_{L^p(\widehat{K})}$$

and again by the equivalence of norms on linear functions on $\widehat{K}$

$$\begin{aligned} \|v_h\|_{L^p(K)} &\sim (\det A_K)^{\frac{1}{p}} \|\hat{v}_h\|_{L^q(\widehat{K})} \\ &= (\det A_K)^{\frac{1}{p}-\frac{1}{q}} \|v_h\|_{L^q(K)} \, . \end{aligned}$$

Now we use that

$$\begin{aligned}(\det A_K) &\sim (\det A_K)\int_{\widehat{K}} 1\\ &= \int_{\widehat{K}}(\det A_K)\\ &= \int_K 1\\ &= \text{volume of } K\\ &\sim h_K^n\end{aligned}$$

This implies (ii). □

We now obtain the main result:

Theorem 5.5. *Let N be the number of nodes in $\mathcal{N}(\Omega)$. Then*

$$\kappa(A) \lesssim \begin{cases} N^{\frac{2}{n}} & \text{for} \quad n=3 \\ N^{\frac{2}{n}+\epsilon} & \text{for} \quad n=2, \epsilon>0 \end{cases}$$

Proof. For $n=3$ (the case $n=2$ is slightly harder [BrSc, p.250]):

Recall from Lemma 5.1 that we need to find bounds for the maximum and minimum of

$$\frac{\mathbf{V}^T A\mathbf{V}}{\mathbf{V}^T\mathbf{V}}$$

over all vectors $\mathbf{V}$ defined on the interior nodes of $\mathcal{T}^h$ in Ω. First, for the upper bound we proceed as follows:

$$\begin{aligned}\mathbf{V}^T A\mathbf{V} &= a(v_h,v_h) \qquad (\text{where} \quad v_h = \sum_{i\in\mathcal{N}(\Omega)} V_i\psi_i)\\ &\lesssim \|v_h\|^2_{H^1(\Omega)}\\ &= \sum_{K\in\mathcal{T}} \|v_h\|^2_{H^1(K)}\\ &\lesssim \sum_{K\in\mathcal{T}} h_K^{-2}\|v_h\|^2_{L^2(K)} \qquad \text{(by Lemma 5.4 (i))}\\ &\sim \sum_{K\in\mathcal{T}} h_K^{n-2}\|v_h\|^2_{L^\infty(K)} \qquad \text{(by Lemma 5.4 (ii) with} \quad q=\infty, p=2)\\ &\sim \sum_{K\in\mathcal{T}}\sum_{x_i\in\overline{K}} V_i^2 \qquad \text{(by Lemma 5.3)}\\ &\lesssim \mathbf{V}^T\mathbf{V}\end{aligned}$$

Hence

$$\lambda_{max}(A) \lesssim 1 .$$

For estimating $\lambda_{min}(A)$, we have

$$\begin{aligned} \mathbf{V}^T\mathbf{V} &\leq \sum_{K\in\mathcal{T}_h}\sum_{x_i\in\overline{K}} V_i^2 \\ &\sim \sum_{K\in\mathcal{T}_h} h_K^{n-2}\|v_h\|^2_{L^\infty(K)} \qquad \text{(by Lemma 5.3) .} \end{aligned}$$

Now by Lemma 5.4 (ii) we get with $p=\infty$, $q=\frac{2n}{n-2}$

$$\|v_h\|_{L^\infty(K)} \sim h_K^{\frac{2-n}{2}}\|v_h\|_{L^{\frac{2n}{n-2}}(K)} .$$

(Note $n\left(\frac{1}{\infty}-\frac{n-2}{2n}\right)=\frac{2-n}{2}$.)

So

$$\begin{aligned} \mathbf{V}^T\mathbf{V} &\lesssim \sum_{K\in\mathcal{T}_h} 1\|v_h\|^2_{L^{\frac{2n}{n-2}}(K)} \\ &\leq \left(\sum_{K\in\mathcal{T}_h} 1^{\frac{n}{2}}\right)^{\frac{2}{n}} \left(\sum_{K\in\mathcal{T}_h}\|v_h\|^{\frac{2n}{n-2}}_{L^{\frac{2n}{n-2}}(K)}\right)^{\frac{n-2}{n}} \end{aligned} \tag{5.11}$$

Here we used Hölder's inequality [Ta, p.6] with $p=\frac{n}{2}$, $q=\frac{n}{n-2}$.

Since the number of elements $\{K : K\in\mathcal{T}_h\}$ $\sim$ the number of nodes in $\mathcal{T}_h$ $= N$, we obtain

$$\mathbf{V}^T\mathbf{V} \lesssim N^{\frac{2}{n}}\|v_h\|^2_{L^{\frac{2n}{n-2}}(\Omega)} .$$

Now we use the Sobolev embedding theorem (see for example [Ci, p.114]) to obtain

$$\begin{aligned} \mathbf{V}^T\mathbf{V} &\lesssim N^{\frac{2}{n}}\|v_h\|^2_{H^1(\Omega)} \\ &\lesssim N^{\frac{2}{n}}\, a(v_h,v_h) \qquad \text{(coercivity)} \\ &= N^{\frac{2}{n}}\,\mathbf{V}^T A\mathbf{V} \end{aligned}$$

Therefore $\lambda_{min}(A)\gtrsim N^{-\frac{2}{n}}$ which yields the result.

The case $n=2$ is more technical, because the embedding theorem ensures only that $H^1(\Omega)$ is embedded in $L^q(\Omega)$ for all $q<\infty$ and not in $L^\infty(\Omega)$. ($L^\infty(\Omega)$ would be required if we were going to repeat the above argument verbatim for $n=2$). This is where the ϵ comes from in the statement of the theorem. However a simpler argument can be applied in the case of quasiuniform meshes (see below) to avoid the ϵ. □

Remark 5.6.

(i) Theorem 5.5 tells us that the number of CG iterations may grow with $O(N^{\frac{1}{n}})$

(ii) $\mathcal{T}_h$ *is called quasiuniform if*

$$h \lesssim h_K \leq h, \quad \text{for all} \quad K \in \mathcal{T}_h \quad \text{as} \quad h \to 0 \ .$$

In this case $N \sim h^{-n}$.

Exercise 5.7. *Prove the result* $\kappa(A) \lesssim h^{-2}$ *for the case of a quasiuniform mesh for* $n = 2$ *or* 3. *Hint: Use boundedness, coercivity and inverse estimates to establish:*

$$\|v_h\|^2_{L^2(\Omega)} \lesssim \mathbf{V}^T A \mathbf{V} \lesssim h^{-2} \|v_h\|^2_{L^2(\Omega)} \ ,$$

where $v_h = \sum_{i \in \mathcal{N}(\Omega)} V_i \phi_i$. *Then show*

$$\|v_h\|^2_{L^2(\Omega)} \sim h^n \sum_{K \in \mathcal{T}_h} \sum_{x_i \in \overline{K}} |V_j|^2$$

and combine these two results.

6 Domain Decomposition Preconditioning

This chapter was not lectured in 2009, so does not form part of the exam

Consider again

$$a(u,v) \;=\; \int_\Omega (\mathcal{A}\nabla u)\,.\nabla v \tag{6.1}$$

where $\mathcal{A}$ is $L^\infty(\Omega)$ elementwise and uniformly positive definite. Consider then the Dirichlet problem approximation in the finite element space

$$V_h \;=\; \text{span}\{\phi_i^h : i \in \mathcal{N}^h(\Omega)\}$$

where ϕ_i^h is the nodal basis without scaling for linear finite elements on simplices on Ω with mesh $\mathcal{T}_h$.

This gives a linear system

$$A\mathbf{U} \;=\; \mathbf{f}\;. \tag{6.2}$$

where $A_{ij} \;=\; a(\phi_j^h, \phi_i^h)$ and A is symmetric positive definite.

In section 5 we saw that A is illconditioned as $h \to 0$. **Preconditioning** of (6.2) involves choosing a symmetric positive definite B^{-1} and solving

$$B^{-\frac{1}{2}} A B^{-\frac{1}{2}} \widetilde{\mathbf{U}} \;=\; \widetilde{\mathbf{f}} \tag{6.3}$$

where

$$\widetilde{\mathbf{U}} \;=\; B^{\frac{1}{2}}\mathbf{U}, \qquad \widetilde{f} \;=\; B^{-\frac{1}{2}}\mathbf{f}\;.$$

The conjugate gradient method for (6.3) can be organised so that only multiplications by A and solves with B are needed. Therefore we require

(i) Solves with B to be "cheap"

(ii) $\kappa(B^{-\frac{1}{2}} A B^{-\frac{1}{2}})$ to be small

Note that since $B^{-\frac{1}{2}} A B^{-\frac{1}{2}}$ is similar to $B^{-1}A$ (and so these two matrices have the same eigenvaues), we have

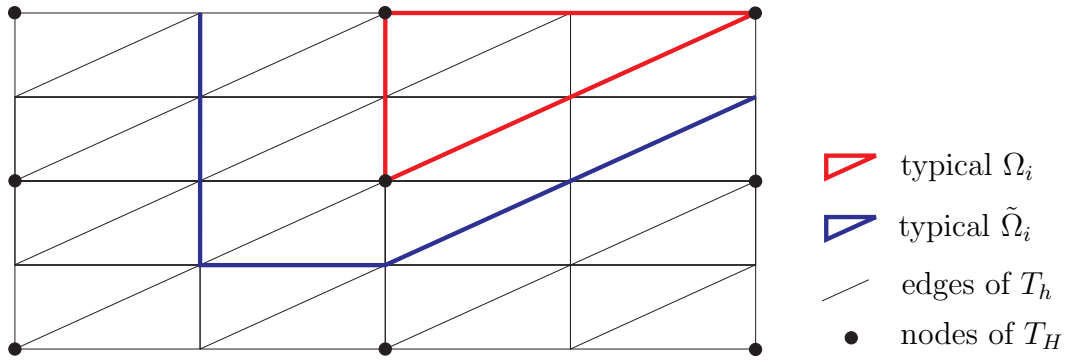
$$\kappa(B^{-\frac{1}{2}} A B^{-\frac{1}{2}}) \;=\; \frac{\lambda_{max}(B^{-1}A)}{\lambda_{min}(B^{-1}A)} \;=:\; \kappa(B^{-1}A) \tag{6.4}$$

In domain decomposition we construct B^{-1} from "partial solves" of A corresponding to finite element approximations in subspaces of V_h.

Suppose $\mathcal{T}_h$ is a refinement of a coarser simplicial mesh $\mathcal{T}_H$ $(H > h)$ and $\{\Omega_i : i = 1, \ldots, s\}$ are elements of $\mathcal{T}_H$. We extend each Ω_i to a connected subdomain $\tilde{\Omega}_i$ with

$$\Omega_i \subset \tilde{\Omega}_i \subset \Omega$$

and require that $\partial\tilde{\Omega}_i$ contains only edges (in 2D) or faces (in 3D) of elemente of $\mathcal{T}_h$. Furthermore we assume $\tilde{\Omega}_i$ overlaps a finite number of $\tilde{\Omega}_j, j \neq i$ as $H, h \to 0$.



Let

$$\delta \;=\; \min_i \Big(\mathrm{dist}(\delta\Omega_i, \delta\tilde{\Omega}_i)\Big) \tag{6.5}$$

With $\mathcal{N}^h(\tilde{\Omega}_i)$ we denote nodes of $\mathcal{T}_h$ in $\tilde{\Omega}_i$. We introduce the restriction matrix

$$(R_i)_{j,j'} \;=\; \delta_{jj'}, \qquad j \in \mathcal{N}^h(\tilde{\Omega}_i),\; j' \in \mathcal{N}^h(\Omega_i)$$

So, for a vector $\mathbf{V}$ defined on nodes of $\mathcal{T}_h$ in Ω, $R_i\mathbf{V}$ is the vector of its values at the nodes of an $\tilde{\Omega}_i$ and

$$A_i \;=\; R_i A R_i^T \tag{6.6}$$

is the subblock of A corresponding to rows and columns in $\tilde{\Omega}_i$.

The first domain decomposition preconditioner is

$$B_1^{-1} \;=\; \sum_{i=1}^{s} R_i^T A_i^{-1} R_i \tag{6.7}$$

Unfortunately, $\kappa(B_1^{-1}A)$ blows up as $H, h \to 0$. To improve the preconditioner we add a global coarse solve on $\mathcal{T}_H$.

Let

$$V_H \;=\; \mathrm{span}\{\Phi_p^H : p \in \mathcal{N}^H(\Omega)\}$$

be a nodal basis for linear finite elements on $\mathcal{T}_H$ (with 0 on $\partial\Omega$).

For $p \in \mathcal{N}^H(\Omega)$, $j \in \mathcal{N}^h(\Omega)$ we set

$$(R_0^T)_{j,p} \;=\; \Phi_p^H(x_j^h)$$

This matrix represents interpolation of nodal values on $\mathcal{T}_H$ to nodal values on $\mathcal{T}_h$ using linear basis functions. Then set

$$A_0 \;=\; R_0 A R_0^T$$

and extend (6.7) to

$$B^{-1} \;=\; \sum_{i=0}^{s} R_i^T A_i^{-1} R_i \tag{6.8}$$

This is called "Additive Schwarz".

We will prove that

$$\kappa(B_1^{-1}A) \lesssim 1$$

independent of H, h, provided $\delta \gtrsim H$.

Note

$$B^{-1}A \;=\; \sum_{i=0}^{s} R_i^T A_i^{-1} R_i A \tag{6.9}$$

To analyse this, let $N^h = \#\mathcal{N}^h(\Omega)$. Then for $\mathbf{V}, \mathbf{W} \in \mathbb{R}^{N^h}$ set

$$\begin{aligned} < \mathbf{V}, \mathbf{W} >_A \;&=\; \mathbf{V}^T A \mathbf{W} \\ &=\; a(v_h, w_h) \end{aligned} \tag{6.10}$$

where

$$v_h \;=\; \sum_{j \in \mathcal{N}^h(\Omega)} \mathbf{V}_j \phi_j^h, \quad w_h \;=\; \sum_{j \in \mathcal{N}^h(\Omega)} \mathbf{W}_j \phi_j^h \tag{6.11}$$

Set

$$\begin{aligned} \mathcal{V}^i \;&=\; \{v_h \in V_h : \mathrm{supp}(v_h) \subset \tilde{\Omega}_i\}, \quad i = 1, \ldots, s \\ \mathcal{V}^0 \;&=\; V_H \;=\; \text{FE space on the coarse mesh} \end{aligned}$$

Lemma 6.1. *For $i = 0, \ldots, s$*

$$< R_i^T A_i^{-1} R_i A \, \boldsymbol{W}, \boldsymbol{V} >_A \;=\; a(\Pi_i w_h, v_h) \tag{6.12}$$

where $\Pi_i : V_h \to \mathcal{V}^i$ and is the orthogonal projection with respect to $a(\cdot, \cdot)$.

Proof. Define $\Pi_i w_h \in V_h$ to have nodal vector $R_i^T A_i^{-1} R_i A \mathbf{W}$. Then $\Pi_i w_h \in \mathcal{V}^i$ and (6.12) holds by construction. Moreover let $v_h \in \mathcal{V}^i$ then

$$\mathbf{V} \;=\; R_i^T \mathbf{Z}^i$$

for some $\mathbf{Z}^i$ and using the symmetry of A and A_i

$$\begin{aligned} a(\Pi_i w_h, v_h) &= < R_i^T A_i^{-1} R_i A \mathbf{W}, R_i^T \mathbf{Z}^i >_A \\ &= \mathbf{W}^T A R_i^T A_i^{-1} \underbrace{R_i A R_i^T}_{=A_i} \mathbf{Z}^i \\ &= \mathbf{W}^T A R_i^T \mathbf{Z}^i \\ &= \mathbf{W}^T A \mathbf{V} \\ &= < \mathbf{W}, \mathbf{V} >_A \\ &= a(w_h, v_h) \end{aligned}$$

□

Appendix: Assumed Material on Functional Analysis

Banach and Hilbert Spaces.

We shall consider normed vector spaces over the field of real numbers. In some cases the extension to complex fields may be needed.

- A *Banach space* X is a normed vector space which is complete (i.e. all Cauchy sequences converge to an element of the space).
- An *inner product space* H is a normed vector space where the norm is derived from an inner product $\|x\|_H = (x,x)_H^{1/2}$.
- A *Hilbert space* is a complete inner product space.

A linear operator T from a Banach space X to a Banach space Y is called *bounded* if there exists a constant $C = C(T)$ such that $\|Tx\|_Y \leq C\|x\|_X$ for all $x \in X$. The norm of T is defined by

$$\|T\| = \sup_{0 \neq x \in X} \frac{\|Tx\|_Y}{\|x\|_X} .$$

A linear operator ϕ which maps X to the (real) scalars is called a *linear functional* on X. The space of all bounded linear functionals on X is called *dual space* and is denoted X'. The norm on X' is

$$\|\phi\|_{X'} = \sup_{0 \neq x \in X} \frac{|\phi(x)|}{\|x\|_X} .$$

Theorem A.1 The Hahn-Banach Theorem [Pr, page 116],[Ta,Thm 4.3-A]

Let f_0 be a bounded linear functional defined on a (linear) subspace M of a normed space X. Then there exists a bounded linear functional f on X such that $f(x) = f_0(x)$, when $x \in M$ and such that $\|f\|_{X'} = \|f_0\|_{M'}$.

Theorem A.2 Banach's Isomorphism Theorem [Pr, page 145],[Ta, Thm 4.2-H].

Suppose T is a bounded linear operator from a Banach space X to a Banach space Y, and that T is a bijection. Then the inverse T^{-1} of T is a bounded linear operator from Y to X. (This is a corollary of The Open Mapping Theorem.)

□

Theorem A.3 Hilbert Space Projection Theorem [Pr, pg 173],[Ta, Thm4.82-A]

Let U be a closed subspace of a Hilbert space H. Then its orthogonal complement $U^\perp = \{v \in H : (v,u)_H = 0 \text{ for all } u \in U\}$ is also a closed subspace of H and

$$H = U \oplus U^\perp \ ,$$

i.e. $U \cap U^\perp = \{0\}$ and each $x \in H$ can be written as $x = u + v$, with $u \in U$ and $v \in U^\perp$. □

Theorem A.4 The Riesz Representation Theorem. [Ha, Thm 6.3.6; Pr, Thm 12.10 Let H be a Hilbert space. Then, for each $\phi \in H'$, there exists a unique $u \in H$ such that

$$\phi(v) = (u,v)_H \ , \text{ for all } v \in H \ .$$

Moreover the map $\tau : \phi \mapsto u$, is linear and is an isometry from H' to H, i.e. $\|\tau\phi\|_H = \|\phi\|_{H'}$. The space H', when endowed with the inner product $(\phi,\psi)_{H'} := (\tau\phi, \tau\psi)_H$ for all $\phi, \psi \in H'$, becomes a Hilbert space and the induced norm coincides with the usual dual norm on H' .

□

References:

- [GRB] G.R. Burton, Sobolev Spaces, Lecture Course in Taught Course Centre (Autumn 2007).
- [BS] S.C. Brenner and L.R. Scott, The Mathematical Theory of Finite Element Methods, Second Edition, Springer, 2002.
- [C] P.G. Ciarlet, The Finite Element Method for Elliptic Problems, North Holland, 1978.
- [ESW] H.C. Elman, D.J. Silvester and A.J. Wathen, Finite Elements and Fast Iterative Solvers, Oxford University Press, 2005
- [EG] A. Ern and J.-L. Guermond, Theory and Practice of Finite Elements, Springer, 2004.
- [GR] P. Grisvard, Elliptic Problems in Nonsmooth Domains, Pitman Press, 1985.
- [Ha] W. Hackbusch, Elliptic Differential Equations, Springer, 1992.
- [I] F. Ihlenburg, Finite Element Analysis of Acoustic Scattering, Springer, 1998.
- [Mc] W. McLean, Strongly Elliptic Systems and Boundary Integral Equations, Cambridge, 2000.
- [M] P. Monk, Finite Elements for Maxwell's Equations, Oxford University Press, 2003.
- [Pr] J.D. Pryce, Functional Analysis, Hutchinson, 1973.
- [Ta] A.E. Taylor, Functional Analysis, Wiley, 1958.
- [ToWi] A. Toselli and O. Widlund, Domain Decomposition Methods - Algorithms and Theory, Springer, 2005.