Mathematatische Methoden der Unsicherheitsquantifizierung

Oliver Ernst

Professur Numerische Mathematik

Sommersemester 2016

Mathematik!
TU Chemnitz

Contents

Contents

Contents

Elliptic Boundary Value Problem

We consider the elliptic boundary value problem of finding the solution of the partial differential equation with Dirichlet boundary condition

$$-\nabla\cdot(a\nabla u) = f \qquad \text{on } D \subset \mathbb{R}^2, \tag{B.1a}$$

$$u = g \qquad \text{on } \partial D, \tag{B.1b}$$

given a convex bounded domain D with sufficiently smooth boundary ∂D, a coefficient function $a : D \to \mathbb{R}^+$, a source term $f : D \to \mathbb{R}$ and boundary data in the form of a function $g : \partial D \to \mathbb{R}$.

The differential operator in (B.1a) is short for

$$\nabla\cdot(a\nabla u) = \sum_{j=1}^{2} \frac{\partial}{\partial x_j}\left(a(\boldsymbol{x})\frac{\partial u(\boldsymbol{x})}{\partial x_j}\right)$$

Equation (B.1a) is a model for diffusion phenomena occurring in, e.g., heat conduction, electrostatics, potential flow and elasticity. Generalizations of (B.1) involve the addition of lower-order terms, other boundary conditions, a matrix-valued coefficient function and dependence of a on u.

If $f \in C(\overline{D})$ and $a \in C^1(\overline{D})$, then a function $u \in C^2(D) \cap C^1(\overline{D})$ which satisfies (B.1) is called a classical solution or a strong solution of the boundary value problem.

There are (theoretical and practical) reasons for generalizing the classical solution concept. The key to this generalization lies in reformulating (B.1) as a variational problem. Multiplying both sides of (B.1a) by an arbitrary function $\phi \in C_0^\infty(D)$, in this context known as a test function, and integrating by parts, we observe that any (classical) solution of (B.1) also satisfies the equation

$$a(u,\phi) = \ell(\phi) \qquad \forall \phi \in C_0^\infty(D), \tag{B.2}$$

with the symmetric bilinear form $a(\cdot,\cdot)$ and linear functional $\ell(\cdot)$ given by

$$a(u,\phi) = \int_D a(\boldsymbol{x}) \nabla u(x) \cdot \nabla \phi(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}, \qquad \ell(\phi) = \int_D f(\boldsymbol{x}) \phi(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}. \tag{B.3}$$

For (B.2) to make sense, it is sufficient that the integrals and derivatives are well-defined.

This is the case if u and ϕ are taken to lie in the Sobolev space

$$H^1(D) := \{v \in L^2(D) : \nabla v \in L^2(D)^2\},$$

which is a Hilbert space with respect to the inner product

$$(u, v)_{H^1(D)} = \int_D (\nabla u \cdot \nabla v + uv)\, \mathrm{d}\boldsymbol{x} = (\nabla u, \nabla v) + (u, v),$$

where we use $(\cdot, \cdot)$ to denote the inner product in $L^2(D)$. The associated norm on $H^1(D)$ is

$$\|u\|^2_{H^1(D)} = \int_D \left(|\nabla u|^2 + u^2\right) \mathrm{d}\boldsymbol{x}.$$

The gradients are in terms of weak derivatives in the sense

$$\left(\frac{\partial u}{\partial x_j}, \phi\right) = -\left(u, \frac{\partial \phi}{\partial x_j}\right) \qquad \forall \phi \in C_0^\infty(D).$$

Stating the boundary condition (B.1b) requires a well-defined notion of evaluating a function from $H^1(D)$ on the lower-dimensional manifold ∂D.

- Functions in $H^1(D)$ satisfying the BC with homogeneous boundary data $g \equiv 0$ are easily defined as lying in the subspace

$$H_0^1(D) := \overline{C_0^\infty(D)}^{\|\cdot\|_{H_1(D)}} \subset H^1(D).$$

- For inhomogeneous boundary data we define the space

$$W := H_g^1(D) := \{v \in H^1(D) : u_{|\partial D} = g\}.$$

  The evaluation on the boundary is understood in the following sense: for a sufficiently smooth boundary there exists a bounded trace operator $\gamma : H^1(D) \to L^2(\partial D)$ such that for all $u \in C^1(\overline{D})$ there holds $\gamma u = u_{|\partial D}$. Since $C^1(\overline{D})$ is dense in $H^1(D)$, we have $\gamma u = \lim_{n\to\infty} u_{|\partial D}$ for any approximating sequence $\{u_n\} \subset C^1(\overline{D})$ converging to u in $H^1(D)$.

Definition B.1

The trace space of $H^1(D)$ for a sufficiently smooth domain D is defined as

$$H^{1/2}(\partial D) := \gamma(H^1(D)) = \{\gamma u : u \in H^1(D)\}.$$

$H^{1/2}(\partial D)$ is a Hilbert space with norm

$$\|g\|_{H^{1/2}(\partial D)} := \inf\{\|u\|_{H^1(D)} : \gamma u = g, u \in H^1(D)\}.$$

Sine in general $H^{1/2}(\partial D) \subsetneq L^2(\partial D)$, boundary data g in (B.1b) must be chosen from $H^{1/2}(\partial D)$.

Lemma B.2

There exists $K_\gamma > 0$ such that, for all $g \in H^{1/2}(\partial D)$, we can find $u_g \in H^1(D)$ with $\gamma u_g = g$ and

$$\|u_g\|_{H^1(D)} \leq K_\gamma \|g\|_{H^{1/2}(\partial D)}$$

We denote the spaces of trial and test functions by

$$W := H^1_g(D), \quad \text{and} \quad V := H^1_0(D).$$

Assumption B.3

The coefficient function $a = a(\boldsymbol{x})$ in (B.1a) satisfies

$$0 < a_{\min} \le a(\boldsymbol{x}) \le a_{\max} < \infty \qquad \text{for almost all } \boldsymbol{x} \in D$$

for positive constants $a_{\min}$ and $a_{\max}$. In particular, $a \in L^\infty(D)$ and a is uniformly bounded away from zero.

By Assumption B.3, the bilinear form $a(\cdot,\cdot)$ is bounded on $H^1(D)$, i.e.,

$$|a(u,v)| \le C\|u\|_{H^1(D)}\|v\|_{H^1(D)}, \qquad \forall u, v \in H^1(D)$$

with a constant $C \le \|a\|_{L^\infty(D)}$.

Definition B.4

A weak solution of (B.1) is a function $u \in W$ such that

$$a(u, v) = \ell(v) \qquad \forall v \in V, \tag{B.4}$$

with $a(\cdot, \cdot)$ and $\ell(\cdot)$ as defined in (B.3).

Definition B.5

A bilinear form $a : H \times H \to \mathbb{R}$ on a Hilbert space H is said to be coercive if there exists a constant $\alpha > 0$ such that

$$a(u, u) \geq \alpha \|u\|_H^2 \qquad \forall u \in H.$$

Lemma B.6 (Lax & Milgram)

Let H be a real Hilbert space with norm $\|\cdot\|$ and let ℓ be a bounded linear functional on H. Let $a : H \times H \to \mathbb{R}$ be a bilinear form that is bounded and coercive. Then there exists a unique $u_\ell \in H$ such that $a(u_\ell, v) = \ell(v)$ for all $v \in H$.

For functions in $H^1(D)$ we introduce the H^1 semi-norm

$$|u|_{H^1(D)} := \left(\int_D |\nabla u|^2 \,\mathrm{d}\boldsymbol{x} \right)^{1/2}.$$

as well as the energy norm associated with the coefficient function a as

$$|u|_a := a(u,u)^{1/2} = \left(\int_D a \nabla u \cdot \nabla u \,\mathrm{d}\boldsymbol{x} \right)^{1/2}.$$

Theorem B.7 (Poincaré-Friedrichs inequality)

For a bounded domain D there exists a constant $C = C_D > 0$ such that

$$\|u\|_{L^2(D)} \le C_D |u|_{H^1(D)} \qquad \forall u \in H_0^1(D).$$

Lemma B.8

Under Assumption B.3 the bilinear form $a : H^1(D) \times H^1_0(D) \to \mathbb{R}$ is bounded and the energy norm is equivalent to the H^1 semi-norm on $H^1(D)$.

Theorem B.9

Let Assumption B.3 hold, $f \in L^2(D)$ and $g \in H^{1/2}(\partial D)$. Then (B.1) has a unique weak solution $u \in W = H^1_g(D)$.

Theorem B.10

Under the conditions of Theorem B.9 the weak solution $u \in W$ satisfies

$$|u|_{H^1(D)} \leq K \left(\|f\|_{L^2(D)} + \|g\|_{H^{1/2}(\partial D)} \right)$$

where $K = \max\{C_D/a_{\min}, K_\gamma(1 + a_{\max}/a_{\min})\}$.

Replacing a und f in (B.1) by approximations $\tilde{a}$ and $\tilde{f}$, leads to the perturbed problem of finding $\tilde{u} \in W$ such that

$$\tilde{a}(\tilde{u}, v) = \tilde{\ell}(v) \qquad \forall v \in V \tag{B.5}$$

with $\tilde{a} : W \times V \to \mathbb{R}$ and $\tilde{\ell} : V \to \mathbb{R}$ defined by

$$\tilde{a}(u, v) = \int_D \tilde{a}(\boldsymbol{x}) \nabla u(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \qquad \tilde{\ell}(\phi) = \int_D \tilde{f}(\boldsymbol{x}) v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{B.6}$$

Theorem B.11

Let Assumption B.3 hold for a as well as for $\tilde{a}$ with constants $\tilde{a}_{\min}$, $\tilde{a}_{\max}$. If, furthermore, $\tilde{f} \in L^2(D)$ and $g \in H^{1/2}(\partial D)$, then problem (B.5) has a unique weak solution $\tilde{u} \in W = H^1_g(D)$.

Theorem B.12

Under the conditions of Theorems B.9 and B.11, if $u, \tilde{u} \in W$ denote the solutions of (B.4) and (B.5), respectively, then

$$|u - \tilde{u}|_{H^1(D)} \leq C_D \tilde{a}_{\min}^{-1} \|f - \tilde{f}\|_{L^2(D)} + \tilde{a}_{\min}^{-1} \|a - \tilde{a}\|_{L^\infty(D)} |u|_{H^1(D)}$$

Contents

Given: linear variational problem of finding $u \in V$, V a Hilbert space with norm $\|\cdot\|$, such that

$$a(u, v) = \ell(v) \qquad \forall v \in V \tag{B.7}$$

with a bilinear form $a(\cdot, \cdot)$ and linear form $\ell(\cdot)$ on V which satisfy the assumptions of the Lax-Milgram lemma.

Galerkin method for finding approximate solutions of (B.7) proceeds by restricting the problem to a finite-dimensional subspace $V_n \subset V$: denote by $u_n \in V_n$ the solution of

$$a(u_n, v) = \ell(v) \qquad \forall v \in V_n. \tag{B.8}$$

Note: The Galerkin approximation u_n of u with respect to the space V_n is uniquely determined since the conditions of the Lax-Milgram lemma are satisfied for Problem (B.8) by inclusion.

The simple structure of a linear variational problem allows its reduction to a problem of best approximation.

Lemma B.13 (Céa)

If the assumptions of the Lax-Milgram lemma apply to Problem (B.7) with solution $u \in V$, then the Galerkin approximation u_n, i.e., the solution of (B.8), satisfies

$$\|u - u_n\| \le \frac{C}{\alpha} \inf_{v \in V_n} \|u - v\|. \tag{B.9}$$

- If the bilinear form $a(\cdot,\cdot)$ is, in addition, symmetric (Hermitian) then, because of coercivity, it defines an inner product on V.
- Galerkin orthogonality then implies u_n is the a-orthogonal projection of u onto V_n and therefore the best approximation to u from V_n with respect to the associated (energy) norm.
- In the energy norm (B.9) is therefore satisfied with $C = \alpha = 1$.
- Coercivity and boundedness also imply that the energy norm is equivalent with $\|\cdot\|$, i.e.,

$$\sqrt{\alpha}\|v\| \le |v|_a \le \sqrt{C}\|v\| \qquad \forall v \in V,$$

which leads to the improved estimate over (B.9)

$$\|u - u_n\| \le \sqrt{\frac{C}{\alpha}} \inf_{v \in V_n} \|u - v\|.$$

We have seen that, for the elliptic BVP (B.1), we have the equivalences

$$\|\cdot\|_{H^1(D)} \asymp |\cdot|_{H^1(D)} \asymp |\cdot|_a.$$

Corollary B.14

Under Assumption B.3, the Galerkin approximation u_n to the solution of the elliptic boundary value problem (B.1), with respect to any subspace V_n of $V = H_0^1(D)$, satisfies

$$|u - u_n|_a = \inf_{v \in V_n} |u - v|_a,$$

$$|u - u_n|_{H^1(D)} \leq \sqrt{\frac{a_{\min}}{a_{\max}}}\, |u - v|_{H^1(D)} \qquad \forall v \in V_n.$$

Given a basis $\{v_1, \dots, v_n\}$ of V_n and the solution $u_n = \sum_{j=1}^n \xi_j v_j$, then the Galerkin variational equation (B.8) is equivalent with

$$\sum_{j=1}^n \xi_j \, a(v_j, v_i) = \ell(v_i), \qquad i = 1, \dots, n,$$

which, when rewritten as a linear system of equation, becomes the Galerkin system

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{B.10}$$

with Galerkin matrix $[\boldsymbol{A}]_{i,j} = a(v_j, v_i)$, unknown vector $[\boldsymbol{x}]_i = \xi_i$ and right-hand side vector $[\boldsymbol{b}]_i = \ell(v_i)$.

- If $a(\cdot, \cdot)$ is symmetric, then so is $\boldsymbol{A}$.
- If $a(\cdot, \cdot)$ is coercive, then $\boldsymbol{A}$ is (uniformly) positive definite.

- Different Galerkin methods result from different choices of subspaces.
- Wavelets.
- Trigonometric functions, global polynomials (spectral methods).
- Radial basis functions.
- The finite element method employs finite dimensional subspaces of the variational spaces (trial and test spaces) consisting of piecewise polynomials with respect to a partition of D.
- We shall assume in the following that D is a polygon (polyhedron), but the finite element method can also be applied to domains with curved boundaries.

Assumptions on the partition of the domain D, denoted by $\mathscr{T}_h$ with elements K:

(Z$_1$) $\overline{D} = \cup_{K \in \mathscr{T}_h} K$.

(Z$_2$) Each $K \in \mathscr{T}_h$ is a closed set with nonempty interor $\mathring{K}$.

(Z$_3$) For two distinct $K_1, K_2 \in \mathscr{T}_h$ there holds $\mathring{K}_1 \cap \mathring{K}_2 = \emptyset$.

(Z$_4$) Each $K \in \mathscr{T}_h$ has a Lipschitz-continuous boundary ∂K.

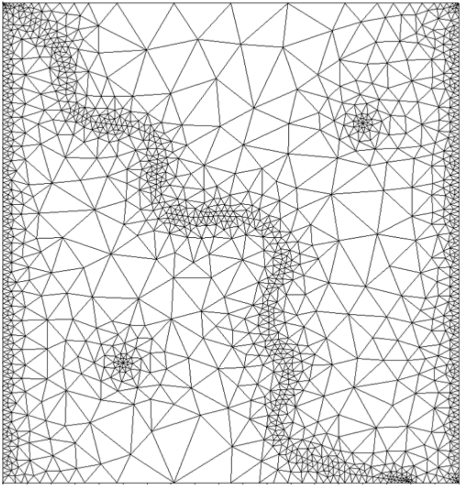
The partition is usually assigned a discretization parameter $h > 0$ given by

$$h := \max_{K \in \mathscr{T}^h} \operatorname{diam} K,$$

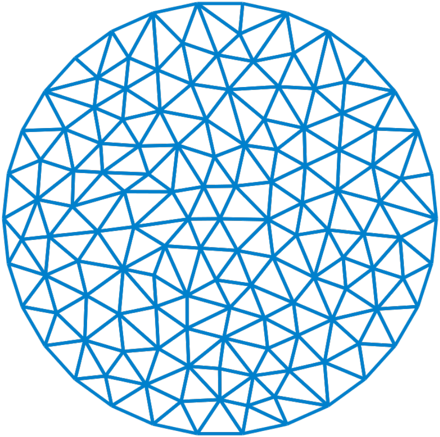
which is a measure of how fine the partition is.

Finite Element Approximation

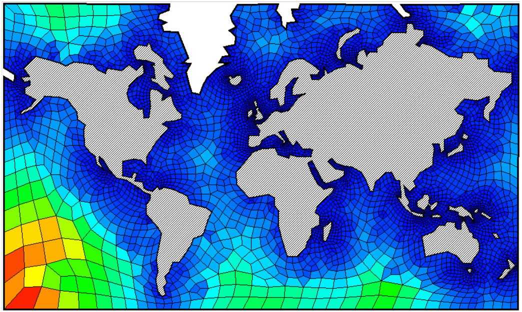
Triangulations



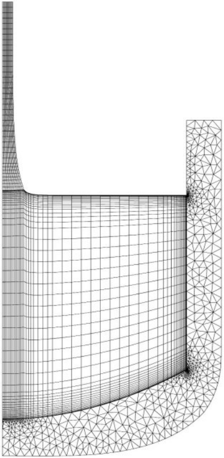
Triangular mesh on a square domain.



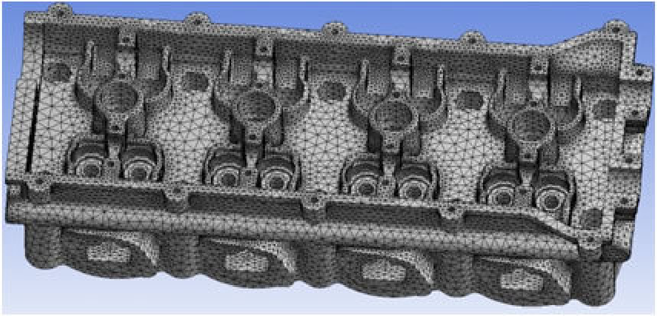
Triangular mesh on a polygonal approximation of a circle.



Quadrilateral mesh on a rectangular (exterior) domain.



Mesh consisting of triangles and quadrilaterals.



Tetrahedral mesh of complex 3D geometry (engine block).

A conforming Galerkin approximation is one which employs finite-dimensional spaces V_n such that $V_n \subset V$.

Let V^h denote a space of piecewise continuous functions $v : \overline{D} \to \mathbb{R}$ with respect to an admissible triangulation $\mathscr{T}_h$ of D, i.e., such that each restriction $v|_K$ to any $K \in \mathscr{T}_h$ is continuous on K.

Theorem B.15

With the notation defined above, there holds $V^h \subset H^1(D)$ if, and only if,

$$V^h \subset C(\overline{D}) \qquad \text{and} \qquad \{v|_K : v \in V^h\} \subset H^1(K).$$

In this case $\{v \in V^h : v = 0 \text{ on } \partial D\} \subset H^1_0(D)$.

According to [Ciarlet, 1978], a finite element is a triple (K, P_K, Ψ_K) such that

(1) K is a nonempty set

(2) P_K is a finite-dimensional space of functions defined on K and

(3) Ψ_K is a set of linearly independent linear functionals ψ on P_K with the property that, for any $p \in P_K$,

$$\psi(p) = 0 \ \ \forall \psi \in \Psi_K \qquad \Rightarrow \qquad p = 0.$$

We shall consider a single finite element, the so-called linear triangle, where

(1) $K \in \mathbb{R}^2$ is a triangle with (non-collinear) vertices $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$,

(2) P_K is the space of all affine functions on K and

(3) Ψ_K consists of the three functionals

$$\Psi_K = \{\psi_j : P_K \to \mathbb{R}, \psi_j(p) = p(\boldsymbol{x}_j), j = 1, 2, 3\}.$$

- To construct a (global) finite element space V^h based on linear triangle elements consider a triangulation $\mathscr{T}^h$ of D consisting of (closed) triangles K which satisfy properties (**Z1**)–(**Z4**).
- The functions in V^h will also lie in $H^1(D)$ if they are continuous on $\overline{D}$, which, for piecewise linear (polynomial) functions, is equivalent with their being continuous across triangle boundaries.
- We thus obtain the space

$$V^h := \{v \in C(\overline{D}) : v|_K \in \mathscr{P}_1 \ \forall K \in \mathscr{T}^h\},$$

  where $\mathscr{P}_k$ denotes the space of (multivariate) polynomials of (complete) degree k.
- A subspace V_0^h of V^h is given by

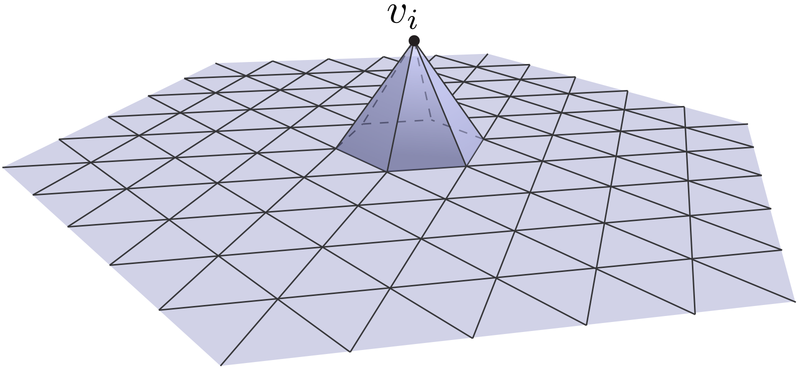
$$V_0^h := \{v \in V^h : v|_{\partial D} = 0\} \subset H_0^1(D).$$

- A continuous piecewise linear function in V^h is completely determined by its values at all triangle vertices.
- Such a (finite) set of parameters which uniquely determine a finite element function is called a set of degrees of freedom (DOF).
- In V_0^h these are the values at all nodes which do not lie on ∂D; denote their number by n.
- A particularly convenient basis $\{\phi_1, \dots, \phi_n\}$ of V_0^h is the so-called nodal basis characterized by

$$\phi_j(\boldsymbol{x}_i) = \delta_{i,j} \qquad i, j = 1, \dots, n.$$

- If $\mathscr{N}^h = \{x_1, \dots, x_n\}$ denotes the set of vertices $x_j \notin \partial D$, then

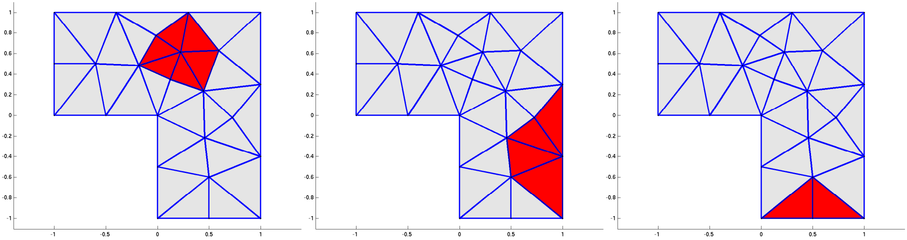
$$\operatorname{supp} \phi_j = \bigcup_{\substack{K \in \mathscr{T}^h \\ \boldsymbol{x}_j \in K}} K.$$



A nodal basis function with its support.

Finite Element Approximation

Nodal basis for linear triangles



Triangulation of an L-shaped domain with the supports of several basis functions.

Implications for Galerkin system (B.10):

$$[\boldsymbol{b}]_i = \ell(\phi_i) = \int_D f\phi_i \,\mathrm{d}\boldsymbol{x} = \int_{\mathrm{supp}\,\phi_i} f\phi_i \,\mathrm{d}\boldsymbol{x},$$

$$\begin{aligned}[\boldsymbol{A}]_{i,j} = a(\phi_j, \phi_i) &= \int_D a(\boldsymbol{x})\nabla\phi_i(\boldsymbol{x}) \cdot \nabla\phi_j(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \\ &= \int_{\mathrm{supp}\,\phi_i \cap \mathrm{supp}\,\phi_j} a(\boldsymbol{x})\nabla\phi_i(\boldsymbol{x}) \cdot \nabla\phi_j(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}.\end{aligned}$$

In particular: Galerkin matrix $\boldsymbol{A}$ is sparse.

Common procedure in assembling the Galerkin system:

(1) Ignore boundary condition initially, i.e., consider all of V^h with nodal basis

$$\{\phi_1, \phi_2, \ldots, \phi_n, \phi_{n+1}, \ldots, \phi_{\tilde{n}}\},$$

$\tilde{n} - n$ the number of vertices on the boundary ∂D.
Yields matrix $\tilde{\boldsymbol{A}} \in \mathbb{R}^{\tilde{n} \times \tilde{n}}$, vector $\tilde{\boldsymbol{b}} \in \mathbb{R}^{\tilde{n}}$.

(2) Then eliminate the DOF associated with boundary vertices.
Yields matrix $\boldsymbol{A}$, vector $\boldsymbol{b}$.

Note:

- Initial approach for step (1): compute $\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{b}}$, entry by entry, i.e., basis function by basis function
- But: shape and connectivity of supports typically very different.
- Simpler: compute $\boldsymbol{A}$, $\boldsymbol{b}$ element by element.

$K \in \mathscr{T}^h$: then for $i, j = 1, 2 \dots, \tilde{n}$:

$$a(\phi_j, \phi_i) = \int_D a \nabla \phi_j \cdot \nabla \phi_i \,\mathrm{d}\boldsymbol{x} = \sum_{K \in \mathscr{T}^h} \int_K a \nabla \phi_j \cdot \nabla \phi_i \,\mathrm{d}\boldsymbol{x} =: \sum_{K \in \mathscr{T}^h} a_K(\phi_j, \phi_i),$$

$$\ell(\phi_i) = \int_D f \phi_i \,\mathrm{d}\boldsymbol{x} = \sum_{K \in \mathscr{T}^h} \int_K f \phi_i \,\mathrm{d}\boldsymbol{x} =: \sum_{K \in \mathscr{T}^h} \ell_K(\phi_i).$$

Setting

$$\begin{aligned} [\tilde{\boldsymbol{A}}_K]_{i,j} &:= a_K(\phi_j, \phi_i) & \qquad & i, j = 1, 2, \dots, \tilde{n}, \\ [\tilde{\boldsymbol{b}}_K]_i &:= \ell_K(\phi_i, & & i = 1, 2, \dots, \tilde{n}, \end{aligned}$$

we obtain

$$\tilde{\boldsymbol{A}} = \sum_{K \in \mathscr{T}^h} \tilde{\boldsymbol{A}}_K, \qquad \tilde{\boldsymbol{b}} = \sum_{K \in \mathscr{T}^h} \tilde{\boldsymbol{b}}_K.$$

Since each element belongs to the support of exactly three basis functions, only (at most) nine entries of $\tilde{\boldsymbol{A}}_K$ and three entries of $\tilde{\boldsymbol{b}}_K$ are nonzero.
Which entries these are can be determined by maintaining an element table:

$[ET(i,j)]_{i=1,2,3;j=1,\dots,n_K}$:

Element	K_1	K_2	$\dots$	K_{n_K}
first vertex	$i_1^{(1)}$	$i_1^{(2)}$	$\dots$	$i_1^{(n_K)}$
second vertex	$i_2^{(1)}$	$i_2^{(2)}$	$\dots$	$i_2^{(n_K)}$
third vertex	$i_3^{(1)}$	$i_3^{(2)}$	$\dots$	$i_3^{(n_K)}$

Here n_K denotes the number of triangles in $\mathscr{T}^h$.

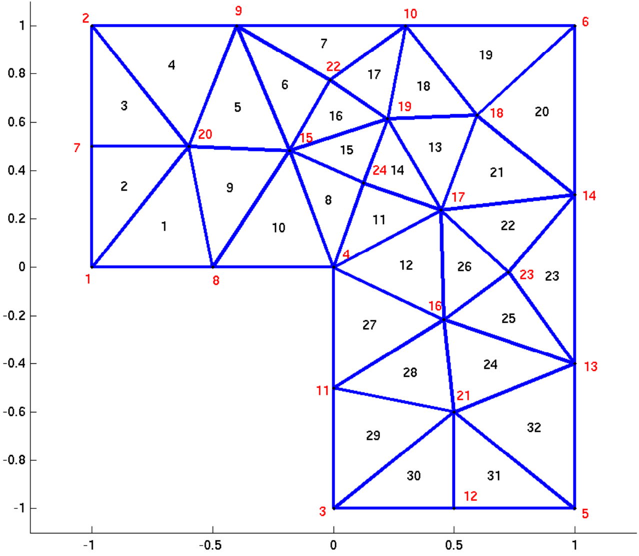
Besides the global vertex numbering

$$x_1, x_2, \dots, x_{\tilde{n}},$$

the element table introduces a second, local vertex numbering

$$x_1^{(K)},\; x_2^{(K)},\; x_3^{(K)}$$

of the vertices (DOFs) associated with K.



Global numbering of vertices (red) and elements (black) in a triangulation of an L-shaped domain.

With this notation the nonzero submatrix $\boldsymbol{A}_K$ of $\tilde{\boldsymbol{A}}_K$ and nonzero subvector $\boldsymbol{b}_K$ of $\tilde{\boldsymbol{b}}_K$ are given by

$$\boldsymbol{A}_K := \begin{bmatrix} a_K(\phi_1^{(K)}, \phi_1^{(K)}) & a_K(\phi_2^{(K)}, \phi_1^{(K)}) & a_K(\phi_3^{(K)}, \phi_1^{(K)}) \\ a_K(\phi_1^{(K)}, \phi_2^{(K)}) & a_K(\phi_2^{(K)}, \phi_2^{(K)}) & a_K(\phi_3^{(K)}, \phi_2^{(K)}) \\ a_K(\phi_1^{(K)}, \phi_3^{(K)}) & a_K(\phi_2^{(K)}, \phi_3^{(K)}) & a_K(\phi_3^{(K)}, \phi_3^{(K)}) \end{bmatrix}, \qquad \boldsymbol{b}_K := \begin{bmatrix} \ell_K(\phi_1^{(K)}) \\ \ell_K(\phi_2^{(K)}) \\ \ell_K(\phi_3^{(K)}) \end{bmatrix}.$$

If K has number k in the enumeration of the elements, then the association of the local numbering $\{\phi_i^{(K)}\}_{i=1,2,3}$ of the three basis functions whose support contains K with the global numbering $\{\phi_j\}_{j=1}^{\tilde{n}}$ of all basis functions is given by

$$\phi_i^{(K)} = \phi_j, \qquad j = ET(i,k), \quad i = 1,2,3.$$

$\boldsymbol{A}_K$ and $\boldsymbol{b}_K$ are sometimes called the element stiffness matrix and element load vector.

We summarize phase (1) of the finite element assembly process in the following algorithm[5]

Algorithm 2: Phase (1) of finite element assembly.

Initialize $\tilde{\boldsymbol{A}} := \boldsymbol{O}$, $\tilde{\boldsymbol{b}} := \boldsymbol{0}$.
foreach $K \in \mathscr{T}_h$ **do**
    Compute $\boldsymbol{A}_K$ and $\boldsymbol{b}_K$
    $k \leftarrow$ [index of element K]
    $i_1 \leftarrow ET(1,k)$, $i_2 \leftarrow ET(2,k)$, $i_3 \leftarrow ET(3,k)$
    $\tilde{\boldsymbol{A}}([i_1\, i_2\, i_3],[i_1\, i_2\, i_3]) \leftarrow \tilde{\boldsymbol{A}}([i_1\, i_2\, i_3],[i_1\, i_2\, i_3]) + \boldsymbol{A}_K$
    $\tilde{\boldsymbol{b}}([i_1\, i_2\, i_3]) \leftarrow \tilde{\boldsymbol{b}}([i_1\, i_2\, i_3]) + \boldsymbol{b}_K$

[5] We use the following Matlab-inspired notation:

$$\boldsymbol{A}([i_1\, i_2\, i_3],[i_1\, i_2\, i_3]) = \begin{bmatrix} a_{i_1,i_1} & a_{i_1,i_2} & a_{i_1,i_3} \\ a_{i_2,i_1} & a_{i_2,i_2} & a_{i_2,i_3} \\ a_{i_3,i_1} & a_{i_3,i_2} & a_{i_3,i_3} \end{bmatrix}, \quad \boldsymbol{b}([i_1\, i_2\, i_3]) = \begin{bmatrix} b_{i_1} \\ b_{i_2} \\ b_{i_3} \end{bmatrix}.$$

Both the numerical integration as well as the error analysis benefit from a change of variables to a reference element $\hat{K} \subset \mathbb{R}^2$. Each element $K \in \mathscr{T}^h$ then has a parametrization $K = F_K(\hat{K})$, where

$$F_K : \hat{K} \to K, \qquad \hat{K} \ni \boldsymbol{\xi} \mapsto \boldsymbol{x} \in K, \quad \boldsymbol{x} = F_K(\boldsymbol{\xi}) = B_K \boldsymbol{\xi} + \boldsymbol{b}_K.$$

Most common for triangular elements: unit simplex

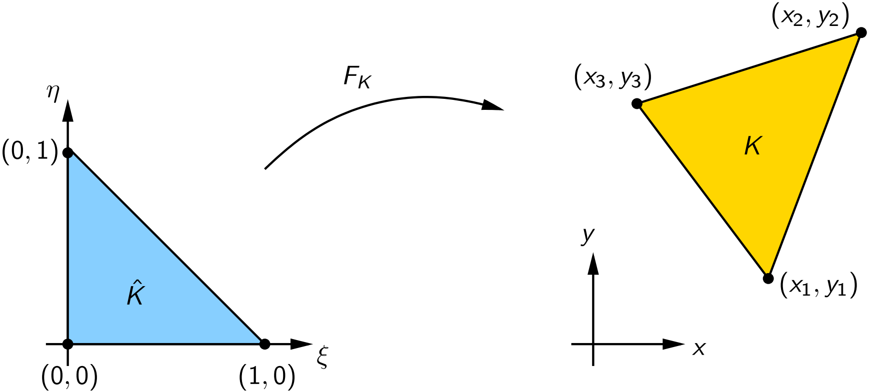
$$\hat{K} = \{(\xi, \eta) \in \mathbb{R}^2 : 0 \le \xi \le 1, 0 \le \eta \le 1 - \xi\}.$$

For each triangle $K \in \mathscr{T}^h$ the affine mapping F_K is determined by prescribing, e.g.,

$$\begin{aligned} (1,0) &\mapsto (x_1, y_1), \\ (0,1) &\mapsto (x_2, y_2), \\ (0,0) &\mapsto (x_3, y_3), \quad \text{i.e.} \end{aligned}$$

Finite Element Approximation

Reference element



$$\begin{bmatrix} x \\ y \end{bmatrix} = \underbrace{\begin{bmatrix} x_1 - x_3 & x_2 - x_3 \\ y_1 - y_3 & y_2 - y_3 \end{bmatrix}}_{B_K} \begin{bmatrix} \xi \\ \eta \end{bmatrix} + \underbrace{\begin{bmatrix} x_3 \\ y_3 \end{bmatrix}}_{\boldsymbol{b}_K}$$

Local (nodal) basis on $\hat{K}$: (dual basis of DOF)

$$\hat{\phi}_1(\xi,\eta) = \xi, \quad \hat{\phi}_2(\xi,\eta) = \eta, \quad \hat{\phi}_3(\xi,\eta) = 1 - \xi - \eta, \qquad (\xi,\eta) \in \hat{K}.$$

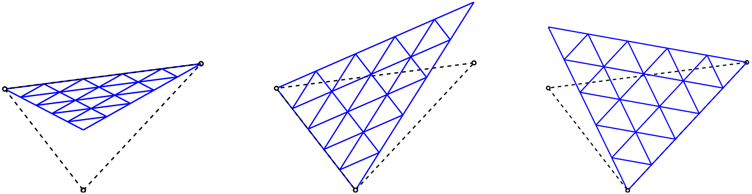
The correspondence

$$\hat{\phi} \mapsto \phi := \hat{\phi} \circ F_K^{-1}, \quad \text{d.h.} \quad \phi(\boldsymbol{x}) := \hat{\phi}(\boldsymbol{\xi}(\boldsymbol{x})) = \hat{\phi}(F_K^{-1}(\boldsymbol{x}))$$

assigns to $\hat{\phi}$ on $\hat{K}$ a unique function ϕ on K.

Local basis functions on K:

$$\phi_j = \hat{\phi}_j \circ F_K^{-1} : K \to \mathbb{R}, \qquad j = 1, 2, 3.$$



The chain rule[6] applied to $\phi(\boldsymbol{x}) = \hat{\phi}(\boldsymbol{\xi}(\boldsymbol{x}))$ gives

$$\nabla\phi = \begin{bmatrix} \phi_x \\ \phi_y \end{bmatrix} = \begin{bmatrix} \hat{\phi}_\xi \xi_x + \hat{\phi}_\eta \eta_x \\ \hat{\phi}_\xi \xi_y + \hat{\phi}_\eta \eta_y \end{bmatrix} = \begin{bmatrix} \xi_x & \eta_x \\ \xi_y & \eta_y \end{bmatrix} \begin{bmatrix} \hat{\phi}_\xi \\ \hat{\phi}_\eta \end{bmatrix} = (DF_K^{-1})^\top \hat{\nabla}\hat{\phi}.$$

Since

$$\begin{aligned} \boldsymbol{x} &= F_K(\boldsymbol{\xi}) = B_K\boldsymbol{\xi} + \boldsymbol{b}_K, && \text{i.e. } DF_K \equiv B_K, \\ \boldsymbol{\xi} &= F_K^{-1}(\boldsymbol{x}) = B_K^{-1}(\boldsymbol{x} - \boldsymbol{b}_K), && \text{i.e. } DF_K^{-1} \equiv B_K^{-1} \end{aligned}$$

we obtain

$$\nabla\phi = B_K^{-\top}\hat{\nabla}\hat{\phi}.$$

[6] $\hat{\nabla}$ indicates differentiation with respect to the variables ξ and η.

This finally gives the element integrals ($\phi_i = \phi_i^{(K)}$, $i = 1, 2, 3$)

$$
\begin{aligned}
a_K(\phi_j, \phi_i) &= \int_K a(\boldsymbol{x})\, \nabla\phi_j(\boldsymbol{x}) \cdot \nabla\phi_i(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \\
&= \int_{\hat{K}} a(\boldsymbol{x}(\boldsymbol{\xi}))\, \left(B_K^{-\top}\hat{\nabla}\hat{\phi}_j(\boldsymbol{\xi})\right) \cdot \left(B_K^{-\top}\hat{\nabla}\hat{\phi}_i(\boldsymbol{\xi})\right) |\det B_K|\, \mathrm{d}\boldsymbol{\xi}.
\end{aligned}
\tag{B.11}
$$

The determinant is given by (note K is a triangle)

$$
|\det B_K| = 2|K|,
$$

$$
B_K^{-\top} = \frac{1}{2|K|} \begin{bmatrix} y_2 - y_3 & x_3 - x_2 \\ y_3 - y_1 & x_1 - x_3 \end{bmatrix},
$$

$$
\begin{bmatrix} \hat{\nabla}\hat{\phi}_1 & \hat{\nabla}\hat{\phi}_2 & \hat{\nabla}\hat{\phi}_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}.
$$

To impose the Dirichlet boundary condition we require that the Galerkin approximation $u^h \in V^h$ satisfy

$$u^h(\boldsymbol{x}_j) = g(\boldsymbol{x}_j) \quad \text{at all boundary vertices } \{\boldsymbol{x}_j\}_{j=n+1}^{\tilde{n}}. \tag{B.12}$$

- We partition the coefficient vector $\boldsymbol{u} \in \mathbb{R}^{\tilde{n}}$ into a first block $\boldsymbol{u}_I \in \mathbb{R}^n$ containing the coefficients associated with the interior vertices $\{\boldsymbol{x}_j\}_{j=1}^{n}$ and a second block $\boldsymbol{u}_B \in \mathbb{R}^{\tilde{n}-n}$ containing the constrained coefficients associated with boundary vertices.
- For the assembled matrix $\tilde{\boldsymbol{A}}$ and vector $\tilde{\boldsymbol{b}}$ this induces the partitionings

$$\tilde{\boldsymbol{A}} = \begin{bmatrix} \tilde{\boldsymbol{A}}_{II} & \tilde{\boldsymbol{A}}_{IB} \\ \tilde{\boldsymbol{A}}_{BI} & \tilde{\boldsymbol{A}}_{BB} \end{bmatrix}, \qquad \tilde{\boldsymbol{b}} = \begin{bmatrix} \tilde{\boldsymbol{b}}_I \\ \tilde{\boldsymbol{b}}_B \end{bmatrix}.$$

- The constraint (B.12) now reads $\boldsymbol{u}_B = \boldsymbol{g}$, where $\boldsymbol{g} \in \mathbb{R}^{\tilde{n}-n}$ contains the boundary data $\{g(\boldsymbol{x}_j)\}_{j=n+1}^{\tilde{n}}$.

This constraint is characterized by there being no coupling of the boundary DOF to either interior DOF or among themselves, resulting in the modified linear system of equations

$$\begin{bmatrix} \tilde{\boldsymbol{A}}_{II} & \tilde{\boldsymbol{A}}_{IB} \\ \boldsymbol{O} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_I \\ \boldsymbol{u}_B \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_I \\ \boldsymbol{g} \end{bmatrix},$$

which gives the reduced system

$$\boldsymbol{A}\boldsymbol{u}_I = \boldsymbol{b}, \qquad \boldsymbol{A} = \tilde{\boldsymbol{A}}_{II}, \quad \boldsymbol{b} = \boldsymbol{b}_I - \tilde{\boldsymbol{A}}_{IB}\boldsymbol{g}$$

for the interior DOF.

Note that this procedure is a discrete variant of the reformulation of the BVP with inhomogeneous Dirichlet boundary conditions to an equivalent one with homogeneous Dirichlet boundary conditions.

Contents

Finite Element Convergence

...in a nutshell

- Céa's lemma characterizes the Galerkin error as one of best appproximation from the FE subspace V^h.
- An upper bound for this error is the distance of the true solution from its interpolant from the FE subspace. This is the uniquely determined function from V^h which possesses the same global DOF as the exact solution.
- The asymptotic behavior of the interpolant is then analyzed on a sequence of meshes $\{\mathscr{T}_{h_n}\}_{n\in\mathbb{N}}$ with $\lim_{n\to\infty} h_n = 0$.
- For the interpolation error to become small, the mesh sequence has to be shape-regular: if ρ_K denotes the radius of the inscribed circle in K and $h_K = \operatorname{diam} K$, then a sequence of meshes is shape-regular provided the ratio

$$\frac{\rho_K}{h_K}, \qquad K \in \mathscr{T}_h$$

  is bounded below uniformly for all $\{\mathscr{T}_{h_n}\}$.
- A priori convergence bounds are obtained by relating the smoothness of the exact solution to the convergence rate h^α of the interpolation error as $h \to 0$.

Finite Element Convergence

Extra regularity

Interpolation estimates for a solution u which is only in $H^1(D)$ do not yield a useful rate h^α with an $\alpha > 0$. For this reason one usually tries to show that the solution possesses more regularity.

Definition B.16

For $r \in \mathbb{N}$ and $D \subset \mathbb{R}^d$ bounded, we denote by $H^r(D)$ the Sobolev space

$$H^r(D) := \{v \in L^2(D) : D^{\boldsymbol{\alpha}} u \in L^2(D) \text{ for all } \boldsymbol{\alpha} \in \mathbb{N}_0^d, |\boldsymbol{\alpha}| \le r\}$$

$H^r(D)$ is a Hilbert space with the inner product

$$(u, v)_{H^r(D)} = \sum_{|\boldsymbol{\alpha}| \le r} \int_D (D^{\boldsymbol{\alpha}} u)(D^{\boldsymbol{\alpha}} v) \,\mathrm{d}\boldsymbol{x}.$$

Finite Element Convergence

Extra regularity, fractional index

For any $r \in \mathbb{R} \setminus \mathbb{N}_0$ we set $r = k + s$, $k \in \mathbb{N}_0$, $s \in (0,1)$ and denote for a bounded domain $D \subset \mathbb{R}^d$ by $|\cdot|_{H^r(D)}$ and $\|\cdot\|_{H^r(D)}$ the Sobolev-Slobodetskii semi-norm and norm defined for $v \in H^k(D)$ by

$$
|v|_{H^r(D)} = \left(\int_{D \times D} \sum_{|\boldsymbol{\alpha}|=k} \frac{[D^{\boldsymbol{\alpha}} v(\boldsymbol{x}) - D^{\boldsymbol{\alpha}} v(\boldsymbol{y})]^2}{|\boldsymbol{x} - \boldsymbol{y}|^{d+2s}} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \right)^{1/2} \quad \text{and}
$$

$$
\|v\|_{H^r(D)} = \left(\|v\|_{H^k(D)}^2 + |v|_{H^r(D)}^2 \right)^{1/2}.
$$

The Sobolev space $H^r(D)$ is then defined as the space of functions $v \in H^k(D)$ such that $|v|_{H^r(D)}^2$ is finite.

- Let V^h denote the space of piecewise linear functions subject to a shape-regular, admissible triangulation $\mathscr{T}_h$ of D.
- Denote by $I_h : C(\overline{D}) \to V^h$ the (global) interpolation operator assigning to each continuous function v the interpolant $v_h \in V^h$ determined by the condition that v_h agrees with v at all vertices of $\mathscr{T}_h$.
- Then the error of best approximation of $u \in C(\overline{D})$ is bounded by the interpolation error

$$\inf_{v \in V^h} |u - v|_{H^1(D)} \leq |u - I_h u|_{H^1(D)}.$$

- If the solution u of (B.4) has additional regularity $u \in H^2(D)$, then the Sobolev imbedding theorem assures that u agrees a.e. with a function in $C(\overline{D})$, so that pointwise evaluation of u and thus the interpolant is well-defined.
- In this case a scaling argument can be used to show

$$|u - I_h u|_{H^1(D)} \leq K\, h\, |u|_{H^2(D)}$$

with a constant K independent of h and u.

Assumption B.17 (H^2 regularity)

There exists a constant $K_2 > 0$ such that, for every $f \in L^2(D)$, the solution of (B.4) belongs to $H^2(D)$ and satisfies
$$|u|_{H^2(D)} \le K_2 \|f\|_{L^2(D)}.$$

Theorem B.18

Under Assumptions B.3 and B.17, the solution u of (B.4) with $f \in L^2(D)$ and the piecewise linear finite element approximation u_h on a sequence of shape-regular meshes satisfy

$$|u - u_h|_a \le K\sqrt{a_{\max}}|u|_{H^2(D)}\, h \le KK_2\sqrt{a_{\max}}\|f\|_{L^2(D)}\, h \tag{B.13}$$

with a constant K independent of h.

Corollary B.19

Under the assumptions of Theorem B.18 there holds

$$|u - u_h|_{H^1(D)} \le K\sqrt{\frac{a_{\max}}{a_{\min}}}|u|_{H^2(D)}\, h \le KK_2\sqrt{\frac{a_{\max}}{a_{\min}}}\|f\|_{L^2(D)}\, h.$$

When the coefficient function a and the source term f are replaced by approximations $\tilde{a} \approx a$ and $\tilde{f} \approx f$, then with the modified bilinear and linear forms defined as in (B.6), we may consider the discrete problem

$$\tilde{a}(\tilde{u}_h, v) = \tilde{\ell}(v) \quad \forall v \in V^h. \tag{B.14}$$

In analogy to Theorem B.11 we obtain

Theorem B.20

Under Assumption B.3 let $\tilde{f} \in L^2(D)$ and $g \in H^{1/2}(\partial D)$. Then (B.14) has a unique solution $\tilde{u}_h \in V^h$.

By the triangle inequality, we have

$$|u - \tilde{u}_h|_{H^1(D)} \leq |u - \tilde{u}|_{H^1(D)} + |\tilde{u} - \tilde{u}_h|_{H^1(D)}.$$

By an obvious extension of Corollary B.14, we obtain the bound

$$|\tilde{u} - \tilde{u}_h|_{H^1(D)} \leq \sqrt{\frac{\tilde{a}_{\max}}{\tilde{a}_{\min}}} \inf_{v \in V^h} |\tilde{u} - v|_{H^1(D)}.$$

Alternatively, if we approximate the data at the discrete level only, we may consider the following splitting as more natural:

$$|u - \tilde{u}_h|_{H^1(D)} \le |u - u_h|_{H^1(D)} + |u_h - \tilde{u}_h|_{H^1(D)}.$$

The second term arises, e.g., if we approximate the Galerkin approximation u_h by approximating the bilinear and linear forms using, e.g., piecewise constant approximations of the coefficient a and source term f.

Straightforward modification of the proof of Theorem B.12 yields

$$|u - \tilde{u}_h|_{H^1(D)} \le C_D \tilde{a}_{\min}^{-1} \|f - \tilde{f}\|_{L^2(D)} + \tilde{a}_{\min}^{-1} \|a - \tilde{a}\|_{L^\infty(D)} |u_h|_{H^1(D)}.$$