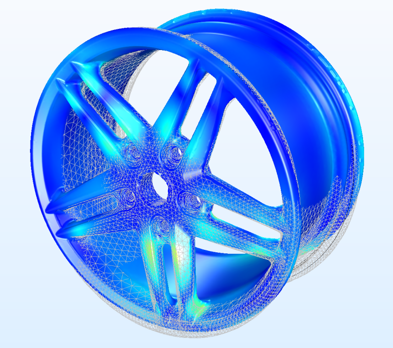
Applied Mathematics 225

Unit 3: Finite element methods

Lecturer: Chris H. Rycroft

Finite element methods

- The finite element method is a framework for discretizing and solving PDEs, especially elliptic PDEs
- Solution is represented as a sum of simple functions (elements)
- Widely used in science and industry, such as for solid mechanics, and thermal analysis.



Wheel rim analysis in COMSOL Multiphysics®. The rim is covered in a triangular mesh, and simple functions are specified on these triangles, from which the solution is constructed.

Comparison with finite-difference methods

Finite element methods

- make it easier to deal with complex boundary conditions,
- provide more mathematical guarantees about convergence.

Finite difference methods

- are simpler to implement, and sometimes more efficient for the same level of accuracy,
- are easier to apply to a wider range of equations (*e.g.* hyperbolic equations).

These are just broad generalizations though—both methods have a large body of literature, with many extensions.

Frequently, the two approaches lead to similar[1] numerical implementations.

[1] And in some cases identical.

Book references

We will make use of the following books:

- Claes Johnson. *Numerical Solution of Partial Differential Equations by the Finite Element Method*. Dover, 2009.[2]
- Thomas J. R. Hughes. *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*. Dover, 2000.[3]
- Dietrich Braess, *Finite elements: Theory, fast solvers, and applications in solid mechanics*, Cambridge University Press, 2007.[4]

[2] A good general introduction.

[3] Comprehensive, with a particular emphasis on solid mechanics.

[4] A more technically rigorous treatment of the subject.

Overview

The main idea is to formulate an elliptic PDE as a variational minimization problem over a suitably-defined function space.

To obtain a numerical method, we approximate this function space by a finite-dimensional subspace, the finite-element space.

There are some subtleties to choosing the correct function space to use. We begin by considering a specific example that motivates the need for a careful treatment.

The need for a careful treatment

Consider a domain Ω that is an open subset of $\mathbb{R}^n$. We can also define the closure[5] $\bar{\Omega}$ and the boundary $\partial\Omega$.[6]

The most natural function spaces to use are $C^k(\Omega)$, the space of all functions on Ω that are differentiable k times.

However, even for some simple cases, these spaces can pose theoretical difficulties, such as a loss of regularity.

[5] This is found by adding all limits of sequences in Ω to it. For example if $\Omega = (0, 1)$, the sequences $x_n = {}^1\!/\!_n$ and $y_n = 1 - {}^1\!/\!_n$ converge to 0 and 1, respectively. Thus $\bar{\Omega} = [0, 1]$.

[6] The boundary is technically defined as $\partial\Omega = \bar{\Omega} \setminus \Omega$. Thus $\partial\Omega = \{0, 1\}$ in this example.

The need for a careful treatment

Consider the domain with reentrant corner

$$\Omega = \{(x, y) \in \mathbb{R}^2 \,:\, x^2 + y^2 < 1 \text{ and } (x < 0 \text{ or } y > 0)\}.$$

Identify $z = x + iy \in \mathbb{C}$ with (x, y). The function $w(z) = z^{2/3}$ is analytic in Ω, and so its imaginary part $u(z) = \operatorname{Im} w(z)$ is a harmonic function that satisfies

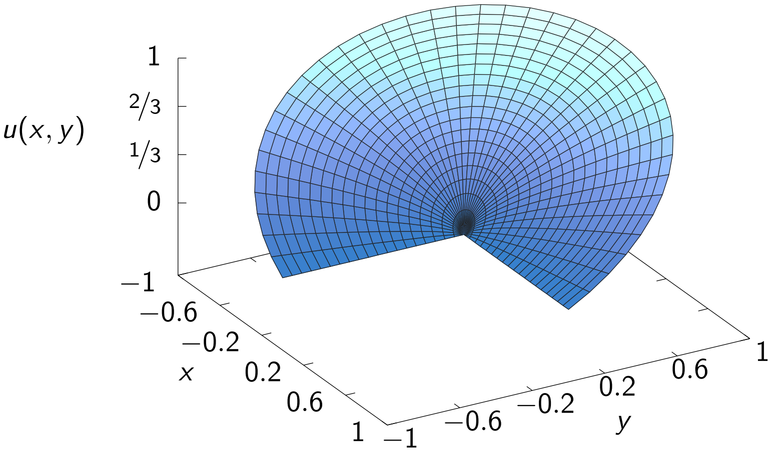
$$\nabla^2 u = 0$$

on Ω with boundary conditions

$$\begin{aligned} u(e^{i\varphi}) &= \sin\frac{2\varphi}{3} && \text{for } 0 \le \varphi \le \frac{3\pi}{2}, \\ u &= 0 && \text{elsewhere on } \partial\Omega. \end{aligned}$$

Note that $w'(z) = \frac{2}{3} z^{-1/3}$ and thus even the first derivatives of u are not bounded as $z \to 0$.

The example function $u(x, y)$



Note how the gradient of u increases without limit near $(x, y) = (0, 0)$. This function is not in $C^1(\bar{\Omega})$, the space of differentiable functions on $\bar{\Omega}$.

The need for a careful treatment

The example described on the previous slides is physically reasonable—$u(x, y)$ could represent the shape that a soap film would take when bounded by a wire in the shape of $\partial\Omega$.

As mentioned, the finite-element method is based on formulating minimization problems over a suitably chosen function space. Even though $C^1(\bar{\Omega})$ appears a natural choice, this example highlights the theoretical difficulties of using it.

We now move onto a model problem, but we bear this issue in mind going forward.

One-dimensional model problem

We now consider a model problem, (D),

$$-u''(x) = f(x) \qquad \text{for } 0 < x < 1,$$

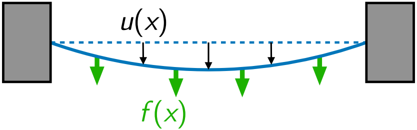
$$u(0) = u(1) = 0,$$

where f is a continuous function. Integrating this equation twice shows that there is a unique solution.

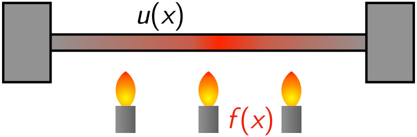
This problem can be used to model several different physical scenarios.

Physical analogs

Scenario 1: Consider an elastic cord under tension between 0 to 1. Let $u(x)$ represent a small downward displacement, and $f(x)$ represent an applied downward force density.



Scenario 2: Let $u(x)$ be the temperature in a bar whose ends are kept at a fixed temperature of zero.[7] Let $f(x)$ be an applied heat source along the bar.



[7] u could be expressed as the temperature relative to some baseline.

Alternative formulations

We now consider two alternative formulations of the model problem. First, introduce an inner product on two functions v and w as

$$(v, w) = \int_0^1 v(x)w(x)dx.$$

Introduce the function space

$$V = \left\{ v \in C([0,1]) \; : \; \begin{array}{l} v' \text{ is piecewise continuous and bounded} \\ \text{on } [0,1] \text{ and } v(0) = v(1) = 0 \end{array} \right\}$$

and the linear functional

$$F(v) = \frac{1}{2}(v', v') - (f, v).$$

Alternative formulations

Formulating the problem as a minimization yields

(M) Find $u \in V$ such that $F(u) \leq F(v)$ for all $v \in V$.

Formulating the problem in variational form yields

(V) Find $u \in V$ such that $(u', v') = (f, v)$ for all $v \in V$.

Comparing problem (V) to problem (M)

Suppose first that u is a solution to (V). Choose $v \in V$ and set $w = v - u$. Then

$$\begin{aligned} F(v) &= F(u + w) \\ &= \frac{1}{2}(u' + w', u' + w') - (f, u + w) \\ &= \frac{1}{2}(u', u') + (u', w') + \frac{1}{2}(w', w') - (f, u) - (f, w). \end{aligned}$$

Since $(u', w') = (f, w)$, this simplifies to

$$F(v) = F(u) + \frac{1}{2}(w', w') \geq F(u)$$

and hence u is a solution to (M).

Comparing problem (M) to problem (V)

Let u be a solution to (M). Let ϵ be a real number, and choose $v \in V$. Then the function

$$g(\epsilon) = F(u + \epsilon v)$$

is a differentiable function with a minimum at $\epsilon = 0$. Writing out g yields

$$g(\epsilon) = \frac{1}{2}(u', u') + \epsilon(u', v') + \frac{\epsilon^2}{2}(v', v') - (f, u) - \epsilon(f, v)$$

and hence

$$0 = g'(0) = (u', v') - (f, v).$$

Since this is true for any $v \in V$, it follows that u is a solution to (V).

Comparing problem (D) to problem (V)

Now suppose that u is a solution to (D). For $v \in V$,

$$-u''v = fv$$

and integrating both sides from zero to one yields

$$-\int_0^1 u''v\,dx = \int_0^1 fv\,dx.$$

Integrating by parts yields

$$\int_0^1 u'v'\,dx = \int_0^1 fv\,dx,$$

where the additional term vanishes due to the boundary conditions. Hence

$$(u', v') = (f, v)$$

and u is a solution to (V).

Uniqueness of the solution to (V)

Let u_1 and u_2 be two solutions to (V), so that

$$(u_1', v') = (f, v), \qquad (u_2', v') = (f, v)$$

for all $v \in V$. Hence $(u_1' - u_2', v') = 0$. Setting $v = u_1 - u_2$ yields

$$0 = (u_1' - u_2', u_1' - u_2') = \int_0^1 (u_1' - u_2')^2 dx.$$

Since u_1' and u_2' are piecewise continuous, it follows that $u_1' - u_2' = 0$ and hence $u_1 - u_2$ is constant. Using the boundary conditions shows that $u_1 = u_2$. Therefore (V) has a unique solution.

Summary

We have now shown that

$$\text{(D)} \implies \text{(V)} \iff \text{(M)}.$$

Does (V) $\implies$ (D)? No, since functions in V are only required to have a piecewise continuous first derivative. They may not have a second derivative, which is required in order to satisfy (D).

However, if we place additional restrictions on a solution u to (V), then we can obtain a solution to (D).

Comparing problem (V) to problem (D)

Suppose that $u \in V$ satisfies (V), and in addition u'' exists and is continuous. Then

$$\int_0^1 u'v'\,dx = \int_0^1 fv\,dx$$

for all $v \in V$ and integrating by parts yields

$$-\int_0^1 u''v\,dx = \int_0^1 fv\,dx.$$

Therefore

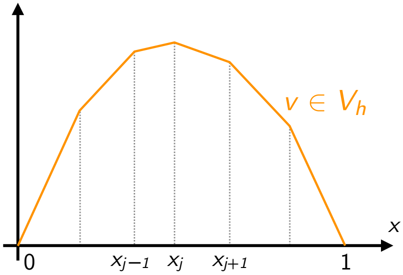
$$\int_0^1 (u'' + f)v\,dx = 0$$

and since this is true for all $v \in V$ it follows that $-u'' = f$. By the construction of V, u satisfies the boundary conditions $u(0) = u(1) = 0$, and hence u is a solution to (D).

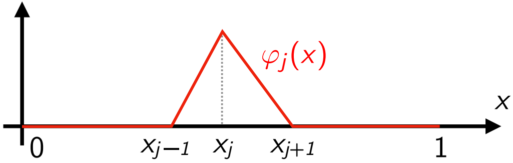
Finite-element implementation for the model problem

We now introduce a finite-dimensional subspace V_h of V on which to formulate the finite-element method on the model problem.

Introduce grid points $0 = x_0 < x_1 < x_2 < \ldots < x_M < x_{M+1} = 1$. Define subintervals $I_j = (x_{j-1}, x_j)$ and grid spacings $h_j = x_j - x_{j-1}$ for $j = 1, \ldots, M+1$ and set $h = \max_j h_j$. Let V_h be the set of functions v such that v is linear on each subinterval, v is continuous on $[0,1]$, and $v(0) = v(1) = 0$. Hence $V_h \subset V$.



Finite-element implementation for the model problem



Introduce basis functions $\varphi_j \in V_h$ for $j = 1, \ldots, M$ that satisfy

$$\varphi_j(x_i) = \delta_{ij}$$

where δ_{ij} is the Kronecker delta. Each basis function is only non-zero over a small part of the interval.

Any function $v \in V_h$ can be represented as a linear combination of these basis functions as

$$v(x) = \sum_{i=1}^{M} \eta_i \varphi_i(x).$$

Finite-element implementation for the model problem

The finite-dimensional equivalent of the minimization problem is then

$$\boxed{(\mathrm{M}_h)\ \text{Find } u \in V_h \text{ such that } F(u) \leq F(v) \text{ for all } v \in V_h.}$$

The finite-dimensional equivalent of the variational problem is

$$\boxed{(\mathrm{V}_h)\ \text{Find } u \in V_h \text{ such that } (u', v') = (f, v) \text{ for all } v \in V_h.}$$

(V_h) is referred to as Galerkin's method and (M_h) is referred to as Ritz' method.

Galerkin's method

Let u_h be a solution to (V_h). Then for any basis function φ_j,

$$(u_h', \varphi_j') = (f, \varphi_j).$$

Write the solution as

$$u_h(x) = \sum_{i=1}^{M} \xi_i \varphi_i(x).$$

Then

$$\sum_{i=1}^{M} \xi_i (\varphi_i', \varphi_j') = (f, \phi_j),$$

which is a linear system, $A\xi = b$, for a vector of unknowns ξ.

Galerkin's method: matrix formulation

Writing out the matrix problem gives

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1M} \\ a_{21} & a_{22} & \dots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \dots & a_{MM} \end{pmatrix}, \ \xi = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_M \end{pmatrix}, \ b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_M \end{pmatrix},$$

where

$$a_{ij} = (\varphi_i', \varphi_j'), \qquad b_i = (f, \varphi_i).$$

A is called the stiffness matrix and b is the load vector. This terminology dates from early work on finite-element methods for structural mechanics.

Galerkin's method for the model problem

We now evaluate A for the model problem. Since the basis functions are localized, most terms in A are zero: $a_{ij} = 0$ if $|i-j| > 1$.

For $j = 1, 2, \ldots, M$,

$$a_{jj} = (\varphi_j', \varphi_j') = \int_{x_{j-1}}^{x_j} \frac{1}{h_j^2} dx + \int_{x_j}^{x_{j+1}} \frac{1}{h_{j+1}^2} dx = \frac{1}{h_j} + \frac{1}{h_{j+1}}$$

and for $j = 2, \ldots, M$,

$$a_{j,j-1} = (\varphi_j', \varphi_{j-1}') = -\int_{x_{j-1}}^{x_j} \frac{1}{h_j^2} dx = -\frac{1}{h_j}.$$

Furthermore $a_{j,j-1} = a_{j-1,j}$ and the stiffness matrix is symmetric.

Properties of the stiffness matrix A

Consider an arbitrary vector $\eta = (\eta_1, \ldots, \eta_M)$ and associated function

$$v = \sum_{i=1}^{M} \eta_i \varphi_i.$$

Then

$$\eta^{\mathsf{T}} A \eta = \sum_{i=1}^{M} \sum_{j=1}^{M} \eta_i a_{ij} \eta_j = \sum_{i=1}^{M} \sum_{j=1}^{M} (\eta_i \varphi_i', a_j \varphi_j') = (v', v') \geq 0.$$

Equality is only achieved when $v' = 0$ everywhere. Since $v(0) = 0$, it implies that $v = 0$ everywhere, and $\eta_j = 0$ for all j.

Hence A is symmetric positive definite (SPD). This is a general property of finite-element stiffness matrices, which increases the available options for solving them numerically (*e.g.* using the Cholesky factorization, conjugate gradient method, *etc.*).

Equal grid spacing

If the grid spacings are equal, so that $h_j = h = 1/(M+1)$, then the matrix problem becomes

$$\frac{1}{h}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix}\begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \vdots \\ \xi_M \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_M \end{pmatrix}$$

This matrix problem is similar to a finite-difference (FD) discretization problem. There are some small differences:

- The b_j terms are evaluated using localized integrals of f, whereas in FD they are pointwise function evaluations.
- The stiffness matrix has a factor of h^{-1}, whereas in FD the differentiation matrix has a factor of h^{-2}. Overall, the h terms balance because the b_j terms incorporate an additional factor of h.

For many problems, finite-element and FD methods will lead to substantially different numerical systems to solve.

Error estimate for the finite-element method

We now aim to find the difference $u - u_h$ where u is the solution of (D) and u_h is the solution of (V$_h$). Since u is also a solution of (V), it follows that

$$(u', v') = (f, v)$$

for all $v \in V_h$, since $V_h \subset V$. Since u_h is a solution of (V$_h$),

$$(u_h', v') = (f, v)$$

for all $v \in V_h$. Subtracting the two yields

$$((u - u_h)', v') = 0$$

for all $v \in V_h$.

Error estimate for the finite-element method

Now, define a norm

$$\|w\| = (w, w)^{1/2} = \sqrt{\int_0^1 w^2 \, dx}.$$

Cauchy's inequality is

$$|(v, w)| \leq \|v\| \, \|w\|.$$

Theorem: For any $v \in V_h$,

$$\|(u - u_h)'\| \leq \|(u - v)'\|.$$

Error estimate for the finite-element method

Proof of theorem: Choose $v \in V_h$ and define $w = u_h - v$. Then

$$\begin{aligned} \|(u-u_h)'\|^2 &= ((u-u_h)', (u-u_h)') + ((u-u_h)', w') \\ &= ((u-u_h)', (u-u_h+w)') = ((u-u_h)', (u-v)') \\ &\leq \|(u-u_h)'\| \, \|(u-v)'\|. \end{aligned}$$

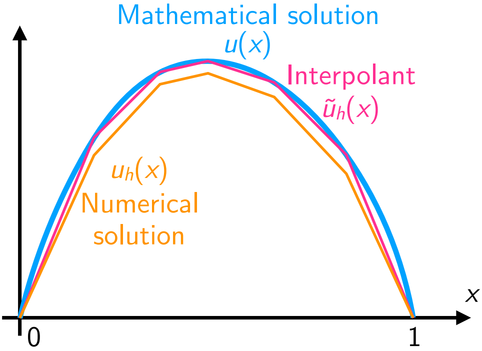
If $\|(u-u_h)'\| = 0$ then the theorem automatically holds.
Otherwise, dividing both sides by $\|(u-u_h)'\|$ gives

$$\|(u-u_h)\| \leq \|(u-v)'\|.$$

Since v is arbitrary, this proves the theorem.

Error estimate for the finite-element method

The theorem allows us to estimate the error $\|(u - u_h)'\|$ by estimating $\|(u - \tilde{u}_h)'\|$ for a suitably chosen $\tilde{u}_h$. Let $\tilde{u}_h$ be the function that linearly interpolates u at the nodes x_j.



We now use the Cauchy approximation theorem to bound the error between u and $\tilde{u}_h$, and hence bound the error between u and u_h.

Error estimate for the finite-element method

Consider x in the subinterval I_j. Then

$$|u'(x) - \tilde{u}_h'(x)| \leq h_j \max_{y \in I_j} |u''(y)|.$$

Similarly

$$|u(x) - \tilde{u}_h(x)| \leq \frac{\max_{y \in I_j} |u''(y)|}{2} |(x - x_{j-1})(x - x_j)|.$$

and since $|(x - x_{j-1})(x - x_j)| \leq h_j^2/4$, it follows that

$$|u(x) - \tilde{u}_h(x)| \leq \frac{h_j^2 \max_{y \in I_j} |u''(y)|}{8}.$$

Taking the maximum bound over all the subintervals shows that

$$|u'(x) - \tilde{u}_h'(x)| \leq h \max_{y \in [0,1]} |u''(y)|, \quad |u(x) - \tilde{u}_h(x)| \leq \frac{h^2}{8} \max_{y \in [0,1]} |u''(y)|.$$

Error estimate for the finite-element method

Using the theorem,

$$\|(u-u_h)'\| \leq \|(u-\tilde{u}_h)'\| \leq h \max_{y\in[0,1]} |u''(y)|.$$

Furthermore, since $u(0) = u_h(0)$,

$$(u-u_h)(x) = \int_0^x (u-u_h)'(y)dy$$

from the fundamental theorem of calculus. Hence

$$\begin{aligned}|u(x)-u_h(x)| &\leq \int_0^x |u'(y)-u_h'(y)|dy \\ &\leq \left(\int_0^x hdy\right) \max_{y\in[0,1]} |u''(y)| \\ &\leq h \max_{y\in[0,1]} |u''(y)|.\end{aligned}$$

Error estimate for the finite-element method

These bounds show that as the grid spacing h decreases, the numerical solution u_h will converge to the mathematical solution u.

The derivation of the bounds shows that the error scales like $O(h)$, which is sufficient to establish convergence.

However, using a more detailed derivation, it is possible to show that the error scales like $O(h^2)$ for this model problem.

Appropriate function spaces for variational problems

In the model problem considered so far, we searched for solutions over the function space

$$V = \left\{ v \in C([0,1]) \, : \, \begin{array}{l} v' \text{ is piecewise continuous and bounded} \\ \text{on } [0,1] \text{ and } v(0) = v(1) = 0 \end{array} \right\}.$$

However, for mathematical analysis it is advantageous to work with a slightly larger space of functions. The condition about requiring a piecewise continuous derivative is stricter than necessary.

We find that it is appropriate to work with Hilbert spaces, which have three requirements detailed on the following slides.

(1) Hilbert spaces are vector spaces

A Hilbert space V is a vector space. It must satisfy basic properties of commutatativity, associativity, and distributivity.[8]

We focus on real vector spaces,[9] where scalar multiplication is done using elements of $\mathbb{R}$.

A key property of a real vector space is linearity, so that for any $\alpha, \beta \in \mathbb{R}$ and $v, w \in V$, the element

$$\alpha v + \beta w$$

is also in V.

[8] See Wolfram MathWorld or Wikipedia for complete details.

[9] There are also, *e.g.*, complex vector spaces where scalar multiplication is done using elements of $\mathbb{C}$.

(2a) A Hilbert space has a scalar product

A linear form is a map $L : V \to \mathbb{R}$ such that for all $v, w \in V$ and $\beta, \theta \in \mathbb{R}$,

$$L(\beta v + \theta w) = \beta L(v) + \theta L(w).$$

A bilinear form is a map $a : V \times V \to \mathbb{R}$ that is linear in each argument, so that for all $u, v, w \in V$ and $\beta, \theta \in \mathbb{R}$,

$$a(u, \beta v + \theta w) = \beta a(u, v) + \theta a(u, w),$$
$$a(\beta u + \theta v, w) = \beta a(u, w) + \theta a(v, w).$$

The bilinear form is symmetric if $a(u, v) = a(v, u)$ for all $u, v \in V$. If

$$a(v, v) > 0 \quad \text{for all } v \in V \text{ with } v \neq 0$$

then a is a scalar product on V.

(2b) A Hilbert space has a scalar product

A Hilbert space has a scalar product. There is an associated norm

$$\|v\|_a = \sqrt{a(v,v)}$$

for all $v \in V$. Any scalar product $\langle \cdot, \cdot \rangle$ will also satisfy Cauchy's inequality,

$$|\langle v, w \rangle| \leq \|v\| \, \|w\|,$$

for all $v, w \in V$.

(3) A Hilbert space is complete

A Hilbert space is complete, so that the limit of any sequence of elements in V is also contained in V.

Specifically, let $v_1, v_2, v_3, \ldots$ of elements in V be a Cauchy sequence. This means that for any $\epsilon > 0$ there is a number $n \in \mathbb{N}$ such that $\|v_i - v_j\| < \epsilon$ for all $i, j > n$.

To be complete, every Cauchy sequence must converge to an element in V, *i.e.*, there exists a $v \in V$ such that for all $\epsilon > 0$, there exists $m \in \mathbb{N}$ such that $\|v - v_i\| < \epsilon$ for all $i > m$.

Completeness is an important property to have, since it allows us to take limits.

Hilbert space example

Let $I = (a, b)$ be an open interval. Then define

$$L_2(I) = \left\{ v \, : \, v \text{ is defined on } I \text{ and } \int_I v^2 \, dx < \infty \right\}.$$

This is the space of all square intergrable functions on I. For $v, w \in L_2(I)$ an appropriate scalar product is

$$(v, w) = \int_I vw \, dx$$

with associated norm

$$\|v\|_{L_2(I)} = \sqrt{\int_I v^2 \, dx} = \sqrt{(v, v)}.$$

Additional Hilbert space

Define

$$H^1(I) = \left\{ v \, : \, v \text{ and } v' \text{ belong to } L_2(I) \right\}.$$

For $v, w \in H^1(I)$ an appropriate scalar product is

$$(v, w)_{H^1(I)} = \int_I (vw + v'w')dx$$

with corresponding norm

$$\|v\|_{H^1(I)} = \int_I \left(v^2 + (v')^2 \right) dx.$$

Hilbert space for model problem

For the model problem, we use the Hilbert space

$$H_0^1(I) = \left\{ v \in H^1(I) \,:\, v(a) = v(b) = 0 \right\}.$$

Even though the members of H^1 are only defined on the open interval (a, b) the requirement that $v' \in L_2(I)$ ensures that there are well-defined limits $v(a)$ and $v(b)$.

For the model problem we specifically set $I = (0, 1)$. Formulating the problem in variational form yields

> (V′) Find $u \in V$ such that $(u', v') = (f, v)$ for all $v \in H_0^1(I)$.

Benefits of the weak formulation

The space $H_0^1(I)$ is larger than the original space V that was considered. $H_0^1(I)$ is specifically tailored to the variational problem, and is the largest space on which the variational problem can be formulated.

Working with $H_0^1(I)$ is frequently useful for proving the existence of solutions.

Furthermore, error estimates are often more natural in the $H^1(I)$ norm that incorporates derivative information.

Generalization to multiple dimensions

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain, with boundary $\Gamma = \partial\Omega$. A generalization of the model problem to Ω is

$$\begin{aligned} -\nabla^2 u &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \Gamma. \end{aligned}$$

Generalizations of our Hilbert spaces are

$$L_2(\Omega) = \left\{ v \,:\, v \text{ is defined on } \Omega \text{ and } \int_\Omega v^2\, dx < \infty \right\},$$

$$H^1(\Omega) = \{ v \,:\, v \in L_2(\Omega) \text{ and } \partial v / \partial x_i \in L_2(\Omega) \text{ for } i = 1, \ldots, d \}.$$

Appropriate scalar products are

$$(v, w) = \int_\Omega vw\, dx, \quad (v, w)_{H^1(\Omega)} = \int_\Omega (vw + \nabla v \cdot \nabla w) dx.$$

Generalization to multiple dimensions

Formulating the problem as a minimization yields

(M′) Find $u \in H_0^1(\Omega)$ such that $F(u) \leq F(v)$ for all $v \in H_0^1(\Omega)$.

Formulating the problem in variational form yields

(V′) Find $u \in H_0^1(\Omega)$ such that $(u', v') = (f, v)$ for all $v \in H_0^1(\Omega)$.

Here $F(v) = \frac{1}{2}a(v, v) - (f, v)$ with

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx, \qquad (f, v) = \int_\Omega fv \, dx.$$

Neumann boundary conditions

The previous examples used Dirichlet boundary conditions, which were straightforward to impose by restricting the function space. However, implementing a Neumann condition requires a different approach.

Consider the Neumann problem

$$\begin{aligned} -\nabla^2 u + u &= f \qquad \text{in } \Omega, \\ \frac{\partial u}{\partial n} &= g \qquad \text{on } \Gamma, \end{aligned}$$

where Ω is a bounded domain and $\partial / \partial n$ is an outward normal derivative.

Neumann boundary conditions

The Neumann problem can be expressed as a variational problem by finding $u \in H^1(\Omega)$ such that

$$a(u, v) = (f, v) + \langle g, v \rangle$$

for all $v \in H^1(\Omega)$ such that

$$a(u, v) = \int_\Omega [\nabla u \cdot \nabla v + uv] \, dx,$$

$$(f, v) = \int_\Omega fv \, dx, \quad \langle g, v \rangle = \int_\Gamma gv \, ds.$$

This is equivalent to minimizing

$$F(v) = \frac{1}{2} a(v, v) - (f, v) - \langle g, v \rangle$$

over $v \in H^1(\Omega)$.

Neumann boundary conditions

To obtain the variational problem, we first multiply the governing equation by a test function $v \in H^1(\Omega)$ and integrate to obtain

$$-\int_\Omega v\nabla^2 u\, dx + \int_\Omega vu\, dx = \int_\Omega fv\, dx.$$

Applying Green's first identity gives

$$\int_\Omega \nabla v \cdot \nabla u\, dx - \int_\Gamma \frac{\partial u}{\partial n} v\, ds + \int_\Omega vu\, dx = \int_\Omega fv\, dx.$$

Substituting $\partial u/\partial n = g$ on Γ yields the variational problem from the previous slide.

Boundary conditions

Hence, the Neumann condition is incorporated into the variational problem itself, rather than by altering the function space that is used. This is called a natural boundary condition.

By contrast, a boundary condition where the function space is restricted is referred to as an essential boundary condition.

Example problem

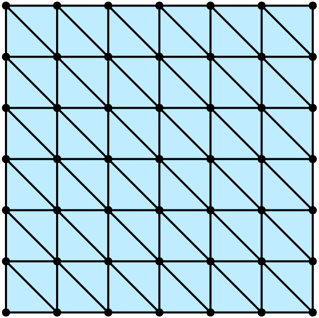
Consider solving the Poisson equation in the unit square $\Omega = (0,1)^2$:

$$\begin{aligned} -\nabla^2 u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Consider a triangulation like that shown, with a mesh size of h. Choose

$$S_h = \{v \in C(\bar{\Omega} : v \text{ is linear in} \\ \text{every triangle and } v = 0\}.$$

In every triangle $v \in S_h$ has the form $v(x, y) = a + bx + cy$.

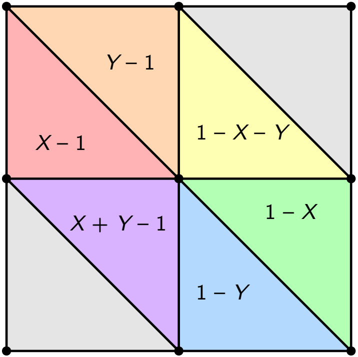


Example problem

Let interior mesh points be labeled (x_j, y_j) for $j = 1, \ldots, N$. $v \in S_h$ is determined by its value of the mesh points. Introduce a basis

$$\psi_i(x_j, y_j) = \delta_{ij}.$$

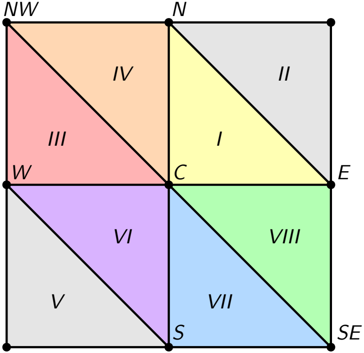
Define the coordinates $X = \frac{x - x_i}{h}$, $Y = \frac{y - y_i}{h}$. Then ψ_i is non-zero in six triangles, and whose values are explicitly shown in the diagram.



Finite-element stencil

Label a group of basis functions

$$
\begin{aligned}
a(\psi_C, \psi_C) &= \int_\Omega (\nabla \psi_c)^2 dxdy \\
&= 2 \int_{I+III+IV} \left[(\partial_x \psi_C)^2 + (\partial_y \psi_C)^2 \right] dxdy \\
&= 2 \int_{I+III} (\partial_x \psi_C)^2 dxdy \\
&\quad + 2 \int_{I+IV} (\partial_y \psi_C)^2 dxdy \\
&= \frac{2}{h^2} \int_{I+III} dxdy + \frac{2}{h^2} \int_{I+IV} dxdy \\
&= 4.
\end{aligned}
$$



Finite-element stencil

For

$$\begin{aligned} a(\psi_C, \psi_S) &= \int_{VI+VII} \nabla\psi_C \cdot \nabla\psi_S \, dxdy \\ &= \int_{VI+VII} (\partial_y\psi_C)(\partial_y\psi_S) dxdy \\ &= -\frac{1}{h^2}\int_{VI+VII} dxdy = -1. \end{aligned}$$

Similarly

$$a(\psi_C, \psi_N) = a(\psi_C, \psi_E) = a(\psi_C, \psi_W) = -1$$

and

$$a(\psi_C, \psi_{NW}) = a(\psi_C, \psi_{SE}) = 0.$$

Finite-element stencil

Thus, for this choice of basis, the associated matrix problem $Az = b$ for the solution $u_h = \sum_k z_k \psi_k$ has a five-point stencil

$$\begin{array}{ccc} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{array}$$

If gridpoints are indexed as $k = (l, m)$ for (x_l, y_m), and the solution is expressed as $u_h = \sum_{l,m} z_{l,m} \psi_{l,m}$, then

$$[Az]_{l,m} = 4z_{l,m} - z_{l-1,m} - z_{l+1,m} - z_{l,m-1} - z_{l,m+1}.$$

where $z_{l,m}$ is treated as zero if it lies on the boundary. This exactly matches our finite-difference stencil for the Poisson equation!

Finite-element stencil

Hence, for this choice of basis, the finite-element (FE) method and the finite-difference (FD) stencil agree. This is not true in general, but highlights the similarities between the two discretization approaches.

Note that the treatment of the source term may differ between FE and FD. In FD, we discretize the field f at gridpoints (l, m) and write

$$[Az]_{l,m} = h^2 f_{l,m}$$

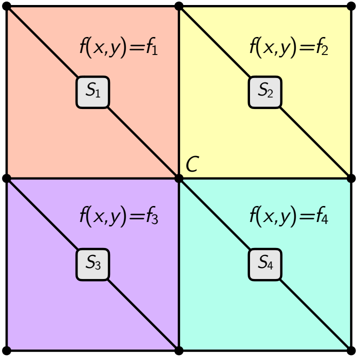
In FE, we have the freedom to specify how f is represented, which affects the $\langle l, \psi_k \rangle$ terms appearing in the Ritz–Galerkin method.

Source term discretization

One approach is to treat f as piecewise constant on grid squares as shown. Then

$$\begin{aligned}\langle l, \psi_c\rangle &= \int_\Omega f\,\psi_C\,dxdy \\ &= \int_{S_1} f_1\psi_C dxdy + \int_{S_2} f_2\psi_C dxdy \\ &\quad + \int_{S_3} f_3\psi_C dxdy + \int_{S_4} f_4\psi_C dxdy \\ &= \frac{h^2(2f_1 + f_2 + f_3 + 2f_4)}{6}.\end{aligned}$$

Note the asymmetry in this formula, which arises because the basis function ψ_k is asymmetric.



Definitions

Define

$$\mathcal{P}_t = \{u(x,y) = \sum_{i+k\leq t} c_{ik}x^i y^k\}$$

to be the set of polynomials of degree $\leq t$. If all polynomials of degree $\leq t$ are used, the finite elements have complete polynomials.

Definitions

A finite element is said to be a C^k element if it is contained in $C^k(\Omega)$. Note that the previous 2D example using right-angled triangles has C^0 elements.

We use the terminology conforming finite element if the functions lie in the Sobolev space in which the variational problem is posed.

Sometimes, nonconforming elements can be useful, *e.g.* to approximate a curved domain with a triangular mesh.

Requirements on the meshes

A partition $\mathcal{T} = \{T_1, T_2, \dots, T_M\}$ of Ω into elements is called admissible if

1. $\bar{\Omega} = \bigcup_{i=1}^{M} T_i$.
2. If $T_i \cap T_j$ consists of exactly one point, it is a common vertex of T_i and T_j.
3. For $i \neq j$, if $T_i \cap T_j$ consists of most than one point, then $T_i \cap T_j$ is a common edge of T_i and T_j.

Properties of the mesh

We write $\mathcal{T}_h$ instead of $\mathcal{T}$ when every element has diameter at most $2h$.

A family of partitions $\{\mathcal{T}_h\}$ is called shape regular provided that there exists a number $\kappa > 0$ such that every T in $\mathcal{T}_h$ contains a circle of radius ρ_T with $\rho_T \geq h_T/\kappa$.

A family of partitions $\{\mathcal{T}_h\}$ is called uniform if there exists a number $\kappa > 0$ such that every element T in $\mathcal{T}_h$ contains a circle with radius $\rho_T \geq h/\kappa$.

Differentiability properties

In the one-dimensional finite element example, we used a piecewise cubic that was continuous but not differentiable.

Theorem (see Braess): Let $k \geq 1$ and suppose Ω is bounded. Then a piecewise infinitely differentiable function $v : \bar{\Omega} \to \mathbb{R}$ belongs to $H^k(\Omega)$ if and only if $v \in C^{k-1}(\Omega)$.

Thus functions in C^0 are in H^1. For second-order elliptic PDEs this allows us to calculate the finite element terms $a(\psi_k, \psi_i)$, since it involves integrals of weak derivatives $\partial \psi_k$.

Fourth-order elliptic problems involve integrals of $\partial^2 \psi_k$ and therefore require C^1 basis functions.

Triangular elements

Any triangle can be transformed into another via an affine transformation $x \mapsto Ax + x_0$ for a matrix A and vector b.

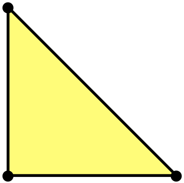
Suppose u is a polynomial of degree t (*i.e.* a member of $\mathcal{P}_t$). If we apply an affine transformation to u we get another polynomial of degree t. Hence $\mathcal{P}_t$ is invariant under affine linear transformations.

We therefore look for different ways to represent polynomials on triangles, from finite element basis functions can be constructed.

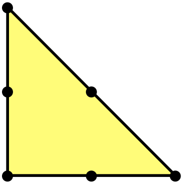
Polynomials on triangles

Let $t \geq 0$. For a triangle T, let $z_1, z_2, \ldots, z_s$ be $s = 1 + 2 + \ldots + (t+1)$ points in T that lie on $t+1$ lines as shown below. Then for every $f \in C(T)$ there is a unique polynomial p of degree $\leq t$ satisfying the interpolation conditions

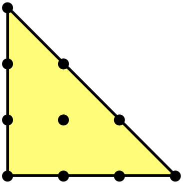
$$p(z_i) = f(z_i).$$



Linear ($t = 1$)



Quadratic ($t = 2$)



Cubic ($t = 3$)

Nodal basis

Suppose that for a given finite element space, there is a set of points such that the function values at those points uniquely determine the function. Then the set of points is called a nodal basis.

Suppose we are given a triangulation of Ω and we place points in each triangle as shown on the previous slide. Points on edges will be common between triangles. Consider a nodal basis made from these points.

Consider two adjacent triangles. The function in each triangle is in $\mathcal{P}_t$. The restriction of the function from either side to the common edge is a polynomial of degree t. Since the restrictions must agree at the $n+1$ nodes along the edge, it follows that the overall function is continuous. Thus we have a C^0 nodal basis.[10]

[10] The finite element example problem uses a one-dimensional version of this basis construction, for the case of cubic elements.

Construction of C^1 elements

Note the nodal basis construction procedure does not lead to C^1 elements, even for $t = 2$ or $t = 3$. The basis functions are only continuous across the edges between triangles.

As shown in the finite element example problem, using C^0 elements can still provide high-order accuracy solutions.

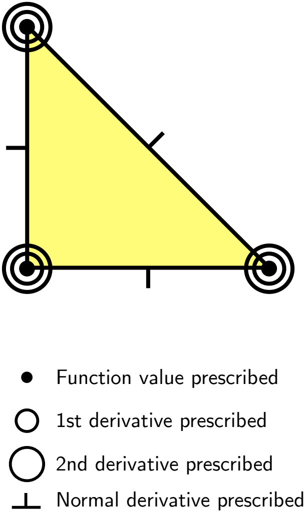
Constructing C^1 elements is more difficult. We provide two examples.

Argyris triangle

We specify the function value, first derivative, and second derivative at each triangular vertex. This gives $1 + 2 + 3 = 6$ constraints per vertex, providing eighteen constraints in total.

To satisfy these constraints we work with $\mathcal{P}_5$, which has 21 degrees of freedom.

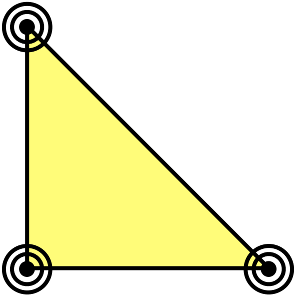
To constrain the remaining three degrees of freedom, we specify the normal derivative at each edge center. This creates a C^1 element.


● Function value prescribed
○ 1st derivative prescribed
◯ 2nd derivative prescribed
⊥ Normal derivative prescribed

Bell triangle

For the Bell triangle we again work with $\mathcal{P}_5$, but restrict to the case when the normal derivatives along each edge are degree 3 instead of degree 4.

These three additional constraints ensure differentiability across edges.



● Function value prescribed

○ 1st derivative prescribed

◯ 2nd derivative prescribed

⊥ Normal derivative prescribed

Bilinear elements

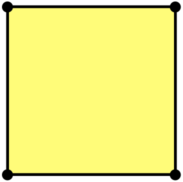
The polynomial families $\mathcal{P}_t$ are not well-suited to rectangular partitions of the domain. Instead we use the tensor product polynomial families

$$\mathcal{Q}_t = \{u(x, y) = \sum_{0 \leq i,k \leq t} c_{ik} x^i y^k\}$$

For $t = 1$, we obtain functions of the form

$$u(x, y) = a + bx + cy + dxy.$$

The four vertices of the square form a nodal basis with C^0 elements.



Finite element definition

A finite element is a triple (T, Π, Σ) with the following properties:

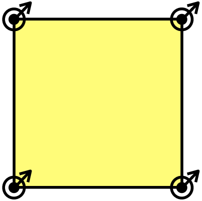
1. T is a polyhedron in $\mathbb{R}^d$.
2. Π is a subspace of $C(T)$ with finite dimension s.
3. Σ is a set of s linear independent functionals on Π. Every $p \in \Pi$ is uniquely defined by the values of the s functionals in Σ.

The Bogner–Fox–Schmit rectangle

The Bogner–Fox–Schmit rectangle leads to C^1 elements on rectangles. The polynomial space is $\Pi = \mathcal{Q}_3$ (with dimension 16). The linear functionals are

$$\Sigma = \{p(a_i), \partial_x p(a_i), \partial_y p(a_i), \partial_{xy} p(a_i), i = 1, 2, 3, 4\}$$

where the a_i are the corners of the rectangles. The element is shown below, with the diagonal arrow indicating a mixed second derivative.



Affine families

A family of finite element spaces S_h for partitions $\mathcal{T}_h$ of Ω is called an affine family if there exists a finite element $(T_{\text{ref}}, \Pi_{\text{ref}}, \Sigma)$ called the reference element such that

4. For every $T_j \in \mathcal{T}_h$ there exists an affine mapping $F_j : T_{\text{ref}} \to T_j$ such that for every $v \in S_h$ its restriction to T_j has the form
$$v(x) = p(F_j^{-1}x)$$
with $p \in \Pi_{\text{ref}}$.