An Introduction to the Finite Element Method for Elliptic Problems

This material is intended as a replacement for section 5.3 in

NUMERICAL ANALYSIS
second course
(Numerisk analys II och Numerisk analys MN2)
Complements to the textbook
and exercises

Martin Berggren

January 8, 2002

UPPSALA UNIVERSITY

Information Technology
Department of Scientific Computing

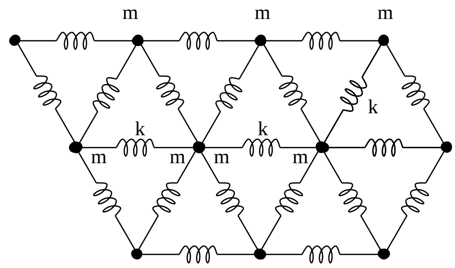

Figure 1: Interconnecting numerous point masses with massless springs gives a simple model of an elastic membrane.

The finite-element method (FEM) developed in the 1950's as a method to calculate elastic deformations in solids. The idea was to model a continuum by an assemblage of "finite elements": as an example, an elastic membrane (such as a drum skin) can be modeled by numerous point masses interconnected with massless springs, as illustrated in figure 1. Fifty years later, the point of view is more abstract, which allows FEM to be used as a general-purpose method, applicable to all kinds of partial differential equations. FEM is the dominating technique for solving solid-mechanics problems such as estimating stresses and strains in elastic material under prescribed loads. Most CAD (Computer Aided Design) systems provide finite-element solvers in a highly integrated fashion. The engineer can typically with a few clicks on the computer screen estimate the deformations and stresses of, say, a machine part during the design. Finite-element methods are also commonly applied to other areas, such as calculations of electromagnetic fields and fluid flows.

To shortly introduce the ideas, this note concentrates on a standard model problem for elliptic boundary-value problems, the Poisson problem. Only homogeneous Dirichlet boundary conditions are covered here.

1 FEM for the Poisson Problem in Two Space Dimensions

We consider the elliptic boundary-value problem

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega, \end{aligned} \tag{1}$$

where Ω is an open, bounded and connected domain in the plane, and $\partial\Omega$ is its boundary. The Laplacian Δ is the sum of second derivatives

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

Letting u represent a temperature field, equation (1) models steady heat conduction in a homogeneous, isotropic material, such as a metal, in which the temperature is held at zero on the boundary. The function f can be used to model heat sources such as electric heaters embedded in the material.

1.1 Vector Calculus and Green's Formula

The finite-element discretization is not applied directly to the Poisson problem in the differential form (1). Instead, a reformulation, the *variational form*, is the basis for discretization. The variational form combines the differential equation and the boundary condition in a single expression. The basic tool used to obtain the variational form is *Green's formula*, a generalization to higher dimensions of the integration-by-parts formula

$$v(x)u'(x)|_0^1 = \int_0^1 v'u'\,dx + \int_0^1 vu''\,dx. \tag{2}$$

To derive Green's formula, we need some definitions and formulas from vector calculus. The "vector" in vector calculus can be thought of as an arrow, or a line segment with a direction. In this section, we will use bold symbols like $\boldsymbol{a}$ to denote such vectors. The components of $\boldsymbol{a}$ will be denoted (a_1, a_2). The dot product between two vectors is defined as

$$\boldsymbol{a} \cdot \boldsymbol{b} = \sum_{i=1}^{2} a_i b_i.$$

The differentiation operator ∇ can be thought of as the "vector operator"

$$\nabla = \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2} \right). \tag{3}$$

This operator may be used in different ways. If it operates on a scalar, differentiable function from $\mathbb{R}^2$ to $\mathbb{R}$, it produces the *gradient vector*

$$\nabla v = \left(\frac{\partial v}{\partial x_1}, \frac{\partial v}{\partial x_2} \right).$$

The components of the gradient vector simply yields the derivative of the function v i the directions of the coordinate axis. Linear combinations

$$a_1 \frac{\partial v}{\partial x_1} + a_2 \frac{\partial v}{\partial x_2}$$

can also be formed. If $\boldsymbol{a} = (a_1, a_2)$ is a vector of *unit length* ($a_1^2 + a_2^2 = 1x1$), we obtain the *directional derivative*, that is, the derivative in the direction of the "arrow" $\boldsymbol{a}$,

$$\frac{\partial v}{\partial a} = a_1 \frac{\partial v}{\partial x_1} + a_2 \frac{\partial v}{\partial x_2} = \boldsymbol{a} \cdot \nabla v. \tag{4}$$

Note that we obtain the derivatives in the coordinate-axes directions by choosing $\boldsymbol{a} = (1, 0)$ and $(0, 1)$, respectively.

If $\boldsymbol{w}$ is a differentiable *vector-valued* function from $\mathbb{R}^2$ to $\mathbb{R}^2$, one may form the dot product between the operator ∇ and the function $\boldsymbol{w}$ to define the *divergence*

$$\nabla \cdot \boldsymbol{w} = \sum_{i=1}^{2} \frac{\partial w_i}{\partial x_i}, \tag{5}$$

The product rule of differentiation says that

$$\frac{\partial}{\partial x_i}(fg) = g\frac{\partial f}{\partial x_i} + f\frac{\partial g}{\partial x_i}.$$

Substituting

$$f = v, \qquad g = \frac{\partial u}{\partial x_i} \tag{6}$$

into formula (5) and summing yields that

$$\sum_{i=1}^{2} \frac{\partial}{\partial x_i}\left(v\frac{\partial u}{\partial x_i}\right) = \sum_{i=1}^{2} \frac{\partial v}{\partial x_i}\frac{\partial u}{\partial x_i} + \sum_{i=1}^{2} v\frac{\partial^2 u}{\partial x_i^2}, \tag{7}$$

for differentiable functions v and twice differentiable functions u. Expression (7) may be written in a "vector form", using the ∇ operator defined in (3) and dot products,

$$\nabla \cdot (v\nabla u) = \nabla v \cdot \nabla u + v\Delta u. \tag{8}$$

The *divergence theorem* (or Gauss' theorem) says that the integral of a vector-field divergence over a domain is equal to the integral of the normal component of the field along the boundaries,

$$\int_\Omega \sum_{i=1}^{2} \frac{\partial w_i}{\partial x_i}\, d\Omega = \int_{\partial\Omega} \sum_{i=1}^{2} n_i w_i\, ds, \tag{9}$$

wher n_i is the components of the outward-directed unit normal vector on $\partial\Omega$. Loosely speaking: the divergence theorem allows the differentiation operator $\partial/\partial x_i$ to be replaced by the normal component n_i at the same time as the integral over the domain Ω is replaced by an integral over the boundary $\partial\Omega$. Again, using the ∇ operator and dot products, expression (9) may be written in the vector form

$$\int_\Omega \nabla \cdot \boldsymbol{w}\, d\Omega = \int_{\partial\Omega} \boldsymbol{n} \cdot \boldsymbol{w}\, ds, \tag{10}$$

with $\boldsymbol{n} = (n_1, n_2)$. The divergence theorem holds for functions $\boldsymbol{w}$ and boundaries $\partial\Omega$ that are sufficiently smooth.

Integrating formula (8) and using the divergence theorem (10) yields

$$\int_\Omega \nabla \cdot (v\nabla u)\, d\Omega = \int_{\partial\Omega} \boldsymbol{n} \cdot \nabla u\, ds = \int_\Omega \nabla v \cdot \nabla u\, d\Omega + \int_\Omega v\Delta u\, d\Omega. \tag{11}$$

Recalling definition (4) of the directional derivative, expression (11) may be rewritten to provide Green's formula in the standard form

$$\int_{\partial\Omega} v \frac{\partial u}{\partial n}\, ds = \int_{\Omega} \nabla v \cdot \nabla u \, d\Omega + \int_{\Omega} v \Delta u \, d\Omega. \tag{12}$$

From this, we see that Green's formula is nothing else than a generalization of the integration-by-parts formula (2) to higher dimensions.

1.2 The Variational Form

A *classical solution* to the Poisson problem (1) is a smooth function u satisfying equation (1). The precise requirements for u to be a classical solution is that it should be twice continuously differentiable, and its first and second derivatives should be functions that can be continuously extended up to the boundary. This assures that Green's formula (12) can be applied on u. Let v be a smooth function from $\overline{\Omega} = \Omega \cup \partial\Omega$ to $\mathbb{R}$ such that $v(x) = 0$ for each $x \in \partial\Omega$. Multiply both sides of equation (1) with v, integrate over Ω, and apply Green's formula (12) to obtain

$$\begin{aligned} \int_{\Omega} v f \, d\Omega &= -\int_{\Omega} v \Delta u \, d\Omega \\ &= -\int_{\partial\Omega} v \frac{\partial u}{\partial n}\, ds + \int_{\Omega} \nabla v \cdot \nabla u \, d\Omega = \int_{\Omega} \nabla v \cdot \nabla u \, d\Omega, \end{aligned} \tag{13}$$

where the fact that v vanishes on the boundary has been used in the last equality. From expression (13) immediately follows

Theorem 1. *If u is a classical solution to the Poisson problem* (1)*, then u satisfies*

$$\int_{\Omega} \nabla v \cdot \nabla u \, d\Omega = \int_{\Omega} v f \, d\Omega, \tag{14}$$

for each smooth function v vanishing on the boundary.

Equation (14) is called the *variational form* of the Poisson equation. Theorem 1 refers to the original problem (1), but the variational form can be used to *define* a function u without reference to the differential equation. For this purpose, we introduce the *function space*

$$V = \left\{ v \mid \int_{\Omega} |\nabla v|^2 \, d\Omega < +\infty \text{ and } v|_{\partial\Omega} = 0 \right\}, \tag{15}$$

where

$$|\nabla v|^2 = \left(\frac{\partial v}{\partial x_1}\right)^2 + \left(\frac{\partial v}{\partial x_2}\right)^2.$$

The condition

$$\int_{\Omega} |\nabla v|^2 \, d\Omega < +\infty$$

corresponds in many applications to demanding that the *energy* should be bounded, for instance when the Poisson equation is used to model steady heat conduction. Note that V is a *linear space*, that is, if $v,\, w \in V$, then $\alpha v + \beta w \in V$ for each $\alpha,\, \beta \in \mathbb{R}$. The space V is a *Sobolev space*, that is, a function space that contains integral or pointwise bounds on the derivatives of functions, and is often denoted $H^1_0(\Omega)$ in the literature.

The variational problem, now formulated without reference to the differential equation (1) is the following.

$$\begin{aligned} &\text{Find } u \in V \text{ such that} \\ &\int_\Omega \nabla v \cdot \nabla u \, d\Omega = \int_\Omega v f \, d\Omega \qquad \forall v \in V. \end{aligned} \tag{16}$$

Solutions to variational problem (16) are called *weak solutions* of the partial differential equation (1). From Theorem 1 follows that classical solutions are weak solutions. As the label "weak" suggests, there are weak solutions that are not classical solutions. However, one can show that weak solutions are classical solutions provided that the function f and the boundary $\partial\Omega$ are sufficiently smooth.

1.3 The Minimization Problem

The variational form above is all that is needed to define a finite-element discretization. However, a classical solution to the particular problem that we consider, equation (1), also satisfies a certain *minimization problem*, that is, the classical solution minimizes the quadratic form

$$F(v) = \frac{1}{2} \int_\Omega |\nabla v|^2 \, d\Omega - \int_\Omega f v \, d\Omega.$$

Similarly as was done for the variational problem, we can also consider the problem of minimizing F within the function space V without reference to classical solutions, that is, consider the problem:

$$\begin{aligned} &\text{find } u \in V \text{ such that} \\ &F(u) \le F(v) \qquad \forall v \in V. \end{aligned} \tag{17}$$

In fact, the variational problem (16) and the minimization problem (17) are equivalent:

Theorem 2. *The element $u \in V$ minimizes F if and only if it is a solution to the variational problem* (16)

Remark 1. The proof below may appear long, but is essentially no more complicated than showing, by differentiation, that the parabola $F(x) = \frac{1}{2}x^2 - xf$ has its minimum at $x = f$.

Proof. For any u, $v \in V$, and $t \in \mathbb{R}$, we have

$$\begin{aligned} F(u+tv) &= \frac{1}{2}\int_\Omega |\nabla u + t\nabla v|^2 \, d\Omega - \int_\Omega f(u+tv)\, d\Omega \\ &= \frac{1}{2}\int_\Omega \left[|\nabla u|^2 + 2t\nabla u\cdot\nabla v + t^2|\nabla v|^2\right] \, d\Omega - \int_\Omega f(u+v)\, d\Omega \\ &= F(u) + t\left(\int_\Omega \nabla v\cdot\nabla u \, d\Omega - \int_\Omega vf\, d\Omega\right) + \frac{t^2}{2}\int_\Omega |\nabla v|^2 \, d\Omega. \end{aligned} \tag{18}$$

(i) Assume that $t = 1$ and that $u \in V$ is a solution to the variational problem (16). Then expression (18) reduces to

$$F(u+v) = F(u) + \underbrace{\frac{1}{2}\int_\Omega |\nabla v|^2 \, d\Omega}_{\geq 0} \geq F(u) \tag{19}$$

for any $v \in V$, which shows that u minimizes F.

(ii) Now assume that $u \in V$ minimizes F. For any $t \in \mathbb{R}$ and $v \in V$, we define the function $f(t) = F(u+tv)$, that is, by perturbing F away from its minimum. Thus, the function f has a minimum for $t = 0$. Expression (18) shows that f is a second-order polynomial in t. The leading-term coefficient is positive for nonzero v, so the polynomial has a minimum when the derivative vanishes. Setting $f'(0) = 0$, we conclude that

$$\int_\Omega \nabla v\cdot\nabla u \, d\Omega - \int_\Omega vf\, d\Omega = 0, \tag{20}$$

for any $v \in V$, that is, u is a solution to the variational problem (16).

□

Remark 2. Variational forms can be defined for practically all elliptic boundary-value problems, but a corresponding minimization form does not always exist, for instance when the differential equation contains first-derivative terms.

Remark 3. In mechanics application the variational form (16) is called *the principle of virtual work*, and the minimization problem (17) is called *the principle of minimum potential energy.*

Remark 4. The terminology used here, "variational" for (16) and "minimization" for (17), is convenient for our purpose, but is not the only existing. Quite commonly the minimization problem is called a variational form. In fact, the notion of variational forms was first attached to minimizations of "functionals" like F in the *calculus of variations.*

1.4 Meshing and Finite-Element Approximation

We introduce a *triangulation* of the domain Ω, that is, Ω will be subdivided into nonoverlapping triangles as illustrated in figures 2 and 4. The triangular corners are called the *nodes* of the triangulation. The *boundary nodes* are the nodes which are located on the boundary, and the *internal nodes* are the nodes which are not boundary nodes. A valid triangulation should not contain "hanging nodes", that is, no node should be located at another triangles side, as in figure 3. The "fineness" of the triangulation is characterized by a parameter $h > 0$, the largest length of any of the triangular sides, for instance.

Now define V_h as the space of all functions that are *continuous* on $\overline{\Omega}$, *linear* on each triangle, and *vanishing* on the boundary $\partial\Omega$. The graph of such a function is a surface composed of triangular-shaped planes, as illustrated in figure 5.

This space is constructed so that $V_h \subset V$, and we define the *finite-element discretization* of the Poisson problem (1) as

$$\begin{aligned}&\text{Find } u_h \in V_h \text{ such that}\\ &\int_\Omega \nabla v_h \cdot \nabla u_h \, d\Omega = \int_\Omega v_h f \, d\Omega \qquad \forall v_h \in V_h.\end{aligned} \tag{21}$$

Note that the discretization is obtained simply by replacing V with the subspace V_h in the variational form (16); this way of discretizing is called a *Galerkin approximation*.

In general, a *finite-element discretization* of a boundary-value problems is a *Galerkin approximations*, based on *piecewise polynomials*, applied to a *variational form* of the boundary-value problem.

1.5 The Algebraic Problem

A function in the above defined space V_h is uniquely defined by its values at the *internal* nodes (we already know that the function is zero at the boundary nodes). To see this, it is enough to note that the planar surface of u_h on each triangle is uniquely defined by the values of u_h at the triangular corners. Let N be the number of internal nodes. Using the *basis functions* $\{\phi_j(\boldsymbol{x})\}_{j=1}^N \subset V_h$, each function $u_h \in V_h$ can be written

$$u_h(\boldsymbol{x}) = \sum_{j=1}^N u_j \phi_j(\boldsymbol{x}), \tag{22}$$

where u_j is the value of u_h at node j, and $\phi_j(\boldsymbol{x})$ is the "tent" function depicted in figure 6. The function ϕ_j is zero everywhere, except that it raises as a "tent" around node j, that is, $\phi_j \in V$ such that

$$\phi_j(\boldsymbol{x}_k) = \begin{cases} 1 & \text{if } k = j, \\ 0 & \text{otherwise,} \end{cases}$$

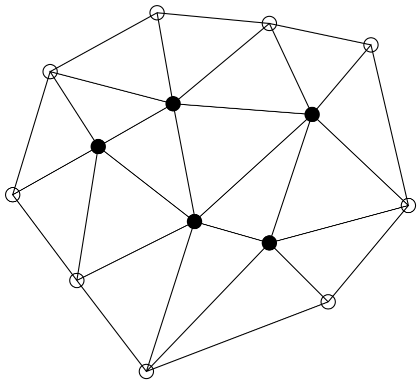

Figure 2: A valid triangulation. Internal nodes are marked by solid dots and boundary nodes by circles.

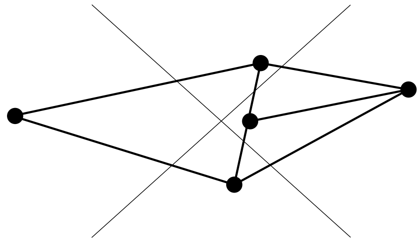

Figure 3: Not a valid triangulation: contains hanging nodes.

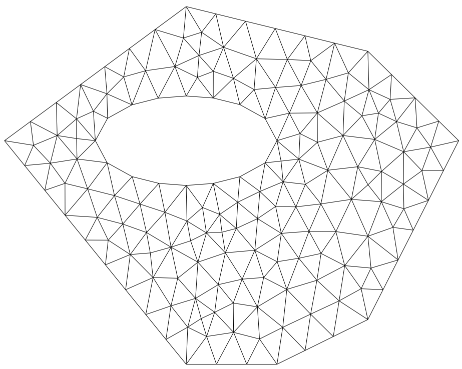

Figure 4: A more complicated triangulated domain (note that the domain may contain holes!)

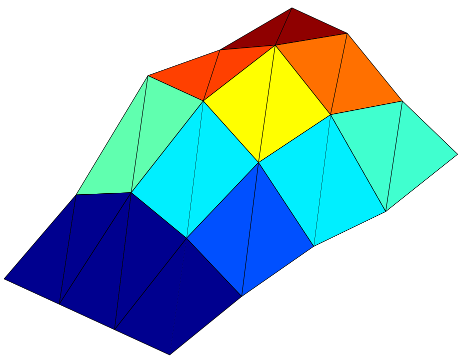

Figure 5: The functions in V_h are continuous and linear on each triangle. (The boundary nodes are not included in this picture.)

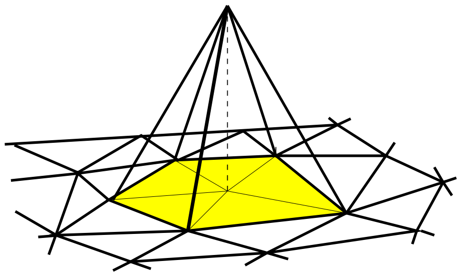

Figure 6: The basis function $\phi_j(\boldsymbol{x})$ is equal to one at node j and zero at all other nodes.

where $\boldsymbol{x}_k$ is the coordinate of node k.

Substituting expansion (22) into equation (21) yields that

$$\sum_{j=1}^{N} u_j \int_\Omega \nabla v_h \cdot \nabla \phi_j \, d\Omega = \int_\Omega v_h f \, d\Omega \qquad \forall v_h \in V_h.$$

Since equation (1.5) should hold for each $v_h \in V_h$, it must in particular hold for $v_h = \phi_i$, $i = 1, \ldots, N$, which means that

$$\sum_{j=1}^{N} u_j \int_\Omega \nabla \phi_i \cdot \nabla \phi_j \, d\Omega = \int_\Omega \phi_i f \, d\Omega \qquad i = 1, \ldots, N. \tag{23}$$

Problem (23) is i system of linear equation in the coefficients u_j, $j = 1, \ldots, N$, that is,

$$\mathbf{A}\mathbf{u} = \mathbf{b}, \tag{24}$$

where the matrix $\mathbf{A}$ has components

$$A_{ij} = \int_\Omega \nabla \phi_i \cdot \nabla \phi_j \, d\Omega,$$

and

$$\mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} \int_\Omega \phi_1 f \, d\Omega \\ \vdots \\ \int_\Omega \phi_N f \, d\Omega \end{pmatrix}.$$

With a terminology borrowed from solid mechanics, the matrix $\mathbf{A}$ is called the *stiffness matrix* and the vector $\mathbf{b}$ the *load vector*. This terminology is used also for cases, like heat conduction, when the PDE we are discretizing has nothing to do with mechanics!

We conclude that a numerical approximation of the Poisson problem with a finite-element method involves setting up and solving the linear system (24).

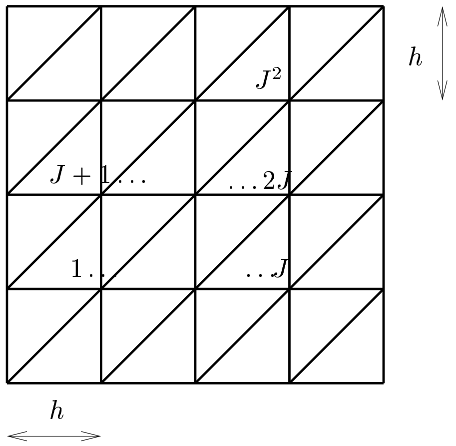

Figure 7: A structured meshing of the unit square.

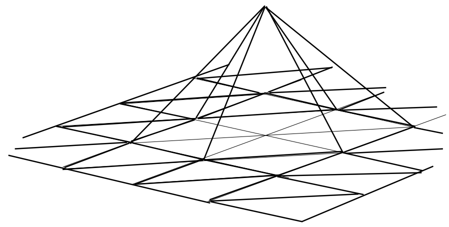

Figure 8: A basis function associated with the mesh in figure 7

1.6 An Example

Let the domain Ω be the unit square, and consider the *structured mesh* of figure 7. There are J internal nodes in both directions and the sides of each triangle are $h = 1/(J+1)$. There is a total of $J^2 = N$ internal nodes, assumed to be numbered in the row-wise direction as indicated in figure 7. The basis functions ϕ_i have the shape indicated in figure 8. The *support* of each basis function, that is, the nonzero region of the function, is on the 6 neighboring triangles which surrounds node i. Note that this means that most of the stiffness matrix elements

$$A_{ij} = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j \, d\Omega$$

are zero. For instance, $A_{i,i+2} = 0$ since there is no overlap in the support for the functions ϕ_i and ϕ_{i+2}; see figure 9. In fact, A_{ij} can be nonzero only when i and j are associated with *nearest-neighboring* nodes (figure 10).

To calculate the stiffness-matrix elements, we need to know the gradients of the basis functions,

$$\nabla\phi_i = \left(\frac{\partial\phi_i}{\partial x}, \frac{\partial\phi_i}{\partial y}\right).$$

The gradient is constant at each triangle since ϕ_i is composed of planar surfaces. Letting the x and y directions be oriented in the horizontal and vertical directions, respectively, the values of the gradient at the support of the basis function are indicated in figure 11. Note that the basis function is equal to one at the filled dot and equal to zero at the open dots, which means that the gradient can simply be read off as the slope of the "tent" function along the sides of the triangles. With the aid of the gradients given in figure 11, we can

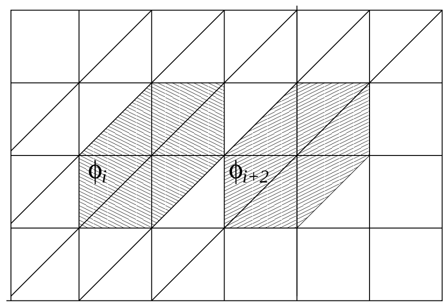

Figure 9: There is no overlap in the support for basis functions ϕ_i and ϕ_{i+2}.

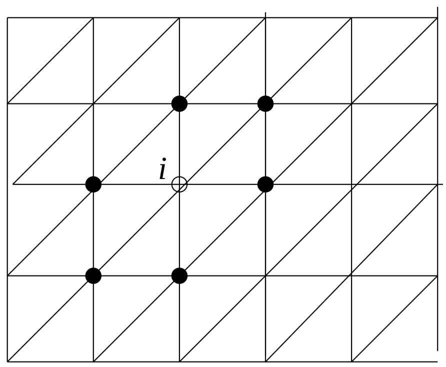

Figure 10: The nearest neighbors to node i are the six nodes marked with black dots. Thus, A_{ij} can be nonzero only when j corresponds to one of the black dots.

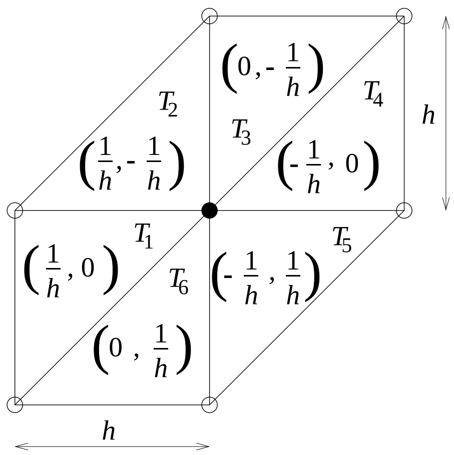

Figure 11: The gradient of basis function ϕ_i is piecewise constant on each triangle. The x- and y-coordinates are given as the pair $(\cdot, \cdot)$ at each triangle of the support of the function.

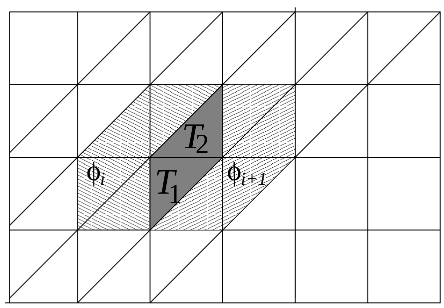

Figure 12: The overlap in the support of basis functions ϕ_i and ϕ_{i+1} are the triangles T_1 and T_2.

compute the diagonal elements in the stiffness matrix,

$$
\begin{aligned}
A_{ii} &= \int_\Omega \nabla\phi_i \cdot \nabla\phi_i \, d\Omega = \sum_{i=1}^{6} \int_{T_k} \nabla\phi_i \cdot \nabla\phi_i \, d\Omega \\
&= \frac{1}{h^2}|T_1| + 2\frac{1}{h^2}|T_2| + \frac{1}{h^2}|T_3| + \frac{1}{h^2}|T_4| + 2\frac{1}{h^2}|T_5| + \frac{1}{h^2}|T_6| \\
&= 8\frac{1}{h^2}\frac{h^2}{2} = 4.
\end{aligned}
$$

To compute $A_{i,i+1}$, note that $\nabla\phi_i \cdot \nabla\phi_{i+1} \neq 0$ only in two triangles (figure 12), thus

$$
\begin{aligned}
&\text{on } T_1: \quad \nabla\phi_i = \left(-\frac{1}{h}, \frac{1}{h}\right) \qquad \nabla\phi_{i+1} = \left(\frac{1}{h}, 0\right) \\
&\text{on } T_2: \quad \nabla\phi_i = \left(-\frac{1}{h}, 0\right) \qquad \nabla\phi_{i+1} = \left(\frac{1}{h}, -\frac{1}{h}\right)
\end{aligned}
$$

and thus

$$
\begin{aligned}
A_{i,i+1} &= \int_\Omega \nabla\phi_i \cdot \nabla\phi_{i+1} \, d\Omega = \sum_{k=1}^{2} \int_{T_k} \nabla\phi_i \cdot \nabla\phi_{i+1} \, d\Omega \\
&= -\frac{1}{h^2}|T_1| - \frac{1}{h^2}|T_2| = -\frac{2}{h^2}\frac{h^2}{2} = -1.
\end{aligned}
$$

Similar calculations yield that

$$
A_{i,i-1} = A_{i,i+J} = A_{i,i-J} = -1, \qquad A_{i,i+J+1} = A_{i,i-J-1} = 0.
$$

Also note that the matrix $\mathbf{A}$ is *symmetric*: $A_{ij} = A_{ji}$. Altogether, we obtain the *block triangular structure* (empty space means zeros!)

$$
\mathbf{A} = \begin{pmatrix}
\mathbf{T} & -\mathbf{I} & & & \\
-\mathbf{I} & \mathbf{T} & -\mathbf{I} & & \\
& \ddots & \ddots & \ddots & \\
& & -\mathbf{I} & \mathbf{T} & -\mathbf{I} \\
& & & -\mathbf{I} & \mathbf{T}
\end{pmatrix}
$$

where $\mathbf{T}$ and $\mathbf{I}$ are the J-by-J matrices

$$
\mathbf{T} = \begin{pmatrix}
4 & -1 & & & \\
-1 & 4 & -1 & & \\
& \ddots & \ddots & \ddots & \\
& & -1 & 4 & -1 \\
& & & -1 & 4
\end{pmatrix}, \quad
\mathbf{I} = \begin{pmatrix}
1 & & & & \\
& 1 & & & \\
& & \ddots & & \\
& & & 1 & \\
& & & & 1.
\end{pmatrix},
$$

Thus, the ith row of the matrix-vector product $\mathbf{Au}$ will be

$$
4u_i - u_{i+1} - u_{i-1} - u_{i+J} - u_{i-J}. \tag{25}
$$

Node $i+1$ and $i-1$ is located to the right and left, respectively, of node i, whereas nodes $i+J$ and $i-J$ are above and below node i. Thus, expression (25) is precisely the classical five-point, finite-difference formula. We reach the remarkable conclusion that the finite-element discretization of the Laplace operator using continuous, piecewise-linear functions on the structured mesh of figure 7 reduces to a standard finite-difference formula for the Laplacian. Note, however, that this does not hold in general; finite-element discretizations are not always easy to interpret as a finite-difference method.

1.7 Properties of the Stiffness Matrix

Consider the stiffness matrix $\mathbf{A}$ with components

$$A_{ij} = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j \, d\Omega,$$

which was obtained by discretizing the Poisson problem (1). This matrix has some very particular properties, which will be discussed in this section: it is *symmetric*, *positive definite*, *sparse*, and *ill conditioned*. All these properties, except the sparsity, reflects the nature of the boundary-value problem (1). Some or all of these properties may change if the equation or the boundary conditions are altered. For instance, if an additional term containing first derivatives of u is added to equation (1), the stiffness matrix will no longer be symmetric. The sparsity is a consequence of the fact that the chosen piecewise-linear approximation allows a representation in a compact basis, the "tent" functions of figure 6.

The symmetry of the matrix is immediate,

$$A_{ij} = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j \, d\Omega = \int_\Omega \nabla\phi_j \cdot \nabla\phi_i \, d\Omega = A_{ji}.$$

Moreover, the matrix is *sparse*, since $A_{ij} = 0$ whenever i and j are not nearest neighbors. The number of neighbors to each point does not increase when the mesh is made finer, as long as the mesh refinements are made in a sensible way, see the discussion in section 1.8. Thus, the number of nonzero elements on each row does not increase with the order of the stiffness matrix, that is, the matrix in a sense becomes sparser and sparser with increasing matrix order.

Recall that a real matrix $\mathbf{A}$ is *positive definite* if $\mathbf{v}^T\mathbf{A}\mathbf{v} > 0$ whenever $\mathbf{v} \neq 0$.

Theorem 3. *The stiffness matrix is positive definite.*

Proof. Let $v_h \in V_h$. Expanding v_h in the "tent" basis functions yields

$$v_h = \sum_{i=1}^{N} v_i \phi_i(\boldsymbol{x}).$$

Setting

$$\mathbf{v} = (v_1, v_2, \ldots, v_N)^T,$$

yields that

$$
\begin{aligned}
\mathbf{v}^T \mathbf{A} \mathbf{v} &= \sum_{i=1}^{N} \sum_{j=1}^{N} v_i \int_\Omega \nabla \phi_i \cdot \nabla \phi_j \, d\Omega v_j \\
&= \int_\Omega \underbrace{\sum_{i=1}^{N} \nabla (v_i \phi_i)}_{=\nabla v_h} \cdot \underbrace{\sum_{j=1}^{N} \nabla (v_j \phi_j)}_{=\nabla v_h} \, d\Omega = \int_\Omega |\nabla v_h|^2 \, d\Omega \geq 0,
\end{aligned}
\tag{26}
$$

with equality if and only if $\nabla v_h = 0$, that is, if v_h is constant. However, since v_h is zero on the boundary (by definition of V_h), it follows that the constant must be zero. Thus expression (26) is zero only if $v_h \equiv 0$, that is, when $\mathbf{v} = \mathbf{0}$. □

One important consequence of Theorem 3 is that equation (24) has a unique solution. This follows from the fact that positive-definite matrices are nonsingular: For a *singular* matrix $\mathbf{A}$, there would be nonzero vector $\mathbf{v}$ so that $\mathbf{Av} = \mathbf{0}$, and thus $\mathbf{v}^T \mathbf{A} \mathbf{v} = 0$. Thus, singular matrices cannot be positive definite, and positive-definite matrices must therefore be nonsingular.

The condition number of the stiffness matrix depends strongly on h. In fact, the growth of the condition number can be estimated to $\text{cond}(\mathbf{A}) = O(h^{-2})$ when h is reduced, provided that the the quotient between the size of the smallest and largest triangle in the mesh is kept bounded as the mesh is refined. The stiffness matrix is thus ill conditioned for fine meshes. However, in practical applications is the condition number typically not large enough to cause problematic amplification of round-off errors. On the other hand, the ill-conditioning is certainly an issue when applying iterative methods (section 5.4) for solving the linear system $\mathbf{Au} = \mathbf{b}$. Iterative techniques will typically converge slowly when applied to linear systems emanating from discretization of elliptic boundary-value problems (regardless of the method used to discretize the equations!). In general, particular techniques have to be used to speed up the convergence rate, so-called "preconditioning". There is also a strategy known as *multigrid* that *exploits* the fact that the matrix is ill conditioned to speed up the convergence rate. Using multigrid, large system of equations can be solved in a very efficient way.

1.8 Accuracy

We have shown how to define a finite-element approximation of the Poisson problem (1), and that this yields the linear system (24) having a unique solution. The question how good the finite-element solution is as an approximation of the original problem will be discussed in this section.

For finite-difference discretizations, accuracy questions are usually addressed indirectly through study of the local truncation error of the difference operators. Stability investigations provides a link between truncation error and error in the solution. The Lax–Richtmyer equivalence theorem (Theorem 4.2) provides this link for time dependent problems. Truncation errors are seldom studied for finite-element discretizations since it is is possible to study the error in the discretization directly. The easiest and most natural way is to work with *integral norms* of the difference between the weak solution u of problem (16) and the finite-element solution u_h of problem (21). The $L^2(\Omega)$ norm of a function,

$$\|v\|_{L^2(\Omega)} = \left(\int_\Omega v^2 \, d\Omega \right)^{1/2},$$

is the analogue for functions of the vector 2-norm. The perhaps most important norm for solutions of the Poisson problem is the *energy norm*

$$\|v\|_V = \left(\int_\Omega |\nabla v|^2 \, d\Omega \right)^{1/2}, \tag{27}$$

that is, the $L^2(\Omega)$-norm of the first derivatives; recall that weak solutions were defined among functions with bounded energy norm (definition (15)). The importance of the energy norm is that the finite-element solution is *optimal* in the energy norm. That is, no other function in V_h yields a smaller error in energy norm:

Theorem 4. *Let u be the solution to variational problem* (16) *and u_h the finite-element solution* (21). *Then*

$$\|u - u_h\|_V \le \|u - v_h\|_V \qquad \forall v_h \in V_h, \tag{28}$$

Proof. By equation (21), the finite-element solution u_h satisfies

$$\int_\Omega \nabla v_h \cdot \nabla u_h \, d\Omega = \int_\Omega v_h f \, d\Omega \qquad \forall v_h \in V_h. \tag{29}$$

From equation (16) follows that the weak solution u satisfies

$$\int_\Omega \nabla v_h \cdot \nabla u \, d\Omega = \int_\Omega v_h f \, d\Omega \qquad \forall v_h \in V_h, \tag{30}$$

since $V_h \subset V$. Subtracting equations (29) and (30) yields that

$$\int_\Omega \nabla v_h \cdot \nabla (u - u_h) \, d\Omega = 0 \qquad \forall v_h \in V_h. \tag{31}$$

Let v_h be an arbitrary element of V_h. Then

$$\begin{aligned}
\|u - u_h\|_V^2 &= \int_\Omega |\nabla(u - u_h)|^2 \, d\Omega = \int_\Omega [\nabla(u - u_h)] \cdot [\nabla(u - u_h)] \, d\Omega \\
&= \int_\Omega \nabla u \cdot \nabla(u - u_h) \, d\Omega - \underbrace{\int_\Omega \nabla u_h \cdot \nabla(u - u_h) \, d\Omega}_{= 0 \text{ by } (31)} \\
&= \int_\Omega \nabla u \cdot \nabla(u - u_h) \, d\Omega - \underbrace{\int_\Omega \nabla v_h \cdot \nabla(u - v_h) \, d\Omega}_{= 0 \text{ by } (31)} \\
&= \int_\Omega \nabla(u - v_h) \cdot \nabla(u - u_h) \, d\Omega \le \|u - v_h\|_V \|u - u_h\|_V,
\end{aligned} \tag{32}$$

where the last inequality follows from the Cauchy–Schwarz inequality. Dividing through with $\|u - u_h\|_V$ yields the conclusion. $\square$

The optimality property (28) does not hold for all elliptic boundary-value problems. For the finite-element solution to be optimal, it is necessary that the variational problem yields a *symmetric* stiffness matrix.

The next step in an analysis of the error is a pure approximation problem. Typically, one considers the *interpolant*, that is, a piecewise-linear function agreeing with u at the node points; note that the interpolant is an element of V_h. The difference between the interpolant and u can be estimated by a type of Taylor expansion. From Theorem 4 it follows that the error in the finite-element solution is smaller or equal to the error in the interpolant. The precise magnitude of this error depends of course on how fine the mesh is, but it also depends on the *quality* of the mesh. Loosely speaking, one should try to avoid very thin triangles.

Altogether, estimating the interpolation error and utilizing Theorem 4, it can be shown that the error in the finite-element solution is of *second order*, that is,

$$\|u_h - u\|_{L^2(\Omega)} = O(h^2). \tag{33}$$

Note that the norm above is not the energy norm; the error is of *first order* if measured in the energy norm. For estimate (33) to hold, assumptions have to be made on the mesh quality and on the smoothness of the solution to the variational problem (16):

(i) (Mesh quality.) The largest angle in any of the triangles should not approach 180° as the mesh is refined. In particular, this means that no triangle successively can become infinitely thin.

(ii) (Smoothness.) The solution needs to be smooth, otherwise the convergence rate will be reduced. Smooth solutions are obtained if f and the boundary Ω are smooth. The solution is also smooth if the boundary is polygonal as long as the domain is *convex*. (If Ω is not polygonal to start with, it is typically approximated with a succession of polygonal domains Ω_h such that $\Omega_h \to \Omega$ as $h \to 0$).

The mesh quality condition above is maintained if the triangles, as the mesh is refined, are subdivided into four triangles in the way indicated in figure 13. Refining each triangle in the mesh in this way reduces all triangular sides with a factor 1/2. The error will thus be reduced with a factor 1/4 (for problems on convex domains at least). Nonsmooth boundaries may cause nonsmooth solutions and a reduced convergence rate. In particular, so-called reentrant corners in the domain, as in figure 14, will cause the convergence rate to be less than second order.

Higher accuracy can thus be obtained through refinement of the mesh ("h method"). This should preferably be done *adaptively*, in the parts of the domain where it is needed, to prevent the size of the stiffness matrix to become too large. There are automatic methods for this. Higher accuracy can also be obtained

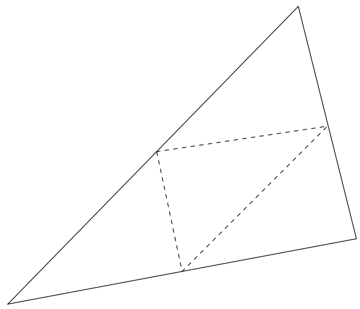

Figure 13: A strategy to maintain mesh quality is to subdivide each triangle into four new triangles by joining the edge midpoints.

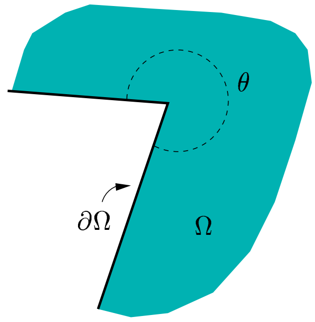

Figure 14: Reentrant corners, for instance polygonal boundaries with an angle $\theta > 180°$ will cause a nonsmooth solution and a convergence rate which is less than second order.

by keeping the mesh fixed and increasing the order of the polynomials on each triangle ("p method").For instance, the error in the sense (33) can be improved to *third order* if V_h consists of continuous functions that are *quadratic* on each element.

1.9 Alternative Elements

Quadrilaterals, that is, a geometric figure obtained by connecting four different points in the plane by straight lines that do not cross, can be used to partition the domain instead of triangles, see figure 15. In this case will the approximating space V_h contain globally continuous functions who vary linearly along the edges of each quadrilateral. However, the functions will no longer be linear *within* the elements. In the special case when the quadrilaterals are rectangles oriented in the coordinate directions, a function $v_h \in V_h$ will be *bilinear*, that is, of the form

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

on each element. The nodal values of v_h (the values of v_h at the four corners of the rectangle) uniquely determine the four coefficients above.

Quadrilateral and, in particular, rectangular elements yields a regular structure that may give high solution accuracy and allow efficient solutions of the associated linear systems. It is, however, harder to generate such meshes automatically on complicated geometries compared to triangular meshes.

For three space dimensions, triangular and quadrilateral meshes generalize to *tetrahedral* and *hexahedral* meshes (figure 16) with advantages and limitations as for corresponding meshes in two space dimensions.

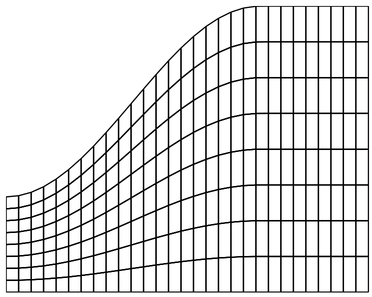

Figure 15: A quadrilateral mesh.

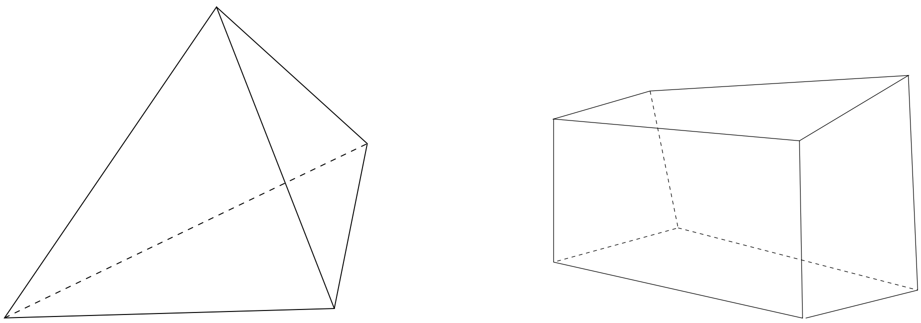

Figure 16: Meshes in three space dimensions can be composed of nonoverlapping tetrahedrals (left) or hexahedrals (right).

Exercises

5.17 For following boundary-value problems, derive weak formulations, define a FE approximation using continuous, piecewise-linear functions on a uniform grid, and specify the linear system associated with the FE approximation.

(a)

$$\begin{aligned} -u'' &= f \qquad \text{in } (0,1), \\ u(0) &= 0, \\ u'(1) &= 0. \end{aligned}$$

(b)

$$\begin{aligned} -u'' &= f \qquad \text{in } (0,1), \\ u(0) &= g, \\ u(1) &= 0. \end{aligned}$$

(c)

$$\begin{aligned} -u'' + au' &= f \qquad \text{in } (0,1), \\ u(0) = u(1) &= 0. \end{aligned}$$

(d)

$$\begin{aligned} -u'' + u &= f \qquad \text{in } (0,1), \\ u'(0) = u'(1) &= 0. \end{aligned} \tag{34}$$

(e)

$$\begin{aligned} -(c(x)u')' &= f \qquad \text{in } (0,1), \\ u(0) = u(1) &= 0, \end{aligned}$$

where $c(x) > 0$ on $[0,1]$.

5.18 Assume that f in equation (34) is a function in the space of approximations, that is,

$$f(x) = \sum_{i=1}^{I} f_i \phi_i(x),$$

where the ϕ_i's are the standard "hat" functions..

(a) Determine the *mass matrix* $\mathbf{M}$ such that the linear system associated with the FE approximation of equation (34) can be written

$$\mathbf{K}\mathbf{u} = \mathbf{M}\mathbf{f},$$

where $\mathbf{f} = (f_1, f_2, \ldots, f_I)^T$.

(b) When computing the mass matrix, use the *trapezoidal rule* to evaluate the integrals involved and compare with above.

5.19 Give a reason why the following boundary-value problem is not well posed in general:

$$\begin{aligned} -u'' &= f \qquad\qquad \text{in } (0,1), \\ u'(0) &= u'(1) = 0. \end{aligned}$$

What happens if a FE discretization is applied and one tries to solve the associated linear system?

5.20 Calculate expressions for the *element stiffness matrix*

$$A^k_{ij} = \int_{T_k} \nabla\phi_i \cdot \nabla\phi_j \, d\Omega$$

and the *element load vector*

$$f^k_i = \int_{T_k} f\phi_i \, d\Omega$$

associated with a generic triangle T_h spanned by the corner points $\boldsymbol{x}_k$, $\boldsymbol{x}_l$, and $\boldsymbol{x}_m$, as in figure 1. The basis functions are the standard "tent" functions for continuous, piecewise-linear functions on a triangular mesh. For the element load vector, use the following quadrature rule:

$$\int_{T_k} g \, d\Omega \approx \frac{g(\boldsymbol{x}_k) + g(\boldsymbol{x}_l) + g(\boldsymbol{x}_m)}{3} \, \text{area}\,(T_k) \qquad \text{(trapezoidal rule)},$$

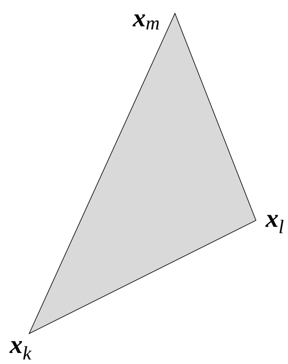

Figure 1. A triangle spanned by the points $\boldsymbol{x}_k$, $\boldsymbol{x}_l$, and $\boldsymbol{x}_m$.

5.21 Consider a finite-element approximation of the boundary-value problem

$$\begin{aligned} -u'' &= f \qquad\qquad \text{in } (0,1), \\ u(0) &= u(1) = 0, \end{aligned}$$

using continuous, piecewise-quadratic functions on a uniform grid. A nodal basis consists of values at the grid points x_i together with the mid-points $x_{i+1/2} = (x_i + x_{i+1})/2$.

(a) Specify and sketch the basis functions.

(b) Specify the sparsity pattern of the stiffness matrix.

5.22 Let Ω be a open, bounded, and connected domain in the plane with boundary $\partial\Omega$. Consider the boundary-value problem

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega, \\ au + \frac{\partial u}{\partial n} &= ag && \text{on } \partial\Omega, \end{aligned}$$

where g is a given function defined on $\partial\Omega$ and $a > 0$. Derive a weak formulation of the problem and define a FE approximation. What happens when a becomes large? Can $a = 0$ be allowed?