Notes on: "The Finite Element Method for Problems in Physics"

Ruiheng Su[a]

[a] *Engineering Physics, UBC*

Contents

1 Linear, elliptic partial differential equations in one dimension. Elasticity, heat conduction and mass diffusion 8

1.1 Linear elliptic partial differential equations - I 8

1.2 Linear elliptic partial differential equations - II 8

1.3 Boundary Conditions 9

1.4 Constitutive relations 10

1.5 Strong form of the partial differential equation. Analytic solution 11

1.6 Weak form of the partial differential equation - I 12

1.7 Weak form of the partial differential equation - II 13

1.8 Equivalence between the strong and weak forms 15

2 Approximation. The finite-dimensional weak form 17

2.1 The Galerkin, or finite-dimensional weak form 17

2.2 Basic Hilbert Spaces - I 18

2.3 Basic Hilbert spaces - II 19

2.4 The finite element method for the one-dimensional, linear, elliptic partial differential equation 19

2.5 Basis Functions - I 22

2.6 Basis functions - II 22

2.7 The bi-unit domain - I 24

2.8 The bi-unit domain - II 25

2.9 The finite dimensional weak form as sum over element subdomains - I 26

2.10 The finite dimensional weak form as a sum over element subdomains - II 27

3 Linear algebra; the matrix-vector form 29

3.1 The matrix-vector weak form - I - I 29

3.2 The matrix-vector weak form - I - II 30

3.3 The matrix-vector weak form - II - I 31

3.4 The matrix-vector weak form - II - II 32
3.5 The matrix-vector weak form - III - I 33
3.6 The matrix-vector weak form - III - II 34
3.7 The final finite element equations in matrix-vector form - I . . 36
3.8 The final finite element equations in matrix-vector form - II . 37

4 More on boundary conditions; basis functions; numerics 41
4.1 The pure Dirichlet problem - I 41
4.2 The pure Dirichlet problem - II 42
4.3 Higher polynomial order basis functions - I 44
4.4 Higher polynomial order basis functions - I - II 46
4.5 Higher polynomial order basis functions - II - I 47
4.6 Higher polynomial order basis functions - III 49
4.7 The matrix-vector equations for quadratic basis functions - I - I 51
4.8 The matrix-vector equations for quadratic basis functions - I - II 53
4.9 The matrix-vector equation for quadratic basis functions . . . 54
4.10 The matrix-vector equations for quadratic basis functions - II - II 56
4.11 Numerical Integration – Gaussian Quadrature 59

5 Analysis of the finite element method 62
5.1 Norms - I 62
5.2 Norms - II 63
5.3 Consistency of the finite element method 65
5.4 The best approximation property 66
5.5 The "Pythagorean Theorem" 68
5.6 Sobolev estimates and convergence of the finite element method 68
5.7 Finite element error estimate 70

6 Variational principles 72
6.1 Functionals. Free energy - I 72

6.2 Functionals. Free energy - II . . . 73
6.3 Extremization of functionals . . . 73
6.4 Derivation of the weak form using a variation principle . . . 74

7 Linear, elliptic partial differential equations for a scalar variable in three dimensions. Heat conduction and mass diffusion at steady state 76

7.1 The strong form of steady state heat conduction and mass diffusion - I . . . 76
7.2 The strong form of steady state heat conduction and mass diffusion - II . . . 77
7.3 The strong form, continued . . . 79
7.4 The weak form . . . 81
7.5 The finite-dimensional weak form - I . . . 83
7.6 The finite-dimensional weak form - II . . . 83
7.7 Three-dimensional hexahedral finite elements . . . 85
7.8 Aside: Insight to the basis functions by considering the two-dimensional case . . . 87
7.9 Field derivatives. The Jacobian - I . . . 88
7.10 Field derivatives. The Jacobian - II . . . 89
7.11 The integrals in terms of degrees of freedom . . . 91
7.12 The integrals in terms of degrees of freedom - continued . . . 92
7.13 The matrix-vector weak form - I . . . 94
7.14 The matrix-vector weak form II . . . 95
7.15 The matrix-vector weak form, continued - I . . . 96
7.16 The matrix-vector weak form, continued - II . . . 97
7.17 The matrix vector weak form, continued further - I . . . 98
7.18 The matrix-vector weak form, continued further - II . . . 100

8 Lagrange basis functions and numerical quadrature in 1 through 3 dimensions 103

8.1 Lagrange basis functions in 1 through 3 dimensions - I . . . 103
8.2 Quadrature rules in 1 through 3 dimensions . . . 104

8.3 Triangular and tetrahedral elements - Linears - I 105
8.4 Triangular and tetrahedral elements - Linears - II 106

9 Linear, elliptic, partial differential equations for a scalar variable in two dimensions 107
9.1 The finite-dimensional weak form and basis functions - I . . . 107
9.2 The finite-dimensional weak form and basis functions - II . . 108
9.3 The matrix-vector weak form 111
9.4 The matrix-vector weak form - II 112

10 Linear, elliptic partial differential equations for vector unknowns in three dimensions (Linearized elasticity) 113
10.1 The strong form of linearized elasticity in three dimensions - I 113
10.2 The strong form of linearized elasticity in three dimensions - II113
10.3 The strong form, continued 114
10.4 The constitutive relations of linearized elasticity 117
10.5 The weak form - I 119
10.6 The finite-dimensional weak form - Basis functions - I 123
10.7 The finite-dimensional weak form - Basis functions - II 125
10.8 Element integrals - I 126
10.9 Element integrals - II 127
10.10 The matrix-vector weak form - I 128
10.11 The matrix-vector weak form - II 130
10.12 Assembly of the global matrix-vector equations - I 131
10.13 Assembly of the global matrix-vector equations - II 133
10.14 Dirichlet boundary conditions - II 136

11 Linear, parabolic partial differential equations for a scalar unknown in three dimensions (Unsteady heat conduction and mass diffusion) 139
11.1 The strong form 139
11.2 The weak form, and finite-dimensional weak form - I 140
11.3 The weak form, and finite-dimensional weak form - II 142

11.4 Basis functions, and the matrix-vector weak form - I 143
11.5 Basis functions, and the matrix-vector weak form - II 145
11.6 Dirichlet boundary conditions; the final matrix-vector equations 145
11.7 Time discretization; the Euler family - I 147
11.8 Time discretization; the Euler family - II 148
11.9 The v-form and d-form 149
11.10 Analysis of the integration algorithms for first order, parabolic equations; modal decomposition - I 151
11.11 Analysis of the integration algorithms for first order, parabolic equations; modal decomposition - II 152
11.12 Modal decomposition and modal equations - I 154
11.13 Modal decomposition and modal equations - II 155
11.14 Modal equations and stability of the time-exact single degree of freedom systems - I 156
11.15 Modal equations and stability of the time-exact single degree of freedom systems - II 157
11.16 Stability of the time-discrete single degree of freedom systems 159
11.17 Behaviour of higher-order modes; consistency - I 161
11.18 Behaviour of higher-order modes; consistency - II 162
11.19 Convergence - I 165
11.20 Convergence - II 167

12 Linear, hyperbolic partial differential equations for a vector unknown in three dimensions (Linear elastodynamics) 169
12.1 The strong and weak forms 169
12.2 The finite-dimensional and matrix-vector weak forms - I . . . 171
12.3 The finite-dimensional and matrix-vector weak forms - II . . 172
12.4 The time-discretized equations 173
12.5 Stability - I 176
12.6 Stability - II 178
12.7 Behaviour of higher-order modes 179
12.8 Convergence 181

Course Description

"This course is an introduction to the finite element method as applicable to a range of problems in physics and engineering sciences. The treatment is mathematical, but only for the purpose of clarifying the formulation. The emphasis is on coding up the formulations in a modern, open-source environment that can be expanded to other applications, subsequently.

"The course includes about 45 hours of lectures covering the material I normally teach in an introductory graduate class at University of Michigan. The treatment is mathematical, which is natural for a topic whose roots lie deep in functional analysis and variational calculus. It is not formal, however, because the main goal of these lectures is to turn the viewer into a competent developer of finite element code. We do spend time in rudimentary functional analysis, and variational calculus, but this is only to highlight the mathematical basis for the methods, which in turn explains why they work so well.

"Much of the success of the Finite Element Method as a computational framework lies in the rigor of its mathematical foundation, and this needs to be appreciated, even if only in the elementary manner presented here.

"A background in PDEs and, more importantly, linear algebra, is assumed, although the viewer will find that we develop all the relevant ideas that are needed. The development itself focuses on the classical forms of partial differential equations (PDEs): elliptic, parabolic and hyperbolic. At each stage, however, we make numerous connections to the physical phenomena represented by the PDEs.

"For clarity we begin with elliptic PDEs in one dimension (linearized elasticity, steady state heat conduction and mass diffusion). We then move on to three dimensional elliptic PDEs in scalar unknowns (heat conduction and mass diffusion), before ending the treatment of elliptic PDEs with three dimensional problems in vector unknowns (linearized elasticity).

"Parabolic PDEs in three dimensions come next (unsteady heat conduction and mass diffusion), and the lectures end with hyperbolic PDEs in three dimensions (linear elastodynamics).

"Interspersed among the lectures are responses to questions that arose

from a small group of graduate students and post-doctoral scholars who followed the lectures live. At suitable points in the lectures, we interrupt the mathematical development to lay out the code framework, which is entirely open source, and C++ based.

"Books: There are many books on finite element methods. This class does not have a required textbook. However, we do recommend the following books for more detailed and broader treatments than can be provided in any form of class: The Finite Element Method: Linear Static and Dynamic Finite Element Analysis, T.J.R. Hughes, Dover Publications, 2000. The Finite Element Method: Its Basis and Fundamentals, O.C. Zienkiewicz, R.L. Taylor and J.Z. Zhu, Butterworth-Heinemann, 2005. A First Course in Finite Elements, J. Fish and T. Belytschko, Wiley, 2007.

"Resources: You can download the deal.ii library at dealii.org. The lectures include coding tutorials where we list other resources that you can use if you are unable to install deal.ii on your own computer. You will need cmake to run deal.ii. It is available at cmake.org."

1 Linear, elliptic partial differential equations in one dimension. Elasticity, heat conduction and mass diffusion

1.1 Linear elliptic partial differential equations - I

Linear elliptic PDEs have the from

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y)$$

Some examples include Laplace's equation and Poisson's equation. Linear elliptic PDEs can describe heat transfer, mass diffusion, and elasticity at steady state.

Consider a 1D beam of length L. At $x = L$, we might specify the following

- u_g: displacement of the rod at $x = L$
- t: "traction" at $x = L$

And acting on every single mass element along the body of the beam might be a distributed body force, f.

Our goal is to find the displacement field, $u(x)$ (or the displacement of every point on the rod as a function of position).

1.2 Linear elliptic partial differential equations - II

The displacement field u is a function of position only. It is not a function of time since we are concerned with elliptic PDEs, which describe elasticity at steady state.

Think of u as a mapping from $(0, L)$ into $\mathbb{R}^1$ which is the 1-D space of real numbers. We will say that this is an open interval. We will give more explanation of this in a couple lessons.

Given $u(0) = u_0$, u_g or t, $f(x)$, and the constitutive relation: $\sigma = Eu_{,x}$ ($u_{,x}$ denotes a spatial derivative on u), we want u such that the following holds

$$\frac{d\sigma}{dx} + f = 0 \in (0, L)$$

This is our linear elliptic PDE.

We want this equation to satisfy the boundary conditions

$$\begin{cases} u(0) = u_0 \\ u(L) = u_g \quad \text{or} \quad \sigma(L) = t \end{cases}$$

We might be given one of the following boundary conditions:

$$\text{Dirichlet boundary conditions } \begin{cases} u(0) = u_0 \\ u(L) = u_g \end{cases}$$

$$\text{Neumann boundary condition } \left\{ \sigma(L) = t \implies Eu_{,x}|_{x=L} = t \right.$$

Dirichlet Boundary Conditions: conditions applied to the primal field. The primal field in our case is u. But it is in general the field for which we are posing the problem.

Neumann Boundary Condition: conditions applied to the spatial derivative of the primal field. Though this is not always true when we get to higher order problems.

1.3 Boundary Conditions

In the current problem, we will always have a Dirichlet boundary condition at $x = 0$. There is a choice of a Neumann or Dirichlet boundary condition at $x = L$. We discard the case that there are Neumann boundary conditions on both ends.

If we had specified the Neumann boundary conditions for both ends $x = 0, L$.

$$\begin{cases} \sigma(0) = Eu_{,x}|_{x=0} = t_0 \\ \sigma(L) = Eu_{,x}|_{x=L} = t_L \end{cases}$$

Consider $u(x)$ these conditions and the differential equation.

$$\frac{d\sigma}{dx} + f = 0 \qquad \forall x \in (0, L)$$

From the constitutive relation we get that

$$\frac{d}{dx} Eu_{,x} + f = 0$$

The solution that we get from solving this equation will not be unique up to a constant displacement field. (If we solved for $u(x)$, $u(x) + \overline{u}$ is also a solution to the equation.) The constant displacement field is also called a rigid body motion.

So we must specify the displacement of the bar at one end to make sure that our solution is unique. This seems to rule out the problem where a bar is tossed along the x axis. But such a problem is no longer at steady state or elliptic. This will become a hyperbolic problem.

Recall the body force, $f(x)$. This is the forcing function in general PDEs. An example of a body force on a 1-D bar might be

$$f(x) = \rho(x)g$$

where ρ is the linear mass density.

1.4 Constitutive relations

We have excluded the end points of our rod from being a subject of the differential equation. This is because we have already specified boundary conditions. Having differential equation also apply their will result in a over-constrained system.

The constitutive equation we have given relatives stress to stain. $u_{,x}$ is sometimes also given the symbol ϵ.

We can also state the same 1-D scalar elliptic problem has a problem of heat or diffusive mass transport.

In a case of heat transport, the primal field we want is a temperature as a function of position. The field remains as a mapping from $(0, L) \to \mathbb{R}^1$, given u_0 and u_g or $\bar{j}$. The constitutive relation is is $j = -\kappa u_{,x}$, such that

$$-\frac{dj}{dx} = f(x) \in (0, L)$$

Here, j represents a heat flux, and $-dj/dx$ is the negative of the divergence of heat flux.

The Dirichlet condition specified the the temperature at the end of the rod, while the Neumann condition specifies the the heat flux.

1.5 Strong form of the partial differential equation. Analytic solution

Recall the 1-D elasticity problem at steady state. If we substitute the the constitutive relation into our ODE, we find that the displacement field we are solving for must be twice differentiable.

We will encounter difficulties in taking derivatives when we have delta functions in the displacement field.

We require a "strong condition" of "smoothness" on the field because the strong form has two spatial derivatives. We also require PDE to hold point wise over the domain.

To obtain an analytical solution, we can simply integrate the ODE.

$$\begin{aligned} -\int_0^y f\,dx = \int_0^y \frac{d\sigma}{dx}\,dx = \\ = \sigma(y) - \sigma(0) \\ = Eu_{,x}|_y - Eu_{,x}|_0 \end{aligned}$$

Let's write this as

$$E\frac{du}{dy} = -\int_0^y f\,dx + \left[E\frac{du}{dx}\right]_{x=0}$$

This is simply an abuse of notation. y is a dummy variable.

Let's integrate once again.

$$\int_0^z E\frac{du}{dy}\,dy = -\int_0^z \left[\int_0^y f\,dx\right] dy + \int_0^z E\left[\frac{du}{dx}\right]_0 dy$$

In the case that E is a constant, we can follow through with the integration.

$$Eu(z) - Eu(0) = -\int_0^z \left[\int_0^y f\,dx\right] dy + E\left[\frac{du}{dx}\right]_0 z$$

Rearranging gives

$$u(z) = \frac{1}{E}\left[-\int_0^z \left[\int_0^y f\,dx\right] dy + \underbrace{E\left[\frac{du}{dx}\right]_0 z}_{\text{applying b.c. at } z=L} + E\underbrace{u(0)}_{u_g}\right]$$

This method of solution is very limited. It can easily get more complex as we increase the dimension of the problem. We can also have complex f and E functions. We can also imagine that the domain of solution will also be rather complicated.

So we want to develop a method of solving general PDEs using data.

1.6 Weak form of the partial differential equation - I

One approach to finding approximate solutions to PDEs is to replace derivatives with finite difference quotients.

The finite difference method is fundamentally different from the finite element approach. The first step is to obtain a weak form of PDE.

We will look at the weak form of a linear elliptic PDE in one dimension.

We want to find $u(x)$ that belongs to a space of functions $\mathcal{S}$. This space of functions is all u such that $u(0) = u_0$.

$$u \in \mathcal{S}$$

where

$$\mathcal{S} = \left\{ u | \underbrace{u(0) = u_0}_{\text{Dirichlet}} \right\}$$

given u_0, t, $f(x)$, and constitutive relation $\sigma = Eu_{,x}$, such that $\forall\ w \in \mathcal{V}$. We are only considering the case of a single Dirichlet boundary condition here.

$$\mathcal{V} = \left\{ w | \underbrace{w(0) = 0}_{\text{homogeneous Dirichlet boundary condition}} \right\}$$

such that the following holds

$$\int_0^L w_{,x}\sigma \, dx = \int_0^L wf \, dx + w(L)t$$

this is the weak form.

The left hand integral can be thought of as some volume in 1-D, when we multiply everything by the cross sectional area A.

$$\int_0^L w_{,x}\sigma A \, dx = \int_0^L wfA \, dx + w(L)At$$

$A\,dx$ is then a volume element, and tA is some boundary force.

1.7 Weak form of the partial differential equation - II

The weak form is the basis of the Finite Element Method and of other *variationally* based numerical methods.

Our claim is that the strong form and the weak form are equivalent.

Strong ⇔ Weak

Consider the strong form

$$\frac{d\sigma}{dx} + f = 0 \in (0, L)$$

with the boundary conditions

$$\begin{cases} u(0) = u_0 \\ \sigma(L) = t \end{cases}$$

where $\sigma = Eu_{,x}$.

To show that this is equivalent to the weak form, we will introduce function w seen in the weak form. In our context, w is called the weighting function. We will multiply w into our strong form and integrate over the domain. We will also multiply by A.

$$\int_0^L w\frac{d\sigma}{dx}A\,dx + \int_0^L wAf\,dx = 0$$

Next, we will integrate by parts. Let $u = w(x)$, $du/dx = w_{,x}$. Let $dv/dx = d\sigma/dx$, $v = \sigma$. So

$$w\sigma A|_0^L - \int_0^L w_{,x}\sigma A\,dx + \int_0^L wAf\,dx = 0$$

Rearranging,

$$\int_0^L w_{,x}\sigma A\,dx = \int_0^L wfA\,dx + w(L)\sigma(L)A - w(0)\sigma(0)A$$

Since w is a function of space $\mathcal{V}$, and we known that $\sigma(L) = t$, we recover our weak form

$$\int_0^L w_{,x}\sigma A\,dx = \int_0^L wfA\,dx + w(L)tA$$

1.8 Equivalence between the strong and weak forms

In the previous step, we demonstrated the equivalence between the strong and weak form of a linear elliptic PDE by converting from the strong form to the weak.

Here, we will learn to go from the weak form to the strong from.

The weak form is such that we want to find a function u belong to $\mathcal{S}$

$$u \in \mathcal{S} = \{u | u(0) = u_0\}$$

such that for all function w belonging to V, which satisfies a homogeneous Dirichlet boundary condition

$$w \in \mathcal{V} = \{w | w(0) = 0\}$$

the following equation holds

$$\int_0^L w_{,x} \sigma A \, dx = \int_0^L wAf \, dx + w(L)\sigma(L)A$$

We will take a reverse approach. Integration by parts transfers the integral from one field to an other.

$$\underbrace{\int_0^L w_{,x} \sigma A \, dx}_{\text{integrate by parts}}$$

Let $w_{,x}$ be dv/dx, so $v = w$. And let $\sigma = u$, $du/dx = \sigma_{,x}$. So

$$-\int_0^L w\sigma_{,x} A \, dx + w\sigma A|_0^L = \int_0^L wAf \, dx + w(L)tA$$

then,

$$-\int_0^L w\sigma_{,x} A \, dx + w(L)\sigma(L)A - w(0)\sigma(0)A = \int_0^L wAf \, dx + w(L)tA$$

Since w belongs to the space of functions that satisfy homogeneous Dirichlet boundary conditions, then, we have

$$-\int_0^L w\sigma_{,x} A \, dx + w(L)\sigma(L)A = \int_0^L wAf \, dx + w(L)tA$$

Rearrange and combine,

$$\int_0^L w(-\sigma_{,x} - f)A\,dx + w(L)A(\sigma(L) - t) = 0$$

To get to this point, we have use the fact that $w \in \mathcal{V}$.

We will claim that since the equation holds for $w \in \mathcal{V}$ the following special case of w also holds. We say $w(x) = \phi(x)(-\sigma_{,x} - f)$, where $\phi(x) > 0$ for $x \in (0, L)$ and $\phi(x) = 0$ at $x = \{0, L\}$. (In words, ϕ is a function that satisfies homogeneous Dirichlet boundary conditions at both $x = 0$ and $x = L$, and is greater than 0 for any point in between 0 and L.)

So assuming this special form of w we have

$$\begin{aligned} 0 &= \int_0^L w(-\sigma_{,x} - f)A\,dx \\ &= \int_0^L \phi(x)(-\sigma_{,x} - f)^2 A\,dx \end{aligned}$$

Let's examine our equation at this point. ϕ is as defined to be greater than 0 over the interior of the bounds of integration. And we have a square term, will will also be greater than or equal to zero.

If the cross sectional area A is nonzero, then it must be that $-\sigma_{,x} - f = 0 \in (0, L)$. This is exactly our PDE in strong form.

Let's return to our equation. By the PDE, we have that

$$w(L)A(\sigma(L) - t) = 0$$

This must also hold for a particular w such that $w(0) = 0$ and $w(L) \neq 0$. This can only be if $\sigma(L) - t = 0$. Which is our Neumann boundary condition of the strong form.

Here, we have shown that since the weak form has to hold for all w that belongs to the space $\mathcal{V}$ it must also hold for certain special choices of w.

From these special choices, we demonstrated that the PDE must hold, and the Neumann boundary condition must hold as well. There is also has Dirichlet boundary condition on u, since $u \in \mathcal{S}$.

2 Approximation. The finite-dimensional weak form

2.1 The Galerkin, or finite-dimensional weak form

In the finite element approach, we will apply approximations to the weak form of the equation. (When we apply approximations to the strong form, we are heading in finite differences direction.)

When we said that $u \in \mathcal{S}$ and $w \in \mathcal{V}$, the spaces $\mathcal{S}$ and $\mathcal{V}$ are infinite dimensional. For example, when $\mathcal{S}$ and $\mathcal{V}$ belong to the space of polynomials, we are considering polynomials of any order.

The difficulty when solving for solutions of the exact statements of the problem is that we are looking for solutions in a very large space. Approximation methods reduces the difficulty in finding a solution by restricting the dimensionality of the space in which we are looking for solutions.

Let's formalize this. We will restrict the solution space and weighting function space.

Finite Dimensional or Galerkin Weak Form: We want to find $u^h(x) \in \mathcal{S}^h \subset \mathcal{S}$, where $\mathcal{S}^h = \left\{u^h \in H^1(0,L) | u^h(0) = u_0\right\}$ ("S sup h contains all functions u sup h, which belongs in a space H one on $(0,L)$"), such for all $w^h \in \mathcal{V}^h \subset \mathcal{V}$, where $\mathcal{V}^h = \{w^h \in H^1(0,L) | w^h(0) = 0\}$ the following holds

$$\int_0^L w^h_{,x} \sigma^h A \, dx = \int_0^L w^L f A \, dx + w^h(L) t A$$

The superscript h denotes that the field is finite dimensional.

This finite dimensional weak form is not equivalent to the strong form in general. When we proved the equivalence of the weak form to the strong form, we use the fact that the weak form holds for all $w \in \mathcal{V}$. Which allowed us to say that since it holds for all $w \in \mathcal{V}$ it must also hold for special classes of w, which then applied the strong form had to hold. But since we are now

restricting the solution and weighting function space to be finite dimensional, we have lost the ability to invoke the argument that the weak form holds for a sufficiently large space of functions.

If it turns out that the solution lives in the smaller space, then the finite dimensional weak form is equivalent to the infinite dimensional strong form.

2.2 Basic Hilbert Spaces - I

We will explore the the function spaces from which we are drawing the finite dimensional functions.

Recall $u^h \in \mathcal{S} = \left\{u^h \in H^1(0,L) | u^h(0) = u_0\right\}$. We will look at the idea behind these function spaces.

Consider a function $v : (0,L) \mapsto \mathbb{R}$. We define the function v to be an L^2 function if

$$\int_0^L v^2(x)\, dx < \infty$$

So $v \in L^2(0,L)$. The function integrated over the domain is square integrable, or bounded. Some examples of such a function v are: a constant, a polynomial, the Heaviside function.

Something that is not L^2 might be $v(x) = \delta(x - x_0)$. One can in general define L^p functions, $p \in \mathbb{R}$.

How about control over the derivatives of v? We say that $v \in H^1(0,L)$ if

$$\int_0^L (v^2 + L^2 v_{,x}^2)\, dx < \infty$$

We introduced the L^2 factor for the purpose of keeping the dimensions the same within the integral.

In general, we use $m(0,L)^{1/d} = L$, which is the measure of our domain. In one dimension, $d = 1$, so the measure is simply the length. Likewise, in three dimensions, the measure is volume, but volume to the power of $1/3$ would give you a notion of length. In $\mathbb{R}^3$, $d = 3$ and the measure of our domain Ω is denoted as $m(\Omega)$.

2.3 Basic Hilbert spaces - II

Using the notion of measures, we say that a function $v \in H^1(0, L)$ if

$$\int_0^L \left(v^2 + (m(0, L))^2 v_{,x}^2\right) \, dx < \infty$$

We restrict v to be a H^1 function since we want it self and it's first derivative be bounded. If we want to control it's higher derivatives too, we would go to higher orders of H space.

Now, when we say $u^h \in \mathcal{S}^h = \{u^h \in H^1(0, L) | u^h(0) = u_0\}$, we are saying here that we expect both the function and its derivative to be square integrable.

Some examples of $H^1(0, L)$ functions are

- constant
- $\sum_{k=0}^{N} a_k x^k$
- A special function that is first linear, then quadratic, then linear, then some polynomial, then linear, then constant. Though there are jumps in the first derivative of this function, it is still square integrable.

There is only a problem when we are trying to require a a function that has discontinuities it self be of H^1. The first derivative of such functions will be infinite, and the square integrability of of first derivative breaks down.

Both the space of square integrable functions and H^n are what are known as Hilbert spaces.

2.4 The finite element method for the one-dimensional, linear, elliptic partial differential equation

We used the idea of function spaces as it is useful to express what is meant by having control over a function and its derivatives.

With this background, we will now introduce the finite element method for the 1-D linear elliptic PDE.

Recall the finite dimensional weak form. We want to find u^h belonging to $\mathcal{S}^h$ which is equal to functions u^h that belong to H^1 on $(0, L)$ such that they satisfy our Dirichlet boundary condition, (we often call u^h trial function), such that for all w^h that belongs in $\mathcal{V}^h$, which equals all functions w^h that belong to H^1 on $(0, L)$, and satisfy homogeneous Dirichlet boundary conditions, the following holds

$$\int_0^L w^h_{,x} \sigma^h A \, dx = \int_0^L w^h f A \, dx + w^h(L) t A$$

In our formulation, we have written σ^h. This means we are saying that σ^h will be obtained from $E u^h_{,x}$ (the gradient of the finite dimensional trial solution). But f will be given as data. We will not approximate the data. So f is not finite dimensional. The traction t is will be a point value. There isn't a question of whether we will approximate it or write it as finite dimensional. This will become something we have to think about as we go to higher dimension though.

How do we obtain these finite dimensional functions u^h and w^h? Or alternatively what are the spaces $\mathcal{S}^h$ and $\mathcal{V}^h$.

The trick is to partition $\Omega = (0, L)$ into *finite elements*, which are disjoint subdomains of Ω.

Each subdomain Ω^e are open. And the union of these subdomains forms our domain Ω.

$$\Omega = \bigcup_{e=1}^{N_{el}} \Omega^e \qquad\qquad \Omega^e = (x^e, x^{e+1}),$$

where N_{el} is the number of elements. There are $N_{el} + 1$ number of nodes.

Just for technical purposes, since a pure union of open subdomains will leave out the points x^e, we will put an over line to signify closure.

$$\overline{\Omega} = \overline{\bigcup_{e=1}^{N_{el}} \Omega^e} \qquad\qquad \overline{\Omega} = \Omega \cup \partial\Omega$$

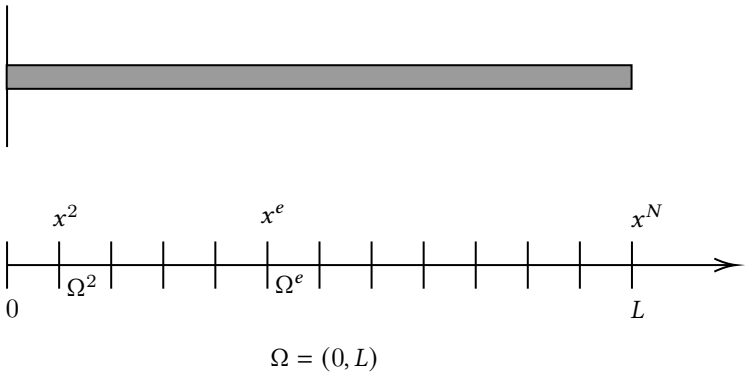

Figure 1: The 1 dimensional rod. The domain is partitioned into subdomains Ω^e.

This makes no difference in the actual computation. These points are what we call a set of zero measure.

We say that the points x^e are the notes of the partition. And Ω^e are the elements. Our weak form was stated as an integral. We can discretize this integral and write it as a sum.

We will first write the weak form as integral over the domain Ω.

$$\int_\Omega w^h_{,x} \sigma^h A \, dx = \int_\Omega w^h f A \, dx + w^h(L) t A$$

We can break the integral over the domain up into the sum of the integral over every subdomain.

$$\sum_{e=1}^{N_{el}} \int_{\Omega^e} w^h_{,x} \sigma^h A \, dx = \sum_{e=1}^{N_{el}} \left[\int_{\Omega^e} w^h f A \, dx \right] + w^h(L) t A$$

Since we have broken the domain up into subdomains, we can focus on how we can define the functions u^h and w^h over each subdomain.

2.5 Basis Functions - I

We have now written our Galerkin weak form in terms of integrals over each subdomain. We will now answer the question of how we can represent our trial solution and weighting function over Ω^e, where $e = 1, 2, \ldots, N_{el}$ - we are looking for a local representation over the subdomain Ω^e.

We can do this by defining local basis functions on Ω^e. There are a finite number of these basic functions over Ω^e and consequently over Ω.

The choice of basis function is up to us. But as an introduction, we will define two basis functions over Ω^e which are linear polynomials.

Let's write out basis functions to be $N^1(x)$ and $N^2(x)$. We can write the trial solution over subdomain e, u_e^h, by

$$u_e^h = \sum_{A=1}^{N_{ne}} N^A(x) d_e^A = N^1(x) d_e^1 + N^2(x) d_e^2$$

In this representation, N_{n_e} represents the number of nodes in the element. Here, we have chosen to represent an element using two nodes. We could have chosen to use more than two, and we will certainly do so when we get to higher dimensions. d_e^A is a "degree of freedom" that is being interpolated on element e.

2.6 Basis functions - II

$N^A(x)$ and d_e^A are also respectively called the nodal basis functions and nodal degrees of freedom.

Similarly, we will also write the expansion for the weighting function, $w_e^h(x)$.

$$w_e^h(x) = \sum_{A=1}^{N_{ne}} N^A(x) c_e^A$$

c_e^A is the degree of freedom for the weighting function. The degrees of freedom for the trial solution and the basis function are different, otherwise they would be the same function.

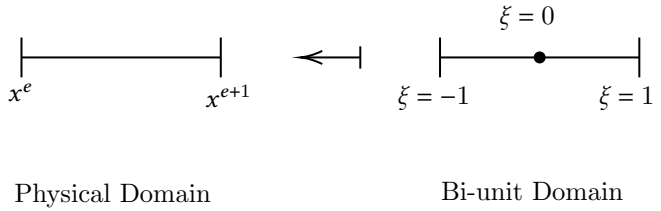

Figure 2: The bi-unit domain ξ

Since we are using the same basis for the weighting function and the trial solution, this is known as a Bubnov-Galerkin method.

We can something more systematic. We can think of the physical subdomain $\Omega^2 = (x^e, x^{e+1})$ as being constructed from a mapping from a different domain, which we call the bi-unit domain.

It has the name bi-unit since it has a length of 2. We do this since it will become very convenient to define our basis functions and carrying out integration of we have this idea.

So we will write our basis functions in the following way, remembering that x was mapped to from ξ.

$$N^1(x) = N^1(x(\xi))$$
$$N^2(x) = N^2(x(\xi))$$

If we abuse the notation here, we can simply write $N^1(\xi)$ and $N^2(\xi)$. We can then write the functions out

$$N^1(\xi) = \frac{1-\xi}{2} \qquad\qquad N^2(\xi) = \frac{1+\xi}{2}$$

The basis function have the kronecker delta property, where

$$N^A(\xi^B) = \delta_{AB} = \begin{cases} 1 & A = B \\ 0 & A \neq B \end{cases}$$

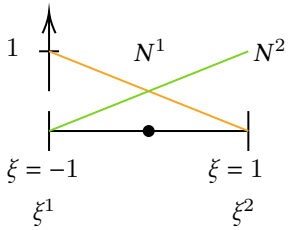

Figure 3: The basis functions on the bi-unit domain Ω^{ξ}. $N^1(-1) = 1, N^1(1) = 0$. $N^2(-1) = 0, N^2(1) = 1$. We can refer to $\xi = -1$ as ξ^1, the first node of our domain, and $\xi = 1$ as ξ^2, the second node of our domain.

2.7 The bi-unit domain - I

Here is another useful property of the linear basis functions on the bi-unit domain. Consider,

$$N^1(\xi) + N^2(\xi) = \frac{1-\xi}{2} + \frac{1+\xi}{2} = 1$$

The sum of the basis functions evaluated at any point on the bi-unit domain is 1. This means we can represent constants.

The basis functions can be generalized to higher order polynomials. They are drawn from a family of polynomials call Lagrange polynomials.

There is another way to look at the basis functions on the physical domain Ω. Let us consider the point x^{e+1} on the physical domain. We have three ways to call this point.

1. Local node number two of subdomain Ω^e
2. Local node number one of subdomain Ω^{e+1}
3. Global node number $e+1$

We recognize that each subdomain on the physical domain is always constructed as a mapping from the same parent bi-unit domain. At global node $e+1$, we can associate the positively sloped basis function from Ω^e and the negatively sloped basis function from Ω^{e+1} as a global basis function for global node $e+1$.

This global basis function is non-zero in the subdomains immediately adjacent to this node. This is called the compact support property of the global basis functions.

2.8 The bi-unit domain - II

We can see that the local definition of basis functions leads to global basis functions associated with each global node, with compact support in elements immediately adjoining that node.

Recall the Galerkin weak form

$$\sum_{e=1}^{N_{el}} \int_{\Omega^e} w^h_{,x} \sigma^h A \, dx = \sum_{e=1}^{N_{el}} \left[\int_{\Omega^e} w^h f(x(\xi)) A \, dx \right] + w^h(L) t A$$

At the very first element of the physical domain (global node number 0), we need to satisfy the Dirichlet boundary condition, $u^h(x^0) = u_0$, and w^h on that point should also be zero (by construction).

As far as w^h is concerned, we only need a single basis function $N^2(x(\xi))$ which is positively sloped for that very first element. But for any other element $e \neq 1$, we have

$$w^h_e(x) = \sum_{A=1}^{N_{ne}} N^A(x) c^A_e$$

as usual.

By looking at the weak form, we see that we need to compute the gradient of the weight function and the gradient of u. We know that we can represent the locate trial solutions and the weight functions as linear combinations of the weight functions.

$$u^h_e(\xi) = \sum_{A=1}^{N_{ne}} N^A(\xi) d^A_e \qquad\qquad w^h_e(\xi) = \sum_{A=1}^{N_{ne}} N^A(\xi) c^A_e$$

So to find the gradient we can just compute the gradient of each sum term by term. So this is now a problem of computing the gradient of our basis

functions.

$$u^h_{e,x}(\xi) = \sum_{A=1}^{N_{ne}} N^A_{,x}(\xi) d^A_e \qquad\qquad w^h_{e,x}(\xi) = \sum_{A=1}^{N_{ne}} N^A_{,x}(\xi) c^A_e$$

The factors d^A_e and c^A_e are degrees of freedom so it has no position dependence. The gradient of our basis functions can be found via the chain rule

$$N^A_{,x}(\xi) = N^A_{,\xi}\xi_{,x}$$

2.9 The finite dimensional weak form as sum over element subdomains - I

We stated that to compute the gradient of trial solutions and the weight functions, we had to be able to compute the gradient ξ with respect to x.

It follows that we need to figure out the mapping from $\Omega^\xi \mapsto \Omega^e$. We will say any point

$$x_e(\xi) = \sum_{A=1}^{N_{ne}} N^A(\xi) x^A_e$$

Here, x^A_e are local nodes. They have equivalent global representations. We have used the same basis functions that was used to represent the trial solutions and the weight functions.

When we use the same basis functions to represent out finite dimensional functions and to interpolate our geometry, we have a isoparameteric formulation.

This lets us say that the derivative of x with respect to ξ at element e is

$$x_{,\xi}|_e = \sum_{A=1}^{N_{ne}} N^A_{,\xi} x^A_e = \frac{x^{e+1} - x^2}{2} = \frac{h^e}{2}$$

we have represented the length of the element e using h^e. There were no requirements that the partition of the element are even. And indeed there are none. We are allowed a non-uniform discretization.

The tangent of the mapping from our bi-unit domain to the physical domain is a constant since we are using linear basis functions.

2.10 The finite dimensional weak form as a sum over element subdomains - II

We have a isoparameteric map that is invertible. It is a one to one and onto mapping. The important fact is that

$$\xi_{,x} = \frac{1}{x_{,\xi}}$$

This fact allows us to say that

$$u^h_{e,x}(\xi) = \sum_{A=1}^{N_{ne}} N^A_{,\xi}\xi_{,x} d^A_e = \sum_{A=1}^{N_{ne}} N^A_{,\xi}\frac{2}{h^e} d^A_e$$

and

$$w^h_{e,x}(\xi) = \sum_{A=1}^{N_{ne}} N^A_{,\xi}\xi_{,x} c^A_e = \sum_{A=1}^{N_{ne}} N^A_{,\xi}\frac{2}{h^e} c^A_e$$

Consider the following integral from the finite dimensional weak form

$$\begin{aligned}\int_{\Omega^e} w^h_{,x}\sigma^h A\,dx &= \int_{\Omega^e} w^h_{,x} EA u^h_{,x}\,dx \\ &= \int_{\Omega^e} \left[\sum_A N^A_{,\xi}\frac{2}{h^e}c^A_e\right] EA u^h_{,x} \left[\sum_B N^A_{,\xi}\frac{2}{h^e}d^A_e\right] dx\end{aligned}$$

Another integral we have is

$$\int_{\Omega^e} w^h f(x(\xi)) A\,dx = \int_{\Omega^e} \left[\sum_{n=1} N^A c^A_e\right] f(x(\xi)) A\,dx$$

Almost everything in these two integrals are parameterized by ξ, yet it is integrated with respect to x.

If we use a change of variables $dx = (dx/d\xi)\,d\xi = (h^e/2)\,d\xi$, we can rewrite these integrals as integrals over the bi-unit domain Ω^ξ.

So we have

$$\int_{\Omega^e} w^h_{,x} EA u^h_{,x}\,dx \implies \int_{\Omega^\xi} \left(\sum_A N^A_{,\xi}\frac{2}{h^e}c^A_e\right) EA \left(\sum_B N^B_{,\xi}\frac{2}{h^e}d^A_e\right)\frac{h^e}{2}\,d\xi$$

and

$$\int_{\Omega^e} w^h f(x(\xi)) A \, dx \implies \int_{\Omega^\xi} \left(\sum_A N^A c_e^A \right) f(\xi) A \frac{h^e}{2} \, d\xi$$

3 Linear algebra; the matrix-vector form

3.1 The matrix-vector weak form - I - I

Recall the following integral in the finite dimensional weak form.

$$\int_{\Omega^e} w^A_{,h} \sigma^h A \, dx$$

for the case that $e = 1$, we can represent w^A using a single basis function N^2.

$$\int_{\Omega^e} \underbrace{N^2_{,\xi} \xi_{,x} c^2_e}_{w^h_{,x}} EA \underbrace{\sum_B N^B_{,\xi} \xi_{,x} d^B_e}_{u^h_{,x}} \, dx = \int_{\Omega^\xi} \left(N^2_{,\xi} \frac{2}{h^e} c^2_e \right) EA \left(\sum_B N^B_{,\xi} \frac{2}{h^e} d^B_e \right) \frac{h^e}{2} \, d\xi$$

Similarly,

$$\int_{\Omega^e} w^h f A \, dx = \int_{\Omega^\xi} \left(N^2 c^2_e \right) f A \frac{h^e}{2} \, d\xi$$

Now, consider the case for a general element Ω^e.

$$\int_{\Omega^e} w^h_{,x} EA u^h_{,x} \, dx \implies \int_{\Omega^\xi} \left(\sum_A N^A_{,\xi} \frac{2}{h^e} c^A_e \right) EA \left(\sum_B N^B_{,\xi} \frac{2}{h^e} d^A_e \right) \frac{h^e}{2} \, d\xi$$

To simplify this integral, we need to recognize that c^A_e and d^B_e are degrees of freedom used to respectively interpolate the weighting function and the trial solution. These degrees of freedom are independent of the independent variables, so we can take them out of the integral. Our integral becomes

$$\sum_{A,B} c^A_e \left(\int_{\Omega^\xi} N^A_{,\xi} \frac{2EA}{h^e} N^B_{,\xi} \, d\xi \right) d^B_e$$

For the integral involving the forcing function

$$\sum_A c^A_e \left(\int_{\Omega^\xi} N^A \frac{f A h^e}{2} \, d\xi \right)$$

For $e = 1$ there is no sum over A. Instead, we only use the index $A = 2$, since $w^h(0) = 0$, and we do not use the $A = 1$ negatively sloped basis function. ("The weighting contribution from the degree of freedom one is not used.")

3.2 The matrix-vector weak form - I - II

The idea is to use a matrix-vector product to eliminate the sums (over A and over B).

Recall the integral

$$\sum_{A,B=1}^{N_{ne}} c_e^A \left(\int_{\Omega^\xi} N_{,\xi}^A \frac{2EA}{h^e} N_{,\xi}^B \, d\xi \right) d_e^B$$

where N_{n_e} in our case is 2.

We can now make some simplifications. Let's assume that the term EA are uniform over the element Ω^e. So we can pull it out of the integral.

We will rewrite the integral as

$$\begin{bmatrix} c_e^1 & c_e^2 \end{bmatrix} \frac{2EA}{h^e} \left(\int_{\Omega^e} \begin{bmatrix} N_{,\xi}^1 N_{,\xi}^1 & N_{,\xi}^1 N_{,\xi}^2 \\ N_{,\xi}^2 N_{,\xi}^1 & N_{,\xi}^2 N_{,\xi}^2 \end{bmatrix} d\xi \right) \begin{bmatrix} d_e^1 \\ d_e^2 \end{bmatrix}$$

Recall that we are using linear basis functions. So we have $N_{,\xi}^1 = -1/2$ and $N_{,\xi}^2 = 1/2$. The integral is trivial.

$$\begin{bmatrix} c_e^1 & c_e^2 \end{bmatrix} \frac{2EA}{h^e} \begin{bmatrix} 1/4 & -1/4 \\ -1/4 & 1/4 \end{bmatrix} \left(\int_{\Omega^e} d\xi \right) \begin{bmatrix} d_e^1 \\ d_e^2 \end{bmatrix}$$

Recall that Ω^ξ is the bi-unit domain. So the integral over domain is 2.

Finally, we have

$$\begin{bmatrix} c_e^1 & c_e^2 \end{bmatrix} \frac{EA}{h^e} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} d_e^1 \\ d_e^2 \end{bmatrix}$$

In a related manner, the following integral can be rewritten into the

matrix-vector weak form using similar assumptions

$$\sum_A c_e^A \left(\int_{\Omega^\xi} N^A \frac{fAh^e}{2}\, d\xi \right) = \begin{bmatrix} c_e^1 & c_e^2 \end{bmatrix} \frac{fAh^e}{2} \int_{-1}^{1} \begin{bmatrix} N^1 = \frac{1-\xi}{2} \\ N_2 = \frac{1+\xi}{2} \end{bmatrix} d\xi$$

$$= \begin{bmatrix} c_e^1 & c_e^2 \end{bmatrix} \frac{fAh^e}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

3.3 The matrix-vector weak form - II - I

For the very first element ($e = 1$), we have

$$c_e^2 \frac{2EA}{h^e} \sum_B \left(\int_{\Omega^\xi} N^2_{,\xi} N^B_{,\xi}\, d\xi \right) d_e^B = c_e^2 \frac{2EA}{h^e} \left(\int_{\Omega^\xi} \begin{bmatrix} N^2_{,\xi} N^1_{,\xi} & N^2_{,\xi} N^2_{,\xi} \end{bmatrix} d\xi \right) \begin{bmatrix} d_e^1 \\ d_e^2 \end{bmatrix}$$

$$= c_e^2 \frac{EA}{h^e} \begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} d_e^1 \\ d_e^2 \end{bmatrix}$$

The integral involving the forcing the function is

$$c_e^2 \left(\int_{\Omega^\xi} N^2 \frac{fAh^e}{2}\, d\xi \right) = c_e^2 \frac{fAh^e}{2}$$

Let's now put everything together. We converted the sums over the degrees of freedom for some element into a matrix-vector product form. We will apply the same idea for to the complete finite dimensional weak form which has been summed over the entire Ω.

$$\sum_{e=1}^{N_{el}} \int_{\Omega^e} w^h_{,x} \sigma^h A\, dx = \sum_{e=1}^{N_{el}} \left[\int_{\Omega^e} w^h f(x(\xi)) A\, dx \right] + w^h(L) tA$$

Rewriting the integral in the matrix-vector weak form we developed,

$$c_1^2 \frac{EA}{h^1} \begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} d_1^1 \\ d_1^2 \end{bmatrix} + \sum_{e=2}^{N_{el}} \begin{bmatrix} c_e^1 & c_e^2 \end{bmatrix} \frac{EA}{h^e} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} d_e^1 \\ d_e^2 \end{bmatrix}$$

$$= c_1^2 \frac{fAh^1}{2} + \sum_{e=2}^{N_{el}} \begin{bmatrix} c_e^1 & c_e^2 \end{bmatrix} \frac{fAh^e}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + c^2_{N_{el}} tA$$

The $w^h(L)$ term has a simply representation, which is $c^2_{N_{el}}$.

3.4 The matrix-vector weak form - II - II

How did we come up with $w^h(L) = c^2_{N_{el}}$? The reasoning lies in the Kronecker delta property of our chosen basis functions when evaluated at the end points.

When we evaluate our trial solution, or our weighting function at a node x^e, it translates into evaluating the functions at ξ^1 in our bi-unit domain.

$$u^h(x^e) = u^h(x(-1)) = \sum_A \underbrace{N^A(\xi^1 = -1)}_{\delta_{A1}} d^A_e = d^1_e$$

$$u^h(x^{e+1}) = u^h(x(1)) = \sum_A \underbrace{N^A(\xi^2 = -1)}_{\delta_{A2}} d^A_e = d^2_e$$

By the same process

$$w^h(x^e) = c^1_e$$
$$w^h(x^{e+1}) = c^2_e$$

The Kronecker delta property of the basis functions ensures that the nodal degrees of freedom of the solution field are the value of solution field at the nodes. This is known as a interpolatary property which happens to apply for the linear basis functions we have chosen: it is not a universal property of arbitrary basis functions.

We have the following relationship between the degree of freedoms indexed by local node numbers and global node numbers.

$$d^A_e = d_{e+A-1}$$

This relationship reads: local degree of freedom A in element e is equal to the global degree of freedom $e + A - 1$.

Let's check that this is the case:

$$d^1_1 = d_1$$

and

$$d^1_{N_{el}} = d_{N_{el}} \qquad\qquad d^2_{N_{el}} = d_{N_{el}+1}$$

there are one more node than there are elements, which is what we expected.

3.5 The matrix-vector weak form - III - I

We wrote the finite dimensional weak form in a matrix-vector manner in the previous segment, using local node numbers. We defined a relationship between the degrees of freedom indexed using local node numbers and indexed using global node numbers. We can rewrite the weak form in terms of the global node numbers.

$$c_1^2 \frac{EA}{h^1} \begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} d_1^1 \\ d_1^2 \end{bmatrix} + \sum_{e=2}^{N_{el}} \begin{bmatrix} c_e^1 & c_e^2 \end{bmatrix} \underbrace{\frac{EA}{h^e} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}}_{K_e} \begin{bmatrix} d_e^1 \\ d_e^2 \end{bmatrix}$$

$$= c_1^2 \frac{fAh^1}{2} + \sum_{e=2}^{N_{el}} \begin{bmatrix} c_e^1 & c_e^2 \end{bmatrix} \underbrace{\frac{fAh^e}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix}}_{F_e} + c_{N_{el}}^2 tA$$

We have put braces under two terms portion of the equation above. This is to introduce some more terminology. K_e is known as the element stiffness matrix and F_e is known as the element force vector.

These names are from the times when finite element analysis is primarily used in structural analysis.

We will proceed with the finite element assembly.

$$\begin{bmatrix} c_2 & c_3 & \dots & c_{N_{el}} & c_{N_{el}+1} \end{bmatrix} EAQ \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{N_{el}} \\ d_{N_{el}+1} \end{bmatrix}$$

The Q matrix is constructed as follows. We first look at the contributions of element $e = 1$. This element uses c_2 only, and involves d_1 and d_2. So it will have contributions at row one, and columns 1 and 2. For the next element, it will use c_2 and c_3, and d_2 and d_3, so it contribute to terms at rows 1 and 2, at columns 2 and 3. This process continues. Q must be a $N_{el} \times N_{el} + 1$ matrix since the c vector has N_{el} columns and the d vector has $N_{el} + 1$ rows.

We can write this in operator notation.

$$\begin{bmatrix} c_2 & c_3 & \dots & c_{N_{el}} & c_{N_{el}+1} \end{bmatrix} EAQ \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{N_{el}} \\ d_{N_{el}+1} \end{bmatrix} = \mathop{\mathsf{A}}_{e=1}^{N_{el}} \begin{bmatrix} c_e^1 & c_e^2 \end{bmatrix} K_e \begin{bmatrix} d_e^1 \\ d_e^2 \end{bmatrix}$$

Here, A is the assembly operator. It assembles over all elements contributions of the form $\langle c_e^1, c_e^2 \rangle$ times the stiffness matrix, times the d column vector.

3.6 The matrix-vector weak form - III - II

For the right hand side of the expression, these terms

$$c_1^2 \frac{fAh^1}{2} + \sum_{e=2}^{N_{el}} \begin{bmatrix} c_e^1 & c_e^2 \end{bmatrix} \frac{fAh^e}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

can be assembled this into

$$\begin{bmatrix} c_2 & c_3 & \dots & c_{N_{el}} & c_{N_{el}+1} \end{bmatrix} \frac{fA}{2} \begin{bmatrix} h^1 + h^2 \\ h^2 + h^3 \\ \vdots \\ h^{N_{el}-1} + h^{N_{el}} \\ h^{N_{el}} \end{bmatrix} = \mathop{\mathsf{A}}_{e=1}^{N_{el}} \begin{bmatrix} c_e^1 & c_e^2 \end{bmatrix} \begin{bmatrix} \frac{fAh^e}{2} \\ \frac{fAh^e}{2} \end{bmatrix}$$

we have pulled the factor $fA/2$ out assuming that they are uniform over the elements.

The term that involves the traction ($c_{N_{el}}^2 tA$) can be written as

$$\begin{bmatrix} c_2 & c_3 & \dots & c_{N_{el}} & c_{N_{el}+1} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ tA \end{bmatrix} = c_{N_{el}+1} tA = c_{N_{el}}^2 tA$$

The matrix vector form of our finite dimensional weak form in all its glory,

$$\begin{bmatrix} c_2 & c_3 & \dots & c_{N_{el}} & c_{N_{el}+1} \end{bmatrix} EAQ \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{N_{el}} \\ d_{N_{el}+1} \end{bmatrix}$$

$$= \begin{bmatrix} c_2 & c_3 & \dots & c_{N_{el}} & c_{N_{el}+1} \end{bmatrix} \frac{fA}{2} \begin{bmatrix} h^1 + h^2 \\ h^2 + h^3 \\ \vdots \\ h^{N_{el}-1} + h^{N_{el}} \\ h^{N_{el}} \end{bmatrix}$$

$$+ \begin{bmatrix} c_2 & c_3 & \dots & c_{N_{el}} & c_{N_{el}+1} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ tA \end{bmatrix}$$

3.7 The final finite element equations in matrix-vector form - I

Let's make the next simplification: h^e is the same for all elements. This allows us to factor out $1/h^2$ within the Q matrix. So we have

$$\begin{bmatrix} c_2 & c_3 & \dots & c_{N_{el}} & c_{N_{el}+1} \end{bmatrix} \frac{EA}{h^e}(h^e \mathrm{Q}) \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{N_{el}} \\ d_{N_{el}+1} \end{bmatrix}$$

$$= \begin{bmatrix} c_2 & c_3 & \dots & c_{N_{el}} & c_{N_{el}+1} \end{bmatrix} \frac{fAh^e}{2} \begin{bmatrix} 2 \\ 2 \\ \vdots \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} c_2 & c_3 & \dots & c_{N_{el}} & c_{N_{el}+1} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ tA \end{bmatrix}$$

The c vector has length N_{el}, the Q matrix must be $N_{el} \times (N_{el}+1)$. The d column vector must be $1 \times (N_{el}+1)$.

The Kronecker delta property of our basis functions tells us that the finite dimensional trial solution and the weighting function evaluated at the nodal points is exactly equal to the nodal degree of freedom. So d_1 is equivalent to $u^{h^e}(x^1 = 0) = u_0$, which is a known quantity (the Dirichlet boundary conditions).

We are free to move the first column of Q to the right hand side, noticing that the first column is zero except at row one, which equals u_0. Let's call

the new matrix H.

$$\begin{bmatrix} c_2 & c_3 & \dots & c_{N_{el}} & c_{N_{el}+1} \end{bmatrix} \frac{EA}{h^e} (h^e \mathrm{H}) \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{N_{el}} \\ d_{N_{el}+1} \end{bmatrix}$$

$$= \begin{bmatrix} c_2 & c_3 & \dots & c_{N_{el}} & c_{N_{el}+1} \end{bmatrix} \left(\frac{fAh^e}{2} \begin{bmatrix} 2 \\ 2 \\ \vdots \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ tA \end{bmatrix} + \frac{EA}{h^e} \begin{bmatrix} u_0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \right)$$

So H is now a $N_{el} \times N_{el}$ matrix.

We will introduce a notation for matrices and vectors we have not been using. Matrix quantities will be denoted with an underline. For example, the element stiffness matrix K_e will now be addressed as $\underline{K_e}$. If the quantity is a a row vector, then we will denote it with the underline as well as a transpose symbol.

We will redefine the notation of the matrix from H to $\underline{K}$. We will denote the c row vector as $\underline{c}^\intercal$. The d vector will be denoted $\underline{d}$. For the vectors multiplied by $\underline{c}$ on the right hand side, we will call the collection $\underline{F}$.

3.8 The final finite element equations in matrix-vector form - II

Using the new notations we defined in the last subsection, we can write the assembled matrix-vector form as follows

$$\underline{c}^\intercal \underline{K} \underline{d} = \underline{c}^\intercal \underline{F} \quad \forall\ \underline{c} \in \mathbb{R}^{N_{el}}$$

There is still one step away from our final formulation of the finite dimensional weak form as matrix vector products. The realization is that the

condition that the above equation must hold for all $\underline{c} \in \mathbb{R}^{N_{el}}$ reflects the condition that $w^h \in \mathcal{V}^h$. We have fixed the functional form of w^h by selecting a polynomial basis function, and the degree of arbitrariness required by this condition is reflected in our matrix vector form by imposing $\underline{c} \in \mathbb{R}^{N_{el}}$. Without proof,

$$\underline{K}\underline{d} = \underline{F}$$

is the final form of the finite element equations.

We have some remarks.

1. The matrix $\underline{K}$ (the stiffness matrix, this differs from the element stiffness matrix, $\underline{K}_e$) is symmetric, positive definite, with banded tri-diagonal structure. The symmetric came from the fact that the term in the weak form which gave rise to our matrix $\underline{K}$ is the following integral

$$\int_\Omega w^h_{,x} \overbrace{E u^h_{,x}}^{\sigma^h} A \, dx$$

   which is a bilinear functional. The integral is unchanged from interchanging the $w^h_{,x}$ and $Eu^h_{,x}$. The equation $\underline{K}\underline{d} = \underline{F}$, resembles Hooke's law, $\underline{K}$ plays the row as the spring constant or spring stiffness. The matrix is positive definite due to E being greater than zero. The banded tri-diagonal structure comes from that there are 2 first derivatives (on w^h and u^h) and we have a set of linear basis functions.

2. The force vector $\underline{F}$

$$\underline{F} = \frac{fAh^e}{2}\begin{bmatrix} 2 \\ 2 \\ \vdots \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ tA \end{bmatrix} + \frac{EA}{h^e}\begin{bmatrix} u_0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$

   The leftmost term in the equation above came from the distributed body force f. It has value fAh^e on every node except $fAh^e/2$ on the last node. The central term came from traction (from the Neumann boundary condition). It only applies to the very last node. The last

term came from the Dirichlet boundary condition. EAu_0/h^e has units of force. It is also called a Dirichlet driven load.

Symmetric Matrix: a matrix $\underline{K}$ is symmetric if $\underline{K} = \underline{K}^\intercal$

Positive Definite Matrix: a matrix $\underline{K}$ is positive definite if $\underline{K}$ is a $n \times n$ matrix, and for all $\underline{d} \in \mathbb{R}^n$, $\underline{d}^\intercal \underline{K} \underline{d} \geq 0$. In particular, $\underline{d}^\intercal \underline{K} \underline{d} > 0$ if $d \neq \underline{0}$. So it follows that $\underline{d}^\intercal \underline{K} \underline{d} = 0$ iff (iff means if and only if) $\underline{d} = \underline{0}$. It is a generalization of a variable being positive.

We have completed the simplest formulation of the one-dimensional problem for linear elliptic PDEs. In principle, we can now find $\underline{d}$ by finding the inverse of $\underline{K}$, so that

$$\underline{K} \underline{d} = \underline{F} \implies \underline{d} = \underline{K}^{-1} \underline{F}$$

From d, we can reconstruct the field at each element by

$$u_e^h = \sum_A N^A d_e^A$$

3.8.1 Coding Assignment 1

What are some `C++` objects and functions we need to implement in our program:

- basis functions/Gradients of the basis functions,
- a function that calculates the L_2 norm of the error,
- generate a mesh,
- create $\underline{F}_e$ and $\underline{K}_e$,
- finite element assembly,
- define and apply our boundary conditions

- Solve for $\underline{d}$
- Output the results into some format

4 More on boundary conditions; basis functions; numerics

4.1 The pure Dirichlet problem - I

We will consider the base where both ends of our rod are subject to Dirichlet boundary conditions.

We want to find $u^h \in \mathcal{S}^h \subset \mathcal{S} = \{u|u(0) = u_0, u(L) = u_g\}$ ($\mathcal{S}^h$ is a subset of $\mathcal{S}$), where $\mathcal{S}^h = \{u^h \in H^1(\Omega)|u^h(0) = u_0, u^h(L) = ug\}$, such that for all $w^h \in \mathcal{V}^h \subset \mathcal{V} = \{w|w(0) = 0, w(L) = 0\}$, this implies $\mathcal{V}^h = \{w^h \in H^1(\Omega)|w^h(0) = 0, w^h(L) = 0\}$, the following holds

$$\int_\Omega w^h_{,x}\sigma^h A\,dx = \int_\Omega w^h f A\,dx$$

There is no contribution from the traction.

The physical picture is a 1D rod, held fixed at $x = 0$ with $u(0) = 0$, and the other has a displacement equal to $u(L) = u_g$. It also has a body force f in the $+x$ direction.

We want ot answer what happens with the homogeneous Dirichlet boundary conditions on weighting function.

Again, we partition our domain into N_{el} open intervals. Because of there is Dirichlet data on both Ω^1 and $\Omega^{N_{el}}$, we will do something special with the basis functions for w^h at these elements.

$$w^h_1 = N^2(\xi)c^2_1 \qquad\qquad w^h_{N_{el}} = N^1 c^1_{N_{el}}$$

recall that N_1 is negatively sloped, and N_2 is positively sloped.

When we work things through, we will find that

$$\int_\Omega w^h_{,x}\sigma^h A\,dx = c^2_1\frac{EA}{h^e}\begin{bmatrix}-1 & 1\end{bmatrix}\begin{bmatrix}d^1_1\\ d^2_1\end{bmatrix} + \sum_{e=2}^{N_{el}-1}\begin{bmatrix}c^1_e & c^2_e\end{bmatrix}\frac{EA}{h^e}\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}\begin{bmatrix}d^1_e\\ d^2_e\end{bmatrix}$$
$$+ c^1_{N_{el}}\frac{EA}{h^e}\begin{bmatrix}1 & -1\end{bmatrix}\begin{bmatrix}d^1_{N_{el}}\\ d^2_{N_{el}}\end{bmatrix}$$

similarly

$$\int_\Omega w^h f A \, dx = c_1^2 \frac{f A h^e}{2} + \sum_{e=2}^{N_{el}-1} \begin{bmatrix} c_e^1 & c_e^2 \end{bmatrix} \frac{f A h^e}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_{N_{el}}^1 \frac{f A h^e}{2}$$

4.2 The pure Dirichlet problem - II

We will carrying out the finite element assembly. Recognize that we have this mapping between local and global node numberings.

$$c_e^A = c_{e+A-1}$$

So the following equalities holds

$$\begin{bmatrix} c_e^1 & c_e^2 \end{bmatrix} = \begin{bmatrix} c_e & c_{e+1} \end{bmatrix}$$
$$\begin{bmatrix} d_e^1 & d_e^2 \end{bmatrix} = \begin{bmatrix} d_e & d_{e+1} \end{bmatrix}$$

Carrying out the assembly process, and making the assumption that the element lengths are uniform and f and A are also uniform over domain. The assembled matrix vector weak form is

$$\underbrace{\begin{bmatrix} c_2 & c_3 & \dots & c_{N_{el}} \end{bmatrix}}_{N_{el}-1} \frac{EA}{h^e} \underbrace{\begin{bmatrix} -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & & 2 & -1 \end{bmatrix}}_{N_{el}-1 \times N_{el}+1} \underbrace{\begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{N_{el}+1} \end{bmatrix}}_{N_{el}+1}$$

$$= \begin{bmatrix} c_2 & c_3 & \dots & c_{N_{el}} \end{bmatrix} \frac{f A h^e}{2} \begin{bmatrix} 2 \\ 2 \\ \vdots \\ 2 \end{bmatrix}$$

By the Kronecker delta property of our basis functions, we know that the trial solution and basis functions evaluated at the nodes is equal to the nodal

degree of freedom. So we can just write d_1 as u_0 and $d_{N_{el}+1} = u_g$. Both u_0 and u_g are known quantities.

The first column of the $N_{el} - 1 \times N_{el} + 1$ matrix has a 1 on row one, and zero for any other rows. Similarly, the last column has zeros for all entries, except on the last row. The first column multiplies with $d_1 = u_0$. The last column multiplies with $d_{N_{el}+1} = u_g$. We can move them to the right hand side. Resulting in

$$\underbrace{\begin{bmatrix} c_2 & c_3 & \dots & c_{N_{el}} \end{bmatrix}}_{N_{el}-1} \frac{EA}{h^e} \underbrace{\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}}_{N_{el}-1 \times N_{el}-1} \underbrace{\begin{bmatrix} d_2 \\ \vdots \\ d_{N_{el}} \end{bmatrix}}_{N_{el}-1} =$$

$$\begin{bmatrix} c_2 & c_3 & \dots & c_{N_{el}} \end{bmatrix} \left(\frac{fAh^e}{2} \begin{bmatrix} 2 \\ 2 \\ \vdots \\ 2 \end{bmatrix} + \frac{EA}{h^e} \begin{bmatrix} u_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \frac{EA}{h^e} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ u_g \end{bmatrix} \right)$$

The $N_{el}-1 \times N_{el}-1$ matrix is the stiffness matrix $\underline{K}$, and the reduced d vector, $\underline{d}$, is the column vector containing $N_{el} - 1$ entries of of the trial solution degrees of freedoms. The c transposed vector $\underline{c}^\intercal$ is the row vector containing the weighting function degrees of freedom. And the quantity in side the large bracket on the right hand side is $\underline{f}$.

So in sum we have

$$\underline{c}^\intercal \underline{K} \underline{d} = \underline{c}^\intercal \underline{F} \; \forall \; \underline{c} \in \mathbb{R}^{N_{el}-1} \implies \underline{K} \underline{d} = \underline{F}$$

The forcing vector is a sum with three contributions.

$$\underline{F} = \underbrace{\frac{fAh^e}{2}\begin{bmatrix}2\\2\\\vdots\\2\end{bmatrix}}_{\text{forcing function}} + \underbrace{\frac{EA}{h^e}\begin{bmatrix}u_0\\0\\\vdots\\0\end{bmatrix}}_{\text{Dirichlet forcing at } x=0} + \underbrace{\frac{EA}{h^e}\begin{bmatrix}0\\0\\\vdots\\u_g\end{bmatrix}}_{\text{Dirichlet forcing at } x=L}$$

4.3 Higher polynomial order basis functions - I

We have developed our problem using linear basis functions. We will start with quadratic basis functions then write a general formula for an arbitrary order polynomial.

When we used polynomial basis functions, our subdomains were an open interval made of two nodes: we required two degrees of freedom to interpolate linear functions. To interpolate quadratic functions, we will require our subdomains to consist of three nodes.

For this reason, our physical domain of length L will consist of $2N_{el}+1$ nodes when we use quadratic basis functions.

Lets consider a physical subdomain Ω^e, made of nodes x_{2e-1} and x_{2e+1} on the ends, and x_{2e} in the middle. It's said to be mapped from a bi-unit domain, with $\xi=-1$ and $\xi=1$ on the ends, and $\xi=0$ in the middle.

Using this convention, we can write our finite dimensional trial solution on some subdomain Ω^e as the expansion of of quadratic basis functions

$$u_e^h = \sum_{A=1}^{N_{el}=3} N^A(x(\xi))d_e^A$$

Similarly, the weighting function can be represented as

$$w_e^h = \sum_{A=1}^{N_{el}} N^A(x(\xi))c_e^A$$

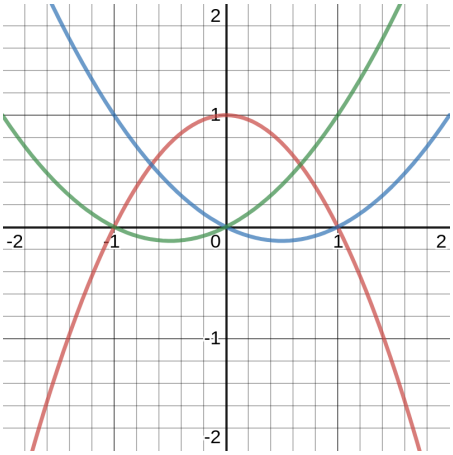

Figure 4: The red curve represents N^2, and blue curve N^1, and the green curve N^3

The basis functions are the following:

$$N^1(\xi) = \frac{\xi(\xi - 1)}{2} \qquad N^2(\xi) = 1 - \xi^2 \qquad N^3(\xi) = \frac{\xi(1 + \xi)}{2}$$

When $\xi^1 = -1$, $\xi^2 = 0$, $\xi^3 = 1$, the Kronecker delta property holds.

$$N^A(\xi^B) = \delta_{AB}$$

Another property, important for representing constants, is that

$$\sum_{A=1}^{N_{el}} N^A(\xi) = 1$$

for any arbitrary point ξ.

The quadratics we have written out are special cases of what are called Lagrange polynomials. We can just use the formula for Lagrange polynomials.

For a polynomial of order $N_{n_e} - 1$ (N_{n_e} is the number of nodes in a single element), the formula for the basis function is

$$N^A(\xi) = \frac{\prod_{B=1, B\neq A}^{N_{ne}} (\xi - \xi^B)}{\prod_{B=1, B\neq A}^{N_{ne}} (\xi^A - \xi^B)} \quad \forall\ A = 1, 2, \ldots, N_{n_e}$$

In fact, the linear basis function (polynomial of order 1) we used before can be completely recovered from our formula for arbitrary order $N_{n_e} - 1$ formula for Lagrange polynomials.

4.4 Higher polynomial order basis functions - I - II

The Lagrange polynomial formula satisfies the Kronecker delta property. Immediately follows by substituting in ξ^A into the formula for $N^A(\xi^A)$. The numerator and denominator cancel to give 1.

For N^A evaluated at a point ξ^C where $\xi \neq A$, then for some term $\xi^C = \xi^B$, and the product goes to zero.

We can check that

$$\sum_A N^A(\xi) = 1$$

holds as well.

We will now develop our finite element formulation with quadratic basis functions. Recall our finite dimensional weak form for the Dirichlet-Neumann problem is

$$\sum_{e=1}^{N_{el}} \int_{\Omega^e} w^h_{,x} \underbrace{\sigma^h}_{Eu^h_{,x}} A \, dx = \sum_{e=1}^{N_{el}} \left[\int_{\Omega^e} w^h f A \, dx \right] + w^h(L) t A$$

We will focus on

$$\int_{\Omega^e} w^h_{,x} E u^h_{,x} A \, dx$$

Recall that we will express our trial solution and weighting function as linear combinations of the basis functions times a corresponding degree of freedom. So the gradient of the trial solution and weighting function can be found by differentiating term by term of the expansion. In this case of quadratic basis functions, the gradients of the trial solution and weighting

functions are

$$u^h_{,x} = \sum_{A=1}^{3} N^A_{,x} d^A_e$$

$$w^h_{,x} = \sum_{A=1}^{3} N^A_{,x} c^A_e$$

Recall that $N^A_{,x} = N^A_{,\xi}\xi_{,x}$. So expanding the gradients inside the integral gives

$$\int_{\Omega^e} \left[\sum_{A=1}^{3} N^A_{,\xi}\xi_{,x} c^A_e\right] EA \left[\sum_{A=1}^{3} N^A_{,\xi}\xi_{,x} d^A_e\right] dx$$

Lets write out $N^A_{,x}$ for each A.

$$N^1_{,\xi} = \frac{d}{d\xi}\left(\frac{\xi(\xi-1)}{2}\right) = \frac{2\xi-1}{2}$$

$$N^2_{,\xi} = \frac{d}{d\xi}(1-\xi^2) = -2\xi$$

$$N^3_{,\xi} = \frac{d}{d\xi}\left(\frac{1+\xi}{2}\right) = \frac{1+2\xi}{2}$$

4.5 Higher polynomial order basis functions - II - I

To find $\xi_{,x}$ in the case of linear basis functions, we used the invertibility of our mapping.

We have a isoparameteric mapping since we are using the same basis functions for the trial solution, weighting function, and the geometry. Instead of linear basis functions, $N^A(\xi)$ will represent quadratic basis functions:

$$x_e(\xi) = \sum_{A=1}^{3} N^A(\xi) x_{e^A}$$

Differentiating term by term gives

$$\begin{aligned} x_{,\xi} &= \sum_{A=1}^{3} N_{,\xi} x_e^A \\ &= \frac{(2\xi - 1)x_e^1}{2} - 2\xi x_e^2 + \frac{(1 + 2\xi)x_e^3}{2} \\ &= \frac{x_e^3 - x_e^1}{2} + \xi \left(x_e^1 - 2x_e^2 + x_e^3 \right) \end{aligned}$$

The first term on the right hand side can be written as the length of subdomain divided by 2. The second term has a surprising simplification. x_e^2 must lie in the center of physical subdomain. In this case,

$$x_e^2 = \frac{x_e^3 + x_e^1}{2}$$

must hold.

Then, the second term on the right must be zero, and we obtain the same relationship we had when the basis functions were linear.

$$x_{,\xi} = \frac{h^e}{2}$$

The tangent of the geometric mapping is constant since we have an affine map.

Using the invertibility of our map, it must be true that

$$\xi_{,x} = \frac{2}{h^e}$$

So the weak form integral can be written as

$$\int_{\Omega^e} \left[\sum_{A=1}^{3} N_{,\xi}^A \frac{2}{h^e} c_e^A \right] EA \left[\sum_{A=1}^{3} N_{,\xi}^A \frac{2}{h^e} d_e^A \right] dx$$

4.6 Higher polynomial order basis functions - III

What about the differential dx? We can write dx as $(dx/d\xi)d\xi$ and we know that $dx/d\xi = x_{,\xi} = h^e/2$. So the integral becomes

$$\int_{\Omega^e} \left[\sum_{A=1}^{3} N^A_{,\xi} \frac{2}{h^e} c^A_e\right] EA \left[\sum_{A=1}^{3} N^A_{,\xi} d^A_e\right] d\xi$$

Assuming that $EA2/h^e$ is independent of ξ, we can write

$$\frac{2EA}{h^e} \sum_{A,B} c^A_e \left(\int_{\Omega^\xi} N_{,\xi} N^B_{,\xi}\, d\xi\right) d^B_e$$

We will use matrix vector notation to get rid of the sum over A and B.

$$\begin{bmatrix} c^1_e & c^2_e & c^3_e \end{bmatrix} \frac{2EA}{h^e} \begin{bmatrix} \int_{-1}^{1} N^1_{,\xi} N^1_{,\xi}\, d\xi & \int_{-1}^{1} N^1_{,\xi} N^2_{,\xi}\, d\xi & \int_{-1}^{1} N^1_{,\xi} N^2_{,\xi}\, d\xi \\ \int_{-1}^{1} N^2_{,\xi} N^1_{,\xi}\, d\xi & \int_{-1}^{1} N^2_{,\xi} N^2_{,\xi}\, d\xi & \int_{-1}^{1} N^2_{,\xi} N^3_{,\xi}\, d\xi \\ \int_{-1}^{1} N^3_{,\xi} N^1_{,\xi}\, d\xi & \int_{-1}^{1} N^3_{,\xi} N^2_{,\xi}\, d\xi & \int_{-1}^{1} N^3_{,\xi} N^3_{,\xi}\, d\xi \end{bmatrix} \begin{bmatrix} d^1_e \\ d^2_e \\ d^3_e \end{bmatrix}$$

Let's calculate the integrals within the 3×3 matrix.

$$\begin{aligned} \int_{-1}^{1} N^1_{,\xi} N^1_{,\xi}\, d\xi &= \int_{-1}^{1} \frac{4\xi^2 - 4\xi + 1}{4}\, d\xi = \frac{1}{4}\left[\frac{4\xi^3}{3} - \frac{4\xi^2}{2} + \xi\right]_{-1}^{1} \\ &= \frac{1}{4}\left(\frac{4}{3} + 1\right) - \frac{1}{4}\left(\frac{-4}{3} - 1\right) = \frac{7}{6} \end{aligned}$$

We did not need to include the terms for 4ξ. This is due to the fact that odd functions over symmetric domains integrate to zero.

$$\int_{-1}^{1} N^1_{,\xi} N^2_{,\xi}\, d\xi = \int_{-1}^{1} \frac{2\xi - 1}{2}(-2\xi)\, d\xi = \frac{1}{2}\int_{-1}^{1} -4\xi^2 + 2\xi\, d\xi = -2\left[\frac{\xi^3}{3}\right]_{-1}^{1} = \frac{-4}{3}$$

$$\int_{-1}^{1} N^1_{,\xi} N^3_{,\xi}\, d\xi = \int_{-1}^{1} \frac{(2\xi - 1)(2\xi + 1)}{4}\, d\xi = \frac{1}{4}\int_{-1}^{1} 4\xi^2 - 1\, d\xi = \frac{1}{4}\left[\frac{4\xi^3}{3} - \xi\right]_{-1}^{1} = \frac{1}{6}$$

$$\int_{-1}^{1} N_{,\xi}^2 N_{,\xi}^2 \, d\xi = \int_{-1}^{1} 4\xi^2 \, d\xi = \left[\frac{4\xi^3}{3} \right]_{-1}^{1} = \frac{8}{3}$$

$$\int_{-1}^{1} N_{,\xi}^2 N_{,\xi}^3 \, d\xi = \int_{-1}^{1} \frac{(-2\xi)(2\xi+1)}{2} \, d\xi = \frac{1}{2} \int_{-1}^{1} -4\xi^2 - 2\xi \, d\xi = \frac{1}{2} \left[\frac{-4\xi^3}{3} \right]_{-1}^{1} = \frac{-4}{3}$$

$$\int_{-1}^{1} N_{,\xi}^3 N_{,\xi}^3 \, d\xi = \frac{1}{4} \int_{-1}^{1} (2\xi+1)^2 \, d\xi = \frac{1}{4} \int_{-1}^{1} 4\xi^2 + 4\xi + 1 \, d\xi = \frac{1}{4} \left[\frac{4\xi^3}{3} + \xi \right]_{-1}^{1} = \frac{7}{6}$$

Inputting our results

$$\begin{bmatrix} c_e^1 & c_e^2 & c_e^3 \end{bmatrix} \underbrace{\frac{2EA}{h^e} \begin{bmatrix} 7/6 & -4/3 & 1/6 \\ -4/3 & 8/3 & -4/3 \\ 1/6 & -4/3 & 7/6 \end{bmatrix}}_{\underline{K_e}} \begin{bmatrix} d_e^1 \\ d_e^2 \\ d_e^3 \end{bmatrix}$$

4.6.1 Coding Assignment

`deal.II` uses an alternate numbering for local and global nodes than we did in class.

Locally, the leftmost node in each subdomain is always numbered 0. The right most node is always 1. The index starts from 2 for the leftmost mid-side node.

Globally, numbering begins at zero. Then goes to 1 at the rightmost node of Ω^1. Then increments by one for every mid-side node in Ω^1. The next index is for the rightmost node of Ω^2.

The formula for the Lagrange polynomials is now

$$\prod_{B=0, B\neq A}^{N_{ne}-1} \frac{\xi - \xi^B}{\xi^A - \xi^B}$$

We have simply shifted the index by 1.

4.7 The matrix-vector equations for quadratic basis functions - I - I

We worked out the form of the stiffness matrix for a general element. We will return to work out integral which will give us the forcing vector.

$$\int_{\Omega^e} w^h f A \, dx = \int_{\Omega^e} \left(\sum_A N^A c_e^A \right) f A \, dx$$

We know that x is ultimately parameterized by ξ. And our formulation is isoparameteric. So can write our integral over the bi-unit domain.

$$= \sum_A c_e^A \int_{\Omega^\xi} N^A f A \frac{dx}{d\xi} \, d\xi = \sum_A c_e^A \int_{\Omega^e} N^A f A \frac{h^e}{2} \, d\xi$$

Again, we are considering the case where f and A are both uniform over Ω^e. So we can pull them out of the integral.

$$\sum_{A=1}^{3} c_e^A \frac{f A h^e}{2} \int_{-1}^{1} N^A d\xi$$

In matrix vector notation:

$$\begin{bmatrix} c_e^1 & c_e^2 & c_e^3 \end{bmatrix} \frac{f A h^e}{2} \begin{bmatrix} \int_{-1}^{1} \frac{\xi(\xi-1)}{2} \, d\xi \\ \int_{-1}^{1} (1-\xi^2) \, d\xi \\ \int_{-1}^{1} \frac{\xi(\xi+1)}{2} \, d\xi \end{bmatrix}$$

Lets compute the integrals within the column vector.

$$\int_{-1}^{1} \frac{\xi(\xi-1)}{2} \, d\xi = \frac{1}{2} \left[\frac{\xi^3}{2} \right]_{-1}^{1} = \frac{1}{3}$$

$$\int_{-1}^{1} (1-\xi^2) \, d\xi = \left[\xi - \frac{\xi^3}{3} \right]_{-1}^{1} = \frac{4}{3}$$

$$\int_{-1}^{1} \frac{\xi(\xi+1)}{2} \, d\xi = \frac{1}{2} \left[\frac{\xi^3}{3} \right]_{-1}^{1} = \frac{1}{3}$$

So we have

$$\begin{bmatrix} c_e^1 & c_e^2 & c_e^3 \end{bmatrix} \frac{fAh^e}{2} \begin{bmatrix} 1/3 \\ 4/3 \\ 1/3 \end{bmatrix}$$

The next step is the assembly of the global matrix vector equations. In this and previous sections, we learned how to write the stiffness matrix and force vector integrals for a general element.

$$\sum_{e=1}^{N_{el}} \int_{\Omega^e} w^h_{,x} \underbrace{\sigma^h}_{Eu^h_{,x}} A \, dx = \sum_{e=1}^{N_{el}} \left[\int_{\Omega^e} w^h f A \, dx \right] + w^h(L) tA$$

Recall in the Dirichlet-Neumann problem, we have $u^h(0) = u_g$, and that $w^h(0) = 0$. In this case, we happen to have $u_g = 0$.

This implies that for element 1, since $w^h(0) = 0$, we can write

$$w_1^h(\xi) = \sum_{A=2}^{3} N^A c_1^A$$

Recall that for the element e, it consists of three nodes, x_{2e-1}, x_e, and x_{e+1}. We can write the sum as

$$w_1^h(\xi) = N^2 c_{2e} + N^3 c_{2e+1} = N^2 c_2 + N^3 c_3, e = 1$$

We write the contribution from element 1 as separate terms. The complete matrix vector weak form written in global node index is

$$\begin{bmatrix} c_2 & c_3 \end{bmatrix} \frac{2EA}{h^1} \begin{bmatrix} -4/3 & 8/3 & -4/3 \\ 1/6 & -4/3 & 7/6 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix}$$

$$+ \sum_{e=2}^{N_{el}} \begin{bmatrix} c_{2e-1} & c_{2e} & c_{2e+1} \end{bmatrix} \frac{2EA}{h^e} \begin{bmatrix} 7/6 & -4/3 & 1/6 \\ -4/3 & 8/3 & -4/3 \\ 1/6 & -4/3 & 7/6 \end{bmatrix} \begin{bmatrix} d_{2e-1} \\ d_{2e} \\ d_{2e+1} \end{bmatrix}$$

$$= \begin{bmatrix} c_2 & c_3 \end{bmatrix} \frac{fAh^1}{2} \begin{bmatrix} 4/3 \\ 1/3 \end{bmatrix} + \sum_{e=2}^{N_{el}} \begin{bmatrix} c_{2e-1} & c_{2e} & c_{2e+1} \end{bmatrix} \frac{fAh^e}{2} \begin{bmatrix} 1/3 \\ 4/3 \\ 1/3 \end{bmatrix} + c_{2N_{el}+1} tA$$

Remember, we can write $w^h(L)$ as $c_{2N_{el}+1}$ is due to Kronecker delta property of our basis functions.

4.8 The matrix-vector equations for quadratic basis functions - I - II

The next step in the process is to get rid of the summation over the elements, and combine those results with the contribution from the first element. So we have

$$\begin{bmatrix} c_2 & c_3 & c_4 & c_5 & c_6 & \dots & c_{2N_{el}-1} & c_{2N_{el}} & c_{2N_{el}+1} \end{bmatrix}$$

$$\frac{2EA}{h^e}\begin{bmatrix} -4/3 & 8/3 & -4/3 & & & & & & \\ 1/6 & -4/3 & (7/6+7/6) & -4/3 & 1/6 & & & & \\ & & -4/3 & 8/3 & -4/3 & & & & \\ & & 1/6 & -4/3 & 7/6+\dots & & & & \\ & & & & & & & & \\ & & & & & \cdots+7/6 & -4/3 & 1/6 & \\ & & & & & -4/3 & 8/3 & -4/3 & \\ & & & & & 1/6 & -4/3 & 7/6 & \end{bmatrix}\begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ \vdots \\ d_{2N_{el}-1} \\ d_{2N_{el}} \\ d_{2N_{el}+1} \end{bmatrix}$$

Aside from c_2 and c_3, which has their own contributions to the matrix. Every c_e contributes to adding an element in the stiffness matrix at row $e-1$, and column e.

This encompasses the left hand side of the integral. For the right hand

side

$$\begin{bmatrix} c_2 & c_3 & c_4 & c_5 & c_6 & \dots & c_{2N_{el}-1} & c_{2N_{el}} & c_{2N_{el}+1} \end{bmatrix} \left(\frac{fAh^e}{2} \begin{bmatrix} 4/3 \\ 1/3+1/3 \\ 4/3 \\ 1/3+1/3 \\ \vdots \\ \cdots + 1/3 \\ 4/3 \\ 1/3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ tA \end{bmatrix} \right)$$

The c degrees of freedom row vectors has length $2N_{el}$. The global stiffness matrix has dimensions $2N_{el} \times 2N_{el}+1$, the d degrees of freedom column vector has dimensions $2N_{el}+1 \times 1$. On the right hand side, each column vector in the bracket has dimension $2N_{el} \times 1$.

4.9 The matrix-vector equation for quadratic basis functions

We accounted for the Dirichlet boundary condition in our global matrix vector form in the d column vector. The Kronecker delta property of our basis functions lets us say that $d_1 = u_g$.

We make the observation that d_1 multiplies with the first column of the stiffness matrix. We will move that to the right hand side. What we have

then is

$$\underbrace{\begin{bmatrix} c_2 & c_3 & c_4 & c_5 & c_6 & \dots & c_{2N_{el}-1} & c_{2N_{el}} & c_{2N_{el}+1} \end{bmatrix}}_{\underline{c}^\intercal}$$

$$\underbrace{\frac{2EA}{h^e}\begin{bmatrix} 8/3 & -4/3 & & & & & & & \\ -4/3 & (7/6+7/6) & -4/3 & 1/6 & & & & & \\ & -4/3 & 8/3 & -4/3 & & & & & \\ & 1/6 & -4/3 & 7/6+\dots & & & & & \\ & & & & & & & & \\ & & & & \cdots+7/6 & -4/3 & 1/6 & & \\ & & & & -4/3 & 8/3 & -4/3 & & \\ & & & & 1/6 & -4/3 & 7/6 & & \end{bmatrix}}_{\underline{K}} \underbrace{\begin{bmatrix} d_2 \\ d_3 \\ d_4 \\ d_5 \\ \vdots \\ d_{2N_{el}-1} \\ d_{2N_{el}} \\ d_{2N_{el}+1} \end{bmatrix}}_{\underline{d}} =$$

$$\underbrace{\begin{bmatrix} c_2 & c_3 & c_4 & c_5 & c_6 & \dots & c_{2N_{el}-1} & c_{2N_{el}} & c_{2N_{el}+1} \end{bmatrix}}_{\underline{c}^\intercal}$$

$$\underbrace{\left(\frac{fAh^e}{2}\begin{bmatrix} 4/3 \\ 1/3+1/3 \\ 4/3 \\ 1/3+1/3 \\ \vdots \\ \cdots+1/3 \\ 4/3 \\ 1/3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ tA \end{bmatrix} + \frac{2EA}{h^e} d_1 \begin{bmatrix} -4/3 \\ 1/6 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \end{bmatrix} \right)}_{\underline{F}}$$

The $\underline{c}^\intercal$ is still $2N_{el}$. $\underline{K}$ is now $2N_{el} \times 2N_{el}$. $\underline{d}$ is now $2N_{el} \times 1$. The $\underline{F}$ vector is still $2N_{el} \times 1$.

Writing it out:

$$\underline{c}^\intercal \underline{K} \underline{d} = \underline{c}^\intercal \underline{F}$$

1. $\underline{K}$ matrix components are different from the linear case.

2. Bandwidth of $\underline{K}$ is 5 due to use of quadratic basis functions. (Recall when we used linear functions, $\underline{K}$ had a banded tri-diagonal structure)

3. The midside nodes have a larger contribution to $\underline{F}$ due to the use of quadratic basis functions.

In how we have written the equations, we have assumed that h^e is the same for all elements. So in effect, it is can be simply written as h.

4.10 The matrix-vector equations for quadratic basis functions - II - II

We will briefly look at what happens when the problem is a Dirichlet-Dirichlet problem.

We want to find $u^h \in \mathcal{S}^h \subset \mathcal{S}, \mathcal{S}^h = \{u^h \in H^1(\Omega) | u^h(0) = u_0, u^h(L) = u_L\}$ such that $\forall\ w^h \in \mathcal{V}^h \subset \mathcal{V}, \mathcal{V}^h = \{w^h \in H^1(\Omega) | w^h(0) = 0, w^h(L) = 0\}$, the following holds

$$\int_{\Omega^e} w^h_{,x} \sigma^L A \, dx = \int_{\Omega} w^h f A \, dx$$

Similar to the Dirichlet-Neumann problem, the first element as no $A = 1$ contribution to the weighting function. Different compared to the Dirichlet Neumann problem, the last element will have no contribution from the $A = 3$ basis function.

The global matrix vector weak form is

$$\begin{bmatrix} c_2 & \dots & c_{2N_{el}} \end{bmatrix} \frac{2EA}{h}$$

$$\begin{bmatrix} -4/3 & 8/3 & -4/3 & & & & & \\ 1/6 & -4/3 & (7/6+7/6) & -4/3 & 1/6 & & & \\ & & -4/3 & 8/3 & -4/3 & & & \\ & & 1/6 & -4/3 & 7/6+\dots & & & \\ & & & & & \cdots+7/6 & -4/3 & 1/6 \\ & & & & & -4/3 & 8/3 & -4/3 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{2N_{el}+1} \end{bmatrix}$$

$$= \begin{bmatrix} c_2 & \dots & c_{2N_{el}} \end{bmatrix} \frac{fAh}{2} \begin{bmatrix} 4/3 \\ 2/3 \\ 4/3 \\ \vdots \\ 4/3 \end{bmatrix}$$

Notice that $\underline{c}^\intercal$ is $1 \times 2N_{el} - 1$. The first and last nodes do not contribute to the weighting function. $\underline{K}$ is $2N_{el} - 1 \times 2N_{el} + 1$, it is missing the last row compared to the $\underline{K}$ matrix for Dirichlet Neumann problem .

Since $d_1 = u_0$ and $d_{2N_{el}+1} = u_L$, which are given by our Dirichlet boundary conditions, and that they respectively multiply the first and last column of the large matrix, we will move it to the right hand side.

$$\begin{bmatrix} c_2 & \dots & c_{2N_{el}} \end{bmatrix} \frac{2EA}{h}$$

$$\begin{bmatrix} 8/3 & -4/3 & & & & & \\ -4/3 & (7/6+7/6) & -4/3 & 1/6 & & & \\ & -4/3 & 8/3 & -4/3 & & & \\ & 1/6 & -4/3 & 7/6+\dots & & & \\ & & & & & \cdots+7/6 & -4/3 \\ & & & & & -4/3 & 8/3 \end{bmatrix} \begin{bmatrix} d_2 \\ d_3 \\ \vdots \\ d_{2N_{el}-1} \\ d_{2N_{el}} \end{bmatrix}$$

$$= \begin{bmatrix} c_2 & \dots & c_{2N_{el}} \end{bmatrix} \left(\frac{fAh}{2} \begin{bmatrix} 4/3 \\ 2/3 \\ 4/3 \\ \vdots \\ 4/3 \end{bmatrix} + \frac{2EA}{h} \underbrace{d_1}_{u_0} \begin{bmatrix} 4/3 \\ 1/6 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \frac{2EA}{h} \underbrace{d_{2N_{el}+1}}_{u_L} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1/6 \\ -4/3 \end{bmatrix} \right)$$

The final $\underline{K}$ matrix is $2N_{el}-1 \times 2N_{el}-1$. $\underline{d}$ is $2N_{el}-1 \times 1$. $\underline{F}$ is $2N_{el}-1 \times 1$.

Concisely,

$$\underline{c}^\intercal \underline{K}\underline{d} = \underline{c}^\intercal \underline{F} \implies \underline{c}^\intercal (\underline{K}\underline{d} - \underline{F}) = 0$$

In the integral statement of the weak form, we said that it must hold for all $w^h \in \mathcal{V}^h$. The refection of this statement is in $c^\intercal$. In the Dirichlet-Neumann case, the matrix-vector form must hold for all $\underline{c}^\intercal \in \mathbb{R}^{2N_{el}}$. In the Dirichlet-Dirichlet case, it must hold for all $\underline{c}^\intercal \in \mathbb{R}^{2N_{el}-1}$.

By standard argument, we can show that

$$\underline{K}\underline{d} = \underline{F}$$

and this is final equation.

4.11 Numerical Integration – Gaussian Quadrature

In the previous sections, we computed the integrals involved in the weak form by hand. But for complicated basis function, this might not be very suitable. Further, if E, f, or A are dependent on position, the integrals can be difficult.

We will take the approach of the Gaussian Quadrature for numerical integration. It has been shown that this method is optimal for integration polynomials; it is possible to exactly integrate polynomials of certain orders.

We want to find what the following integral evaluates to:

$$\int_{-1}^{1} g(\xi)\, d\xi$$

The approach is to rewrite integral as a sum

$$\sum_{\ell=1}^{N_{int}} g(\xi_\ell) w_\ell$$

N_{int} represents the number of integration points. g is evaluated at a point ξ that is indexed by ℓ. ξ_ℓ is an integration point. w_ℓ is a weight ascribed to the integration point.

The sum we have written is the general form of a quadrature rule.

There is a condition to the weights. We require that

$$\sum_{\ell=1}^{N_{int}} w_\ell = 2$$

The reason is: if g is a constant C, then the integral over -1 to 1 will result in $2g$. So

$$\int_{-1}^{1} C\, d\xi = C \sum_{\ell=1}^{N_{int}} w_\ell = 2C$$

This ensures that we can at least integrate constants exactly!

Here are the integration rules:

$$
\begin{array}{lll}
N_{int} = 1 & \xi_1 = 0 & w_1 = 2 \\
N_{int} = 2 & \xi_1 = \dfrac{-1}{\sqrt{3}} & w_1 = 1 \\
 & \xi_2 = \dfrac{1}{\sqrt{3}} & w_2 = 1 \\
N_{int} = 3 & \xi_1 = -\sqrt{\dfrac{3}{\sqrt{5}}} & w_1 = \dfrac{5}{9} \\
 & \xi_2 = 0 & w_2 = \dfrac{8}{9} \\
 & \xi_3 = \sqrt{\dfrac{3}{5}} & w_3 = \dfrac{5}{9} \\
N_{int} = \ldots & &
\end{array}
$$

The Gaussian quadrature with N_{int} points exactly integrates polynomials of order $\leq 2N_{int} - 1$. For example, $N_{int} = 3$ can at most exactly integrate a polynomial of degree 5.

4.11.1 Coding assignment 1 (functions "assemble_system")

Here we give some clarification on how `Flocal` in `assemble_system` will be computed. Recall the definition of the local force vector. Here, the only assumption we are making is that the cross section of the rod is uniform, and that each element length is uniform.

$$\texttt{Flocal}[A] = \frac{Ah^e}{2} \int_{-1}^{1} N^{\texttt{A}}(\xi) f(x(\xi))\, d\xi$$

The forcing function was written as $f(x(\xi))$. We know that f is a function of x. We can interpolate x from ξ by

$$x(\xi_{\texttt{q}}) = \sum_{A=0}^{\texttt{dofs\_per\_elem}} N^A(\xi_{\texttt{q}}) x^A$$

where x^A is the value of x at node `A`. To find x given A, use `nodeLocations[local_dof_indices[A]]`. `local_dof_indices` maps a local

node number to a global node number. `nodeLocation` maps a global node number to an $x-$coordinate value.

In practice, we will perform this integral using the Gaussian quadrature method.

$$\texttt{Flocal[A]} = \frac{Ah^e}{2} \sum_{\texttt{q=0}}^{\texttt{quadRule}-1} N^A(\xi_\texttt{q}) f(x(\xi_\texttt{q})) w(\xi_\texttt{q})$$

For `Klocal[A][B]`, recall the definition

$$\texttt{Klocal[A][B]} = \frac{2EA}{h^e} \int_{-1}^{1} N^\texttt{A}_{,\xi} N^\texttt{B}_{,\xi} \, d\xi$$

Lastly, we will perform the process of finite element assembly.

```
F[local_dof_indices[A]]+=Flocal[A];
```

For the global stiffness matrix, `K` is as defined a sparse matrix. So we will need to do

```
K[local_dof_indices[A]][local_dof_indices[B]]+=Klocal[A][B];
```

5 Analysis of the finite element method

5.1 Norms - I

In the past weeks, we have stated the 1D linear elliptic PDE in completeness. Before we move on to high dimensional problems, we will use this week to understand more about the mathematical basis of the finite element method.

We will understand why the finite element method works, what its special properties are, and obtain a high level view of why finite element convergence works.

We will talk about norms. But before that, consider the finite dimensional trial solution, u^h. We know that the trial solution over an element can be interpolated using

$$u_e^L = \sum_A N^A d_e^A$$

Figure 5: Quadratic Lagrange polynomial basis functions plotted over each physical element Ω^e. Its clear that $d_A = 1$ for this case (d_A is in global numbering).

The finite dimensional trial solution is continuous. We say that it is "C zero on omega": $C^0(\Omega)$.

The derivative of the trial solution within each element is continuous. We can say that $u_{e,x}^A$ is "C zero on omega e": $C^0(\Omega^e)$. The derivative in general is discontinuous over the entire physical domain. There are points of discontinuity of over element boundaries. So $u_{,x}^h$ is not $C^0(\Omega)$.

The Lagrange polynomial basis function have been constructed to be only $C^0(\Omega)$; they are not $C^n(\Omega)$ for $n > 0$.

In general, a function is in $C^n(\Omega)$ if its derivatives up to order n are con-

tinuous on Ω.

However, is u^h and its first derivative bounded? In other words, is $u^h \in H^1(\Omega)$?

$$\int_\Omega \left(u^h\right)^2 m(\Omega)^{2/n_{sd}} \left(u^h_{,x}\right)^2 \, dx < \infty$$

where n_{sd} is the number of spatial dimensions.

It turns out that while the first derivative is discontinuous, its remains square integrable.

5.1.1 Coding assignment 1 (functions: "solve" to "l2norm_of_error")

By definition, the L_2 norm of the error is

$$\|e\| = \left[\int_\Omega e^2 \, dx\right]^{1/2}$$

In the actual `C++` implementation,

$$\begin{aligned}
\|e\| &= \left[\sum_{\texttt{elem}} \int_{\Omega^e} e^2 \, dx\right]^{1/2} \\
&= \left[\sum_{\texttt{elem}} \int_{-1}^{1} e^2 \frac{h^e}{2} \, d\xi\right]^{1/2} \\
&= \left[\sum_{\texttt{elem}} \sum_{\texttt{q}} \left(u(x(\xi_q)) - u^h(x(\xi_q))\right)^2 \frac{h^e}{2} w(\xi_q)\right]^{1/2}
\end{aligned}$$

5.2 Norms - II

We will denote the H^1-norm of a function v to be

$$\|v\|_1 = \left[\frac{1}{m(\Omega)^{1/n_{sd}}} \int_\Omega v^2 + m(\Omega)^{2/n_{sd}} \left(v_{,x}\right)^2 \, dx\right]^{1/2}$$

We have defined the H^1-Hilbert norm of v. This is an example of more general norms called Sobolev norms. We can extend this to define the H^n-norm by including the first n derivatives.

The H^0-norm of the function v is then

$$\|v\|_0 = \left[\frac{1}{m(\Omega)^{1/n_{sd}}} \int_\Omega v^2 \, dx \right]^{1/2}$$

This is completely equivalent to the L^2-norm (L^2 is the space of square integrable functions).

We define the energy norm of v to be

$$\left[\int_\Omega v_{,x} E v_{,x} \, dx \right]^{1/2}$$

The name comes from structural mechanics. It resembles the formula for the strain energy of v.

The result of the equivalence of norm is as follows: for constants c_1 and c_2, and energy norm of the function v can be bounded by the H^1 norm of v.

$$c_1 \|v\|_1 \leq \left[\int_\Omega v_{,x} E v_{,x} \, dx \right]^{1/2} \leq c_2 \|v\|_1$$

This tells us that the H^1 and the energy norm of v behaves in the same away. It is in this way we say that the two norms are equivalent.

We define the inner product notation.

$$(w, f) = \int_\Omega w f \, dx$$

This is the L^2 inner product of w and f. This is the sort of inner product that appears in the force vector.

We also define the bilinear form notation.

$$a(w, u) = \int_\Omega w_{,x} E u_{,x} \, dx$$

This is the form of the integral that appears in the stiffness matrix. It is "bilinear" in that it is linear in w and u.

When $w = u = v$, then $a(v, v)$ is the square of the energy norm. We will use the notation of $a(v, v)$ of the energy norm squared.

We can always establish the equivalence between the H^1 and energy norm. This relies on two facts. The H^1 norm and the energy norm exists iff

$$\|v\|_1 < \infty$$
$$a(v, v) < \infty$$

these conditions would implies the second fact that

$$m(\Omega) < \infty$$

so Ω is finite.

5.3 Consistency of the finite element method

We will look at the consistency and best approximation property of the finite element method.

Recall the weak form

$$\underbrace{\int_\Omega w_{,x} E u_{,x} A \, dx}_{a(w,u)} = \underbrace{\int_\Omega w f A \, dx}_{(w,f)} + w(L) t A$$

We can recognize that the left hand term is the the energy norm squared, and the first right hand term is the inner product between w and f. There are also additional factors of A within the integrals. Our original definitions did not include these factors but we could think of A as a part of dx, and define the norms this way.

For the right most term, we can define it as $(w, t)_L$. This is the inner product of w and t evaluated at L.

So we can write the weak form in an abstract notation:

$$a(w, u) = (w, f) + (w, t)_L$$

We can write the finite dimensional weak form in the same manner.

$$a(w^h, u^h) = (w^h, f) + (w^h, t)_L$$

We stated that the infinite dimensional weak form holds for all $w \in \mathcal{V}$. Since $w^h \in \mathcal{V}^h \subset \mathcal{V}$, so the infinite dimensional weak form relation also holds w^h (u remains infinite dimensional). So,

$$a(w^h, u) = (w^h, f) + (w^h, t)_L$$

This is called the consistency condition. The equation looks as if we had simply substituted the the exact solution into the finite dimension weak form. This condition says that the finite element method admits the exact solution that it can find or accept. This is not the case with finite differences. We are no guaranteed to satisfy the finite difference equations if we substituted the exact solutions.

If we subtract the infinite dimensional weak form (with finite dimensional weighting functions) from the finite dimensional weak form, we get

$$\begin{aligned} a(w^h, u^h) - a(w^h, u) &= (w^h, f) - (w^h, f) + (w^h, t)_L - (w^h, t)_L \\ &= 0 \end{aligned}$$

Since the $a(w, u)$ is bilinear, this lets us say that

$$a(w^L, u^h - u) = 0$$

The expression $u^h - u$ gives the error, e, in the finite dimensional solution.

The idea of the bilinear form is a general idea of taking one function and projecting it onto another. $a(w^h, e) = 0$ says that the project of the error on the space $\mathcal{V}^h$ is zero. "The error lies outside of the space $\mathcal{V}^h$ since it has no projection left in $\mathcal{V}^h$." We can also say that the error is orthogonal to the space $\mathcal{V}^h$.

5.4 The best approximation property

Let $u^h \in \mathcal{S}^h$ be the finite element solution, $w^h \in \mathcal{V}^h$ be a weighting function, and $U^h \in \mathcal{S}^h = \{U^h \in H^1(\Omega) | U^h(0) = u_0\}$. Note that

$$U^h = u^h + w^h.$$

This is true since both u^h and w^h are H^1 functions by definition. u^h satisfies the Dirichlet boundary condition, and $w^h(0) = 0$.

Best approximation property:

$$a(e, e) \leq a(U^h - u, U^h - u) \ \forall \ U^h \in \mathcal{S}^h$$

The finite element solution minimizes the energy norm of $U^h - u$ over all members $U^h \in \mathcal{S}^h$. When we use the the energy norm as an error estimate, the finite element picks the solution that minimizes this error.

We can now prove this theorem. Consider $a(e+w^h, e+w^h)$. We can expand this as follows

$$a(e, e) + a(e, w^h) + a(w^h, e) + a(w^h, w^h).$$

This looks just like the expansion of a perfect square. Since the bilinear form is also symmetric, the two middle terms are equivalent. Thus,

$$a(e, e) + \underbrace{2a(w^h, e)}_{=0} + a(w^h, w^h)$$

The central terms are zero due to the consistency relation. So,

$$a(e + w^h, e + w^h) = a(e, e) + a(w^h, w^h)$$

A fundamental property of norms is that they must be ≥ 0. $a(e + w^h, e + w^h)$ is the sum of two non-negative terms. So

$$a(e, e) \leq a(e + w^h, e + w^h)$$

The term $a(e + w^h, e + w^h)$ is $a(u^h - u + w^h, u^h - u + w^h)$ by definition. We can combine the terms u^h and w^h to result in U^h. We recover

$$a(e, e) \leq a(U^h - u, U^h - u)$$

5.5 The "Pythagorean Theorem"

We will state a corollary of the consistency property.

$$a(u,u) = a(u^h,u^h) + a(e,e),$$

which holds for when $\mathcal{S}^h = \mathcal{V}^h$. This is true when Dirichlet boundary conditions are homogeneous.

We will prove this corollary. By definition, $u = u^h - e$. So

$$\begin{aligned} a(u,u) &= a(u^h - e, u^h - e) \\ &= a(u^h,u^h) - a(u^h,e) - a(e,u^h) + a(e,e) \\ &= a(u^h,u^h) - 2a(u^h,e) + a(e,e) \end{aligned}$$

By the consistency relation, which holds for all functions in the space $\mathcal{V}^h$, the middle term is zero (since u^h is and w^h belong to the same function space $\mathcal{V}^h$).

$$\begin{aligned} a(u,u) &= a(u^h - e, u^h - e) \\ &= a(u^h,u^h) - a(u^h,e) - a(e,u^h) + a(e,e) \\ &= a(u^h,u^h) + a(e,e) \end{aligned}$$

This implies a second corollary, that the finite element solution underestimates energy norm of the problem.

$$a(u^h,u^h) \leq a(u,u)$$

This property offers a sense the control over the solution. In the case of an elasticity property, the physical interpretation is that the finite element strain energy is an underestimate.

5.6 Sobolev estimates and convergence of the finite element method

Before we can talk about the convergence of the finite element method, we need to know about Sobolev estimates.

We will reuse the symbol U^h. However, it will no longer belong to H^1 only.

$$U^h \in \mathcal{S}^h = \{U^h \in H^N(\Omega) | U^h(0) = u_0\}$$

U^h does not necessarily represent u^h.

Consider a special class of the functions U^h, such that at any node x_A where A is a global node numbering

$$U^h(x_A) = u^h(x_A) = d_A,$$

by the Kronecker delta property. However, this would mean that U^h is exactly equal to u^h.

Consider $\tilde{U}^h$, such that

$$\tilde{U}^h = u(x_A)$$

This means that $\tilde{U}^h$ as parameterized by x is nodally exact. $\tilde{U}^h$ belongs to $\mathcal{S}^h$. Thus over each element, $\mathcal{U}^h$ is represented by the basis functions we choose. It is also called an interpolate since we are taking the exact solution at each point, and using the basis functions to interpolate between the nodal values to get $\tilde{U}^h$.

The analysis of Sobolev spaces gives us the following result of the interpolation error estimate in Sobolev spaces, for a uniform element length h^e:

$$\|\tilde{U}^h - u\|_m \leq c\,(h^e)^\alpha\,\|u\|_r$$

Even though $\tilde{U}^h$ is nodally exact, we fail to the hit the exact solution over element. So $\tilde{U}^h - u$ is the interpolation error. This statement says that the H^m norm of the interpolation error has an upper bound given by the quantity on the right.

On the right hand side, c is a constant, r is the regularity of the exact solution, and α is an exponent that satisfies

$$\alpha = \min(k + 1 - m, r - m),$$

where k is the polynomial order of the finite dimensional basis functions.

The H^r norm of the exact solution u gives a measure of regularity (smoothness) of u. The sum of the integral of u and its r^{th} derivative over Ω is bounded.

If r is large, then the exact solution is very smooth. So α is essentially $k+1-m$.

$$\|\tilde{U}^h - u\|_m \leq c\,(h^e)^{k+1-m}\,\|u\|_r$$

Provided that $k+1-m > 0$, then $\|\tilde{U}^h - u\|_m$ tends to zero at a rate of $k+1-m$ as h^e tends to zero.

We can make $k+1-m$ positive by considering higher order basis functions, or only requiring a small m.

This is a property that comes entirely from Sobolev spaces.

5.7 Finite element error estimate

Recall the equivalence of the energy norm and the H^1 norm. Using abstract notation

$$c_1\|v\|_1 \leq a(v,v)^{1/2} \leq c_2\|v\|_1$$

This equivalence extends to general H^n norms.

$$c_1\|v\|_n \leq a(v,v)^{1/2} \leq c_2\|v\|_n$$

The result is that

$$\|e\|_n \leq \overline{c}\,(h^e)^{\alpha}\,\|u\|_r$$

Here, we are talking about the the finite element error as opposed to the interpolation error.

We will prove this result. By the equivalence of norms, the energy norm has a lower bound given by

$$c_1\|e\|_n \leq a(e,e)^{1/2}$$

Then, by the best approximation property, we can say that

$$c_1\|e\|_n \leq a(e,e)^{1/2} \leq a(U^h - u, U^h - u)^{1/2}$$

Recall that $\tilde{U}^h$ is a special instance of U^h. So the following also holds

$$c_1\|e\|_n \leq a(e,e)^{1/2} \leq a(\tilde{U}^h - u, \tilde{U}^h - u)^{1/2}$$

Applying the equivalence property again yields

$$c_1\|e\|_n \leq a(e,e)^{1/2} \leq a(\tilde{U}^h - u, \tilde{U}^h - u)^{1/2} \leq c_2\|\tilde{U}^h - u\|_n$$

Invoking the Sobolev interpolation error estimate, we have

$$c_2\|\tilde{U}^h - u\|_n \leq c_2 c\,(h^e)^\alpha\,\|u\|_r$$

We can collect these results to say that

$$\|e\|_n \leq \frac{c_2 c}{c_1}\,(h^e)^\alpha\,\|u\|_r$$

We can write the constants c_2c/c_1 as $\bar{c}$. We have recovered our original result.

$$\|e\|_n \leq \bar{c}\,(h^e)^\alpha\,\|u\|_r$$

1. For sufficiently smooth u, then $\alpha = \min(k+1-n, r-n) = k+1-n$.
2. Consider the case of $n = 1$. Then

   $$\|e\|_1 \leq \bar{c}\,(h^e)^k\,\|u\|_r$$

   the order of convergence of the H^1 norm of the error is equal to the polynomial order k as h^e tends to zero.
3. If we wanted to compute the L^2 norm of the error (equivalent to the H^0 norm), this requires the Aubin-Nitsche method. But here is the result

   $$\|e\|_{L_2} \leq C\,(h^e)^{k+1}\,\|u\|_r$$

We have summarized the key results in conventional finite element analysis. As we refine the mesh, our solution converges to the exact solution. In each of these error estimates, we have used an integral over the entire domain since we want to make sure we have control over the error over the entire domain Ω.

6 Variational principles

6.1 Functionals. Free energy - I

We will look at a method to derive the weak form for the 1D linear elliptic equation.

Consider the following motivation. The integral I,

$$I(u) = \int_\Omega \frac{1}{2} EA \left(u_{,x}\right)^2 \, dx$$

E represents the modulus, $u_{,x}$ is the strain, and the entire integral gives the strain energy in 1D linearized elasticity.

This allows us to construct the following integral, which we will denote as $\pi[u]$:

$$\pi[u] = \underbrace{\int_\Omega \frac{1}{2} EA \left(u_{,x}\right)^2 \, dx}_{\text{Strain energy}} - \underbrace{\int_\Omega fuA \, dx}_{\text{Work done by body force}} - \underbrace{tAu(L)}_{\text{Work done by traction}}$$

, where $u \in \mathcal{S} = \{u | u(0) = u_0\}$ and f, t, and constitutive relation $\sigma = Eu_{,x}$ are given.

Recognize that $\pi[u]$ is the Gibb's free energy for purely mechanical systems (or problems). It is also called the potential in the context of mechanics.

u was written in rectangular brackets. We want to think of the quantity $\pi[u]$ as no a function.

A function takes a point value of its argument, and returns another point value. A function $g(x) : \mathbb{R} \mapsto \mathbb{R}$. The function is what we get by connecting all the dots (in 1D).

π in our case is a mapping of a field, u, to the real numbers. π is an integral, so the input is actually the entire field u. Since we draw our function u from the space $\mathcal{S}$, π is then

$$\pi[u] : \mathcal{S} \mapsto \mathbb{R}$$

6.2 Functionals. Free energy - II

So for different displacement fields, π evaluates to different values.

π is a functional. Anything that takes an entire field as an argument is a functional. Some examples are integrals and derivatives.

π is our Gibb's free energy functional. Extrema of free energies characterize equilibrium states of systems. We are at an extrema whenever the derivative of free energy is equal to zero. If the second derivative of the free energy is equal to zero, we have a stable equilibrium. If the second derivative is less than zero, we have an unstable equilibrium. If the second derivative is equal to zero, then it is a neutrally stable equilibrium.

To be able to find an extrema of the functional $\pi[u]$, π must be sufficiently smooth. But what is an appropriate notion of a derivative of pi with respect to u? How can you differentiate with respect to an entire field?

6.3 Extremization of functionals

The method we will use is a variational method. It is called a variational method because we are considering variations.

w are always chose to vanish at the Dirichlet boundary so that the perturbed field u_ϵ does not violate the Dirichlet boundary condition. This is a principle in variational methods.

Consider the perturbed functional $\pi[u_\epsilon]$:

$$\frac{d}{d\epsilon}\pi[u_\epsilon]$$

this is the amount of variation in π for variation in u, having chosen the "form" of the variation coming from w. π is a function of ϵ.

How can we know how π varies around u? We can figure this out by setting $\epsilon = 0$. This becomes the variation in π with respect to u, at u. This is a functional derivative.

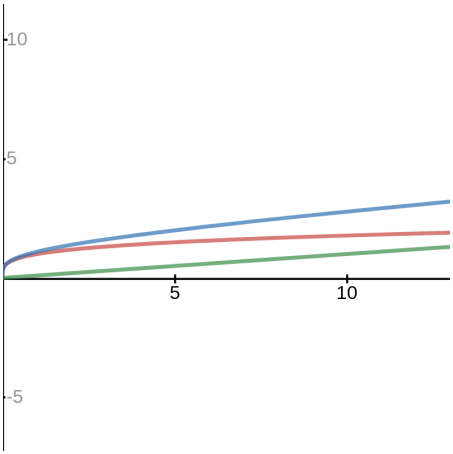

Figure 6: The red function represents the displacement field u, $u_\epsilon = u + \epsilon w(x)$ is in blue. The parameter ϵ is a real number. This implies that w, shown in green, is also a field, satisfying $w(0) = 0$, and belongs to the space $w \in \mathcal{V} = \{w|w(0) = 0\}$. The role of ϵ is the control the amount of variation we are applying to u.

6.4 Derivation of the weak form using a variation principle

We have defined what functions are; and the notion of a function derivative. We required this so we can talk about equilibrium of the free energy functional.

We want to find $u \in \mathcal{S} = \{u|u(0) = u_0\}$ such that for all variations of $w \in \mathcal{V} = \{w|w(0) = 0\}$

$$\left[\frac{d}{d\epsilon}\pi[u_\epsilon]\right]_{\epsilon=0} = 0$$

This means that

$$\frac{d}{d\epsilon}\left[\int_\Omega \frac{1}{2}EA\,(u+\epsilon w)_{,x}^2\,dx - \int_\Omega f(u+\epsilon w)A\,dx - tA(u+\epsilon w)(L)\right]_{\epsilon=0} = 0$$

Notice that there are no problem in take the derivative with respect to ϵ.

Integration results in a real number involving some contribution of ϵ.

$$\left[\int_\Omega \frac{1}{2} EA2\left(u+\epsilon w\right)_{,x} w_{,x}\, dx - \int_\Omega fwA\, dx - tAw(L)\right]_{\epsilon=0} = 0$$

Setting $\epsilon = 0$,

$$\int_\Omega EAu_{,x} w_{,x}\, dx - \int_\Omega fwA\, dx - tAw(L) = 0$$

Lets rearrange this

$$\int_\Omega w_{,x} \underbrace{Eu_{,x}}_{\sigma} A\, dx - \int_\Omega wfA\, dx - w(L)tA = 0$$

$$\int_\Omega w_{,x}\sigma A\, dx - \int_\Omega wfA\, dx - w(L)tA = 0$$

We recover the weak form (the statement must hold for all w).

When an extremization principle is available (extrema of free energy functional in this case), the weak form can be obtained using variational calculus. A variational principle exists.

Note how this variation is not appropriate for the physics of heat conduction and mass diffusion. The mathematics however works.

Variational principles exists for

- Elasticity at steady state
- Schrodinger equation at steady state

7 Linear, elliptic partial differential equations for a scalar variable in three dimensions. Heat conduction and mass diffusion at steady state

7.1 The strong form of steady state heat conduction and mass diffusion - I

We will now look at 3-D linear elliptic PDEs. The quantity we are solving for is a scalar. The canonical problems described include

- Steady state heat conduction
- Steady state mass diffusion

We will jump into it by stating the strong form. Let's visualize the domain in which are solving the problem.

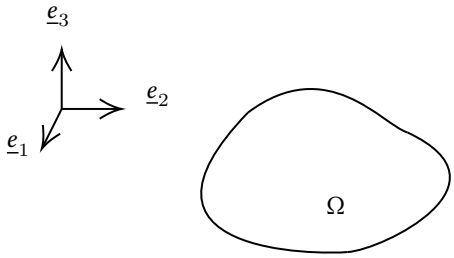

Figure 7: Shown are three Cartesian orthonormal basis vectors, $\{e_i\}$, $i = 1, 2, 3$. Ortho means that the three vectors are orthogonal to each other, and normal means that they are of normal magnitude. Ω is a region in three dimensional space, often named the continuum potatoe. Ω does no include its boundary, which we will denote as $\partial\Omega$.

The three orthonormal basis vectors are such that

$$\underline{e}_i \cdot \underline{e}_j = \delta_{ij}$$

For our purposes, Cartesian simply means that they are fixed.

Ω is open in $\mathbb{R}^3$.

In this setting, we want to find a field u, given $f(\underline{x})$, u_g, j_n, and the constitutive relation $j_i = -\kappa_{ij} u_{,j}$ $(i, j = 1, 2, 3)$, such that

$$-j_{i,i} = f$$

in Ω. The boundary conditions are that

$$u = u_g(\underline{x})$$

on $\partial\Omega_u$ (Dirichlet), and

$$-\underline{j} \cdot \underline{n} = j_n$$

on $\partial\Omega_j$ (Neumann).

The forcing function f in the 1-D case was a function of position. Now, it is parameterized by the position vector $\underline{x}$ as measured from the origin of our Cartesian basis.

What do $\partial\Omega_u$ and $\partial\Omega_j$ represent. They are two regions of the boundary. We will say that $\partial\Omega_u$ is open, and $\partial\Omega_j$ is the complement of $\partial\Omega_u$. So

$$\partial\Omega_u \cap \partial\Omega_j = \emptyset$$
$$\partial\Omega = \partial\Omega_u \cup \partial\Omega_j$$

(The first line says that the intersection between the two sets is the empty set. The second says that the boundary is the union between the two regions.)

7.2 The strong form of steady state heat conduction and mass diffusion - II

j_i (j sub i), is the flux vector in coordinate notation, where $i = 1, 2, 3$. The subscript simply denotes the components of the vector for our purposes.

In direct notation,

$$\underline{j} = \begin{bmatrix} j_1 \\ j_2 \\ j_3 \end{bmatrix}$$

To say that $\underline{j}$ is a vector, we can say that

$$\underline{j} \in \mathbb{R}^3$$

To state more about the PDE, consider heat conduction at steady state in 3-D. u then represents the temperature, $\underline{j}$ is the heat flux vector (the amount of heat crossing perpendicular to a unit area per unit time).

The constitutive relation

$$j_i = -\kappa_{ij} \underbrace{u_{,j}}_{\partial u / \partial x_j}$$

this is the Fourier law of heat conduction. $u_{,j}$ is a component of the temperature gradient. The coefficient in front is the heat conductivity tensor. A tensor, in our setting can be thought of the generalization of a vector.

In direct notation

$$\underline{\underline{\kappa}} = \begin{bmatrix} \kappa_{11} & \kappa_{12} & \kappa_{13} \\ \kappa_{21} & \kappa_{22} & \kappa_{23} \\ \kappa_{33} & \kappa_{32} & \kappa_{33} \end{bmatrix}$$

We will usually consider κ is symmetric.

$$\underline{\underline{\kappa}} = \underline{\underline{\kappa}}^{\mathsf{T}} \implies \kappa_{ij} = \kappa_{ji}$$

Kappa also has the property that it is positive semi-definite. if $\underline{\xi} \in \mathbb{R}^3$, then $\underline{\xi} \cdot \underline{\underline{\kappa}} \underline{\xi} \geq 0$ (it is semi-definite since the we allow $\underline{\xi} \cdot \underline{\underline{\kappa}} \underline{\xi} = 0$). The physical interpretation is that in some directions, this material is allowed to act like an insulator, where there is no conduction.

So if $\underline{\xi} \cdot \underline{\underline{\kappa}} \underline{\xi} \geq 0$, for $\underline{\xi} \neq \underline{0}$, then there is no heat conduction along ξ.

Another remark is negative sign in the statement of the Fourier law. This ensures that the heat flux vector is in the opposite direction of the temperature gradient. Since heat flows from hot to cold.

The boundary condition is such that

$$u = u_g(\underline{x})$$

on $\partial\Omega_u$. We are setting the temperature on this region $\partial\Omega_u$, which can be a non-uniform field.

For the Neumann condition,

$$-\underline{j} \cdot \underline{n} = j_n$$

This is the heat flux vector dotted with the negative of the unit outward normal. We are controlling normal component of the heat influx over the compliment of the Dirichlet boundary, or the Neumann boundary. In coordinate notation, we have

$$-j_i n_i = j_n$$

Even though we set up our 1-D problem in the context of elasticity, we did not jump straight into an example for three dimensional elasticity as the unknown quantity (displacement) is a vector field.

7.3 The strong form, continued

For completeness, let's look at the problem of mass diffusion.

- u: concentration, in mass per unit volume, or number of particles per unit volume. Or it may have been normalised with respect to a reference concentration. It which case it will be called a composition.
- $\underline{j}$: represent the mass (or number) flow perpendicular to a unit area per unit time. It's called mass flux or the number flux.
- $\underline{j} = -\underline{\kappa}\nabla u = -\underline{\kappa}\partial u/\partial \underline{x}$: $\underline{\kappa}$ is the diffusivity tensor, with the same property that as the heat conductivity tensor.
- $u = u_g$ on $\partial\Omega_u$: concentration boundary condition
- $-\underline{j} \cdot \underline{n} = j_n$ on $\partial\Omega_j$: mass influx boundary condition

Everything is really the same between the heat conduction and mass diffusion in 3 dimensions.

Lastly, let's write out the strong form fully in direct notation. We want to find the scalar u given u_g, j_n, f, the constitutive relation $\underline{j} = -\underline{\kappa}\underline{\nabla}u$, such that

$$-\underline{\nabla} \cdot \underline{j} = f$$

on Ω, with boundary conditions

$$u = u_g$$

on $\partial\Omega_u$, and

$$-\underline{j} \cdot \underline{n} = j_n$$

on $\partial\Omega_j$. Substituting the constitutive relation into the PDE, we get

$$-\underline{\nabla} \cdot \left(-\underline{\kappa}\underline{\nabla}u\right) = f$$

in Ω. If $\underline{\kappa}$ is spatially uniform (which means that $\underline{\kappa}$ is not a function of position), then f is equal to kappa contracted with the hessian of u:

$$\underline{\kappa} : \underline{\nabla}^2 u = f$$

Hessian matrix: If all second partial derivatives of a function $f(\underline{x})$ exists, then the Hessian matrix, $\underline{H}$, of f in coordinate notation is

$$H_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j} = f_{,x_i x_j}$$

The double dot denotes a contraction. It is an extension of the dot product to tensors.

$$\kappa_{ij} u_{,ij} = f$$

In the special case that $\kappa_{ij} = \kappa\delta_{ij}$ (the tensor is represented as scalar multiplied by the Kronecker delta tensor), then

$$\kappa\delta_{ij} u_{,ij} = f \implies \kappa u_{,ii} = f$$

in Ω, this is often called the Poisson equation. The same thing in direct notation is

$$\kappa \nabla^2 u = f$$

We did not write ∇^2 with an underline in this case. This means that it is simply the Laplacian operator and not a tensor. The Neumann boundary condition in this case is

$$-\underline{j} \cdot \underline{n} = j_n \implies \kappa \underline{\nabla} u \cdot \underline{n} = j_n$$

This is then a requirement on the normal gradient of temperature.

$\kappa_{ij} = \kappa \delta_{ij}$ is a common simplification. What we have here is called isotropic heat conduction. We are saying that the body is such that heat flows only in the direction of the temperature gradient. If we introduce a temperature in one direction, it does not induce heat flow in any other direction. It also says that the amount of heat flow we get by introducing a temperature in any direction is the same. The heat flow vector will align with the temperature gradient it self, given that we specified the same temperature gradient in the different direction.

7.4 The weak form

We are now ready to state the weak form for the 3-D linear elliptic PDE for a scalar unknown.

We want to find $u \in \mathcal{S} = \{u | u = u_g \text{ on } \partial\Omega_u\}$ giben u_g, j_n, f and the constitutive relation $j_i = -\kappa_{ij} u_{,j}$ such that for all $w \in \mathcal{V} = \{w | w = 0 \text{ on } \partial\Omega_u\}$ so that the following holds

$$\int_\Omega w_{,i} j_i \, dV = \int_\Omega w f \, dV - \int_{\partial\Omega_j} w j_n \, dS$$

We will show how we can obtain weak form starting from the strong form. Find u, given $f(\underline{x})$, u_g, j_n, $j_i = -\kappa_{ij} u_{,j}$ $(i, j = 1, 2, 3)$, such that

$$-j_{i,i} = f$$

in Ω. The boundary conditions are that

$$u = u_g(\underline{x})$$

on $\partial\Omega_u$, and

$$-\underline{j} \cdot \underline{n} = j_n$$

on $\partial\Omega_j$.

Consider $w \in \mathcal{V} = \{w | w = 0 \text{ on } \partial\Omega_u\}$. We will multiply the strong form of the PDE by w, and integrate by parts.

$$\int_\Omega -w j_{i,i}\, dV = \int_\Omega w f\, dV$$

Integration by parts is the application of the product rule and the divergence theorem.

$$\int_\Omega \underbrace{\left(-(w j_i)_{,i} + w_{,i} j_i\right)}_{-w j_{i,i}}\, dV = \int_\Omega w f\, dV$$

See that $w j_i$ is really a vector. The term $(w j_i)_{,i}$ is essentially the divergence of $w\underline{j}$. So we can apply the divergence theorem to that term and convert the volume integral to a surface integral.

$$-\int_{\partial\Omega} w j_i n_i\, dS + \int_\Omega w_i j_i\, dV = \int_\Omega w f\, dV$$

Taking the surface integral to the right hand side:

$$\int_\Omega w_i j_i\, dV = \int_\Omega w f\, dV + \int_{\partial\Omega_u} w j_i n_i\, dS + \int_{\partial\Omega_j} w j_i n_i\, dS$$

Since we have picked w to live in the space $\mathcal{V}$, satisfying homogeneous Dirichlet boundary conditions the surface integral over $\partial\Omega_u$ is zero.

Over $\partial\Omega_j$ we defined $j_i n_i$ to be equal to $-j_n$. Putting everything together, we arrive at the weak form.

$$\int_\Omega w_{,i} j_i\, dV = \int_\Omega w f\, dV - \int_{\partial\Omega_j} w j_n\, dS$$

We have shown that the weak form is implied by the strong form. We can show that the weak form also implies the strong form.

Recall the partial differential equation of the strong form, where $-j_{i,i} = f \implies -\underline{\nabla} \cdot \underline{j} = f$ in Ω. The interpretation of this equation is that the net influx over any point $\underline{x}$ in Ω is f.

7.5 The finite-dimensional weak form - I

We have stated the infinite dimensional weak form. We will now state the finite-dimensional weak form.

We want to find $u^h \in \mathcal{S}^h \subset \mathcal{S}$, where $\mathcal{S}^h$ is a finite dimensional function space,

$$\mathcal{S}^h = \{u^h \in H^1(\Omega) | u^h = u_g \text{ on } \partial\Omega_u\},$$

given u_g, j_n, f, $j_i^h = -\kappa_{ij} u^h_{,j}$

$$\int_\Omega w^h_{,i} j_i^h \, dV = \int_\Omega w^h f \, dV - \int_{\partial\Omega_j} w^h j_n \, dS$$

7.6 The finite-dimensional weak form - II

We will define the spaces $\mathcal{S}^h$ and $\mathcal{V}^h$ by partitioning Ω into subdomains $\Omega^e \subset \Omega \subset \mathbb{R}^3$, where $e = 1, \ldots, N_{el}$.

We define each Ω^e to be a open subset. So

$$\overline{\Omega} = \overline{\bigcup_{e=1}^{N_{el}} \Omega^e}$$

this equality holds when the over line represents applying closure.

In this sense, the intersection between any two subdomains is the empty set.

To define these partitions, will use the one that is easiest to get to given what we know about the one dimensional case. We will consider hexahedral

element subdomains, Ω^e, $e = 1, \ldots, N_{el}$. (Hexahedral means that these the elements resemble hexahedrons, which are polyhedrons with 6 faces.)

> **Polyhedron:** In geometry, a polyhedron (plural polyhedra or polyhedrons) is a three-dimensional shape with flat polygonal faces, straight edges and sharp corners or vertices.

We will consider trilinear basis functions over the subdomains. To do this, we will introduce nodes over the hexahedron, for a total of eight nodes. This is why they are sometimes called "eight node bricks".

We can say that

$$u_e^h = \sum_{A=1}^{N_{ne}=8} N^A(\underline{x}) d_e^A$$

and that

$$w_e^h = \sum_{A=1}^{N_{ne}=8} N^A(\underline{x}) c_e^A$$

In the 1-D case, we regarded every subdomain as a mapping from a parent domain, Ω^ξ. We can do the same here. We will do the same in this 3-D case.

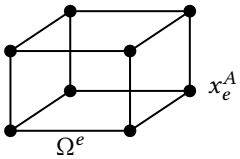

Figure 8: To identify each node, we can address each node as x_e^A, $A = 1, \ldots, 8$. This is a local numbering of nodes.

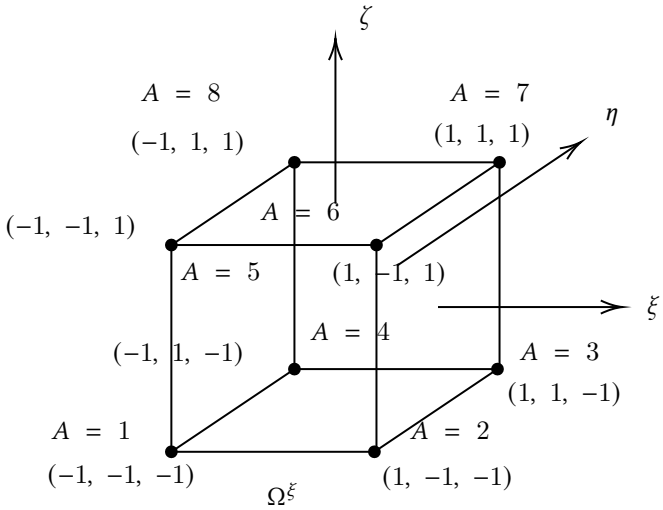

Figure 9: The parent domain is a bi-unit domain. Each node can be referred to as (ξ^A, η^A, ζ^A), where $A = 1, \ldots, 8$

7.7 Three-dimensional hexahedral finite elements

The following are our trilinear basis function

$$N^A(\xi, \eta, \zeta) = \frac{1}{8}(1 + \xi\xi^A)(1 + \eta\eta^A)(1 + \zeta\zeta^A)$$

$N^A(\xi, \eta, \xi)$ formulated this way by multiplying 1-D basis function to extend into multiple dimensions is called a tensor product function.

Explicitly:

$$N^1(\xi,\eta,\zeta) = \frac{1}{8}(1-\xi)(1-\eta)(1-\zeta)$$
$$N^2(\xi,\eta,\zeta) = \frac{1}{8}(1+\xi)(1-\eta)(1-\zeta)$$
$$N^3(\xi,\eta,\zeta) = \frac{1}{8}(1+\xi)(1+\eta)(1-\zeta)$$
$$N^4(\xi,\eta,\zeta) = \frac{1}{8}(1-\xi)(1+\eta)(1-\zeta)$$
$$N^5(\xi,\eta,\zeta) = \frac{1}{8}(1-\xi)(1-\eta)(1+\zeta)$$
$$N^6(\xi,\eta,\zeta) = \frac{1}{8}(1+\xi)(1-\eta)(1+\zeta)$$
$$N^7(\xi,\eta,\zeta) = \frac{1}{8}(1+\xi)(1+\eta)(1+\zeta)$$
$$N^8(\xi,\eta,\zeta) = \frac{1}{8}(1-\xi)(1+\eta)(1+\zeta)$$

These functions have the Kronecker delta property, where

$$N^A(\xi^B,\eta^B,\zeta^B) = \delta_{AB}$$

Another property allows us to represent constants

$$\sum_{A=1}^{N_{ne}} N^A(\xi,\eta,\zeta) = 1$$

These are lagrange polynomial basis functions in three dimensions. We see why they are called tri-linear, as they are linear in each of the three coordinate directions.

The map from Ω^ξ to Ω^e is obtained by interpolating $\underline{x}(\underline{\xi})$, where $\xi_1 = \xi$, $\xi_2 = \eta$, $\xi_3 = \zeta$.

$$\underline{x}_e(\underline{\xi}) = \sum_{A=1}^{N_{ne}} N^A(\underline{\xi})\underline{x}_e^A$$

$\underline{x}_e^A$ are the nodal coordinates in the physical domain.

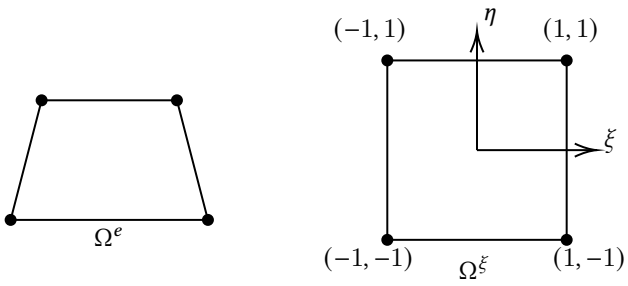

Figure 10: An arbitrary quadrilateral physical subdomain and a parent bi-unit domain. Any node in the bi-unit domain can be addressed as (ξ^A, η^A), where $A = 1, \ldots, N_{n_e}$.

Since we are interpolating our geometry, trial solution, and weighting function using the same basis functions, this is an isoparameteric formulation.

7.8 Aside: Insight to the basis functions by considering the two-dimensional case

We will try to gain some insight by looking at the bi-linear lagrange polynomrial basis functions for 2-D domains.

We will partition the domain into quadrilateral element subdomains, Ω^e, $= 1, \ldots, N_{n_e}$. (A quadrilateral is a closed 2-D shape with 4 sides.)

The basis functions parameterized by ξ and η are

$$N^1(\xi,\eta) = \frac{1}{4}(1-\xi)(1-\eta)$$
$$N^2(\xi,\eta) = \frac{1}{4}(1+\xi)(1-\eta)$$
$$N^3(\xi,\eta) = \frac{1}{4}(1+\xi)(1+\eta)$$
$$N^4(\xi,\eta) = \frac{1}{4}(1-\xi)(1+\eta)$$

All the basis functions take on the value of 1/4 in the origin of the bi-unit domain. These functions also satisfy the Kronecker delta property, and the

property that

$$\sum_{A=1}^{N_{ne}} N^A(\xi,\eta) = 1$$

The gradients of the bilinear basis functions are

$$\nabla N^1 = \begin{bmatrix} -(1-\eta)/4 & -(1-\xi)/4 \end{bmatrix}$$
$$\nabla N^2 = \begin{bmatrix} (1-\eta)/4 & -(1+\xi)/4 \end{bmatrix}$$
$$\nabla N^3 = \begin{bmatrix} (1+\eta)/4 & (1+\xi)/4 \end{bmatrix}$$
$$\nabla N^4 = \begin{bmatrix} -(1+\eta)/4 & (1-\xi)/4 \end{bmatrix}$$

7.9 Field derivatives. The Jacobian - I

We want to find $u^h \in \mathcal{S}^h$,

$$\mathcal{S}^h = \{u^h \in H^1(\Omega) | u^h = u_g \text{ on } \partial\Omega_u\},$$

given u_g, j_n, f, $j_i^h = -\kappa_{ij} u^h_{,j}$

$$\int_\Omega w^h_{,i} j_i^h \, dV = \int_\Omega w^h f \, dV - \int_{\partial\Omega_j} w^h j_n \, dS$$

Having defined a partition for the domain allows us to write the integral over the entire domain Ω into sum of integrals over subdomains.

$$\sum_e \int_{\Omega^e} w^h_{,i} j_i^h \, dV = \sum_e \int_{\Omega^e} w^h f \, dV - \sum_{e \in E} \int_{\partial\Omega_j^e} w^h j_n \, dS$$

Element $e \in E$ is such that

$$\partial\Omega^e \cap \partial\Omega_j \neq \emptyset$$

this simply means that E is the numbering for the set of elements whose surfaces coincide with the Neumann boundary.

Similar to the 1-D case, we also need to compute gradients here.

We will introduce a new notation that will make our formulas more concise.

$$\begin{bmatrix} \xi \\ \eta \\ \zeta \end{bmatrix} = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix}$$

We can write

$$u^h_{e,i} = \sum_{A=1}^{N_{ne}} N^A_{,i} d^A_e$$
$$w^h_{e,i} = \sum_{A=1}^{N_{ne}} N^A_{,i} c^A_e$$

To denote components of the $\underline{\xi}$, we will use an upper case index. ξ_I, for $I = 1, 2, 3$ denotes the 3 components of $\underline{\xi}$.

We have parameterized N with ξ. So

$$N^A_{,i} = \frac{\partial N^A}{\partial x_i} = \frac{\partial N^A}{\partial \xi_I} \frac{\partial \xi_I}{\partial x_i}$$

We are implying a summation over the index I. This is because N^A is parameterized by ξ_1, ξ_2, ξ_3. This is the so called Einstein summation convention.

It is simple to figure out the term $\partial N^A / \partial \xi_I$. The term $\partial \xi_I / \partial x_i$ is a little less trivial.

7.10 Field derivatives. The Jacobian - II

Recall the mapping

$$\underline{x}_e(\underline{\xi}) = \sum_{A=1}^{N_{ne}} N^A(\underline{\xi}) \underline{x}^A_e$$

In coordinate notation, it is

$$x_i(\underline{\xi}) = \sum_{A=1}^{N_{ne}} N^A(\underline{\xi}) x_{e_i}^A$$

Differentiating this term gives

$$\frac{\partial x_i}{\partial \xi_I} = \sum_{A=1}^{N_{ne}} N_{,I}^A x_{e_i}^A$$

This term does not immediately help us. Since what we want is actually the inverse. The mapping $\underline{x}(\underline{\xi})$ we have takes point $\underline{\xi}$ in the bi-unit domain, to a point $\underline{x}$ in the parent subdomain. This is a point-to-point vector map.

We can compute the gradient of the map, (as in, find the tangent map), by finding the Jacobian of the map, which is a tensor.

$$\underline{\underline{J}} = \frac{\partial \underline{x}}{\partial \underline{\xi}}$$

In coordinate notation, the Jacobian is

$$J_{iI} = \frac{\partial x_i}{\partial \xi_I}$$

We can represent $\underline{\underline{J}}$ as a matrix.

$$\underline{\underline{J}} = \begin{bmatrix} \frac{\partial x_1}{\partial \xi_1} & \frac{\partial x_1}{\partial \xi_2} & \frac{\partial x_1}{\partial \xi_3} \\ \cdots & \cdots & \frac{\partial x_2}{\partial \xi_3} \\ \cdots & \cdots & \frac{\partial x_3}{\partial \xi_3} \end{bmatrix}$$

The map $\underline{x}(\underline{\xi}) : \Omega^{\xi} \mapsto \Omega^e$ is a C^{∞} map. We can show that because of this property, there exists $\underline{\underline{J}}^{-1} = \partial \underline{\xi} / \partial \underline{x}$. Since we know the representation of every single entry into the $\underline{\underline{J}}$, we can compute $\underline{\underline{J}}^{-1}$.

$$\underline{\underline{J}}^{-1} = \begin{bmatrix} \frac{\partial \xi_1}{\partial_{x_1}} & \frac{\partial \xi_1}{\partial_{x_2}} & \frac{\partial \xi_1}{\partial_{x_3}} \\ \cdots & \cdots & \frac{\partial \xi_2}{\partial_{x_3}} \\ \cdots & \cdots & \frac{\partial \xi_3}{\partial_{x_3}} \end{bmatrix}$$

The components to the Jacobian is what we need.

$$J_{Ii}^{-1} = \frac{\partial \xi_I}{\partial x_i}$$

7.11 The integrals in terms of degrees of freedom

We will now consider the three integrals in the finite dimensional weak form explicitly.

Consider this first integral. We will substitute in the constitutive relation.

$$\int_{\Omega^e} w_{,i}^h j_i^h \, dV = -\int_{\Omega^e} w_{,i}^h \kappa_{ij} u_{,j}^h \, dV$$

Expanding the gradients of w and u in terms of our basis function and also expanding the divergence using chain rule:

$$-\int_{\Omega^e} \left(\sum_{A=1}^{N_{n_e}} N_{,i}^A c^A \right) \kappa_{ij} \left(\sum_{B=1}^{N_{n_e}} N_{,j}^B d_e^B \right) dV = -\int_{\Omega^e} \left(\sum_A N_{,I}^A \xi_{I,i} c_e^A \right) \kappa_{ij} \left(\sum_B N_{,J}^B \xi_{J,j} d_e^B \right) dV$$

Since the degrees of freedom are independent of position, we can pull them out of the integral.

$$-\sum_{A,B} c_e^A \left[\int_{\Omega^e} N_{,I}^A \xi_{I,i} \kappa_{ij} N_{,J}^B \xi_{J,j} \, dV \right] d_e^B$$

In this form, while we us an explicit summation symbol to run through the degrees of freedom, Einstein summation convention is implied in the integrand (for summing over indices of physics/parent domain coordinates).

The next step is to change variables into the parent subdomain to turn our integral into an integral within the parent subdomain.

It is a conventional result the following holds:

$$dV = \det \left[\underline{J}(\underline{\xi}) \right] \, dV^\xi$$

7.12 The integrals in terms of degrees of freedom - continued

Following our work from the last subsection, we realized the following is equivalent

$$\begin{aligned}\int_{\Omega^e} w^h_{,i} j^h_i \, dV &= -\sum_{A,B} c^A_e \left[\int_{\Omega^e} N^A_{,I} \xi_{I,i} \kappa_{ij} N^B_{,J} \xi_{J,j} \, dV \right] d^B_e \\ &= -\sum_{A,B} c^A_e \left[\int_{\Omega^\xi} N^A_{,I} \xi_{I,i} \kappa_{ij} N^B_{,J} \xi_{J,j} \det \left[\underline{J}\right] \, dV^\xi \right] d^B_e \\ &= -\sum_{A,B} c^A_e \left[\int_{\xi_1=-1}^{1} \int_{\xi_2=-1}^{1} \int_{\xi_3=-1}^{1} N^A_{,I} \xi_{I,i} \kappa_{ij} N^B_{,J} \xi_{J,j} \det \left[\underline{J}\right] \, d\xi_1 d\xi_2 d\xi_3 \right] d^B_e\end{aligned}$$

In the integrand on the right, the $\underline{\xi}$ dependence on the terms is as follows:

- $N^A_{,I}$: N is parameterized by $\underline{\xi}$, and is tri-linear so the derivative with respect to one of the component still guarantees $\underline{\xi}$ dependence.
- $\xi_{I,i}$: is a component of the inverse of the tangent map. So the it will still depend on ξ.
- κ_{ij}: if we were not considering uniform conductivity or diffusivity, then $\underline{\kappa}$ can be dependent on position. Since we know that we can interpolate from the parent domain to the physical domain by using our basis functions, it too may be a function of $\underline{\xi}$
- $N^B_{,J}$, $\xi_{J,j}$ both have $\underline{\xi}$ dependence for the same reasons as given above

In the case that κ_{ij} is indeed independent of position, it isn't so difficult to evaluate the integral analytically. But we will not do so. We will come back to how to evaluate this using numerical integration.

We expect the integral to give a scalar value as "all the little is, little js will be contracted away", but it will indexed by A and B, corresponding to the local degrees of freedom. We will denote the result to be K^{AB}_e.

So the result is

$$\sum_{A,B} c_e^A K_e^{AB} d_e^B$$

The next step is to get rid of the explicit sum over A and B by using matrix vector notation for local degrees of freedom numbering.

$$-\begin{bmatrix} c_e^1 & c_e^2 & \dots & c_e^{N_{ne}} \end{bmatrix} \begin{bmatrix} K_e^{AB} & \dots & K_e^{1N_{ne}} \\ \vdots & \ddots & \vdots \\ \dots & \dots & K_e^{N_{ne}N_{ne}} \end{bmatrix} \begin{bmatrix} d_e^1 \\ d_e^1 \\ \vdots \\ d_e^{N_{ne}} \end{bmatrix} = -\underline{c}_e^\intercal \underline{\underline{K}}_e \underline{d}_e$$

The matrix $\underline{\underline{K}}_e$ is called the element conductivity or diffusivity matrix.

Now we can consider the second finite dimensional weak form integral.

$$\begin{aligned}
\int_{\Omega^e} w^h f \, dV &= \int_{\Omega^e} \left(\sum_A N^A c_e^A \right) f \, dV \\
&= \sum_A c_e^A \int_{\Omega^\xi} N^A f \det \left[\underline{\underline{J}}(\xi) \right] \, dV^\xi \\
&= \begin{bmatrix} c_e^1 & c_e^2 & \dots & c_e^{N_{ne}} \end{bmatrix} \int_{-1}^1 \int_{-1}^1 \int_{-1}^1 \begin{bmatrix} N^1 \\ N^2 \\ \vdots \\ N^{N_{ne}} \end{bmatrix} f(\underline{\xi}) \det \left[\underline{\underline{J}}(\underline{\xi}) \right] \, d\xi_1 d\xi_2 d\xi_3 \\
&= \begin{bmatrix} c_e^1 & c_e^2 & \dots & c_e^{N_{ne}} \end{bmatrix} \begin{bmatrix} F_e^{int_1} \\ F_e^{int_2} \\ \vdots \\ F_e^{int_{N_{ne}}} \end{bmatrix} \\
&= \underline{c}_e^\intercal \underline{F}_e^{int}
\end{aligned}$$

7.13 The matrix-vector weak form - I

We have been looking at the element level integrals in their matrix vector weak form. We are still missing the integral that imposes the Neumann boundary condition. So let us consider that integral.

$$-\int_{\partial\Omega_j^e} w^h j_n \, dS$$

Expanding this integral using the finite element basis functions

$$-\int_{\partial\Omega_j^e} \left(\sum_{A=1}^{N_{ne}} N^A c_e^A \right) j_n \, dS$$

Our summation is over all possible nodes within the element. But not all these nodes lie on the Neumann boundary.

Let's define a set $\mathcal{A}_N$:

$$\mathcal{A}_N = \{A | \underline{x}_e^A \in \partial\Omega_j^e\}$$

This is the set that contains the local node number A for all nodes $\underline{x}_e^A$ which lies on the Neumann boundary. The subscript N suggests that the set consists of nodes for the Neumann boundary.

So we can write the integral instead as

$$-\sum_{A\in\mathcal{A}_N} c_e^A \int_{\partial\Omega_j^e} N^A j_n \, dS$$

The last step is the convert this integral to one in the bi-unit domain. We will write

$$-\sum_{A\in\mathcal{A}_N} c_e^A \int_{\partial\Omega_j^\xi} N^A j_n \det\left[\underline{J}_s\right] \, dS^\xi$$

$\underline{J}_s$ is a lower dimensional tangent map of the map from the parent domain to the physical domain, $\widetilde{x}(\xi_2, \xi_3)$,

$$\underline{J}_s = \frac{d\underline{\widetilde{x}}}{d\underline{\xi}} = \begin{bmatrix} \widetilde{x}_{1,\xi_2} & \widetilde{x}_{1,\xi_3} \\ \widetilde{x}_{2,\xi_2} & \widetilde{x}_{1,\xi_3} \end{bmatrix}$$

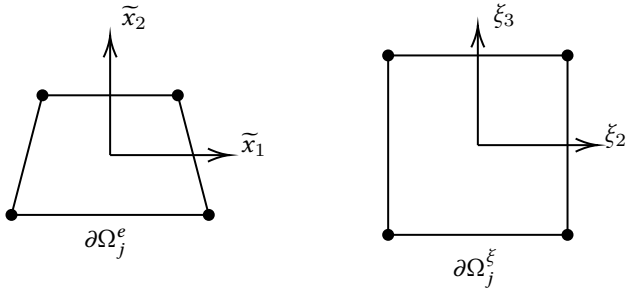

Figure 11: Schematic of the surface $\partial\Omega_j^e$ which is mapped to from $\partial\Omega_j^\xi$. $\widetilde{x}_2$ and $\widetilde{x}_1$ are the local coordinates of the surface in the physical domain. The choice of ξ_2 and ξ_3 was arbitrary in this case.

7.14 The matrix-vector weak form II

It may be difficult to construct the mapping $\widetilde{x}(\xi_2, \xi_3)$ in the case that $\widetilde{x}_2$ and $\widetilde{x}_1$ do not coincide with the global physical coordinates axes. In the case that it does, then the mapping is

$$\underline{\widetilde{x}}(\underline{\xi}) = \begin{bmatrix} x_1(\xi_2, \xi_3) \\ x_2(\xi_2, \xi_3) \end{bmatrix}$$

With the mapping, we can write the integral as a double integral over the surface $\partial\Omega_j^\xi$ in the parent domain:

$$-\sum_{A\in\mathcal{A}_N} c_e^A \int_{\partial\Omega_j^e} N^A j_n \, dS = -\sum_{A\in\mathcal{A}_N} c_e^A \int_{\xi_i=-1}^{1}\int_{\xi_j=-1}^{1} N^A j_n \det\left(\underline{J}_s\right) d\xi_i d\xi_j$$

we wrote ξ_i and ξ_j since the face could be mapped from any of the 6 faces of the bi-unit domain.

Now we will replace the summation over the local node numbers as matrix-

vector products.

$$-\begin{bmatrix} c_e^{A_1} & c_e^{A_1} & c_e^{A_3} & c_e^{A_4} \end{bmatrix} \begin{bmatrix} F^{jA_1} \\ F^{jA_2} \\ F^{jA_3} \\ F^{jA_4} \end{bmatrix},$$

which holds for $A_1, \ldots, A_4 \in \mathcal{A}_N$. The components of the column vector have superscript j to signify that they are coming from the influx boundary condition.

There is one more step we can take. We can expand the matrix vector weak form to also include the nodes that do not lie on the Neumann boundary.

In our eight node brick subdomain, say that nodes $A = 1, 4, 5, 8$ lie on the Neumann boundary, and the other ones do not. We can write the integral as

$$-\int_{\partial\Omega_j^e} w^h j_n \, dS = -\begin{bmatrix} c_e^{A_1} & c_e^2 & c_e^3 & c_e^{A_2} & c_e^{A_3} & c_e^6 & c_e^7 & c_e^{A_4} \end{bmatrix} \begin{bmatrix} F^{jA_1} \\ 0 \\ 0 \\ F^{jA_2} \\ F^{jA_3} \\ 0 \\ 0 \\ F^{jA_4} \end{bmatrix}$$

$$= -\underline{c}_e^{\mathsf{T}} \underline{F}_e^j$$

7.15 The matrix-vector weak form, continued - I

We are ready to assemble the element matrix vector representations into global matrix vector weak form.

We will use the concise notation for all the element matrix vector weak

forms.

$$-\sum_{e=1}^{N_{el}} \underline{c}_e^\intercal \underline{K}_e \underline{d}_e = \sum_{e=1}^{N_{el}} \underline{c}_e^\intercal \underline{F}_e^{int} - \sum_{e \in E} \underline{c}_e^\intercal \underline{F}_e^j$$

We can multiply through -1 and place the negative sign on the forcing vector term.

$$\sum_{e=1}^{N_{el}} \underline{c}_e^\intercal \underline{K}_e \underline{d}_e = -\sum_{e=1}^{N_{el}} \underline{c}_e^\intercal \underline{F}_e^{int} + \sum_{e \in E} \underline{c}_e^\intercal \underline{F}_e^j$$

As an aside, we could have arrived at our steady state PDE by considering the time dependent case.

$$\underbrace{c\frac{\partial u}{\partial t}} = \underbrace{-j_{i,i}} \underbrace{-f}$$

and considering $u_{,t}$ to be zero. The left term is the rate of change in temperature, the first term on the right is net heat influx, and the second term is the local distributed heating. (c is the specific heat of the medium.)

So let's define $\overline{f} = -f$. Writing the weak form using $\overline{f}$ removes the negative sign in front of the second integral.

$$\sum_{e=1}^{N_{el}} \underline{c}_e^\intercal \underline{K}_e \underline{d}_e = \sum_{e=1}^{N_{el}} \underline{c}_e^\intercal \overline{\underline{F}}_e^{int} + \sum_{e \in E} \underline{c}_e^\intercal \underline{F}_e^j$$

We are now ready to move on to the assembly global matrix vector weak forms. The assembly was simple to do in one dimension, but less trivial in three dimensions. We need to consider about mesh connectivity.

7.16 The matrix-vector weak form, continued - II

Each node within a hexahedral element is assigned an integer from 1 to 8. The question of how we represent a general internal global node $\overline{A}$ as a node

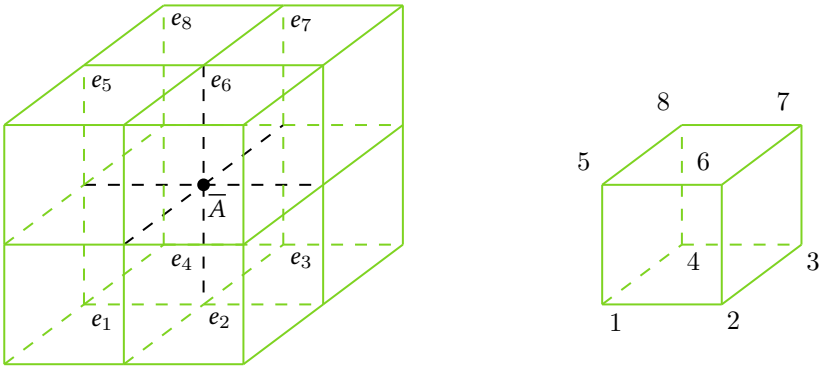

Figure 12: A general internal node, shown as $\overline{A}$, belongs to 8 elements in the case of hexahedral elements, shown on the left. We will follow the nodal numbering scheme as shown on the right.

in each of the eight elements locally is mesh connectivity. This information is provided typically in an input file. If the mesh is fairly regular, then these things can be generated on the fly.

In such a file (table 1), there may be a listing of elements which contain $\overline{A}$ (they do not have to be sequentially numbered).

This information helps us during global finite element assembly.

7.17 The matrix vector weak form, continued further - I

In a previous subsection, we stated the matrix-vector weak form as summations over all local matrix-vector equations.

$$\sum_{e=1}^{N_{el}} \underline{c}_e^{\mathsf{T}} \underline{K}_e \underline{d}_e = \sum_{e=1}^{N_{el}} \underline{c}_e^{\mathsf{T}} \overline{\underline{F}}_e^{int} + \sum_{e \in E} \underline{c}_e^{\mathsf{T}} \underline{F}_e^{j}$$

We will state the global matrix vector equation as

$$\underline{c}^{\mathsf{T}} \overline{\underline{K}} \overline{\underline{d}} = \underline{c}^{\mathsf{T}} \overline{\underline{F}}^{int} + \underline{c}^{\mathsf{T}} \underline{F}^{j} \tag{1}$$

Table 1: A local destination array. It is a listing of each Ω^e. The global representation of the i local node is at i position i on each row (assuming i begins from 1).

Element	$A=1$	$A=2$	$A=3$	$A=4$	$A=5$	$A=6$	$A=7$	$A=8$
e_1	_	_	_	_	_	_	$\overline{A}$	_
$\vdots$	_	_	_	_	_	_	_	_
e_2	_	_	_	_	_	_	_	$\overline{A}$
e_3	_	_	_	_	_	$\overline{A}$	_	_
e_4	_	_	_	_	$\overline{A}$	_	_	_
$\vdots$	_	_	_	_	_	_	_	_
e_5	_	_	$\overline{A}$	_	_	_	_	_
e_6	_	_	_	$\overline{A}$	_	_	_	_
e_7	$\overline{A}$	_	_	_	_	_	_	_
$\vdots$	_	_	_	_	_	_	_	_
e_8	_	$\overline{A}$	_	_	_	_	_	_

We introduced bars over the K and d. We still need to handle the Dirichlet boundary conditions.

The terms $\underline{\overline{K}}$ and $\underline{\overline{F}}^{int}$ are abstract representations of the result of assembly over all individual element conductivity/diffusivity matrices and forcing vectors.

$$\underline{\overline{K}} = \bigwedge_{e=1}^{N_{el}} \underline{K}_e \qquad\qquad \underline{\overline{F}}^{int} = \bigwedge_{e=1}^{N_{el}} \underline{\overline{F}}_e^{int}$$

For each entry in $\underline{\overline{K}}$, each row and column corresponds to a global node. If we follow table 1, then at entry $\overline{A}, \overline{A}$, we have contributions from the local conductivity/diffusivity matrix of elements $e_{1\to 8}$. The total contribution is

$$K_{e_1}^{77} + K_{e_2}^{88} + K_{e_3}^{55} + K_{e_4}^{66} + K_{e_5}^{33} + K_{e_6}^{44} + K_{e_7}^{11} + K_{e_8}^{22}$$

For any row $\overline{A}$ and column $\overline{B}$, the total contribution will be from each

element containing both nodes $\overline{A}$ and $\overline{B}$. In the local $\underline{K_e}$ matrices, the row and column which contributes corresponds to the local node numbers that $\overline{A}$ and $\overline{B}$ is equal to.

For $\underline{\overline{F}}^{int}$, the row $\overline{A}$ will be the sum of entries at the row of the local forcing vector corresponding to the local node numbering of $\overline{A}$ of all elements which contain $\overline{A}$.

$$F_{e_1}^{int^7} + F_{e_2}^{int^8} + F_{e_3}^{int^5} + F_{e_4}^{int^6} + F_{e_5}^{int^3} + F_{e_6}^{int^4} + F_{e_7}^{int^1} + F_{e_8}^{int^2}$$

A similar process is applied to the Neumann forcing vector.

7.18 The matrix-vector weak form, continued further - II

We have looked at how we construct the global matrices and vectors from the mesh connectivity information. Now we will deal with accounting for the Dirichlet boundary conditions.

Suppose that the global degrees of freedom $\overline{A}$, $\overline{B}$, ... belong to the set $\overline{\mathcal{A}}_D$ (D for Dirichlet) of the global degrees of freedom on which the Dirichlet boundary conditions are specified.

$$\{\overline{A}, \overline{B}, \dots\} \in \overline{\mathcal{A}}_D$$

If A is the local degree of freedom in element e_i corresponding to the global degree of freedom $\overline{A} \in \overline{\mathcal{A}}_D$, then the weighting function is constructed so that

$$w_{e_i}^h = \sum_{B=1, B\neq A}^{N_{ne}} N^B c_e^B$$

This means that the vector $\underline{c}^\intercal$ will not contain $\overline{A}$. Its dimensionality will be reduced. It will be missing all global degrees of freedom corresponding to nodes that lie on the Dirichlet boundary.

In general, the measure of the set $m\left(\overline{\mathcal{A}}_D\right) = \mathcal{N}_D$ (fancy term for number of elements in the set), determines the number of entries in $\underline{c}^\intercal$. The dimensions

of $\underline{c}^\intercal$ is

$$N_{sd}t \times N_{nodes} - \mathcal{N}_D$$

The dimensions of $\underline{\overline{d}}$ is

$$N_{sd} \times N_{nodes}$$

So the dimensions of $\underline{\overline{K}}$ is

$$(N_{sd} \times N_{nodes} - \mathcal{N}_D) \times (N_{sd} \times N_{nodes})$$

N_{sd} is the number of spatial dimensions, and N_{nodes} is the number of nodes per spatial dimension. Since we have been using tri-linear basis functions so far, $N_{node} = 2$.

Some of the entries of $\underline{\overline{d}}$ are known values from the Dirichlet boundary condition. We will remove theses degrees of freedom in $\underline{\overline{d}}$ and also remove columns of the $\underline{\overline{K}}$ which gets multiplied to the known d degrees of freedom, and move them to right hand side. So $\underline{\overline{K}}$ and $\underline{\overline{d}}$ will have reduced dimensionality and are written without the overline. Written out

$$\underline{c}^\intercal \underline{Kd} = \underline{c}^\intercal \left(\underline{\overline{F}}^{int} + \underline{F}^j \right) - \left(\underline{c}^\intercal \underline{\overline{K}}_{\overline{A}} d^{\overline{A}} + \dots \right),$$

where $\underline{\overline{K}}_{\overline{A}}$ represents the column of $\underline{\overline{K}}$ which corresponded to $\overline{A} \in \overline{\mathcal{A}}_D$

The dimensions of $\underline{K}$ is

$$(N_{sd} \times N_{nodes} - \mathcal{N}_D)^2$$

Rearranging terms,

$$\underline{c}^\intercal \left[\underbrace{\underline{Kd}}_{} \underbrace{- \left(\underline{\overline{F}}^{int} + \underline{F}^j \right) + \left(\underline{\overline{K}}_{\overline{A}} d^{\overline{A}} + \dots \right)}_{} \right] = 0$$

which has to hold for all $\underline{c} \in \mathbb{R}^{N_{sd} \times N_{node} - \mathcal{N}_D}$ (this requirement comes from that $\underline{c}$ is the degrees of freedom we used to interpolate the weighting function). Due to this condition, the following must hold

$$\underline{Kd} = \underbrace{\left(\underline{\overline{F}}^{int} + \underline{F}^j \right) - \left(\underline{\overline{K}}_{\overline{A}} d^{\overline{A}} + \dots \right)}_{\underline{F}}$$

This is our final matrix vector equations.

7.18.1 Unit 7 Quiz

1. Consider an element that has none of its faces lying on the boundary of the domain. What can you say about the element conductivity matrix?

    - It is square, symmetric, but not positive definite.

8 Lagrange basis functions and numerical quadrature in 1 through 3 dimensions

8.1 Lagrange basis functions in 1 through 3 dimensions - I

This week, we will present a view of the basis functions in a single dimension and extend them to multiple dimensions using a process called a tensor product.

Recall that in 1-D, the number of nodes in an element was one more than the basis function order and the basis functions were lagrange polynomials. For these 1-D basis functions, let's we will put a tilde over it.

In two dimensions, we used bilinear functions, and our quadrilateral element consisted of 4 nodes, which is the number of nodes we had in 1-D raised to the power of 2. In general, the basis functions where given by

$$N^A(\xi_1, \xi_2) = \widetilde{N}^B(\xi_1)\widetilde{N}^C(\xi_2)$$

for $B, C = 1, \ldots, N_{n_e 1D}$, and $A = 1, \ldots N_{n_e}$. In this case, $N_{n_e 1D} = 2$ for linears, and $N_{n_e} = 4$.

In three dimensions, the trilinears can be constructed using

$$N^A(\xi_1, \xi_2, \xi_3) = \widetilde{N}^B(\xi_1)\widetilde{N}^C(\xi_2)\widetilde{N}^D(\xi_2)$$

where $B, C, D \in \{1, \ldots, N_{n_e 1D}\}$, and $A \in \{1, \ldots, N_{n_e}\}$.

How to set to values of B, C, D are dependent on the user.

For any node A, the general formula for triquadratic basis functions is the same as for trilinears. We simply have to replace the linear basis functions multiplied together by quadratic 1-D basis functions. Using the projection shown in figure 13 can figure out the value of B, C, D by relating nodes along $\xi_{1\to 3}$ and take them to be the node's position relative to their respective axes if we only regarded a single axis at a time.

For example, the triquadratic basis function mid-element node is

$$\widetilde{N}^2(\xi_1)\widetilde{N}^2(\xi_2)\widetilde{N}^2(\xi_3)$$

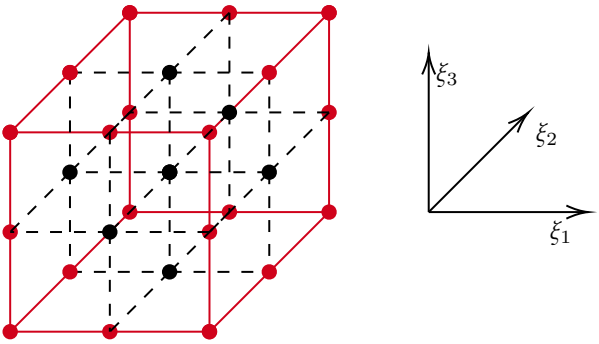

Figure 13: There are 27 nodes in total for a triquadratic element. This comes from the fact that $N_{n_e 1D} = 3$, so there will be 3^3 number of nodes in 3-D.

It is the midside node in all the $\xi_{1\to 3}$ axis.

8.1.1 Coding assignment 2 (2D problem) - I

The main function remains relatively similar to the 1-D problem, with the notable difference that we will now pass in vector, as opposed to an integer to generate a quadrilateral element mesh.

In the 2-D case, we numbered the nodes of the quadrilateral elements starting from 1 at the the bottom left corner, and went counterclockwise. `Dealii` does its numbering differently. The first bottom left node starts with 0, then we move to the node on the right, the top left node, and finally the top right node. The mappings from the numbering we used in class to the `Dealii` number is: $1 \mapsto 0$, $2 \mapsto 1$, $3 \mapsto 3$, $4 \mapsto 2$.

8.2 Quadrature rules in 1 through 3 dimensions

Recall that in one dimensions, we performed the integrals within the stiffness matrix and forcing vectors using Gaussian quadrature. We converted

integration into evaluating quadrature rules.

$$\int_{-1}^{1} g(\xi_1)\, d\xi_1 = \sum_{\ell=1}^{N_{int}} g(\xi_1^{\ell}) w_{\ell}$$

In two dimensions, we would need ot integrate terms of the following type:

$$\int_{\xi_2=-1}^{1} \int_{\xi_1=-1}^{1} g(\xi_1, \xi_2)\, d\xi_1 d\xi_2$$

We can numerically integrate the double integral by doing the integrals over ξ_1 and ξ_2 sequentially.

$$\int_{\xi_2=-1}^{1} \left[\sum_{\ell_1=1}^{N_{int}^1} g(\xi_1^{\ell_1}, \xi_2) w_{\ell_1} \right] d\xi_2 = \sum_{\ell_2=1}^{N_{int}^2} \sum_{\ell_1=1}^{N_{int}^1} g\left(\xi_1^{\ell_1}, \xi_2^{\ell_2}\right) w_{\ell_1} w_{\ell_2}$$

We added superscripts to N_{int} to give us the generality that we can chose a different number of integration points along each axis.

Recall that the sum of the weights in 1-D was equal to 2. So in 2-D, the sum of the weights will be equal to 4.

In 3-D, we are trying to integrate:

$$\int_{\xi_3=-1}^{1} \int_{\xi_2=-1}^{1} \int_{\xi_1=-1}^{1} g(\xi_1, \xi_2, \xi_3)\, d\xi_1 d\xi_2 d\xi_3 = \sum_{\ell_3=1}^{N_{int}^3} \sum_{\ell_2=1}^{N_{int}^2} \sum_{\ell_1=1}^{N_{int}^1} g(\xi_1^{\ell_1}, \xi_2^{\ell_2}, \xi_3^{\ell_3}) w_{\ell_1} w_{\ell_2} w_{\ell_3}$$

8.2.1 Coding assignment 2 (2D problem) - II

The Einstein summation notation is that whenever a index variable appears twice in a single term, and is not otherwise defined, it implies summation of that term over all values of the index.

8.3 Triangular and tetrahedral elements - Linears - I

The quadrilateral and hexahedron elements are not the simplest element that can be used to fill space. We will look at triangular and tetrahedron elements.

These are also called simplex elements in general.

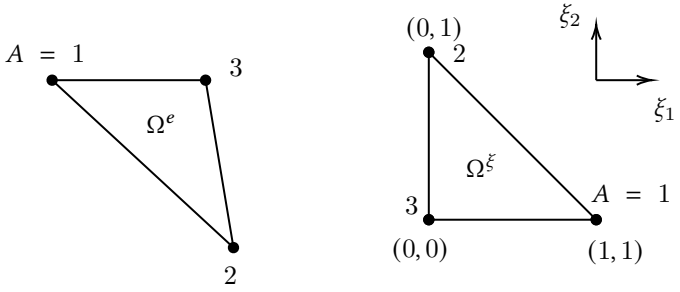

Figure 14: Ω^e in general is a scalene triangle in 2-D. We say that it is mapped from a unit (as opposed to bi-unit) parent domain, with the the local numbering conventions as shown. We will also define a third coordinate ξ_3, which is equal to $1 - \xi_1 - \xi_2$.

8.4 Triangular and tetrahedral elements - Linears - II

We can write out our linear basis functions.

$$N^1 = \xi_1$$
$$N^2 = \xi_2$$
$$N^3 = \xi_3$$

We might choose to use the simplex element for its simplicity. But it has its draw backs.

The basis functions 3-D for simplex elements are

$$N^A(\xi_1, \xi_2, \xi_3, \xi_4) = \xi^A$$

for $A = 1, 2, 3, 4$.

When we take the gradient of these functions, we will end up with constant gradients, which limits how well we can express higher order functions.

One can define higher order basis functions for simplex elements, but they will require other numerical integration schemes as Gaussian quadrature is no longer optimal.

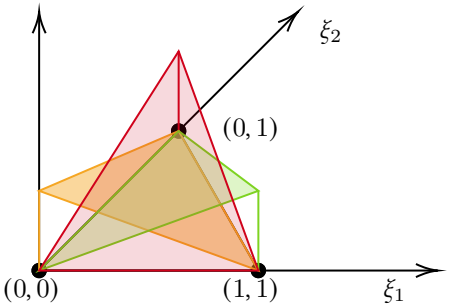

Figure 15: A perspective view of the parent domain with the basis functions visualized. The green surface is N^1, the red surface is N^2, and the orange surface is N^3.

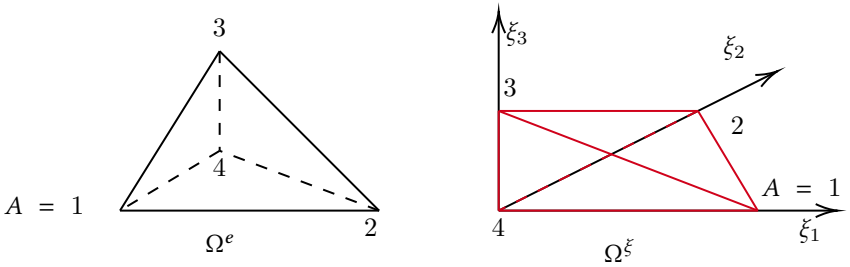

Figure 16: An arbitrary tetrahedron mapped from a right tetrahedron in the parent domain. We define a four coordinate $\xi_4 = 1 - \xi_1 - \xi_2 - \xi_3$.

9 Linear, elliptic, partial differential equations for a scalar variable in two dimensions

9.1 The finite-dimensional weak form and basis functions - I

We will take a step back and investigate the linear elliptic PDE problem for a scalar field in two dimensions.

Lets state the strong and weak forms. Given g_g, j_n, f, the constitutive relation $j_i = -\kappa_{ij} u_{,j}$, for $i, j = 1, 2$.

- Strong form: find u such that

$$-j_{i,i} = \underbrace{f}_{-\overline{f}}$$

in $\Omega \subset \mathbb{R}^2$, with boundary conditions $u = u_g$ on $\partial\Omega_u$, and $-j_i n_i = j_n$ on $\partial\Omega_j$.

- Finite dimensional weak form: find $u^h \in \mathcal{S}^g \subset \mathcal{S}$, where $\mathcal{S}^h = \{u^h \in H^1(\Omega) | u^h = u_g \text{ on } \partial\Omega_u\}$ such that for all $w^h \in \mathcal{V}^h \subset \mathcal{V}$; $\mathcal{V}^h = \{w^h \in H^1(\Omega) | w^h = 0 \text{ on } \partial\Omega_u\}$ such that the following holds

$$-\int_\Omega w^h_{,i} j^h_i \, dA = \int_\Omega w^h \overline{f} \, dA + \int_{\partial\Omega_j} w^h j_n \, dS$$

for $i = 1, 2$.

We know that we can turn the integral over the domain Ω into sum of integrals over the element subdomains. Let's work with quadrilateral domains. In this case, any element Ω^e is said to be mapped from a bi-unit parent domain. We can interpolate the physical geometry using the relationship

$$\underline{x}_e(\underline{\xi}) = \sum_{A=1}^{N_{ne}} N^A(\underline{\xi}) \underline{x}_e^A$$

Similarly, we can work with triangular elements which are mapped from a unit parent domain.

In either case, we will stick to an isoparametric formulation for simplicity.

9.2 The finite-dimensional weak form and basis functions - II

Having decide on the order of polynomial basis functions, we can write the finite dimensional trial solution and weighting function over each subdomain as linear combinations of the nodal dof and the basis function. Let's compute

the gradient of the expansions

$$u^h_{e,i} = \sum_{A=1}^{N_{ne}} N^A_{,i} d^A_e = \sum_{A=1}^{N_{ne}} \left[N^A_{,\xi_1} \frac{\partial \xi_1}{\partial x_i} + N^A_{,\xi_2} \frac{\partial \xi_2}{\partial x_i} \right] d^A_e$$

The derivatives $\partial\xi_1/\partial x_i$, $\partial\xi_2/\partial x_i$ can be found by inverting the Jacobian.

Since we have a map $x_i(\underline{\xi})$, In coordinate notation, it is

$$x_i(\underline{\xi}) = \sum_{A=1}^{N_{ne}} N^A(\underline{\xi}) x^A_{e_i}$$

Differentiating gives

$$\frac{\partial x_i}{\partial \xi_I} = \sum_{A=1}^{N_{ne}} N^A_{,I} x^A_{e_i}$$

And the Jacobian is a 2 by 2 matrix:

$$\underline{J}(\underline{\xi}) = \begin{bmatrix} \frac{\partial x_1}{\partial \xi_1} & \frac{\partial x_1}{\partial \xi_2} \\ \frac{\partial x_2}{\partial \xi_1} & \frac{\partial x_2}{\partial \xi_2} \end{bmatrix}$$

Which has an inverse defined by

$$\underline{J}^{-1}(\underline{\xi}) = \begin{bmatrix} \frac{\partial \xi_1}{\partial x_1} & \frac{\partial \xi_1}{\partial x_2} \\ \frac{\partial \xi_2}{\partial x_1} & \frac{\partial \xi_2}{\partial x_2} \end{bmatrix}$$

And we can use to compute

$$u^h_{e,i} = \sum_{A=1}^{N_{ne}} \frac{\partial N^A}{\partial \xi_I} \frac{\partial \xi_I}{\partial x_i} d^A_e$$

Remember that there is a sum implied over the the index variable I by the Einstein summation convention.

With this, we can write the integral over the domain into sum of integrals over each subdomain.

$$-\int_\Omega w^h_{,i} j^h_i \, dA = -\sum_e \int_{\Omega^e} w^h_{,i} j^h_i \, dA$$

Change from integrating over the physical subdomain to the parent subdomain:

$$\int_{\Omega^e} w^h_{,i} j^h_i \, dA = -\sum_{A,B=1}^{N_{ne}} c_e^A \left[\int_{-1}^{1} \int_{-1}^{1} N^A_{,i} \left(-\kappa_{ij}\right) N^B_{,j} \det(\underline{J}) \, d\xi_1 d\xi_2 \right] d_e^A$$

$$= -\sum_{A,B=1}^{N_{ne}} c_e^A \left[\int_{-1}^{1} \int_{-1}^{1} N^A_{,I} \xi_{,i} \left(-\kappa_{ij}\right) N^B_{,J} \xi_{,j} \det(\underline{J}) \, d\xi_1 d\xi_2 \right] d_e^A$$

We can include the negative sign infront to be a part of the double integral; it will cancel with the negative sign in front of κ_{ij}. Let's expand the the summation convention.

$$\sum_{A,B=1}^{N_{ne}} c_e^A \left[\int_{-1}^{1} \int_{-1}^{1} \sum_{I=1}^{N_{sd}} \sum_{i=1}^{N_{sd}} \sum_{J=1}^{N_{sd}} \sum_{j=1}^{N_{sd}} N^A_{,I} \xi_{I,i} \left(\kappa_{ij}\right) N^B_{,J} \xi_{J,j} \det(\underline{J}) \, d\xi_1 d\xi_2 \right] d_e^A$$

Let's write the integral within the summation over A and B in Gaussian quadrature.

$$\sum_{\ell_1=1}^{N^1_{int}} \sum_{\ell_2=1}^{N^2_{int}} \sum_{I,i,J,j=1}^{N_{sd}} \left[N^A_{,I} \xi_{,i} \kappa_{ij} N^B_{,J} \xi_{,j} \det\left(\underline{J}\right) \right] \left(\xi_1^{\ell_1}, \xi_2^{\ell_2}\right) w_{\ell_2} w_{\ell_1}$$

This term is actually K_e^{AB}.

The integral involving the source term is

$$\int_{\Omega} w^h \overline{f} \, dA = \sum_e \int_{\Omega^e} w^h \overline{f} \, dA$$

We will expand the subdomain integral

$$\int_{\Omega^e} w^h \overline{f} \, dA = \sum_A c_e^A \int_{-1}^{1} \int_{-1}^{1} N^A \overline{f} \det\left(\underline{J}(\underline{\xi})\right) \, d\xi_1 d_{\xi_2}$$

$$= \sum_A c_e^A \sum_{\ell_1=1}^{N^1_{int}} \sum_{\ell_2=1}^{N^2_{int}} \left[N^A \overline{f} \det\left(\underline{J}\right) \right] \left(\xi_1^{\ell_1}, \xi_2^{\ell_2}\right) w_{\ell_2} w_{\ell_1}$$

9.3 The matrix-vector weak form

The Neumann condition integral is

$$\int_{\partial\Omega_j} w^h j_n \, dS$$

In two dimensions, this integral is over a curve.

Not all elements will coincide with the boundary. The summation over e will only apply to $e \in E$. E is the set of Indices of elements that coincide with the Neumann boundary.

$$\int_{\partial\Omega_j} w^h j_n \, dS = \sum_{e \in E} \int_{\partial\Omega_j^e} w^h j_n \, dS$$

w^h within the summation can be further expanded using basis functions. But since not all nodes lie on the Neumann boundary, we restrict the summation to be over the set of node Indices whose corresponding node lies on the Neumann boundary. On the physical domain, a vector $\widetilde{x}$ will point along the direction of curve element. For whatever $\widetilde{x}$, the curve element is mapped to from one of the axes of the parent domain.

$$\int_{\partial\Omega_j^e} w^h j_n \, dS = \sum_{A \in \mathcal{A}_N} c_e^A \underbrace{\int_{\partial\Omega_j^\xi} N^A j_n \det(\underline{J}_s) \, d\xi_1}_{F_e^{jA}}$$

where

$$\underline{J}_s = \frac{\partial \widetilde{x}}{\partial \xi_1}$$

Recall in 3 dimensions, $\underline{J}_s$ was a two by two matrix.

In the case of triangular elements with linear basis functions:

- the Jacobian from the parent domain to the physical domain will be constant in all entries. Since the basis functions are linear, not bilinear.

- The determinant of the Jacobian is exactly twice the area of the triangular element.

$$\det\left[\underline{\underline{J}}(\underline{\xi})\right] = 2m(\Omega^e)$$

The next step is assembly into local matrix-vector form and and global matrix-vector form. The principle behind this process is exactly the same as we did in 3-D.

9.4 The matrix-vector weak form - II

The assembly process is near identical to what has been outlined in subsections 7.17 and 7.18.

10 Linear, elliptic partial differential equations for vector unknowns in three dimensions (Linearized elasticity)

10.1 The strong form of linearized elasticity in three dimensions - I

We will now turn to look at solving three dimensional linear elliptic PDEs for a vector field. The model problem here will be linearized elasticity. Non-linear elasticity has other technical issues and it not necessarily elliptical.

The domain Ω is a volume in three dimensions. Let's state the data we have. Given u_i^g (g denotes "given"), $\bar{t}_i$, f_i ($\underline{u}^g, \underline{\bar{t}}, \underline{f} \in \mathbb{R}^3$, they are all vector data) and the constitutive relation

$$\sigma_{ij} = \mathbb{C}_{ijkl}\epsilon_{kl}$$

σ_{ij} is the coordinate notation of the Cauchy stress tensor ($\underline{\sigma}$). $\mathbb{C}_{ijkl}$ is the coordinate notation of the elasticity tensor ($\underline{\underline{\mathbb{C}}}$), and ϵ_{kl} is the coordinate notation of the infinitesimal strain tensor. In direct notation,

$$\underline{\sigma} = \underline{\underline{\mathbb{C}}} : \underline{\epsilon}$$

Since the indices k and l are being contracted out, we used the colon to represent contraction.

10.2 The strong form of linearized elasticity in three dimensions - II

We are also given the kinematic relation

$$\epsilon_{kl} = \frac{1}{2}\left(\frac{\partial u_k}{\partial x_l} + \frac{\partial u_l}{\partial x_k}\right)$$

In direct notation, this is written as

$$\underline{\epsilon} = \mathrm{sym}\underline{\nabla u}$$

This is saying that epsilon equals the symmetric part of the gradient of u.

The task is to find u_i, where $\underline{u} \in \mathbb{R}^3$, such that the following holds

$$\sigma_{ij,j} + f_i = 0$$

in Ω, where $i, j = 1, 2, 3$. (This is saying that the stress divergence plus f equals zero.) This is the quasi-static stress equilibrium condition.

Since i, j runs over the three spatial dimensions, what we really have is a single PDE for each component of the displacement field. This means that we require boundary conditions for each component of the field, and the surface over which the Dirichlet boundaries are defined can be different for different components of the PDE. They are allowed to overlap.

$$u_i = u_i^g$$

on $\partial\Omega_{u_i}$.

Since the complement of each Dirichlet boundary is the Neumann boundary,

$$\partial\Omega_{u_i} \cap \partial\Omega_{\bar{t}_i} = \emptyset$$
$$\partial\Omega = \partial\Omega_{u_i} \cup \partial\Omega_{\bar{t}_i}$$

indeed the Neumann boundary can be different for different components of the solution field. The Neumann condition is that

$$\sigma_{ij} n_j = \bar{t}_i$$

on $\partial\Omega_{\bar{t}_i}$.

The consequence of this is that on a given boundary, we may often not control all components of the displacement.

10.3 The strong form, continued

For completeness we will write the strong form of the 3-D linearized elasticity problem in direct notation.

Given u_i^g, $\bar{t}_i$, $\underline{f}$, the constitutive relation $\underline{\sigma} = \underline{\underline{\mathbb{C}}} : \underline{\epsilon}$, and the kinematic relation $\underline{\epsilon} = \text{sym}(\underline{\overline{\nabla} u})$, find $\underline{u}$ such that

$$\begin{cases} \underline{\nabla} \cdot \underline{\sigma} + \underline{f} = \underline{0} & \text{in } \Omega \\ u_i = u_i^g & \text{on } \partial\Omega_{u_i} \\ \sigma_{ij} n_j = \bar{t}_i & \text{on } \partial\Omega_{\bar{t}_i} \end{cases}$$

The term

$$\underline{\nabla} \cdot \underline{\sigma} = \text{div}(\underline{\sigma}) = \frac{\partial}{\partial \underline{x}} \cdot \underline{\sigma}$$

is the divergence of sigma, or dotting the gradient operator with sigma.

Let's go over the constitutive relation.

$$\sigma_{ij} = \mathbb{C}_{ijkl} \epsilon_{kl}$$

- $\mathbb{C}_{ijkl}$ is a fourth order elasticity tensor, which is constant with respect to $\underline{\epsilon}$ in linearized elasticity. This makes the relationship between strain and stress a linear relationship

$\underline{\underline{\mathbb{C}}}$ has the property of major symmetry, which means that

$$\mathbb{C}_{ijkl} = \mathbb{C}_{klij}$$

It follows from the fact that in problems of linearized elasticity, we have the idea that there exists a function $\psi(\underline{\epsilon})$ which is a mapping from the space of symmetric, second order tensors, $S(3)$, to real numbers. This function ψ is the strain energy density function.

It can be shown that

$$\mathbb{C}_{ijkl} = \frac{\partial^2 \psi}{\partial \epsilon_{ij} \partial \epsilon_{kl}}$$

Since $\mathbb{C}_{ijkl}$ is by definition constant with respect to strain, its clear that ψ will be a quadratic function. In this case, assuming ψ is smooth, we have

$$\mathbb{C}_{ijkl} = \frac{\partial^2 \psi}{\partial \epsilon_{ij} \partial \epsilon_{kl}} = \frac{\partial^2 \psi}{\partial \epsilon_{kl} \partial \epsilon_{ij}} = \mathbb{C}_{klij}$$

It turns out that $\underline{\underline{\mathbb{C}}}$ has other minor symmetries as well.

$$\mathbb{C}_{ijkl} = \mathbb{C}_{jikl}$$

This follows from the definition that the Cauchy stress tensor is symmetric.

$$\sigma_{ij} = \sigma_{ji}$$

and this is a result which follows from the static equilibrium and or conservation of angular momentum. Applying the constitutive relationship,

$$\sigma_{ij} = \mathbb{C}_{ijkl}\epsilon_{kl} = \sigma_{ji} = \mathbb{C}_{jikl}\epsilon_{kl}$$

which recovers the first stated minor symmetry. The second minor symmetry is that

$$\mathbb{C}_{ijkl} = \mathbb{C}_{ijlk}$$

This is due to the kinematic relation

$$\epsilon_{kl} = \frac{\left(u_{k,l} + u_{l,k}\right)}{2}$$

Since the addition operation can be interchanged, we have

$$\epsilon_{kl} = \epsilon_{lk}$$

Invoke the constitutive relation,

$$\sigma_{ij} = \mathbb{C}_{ijkl}\epsilon_{kl} = \mathbb{C}_{ijkl}\epsilon_{lk}$$

Since l and k are dummy indices, we can equally write the last term as

$$\sigma_{ij} = \mathbb{C}_{ijkl}\epsilon_{kl} = \mathbb{C}_{ijlk}\epsilon_{kl}$$

The middle and rightmost term imply

$$\mathbb{C}_{ijkl} = \mathbb{C}_{ijlk}$$

, which is the second minor symmetry relation we state.

10.4 The constitutive relations of linearized elasticity

Another property of the linear elasticity tensor, $\underline{\mathbb{C}}$, is that it is positive definite.

For all general tensor $\underline{\Theta} \in GL(3)$, which is the group of general linear transformations in $\mathbb{R}^3$ (this is just a fancy way of saying matrices in 3D if you regard every tensor as being a square matrix),

$$\underline{\Theta} : \underline{\mathbb{C}} : \underline{\Theta} \geq 0$$

Using direct notation,

$$\Theta_{ij}\mathbb{C}_{ijkl}\Theta_{kl} \geq 0$$

Further,

$$\underline{\Theta} : \underline{\mathbb{C}} : \underline{\Theta} = 0$$

iff $\underline{\Theta}$ is the zero tensor.

The implication is that linearized elasticity is talking about materials that are not subjected to material instabilities (such as fractures, shear bands, slip bands), which require other theories to describe.

The elasticity tensor takes on a special form when the material in question is isotropic (we are saying that the physical properties of the material is independent of the orientation of the system, this is different from homogeneity, where the material properties are independent of position). Isotropic materials can be specified elastically with just two constants.

In coordinate notation:

$$\mathbb{C}_{ijkl} = \lambda\delta_{ij}\delta_{kl} + 2\mu \underbrace{\frac{1}{2}\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right)}_{\mathbb{I}_{ijkl}}$$

The term $\mathbb{I}_{ijkl}$ is known as a fourth-order symmetric identity tensor. This tensor has the property such that $\mathbb{I}_{ijkl}$ acting on any second order tensor Θ_{kl} returns a second order tensor with only the symmetric part.

$$\mathbb{I}_{ijkl}\Theta_{kl} = \frac{1}{2}(\Theta_{ij} + \Theta_{ji})$$

In direction notation,

$$\underline{\underline{\mathbb{C}}} = \lambda \underline{\underline{\mathbb{1}}} \otimes \underline{\underline{\mathbb{1}}} + 2\mu \underline{\underline{\mathbb{I}}}$$

In direct notation, the second-order Kronecker delta tensor is denoted as $\mathbb{1}_{ij}$. The constants λ and μ are called Lamé constants.

The symbol $\otimes$ denotes the tensor product.

Let E be young's modulus, ν be Poisson's ratio.

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}$$
$$\mu = \frac{E}{2(1+\nu)}$$

μ is called the shear modulus. Also, the bulk modulus, K, can be written as

$$K = \frac{E}{3(1-2\nu)}$$

Poisson's ratio ν is bounded such that

$$-1 < \nu < 1/2$$

The $1/2$ limit represents elastic incompressibility (this makes the bulk modulus approach infinity). The -1 limit represents a shear unstable material (this makes the shear modulus infinity).

In this case of the elasticity tensor for isotropic materials, $\mathbf{C}$ begin positive definite implies that

$$\lambda + 2\mu > 0$$
$$\mu > 0$$

These conditions mean that there can be propagating longitudinal and shear waves in the elastic material. Since the longitudinal wave speed is given by

$$c_{log} = \sqrt{\frac{\lambda + 2\mu}{\rho}},$$

where ρ is mass density; the shear wave speed is

$$c_{shear} = \sqrt{\frac{\mu}{\rho}}$$

In most materials of interest, we can expect $c_{\log} > c_{shear}$.

10.5 The weak form - I

We will now move on to write the weak form for the linear elasticity equation. We will first write down the weak form, and show how we can obtain it from the strong form. We know that the weak form implies the strong form, and the strong form implies the weak form.

We will also using coordinate notation only from now on.

Infinite-dimensional weak form for 3-D linear elliptic PDE for a vector variable: Given u_i^g, $\bar{t}_i$, f_i, the constitutive relation

$$\sigma_{ij} = \mathbb{C}_{ijkl}\epsilon_{kl}$$

and the kinematic relation

$$\epsilon_{kl} = \frac{1}{2}\left(\frac{\partial u_k}{\partial x_l} + \frac{\partial u_l}{\partial x_k}\right)$$

find $u_i \in \mathcal{S} = \{u_i | u_i = u_i^g \text{ on } \partial\Omega_{u_i}\}$, $i = 1, 2, 3$, such that for all $w_i \in \mathcal{V} = \{w_i | w_i = 0 \text{ on } \partial\Omega_{u_i}\}$, $i = 1, 2, 3$ the following holds:

$$\int_\Omega w_{i,j}\sigma_{ij}\, dV = \int_\Omega w_i f_i\, dV + \sum_{i=1}^{N_{sd}} \int_{\partial\Omega_{\bar{t}_i}} w_i \bar{t}_i\, dS$$

There is no summation implied over the repeating indices in the right-most term. This is due to the fact that there are different traction conditions for different Neumann boundaries. The summation symbol on that term iterates over all spatial dimensions to account for each component of the traction condition.

With the appropriate given data and relationships, the strong form is that we want to find u_i such that

$$\begin{cases} \sigma_{ij,j} + f_i = 0 & \text{in } \Omega \\ u_i = u_i^g & \text{on } \partial\Omega_{u_i} \\ \sigma_{ij} n_j = \bar{t}_i & \text{on } \partial\Omega_{\bar{t}_i} \end{cases}$$

Now, let's consider $w_i \in \mathcal{V} = \{w_i | w_i = 0 \text{ on } \partial\Omega_{u_i}\}$, we multiply the PDE by w_i and integrate over Ω.

$$\int_\Omega w_i \sigma_{ij,j} \, dV + \int_\Omega w_i f_i \, dV = 0$$

Now, we apply integration by parts to the leftmost term. First, we will apply the product rule:

$$\int_\Omega \left((w_i \sigma_{ij})_{,j} - w_{i,j} \sigma_{ij} \right) \, dV + \int_\Omega w_i f_i \, dV = 0$$

Let's separate the leftmost integral,

$$\int_\Omega (w_i \sigma_{ij})_{,j} \, dV - \int_\Omega w_{i,j} \sigma_{ij} \, dV + \int_\Omega w_i f_i \, dV = 0$$

In the leftmost integral, the summation convention tells us that the index i is contracted out. So what is left is the divergence of a vector. This allows us to apply the divergence theorem.

$$\int_{\partial\Omega} w_i \sigma_{ij} n_j \, dS - \int_\Omega w_{i,j} \sigma_{ij} \, dV + \int_\Omega w_i f_i \, dV = 0$$

Let's rearrange the integral.

$$\int_\Omega w_{i,j} \sigma_{ij} \, dV = \int_\Omega w_i f_i \, dV + \int_{\partial\Omega} w_i \sigma_{ij} n_j \, dS$$

We can write the boundary as

$$\partial\Omega = \partial\Omega_{u_i} \cup \partial\Omega_{\bar{t}_i}, i = 1, 2, 3$$

Also, we can regard the term $w_i \sigma_{ij} n_j$ as a dot product of $\underline{w}$ and the vector that results from $\sigma_{ij} n_j$ when j is contracted away.

$$\int_\Omega w_{i,j} \sigma_{ij} \, dV = \int_\Omega w_i f_i \, dV + \sum_{i=1}^{N_{sd}} \int_{\partial\Omega_{u_i} \cup \partial\Omega_{\bar{t}_i}} w_i \left(\sigma_{ij} n_j \right) \, dS$$

In the rightmost term, we are not having an implicit sum over the i index. There is an implicit sum over j.

We will break up the integral over the union of the Dirichlet and Neumann boundaries into a sum. Then, the next step is the invoke the boundary conditions on w_i and $\sigma_{ij} n_j$.

$$\int_\Omega w_{i,j} \sigma_{ij} \, dV = \int_\Omega w_i f_i \, dV + \sum_{i=1}^{N_{sd}} \left(\int_{\partial\Omega_{u_i}} w_i \left(\sigma_{ij} n_j \right) \, dS + \int_{\partial\Omega_{\bar{t}_i}} w_i \left(\sigma_{ij} n_j \right) \, dS \right)$$

$$\int_\Omega w_{i,j} \sigma_{ij} \, dV = \int_\Omega w_i f_i \, dV + \sum_{i=1}^{N_{sd}} \left(0 + \int_{\partial\Omega_{\bar{t}_i}} w_i \bar{t}_i \, dS \right)$$

By definition, w_i over the Dirichlet boundary is zero, and $\sigma_{ij} n_j = \bar{t}_i$ on the Neumann boundary.

So we recover the infinite-dimensional weak form:

$$\int_\Omega w_{i,j} \sigma_{ij} \, dV = \int_\Omega w_i f_i \, dV + \sum_{i=1}^{N_{sd}} \int_{\partial\Omega_{\bar{t}_i}} w_i \bar{t}_i \, dS$$

As we did in 1-D, we can also obtain the weak form as the Euler-Lagrange conditions of a variational principle on extremization a free energy functional in 3-D. This is very powerful principle as it lets us circumvent the problem that the methods we have been developing do not work for constrained problems.

Lets state the finite-dimensional weak form.

Finite-dimensional weak form for 3-D linear elliptic PDE for a vector variable: Given u_i^g, $\bar{t}_i$, f_i, the constitutive relation

$$\sigma_{ij} = \mathbb{C}_{ijkl}\epsilon_{kl}$$

and the kinematic relation

$$\epsilon_{kl} = \frac{1}{2}\left(\frac{\partial u_k}{\partial x_l} + \frac{\partial u_l}{\partial x_k}\right)$$

find $u_i^h \in \mathcal{S}^h \subset \mathcal{S}$, where $\mathcal{S}^h = \{u_i^h \in H^1(\Omega) | u_i^h = u_i^g \text{ on } \partial\Omega_{u_i}\}$ such that for all $w_i^h \in \mathcal{V}^h \subset \mathcal{V}$, where $\mathcal{V}^h = \{w_i^h \in H^1(\Omega) | w_i^h = 0 \text{ on } \partial\Omega_{u_i}\}$, the following holds:

$$\int_\Omega w_{i,j}^h \underbrace{\sigma_{ij}}_{\mathbb{C}_{ijkl}\epsilon_{kl}^h} dV = \int_\Omega w_i^h f_i \, dV + \sum_{i=1}^{N_{sd}} \int_{\partial\Omega_{\bar{t}_i}} w_i^h \bar{t}_i \, dS$$

We did not place superscript h over the forcing and traction since they are given as data. And, ϵ_{kl}^h is

$$\epsilon_{kl}^h = \frac{1}{2}\left(\frac{\partial u_k^h}{\partial x_l} + \frac{\partial u_l^h}{\partial x_k}\right)$$

10.5.1 Handling summation over spatial dimensions

We want to make some clarifications how summation over spatial dimensions are computed. Let's start out by looking at the strong form.

$$\sigma_{ij,j} + f_i = 0,$$

for $i = 1, 2, 3$. There is sum implied over the j, but not over i. This PDE is actually three equation for the three components of the displacement field.

$$\begin{aligned}
\sigma_{1j,j} + f_1 &= 0 \\
\sigma_{2j,j} + f_2 &= 0 \\
\sigma_{3j,j} + f_3 &= 0
\end{aligned}$$

When we multiply the PDE by w_i and integrate over the domain Ω, we are implying sums over both i and j. By doing so, we collapse the three equations into a single equation.

$$\int_\Omega w_1\sigma_{1j,j} + w_2\sigma_{2j,j} + w_3\sigma_{3j,j} + w_1 f_1 + w_2 f_2 + w_3 f_3 \, dV = 0$$

The summation over j is still implied in the equation we have written above.

For the weak form,

$$\int_\Omega w^h_{i,j}\sigma_{ij} \, dV = \int_\Omega w^h_i f_i \, dV + \sum_{i=1}^{N_{sd}} \int_{\partial\Omega_{\bar{t}_i}} w^h_i \bar{t}_i \, dS$$

if we write out all the terms explicitly, we have

$$\begin{aligned}
&\int_\Omega w^h_{1,1}\sigma_{11} + w^h_{1,2}\sigma_{12} + w^h_{1,3}\sigma_{13} + w^h_{2,1}\sigma_{21} + w^h_{2,2}\sigma_{22} + w^h_{2,3}\sigma_{23} \\
&\quad + w^h_{3,1}\sigma_{31} + w^h_{3,2}\sigma_{3,2} + w^h_{3,3}\sigma_{33} \, dV \\
&\quad = \int_\Omega w^h_1 f_1 + w^h_2 f_2 + w^h_3 f_3 \, dV + \int_{\partial\Omega_{\bar{t}_1}} w^h_1 \bar{t}_1 \, dS + \int_{\partial\Omega_{\bar{t}_2}} w^h_2 \bar{t}_2 \, dS + \int_{\partial\Omega_{\bar{t}_3}} w^h_3 \bar{t}_3 \, dS
\end{aligned}$$

10.6 The finite-dimensional weak form - Basis functions - I

We have written the finite dimensional weak form for our problem. As we have done in the past, the next step is to partition our domain into open subdomains, and define basis functions over these partitions.

Let's consider the case of hexahedron elements, Ω^e. Recall that the original domain recovered by applying closure to the sum of the open subdomains.

$$\overline{\Omega} = \overline{\bigcup_e \Omega^e}$$

The subdomains are open, in that for any element e_1 and e_2, we have

$$\Omega^{e_1} \cap \Omega^{e_2} = \emptyset$$

There is one difference in the vector problem versus the scalar problem. The trial solution is a vector field, and in coordinate notation, we have

$$u_{i_e}^h = \sum_{A=1}^{N_{n_e}} N^A d_{i_e}^A$$

We wrote the subscript i before e since e_i may be confused for talking about the i^{th} element.

Since the trial solution is a vector quantity, the there will a number of nodal degrees of freedom equal to the number of spatial dimensions at each node. In direct notation, the trial solution over element e is

$$\underline{u}_e^h = \sum_{A=1}^{N_{n_e}} N^A \underline{d}_e^A$$

It's clear that for each element, there are $N_{n_e} \times N_{sd}$ number of degrees of freedom. And

$$\underline{u}_e^h, \underline{d}_e^A \in \mathbb{R}^3$$

Our approach is to use scalar basis functions, and vector degrees of freedom. But this is not a universal approach. In electromagnetic problems, the basis functions are often vectors, and the degrees of freedom are vectors.

Similarly, the weighting function is also a vector quantity, and we can expand it in terms of scalar basis functions and vector degrees of freedoms.

$$w_{i_e}^h = \sum_A N^A c_{i_e}^A$$
$$\underline{w}_e^h = \sum_A N^A \underline{c}_e^A$$

where

$$\underline{w}_e^h, \underline{c}_e^A \in \mathbb{R}^3$$

As before we say that every element in the physical domain is mapped to from a parent bi-unit domain. The basis functions in 3-D comes from multiplying Lagrange polynomial basis functions in 1-D in each of the ξ coordinate directions. Refer to section 8.1.

Lastly, we are still considering a isoparametric formulation.

10.7 The finite-dimensional weak form - Basis functions - II

Recall the isoparameteric map from the parent domain to the physical domain.

$$x_{i_e}(\underline{\xi}) = \sum_A N^A(\underline{\xi}) x_{i_e}^A$$

In our weak form, we are required to compute the following gradients:

$$w_{i,j_e}^h = \sum_A N_{,j}^A c_{i_e}^A$$
$$u_{i,j_e}^h = \sum_A N_{,j}^A d_{i_e}^A$$

These gradients show up in the weak form.

$$\int_\Omega w_{i,j}^h \sigma_{ij}\, dV = \int_\Omega w_i^h f_i\, dV + \sum_{i=1}^{N_{sd}} \int_{\partial\Omega_{\bar{t}_i}} w_i^h \bar{t}_i\, dS$$

For the first term, we can invoke the constitutive relation

$$\sigma_{ij}^h = \mathbb{C}_{ijkl}\epsilon_{kl}^h$$

and the kinematic relation

$$\epsilon_{kl}^h = \frac{1}{2}\left(\frac{\partial u_k^h}{\partial x_l} + \frac{\partial u_l^h}{\partial x_k}\right),$$

which says the strain is the symmetric part of the displacement gradient.

Since the the elasticity tensor has the property of minor symmetry, we are free to not worry about the the symmetrization of the displacement gradient. As long as we ensure minor symmetry, we can say that

$$\mathbb{C}_{ijkl}\epsilon_{kl}^h = \mathbb{C}_{ijkl}u_{k,l}^h$$

Within the gradients to finite dimensional trial solution and weighting function, we are differentiating with respect to x_i. We have established a mapping from $\underline{\xi}$ to $\underline{x}$. If we consider the inverse mapping $\underline{\xi}(\underline{x})$ then the following holds by the chain rule of differentiation

$$N^A_{,j} = N_{,I}\xi_{I,j}$$

The capital indices denote components of $\underline{\xi}$.

To obtain the terms $\xi_{I,j}$, we will use the inverse of the Jacobian, $J^{-1}(\underline{\xi})$.

10.8 Element integrals - I

Since we are partitioning the domain and boundaries and subdomains and sub-boundaries, we can write the finite dimensional weak form as summation of integrals over the partitions.

$$\sum_e \int_{\Omega^e} w^h_{i,j}\sigma^h_{ij}\,dV = \sum_e \int_{\Omega^e} w^h_i f_i\,dV + \sum_{i=1}^{N_{sd}} \sum_{e\in E_i} \int_{\partial\Omega^e_{t_i}} w^h_i \bar{t}_i\,dS$$

The left most integral is the term for the Neumann boundary condition. The first summation iterates over the spatial dimensions. For the spatial dimension indexed by i, we account for the contributions from all elements Ω^e, where $e \in E_i$.

Let's look at the first left hand side integral.

$$\begin{aligned}\int_{\Omega^e} w^h_{i,j}\sigma^h_{ij}\,dV &= \int_{\Omega^e} w^h_{i,j}\mathbb{C}_{ijkl}u^h_{k,l}\,dV \\ &= \int_{\Omega^e} \left(\sum_A N^A_{,j}c^A_{i_e}\right)\mathbb{C}_{ijkl}\left(\sum_B N^B_{,l}d^B_{k_e}\right)dV\end{aligned}$$

Since the degrees of freedom do not depend on position, we will pull them out of the integral.

$$\sum_{A,B=1}^{N_{ne}} c^A_{i_e}\left(\int_{\Omega^e} N^A_{,j}\mathbb{C}_{ijkl}N^B_{,l}\,dV\right)d^B_{k_e}$$

The next step is to do a change of variables into the bi-unit parent domain.

$$\sum_{A,B=1}^{N_{ne}} c_{i_e}^A \left(\int_{\Omega^\xi} N_{,j}^A \mathbb{C}_{ijkl} N_{,l}^B \det \left[\underline{\underline{J}}(\underline{\xi}) \right] \, dV^\xi \right) d_{k_e}^B$$

We can perform the integral by using Gaussian quadrature. Remember that if $\mathbb{C}$ is dependent on position in a non-polynomial relationship, then Gaussian quadrature is sub-optimal.

Let's look at the integral. It's clear that the indices j and l will be contracted away, but ik will remain along the superscripts A, B. We will denote the result of that integral as $K_{ik_e}^{AB}$, which is a component of a second order tensor. So we can write the result as

$$\sum_{A,B} c_{i_e}^A K_{ik_e}^{AB} d_{k_e}^B,$$

which gives a scalar quantity, as we expect.

10.9 Element integrals - II

Let's write the result of the last subsection in direct notation.

$$\sum_{A,B} \underline{c}_e^{A^\intercal} \underline{\underline{K}}_e^{AB} \underline{d}_e^B$$

We have

$$\underline{c}_e^A, \underline{d}_e^B \in \mathbb{R}^3$$

and

$$\underline{\underline{K}}_e^{AB} \in GL(3)$$

We said that $\underline{\underline{K}}_e^{AB}$ is a general second order tensor in three dimensions. It is not necessarily symmetric, though we will find that it is symmetric for $A = B$.

Next, consider the integral

$$\int_{\Omega^e} w_i^h f_i \, dV$$

We know that the term w_i^h can be expanded using our basis functions.

$$\int_{\Omega^e} w_i^h f_i \, dV = \int_{\Omega^e} \left(\sum_A N^A c_{i_e}^A \right) f_i \, dV$$

We can pull the summation and nodal degree of freedom out of the integral, and also apply a change of variables.

$$\sum_A c_{i_e}^A \int_{\Omega^\xi} N^A f_i \det \left[\underline{\underline{J}}(\underline{\xi}) \right] \, dV^\xi$$

We will denote the result of the integral as $F_{i_e}^{int^A}$. It's clear that the integral evaluates to a vector quantity since i and A are left to be free indices.

In direct notation, we have

$$\sum_A \underline{c}_e^{A^\intercal} \underline{F}_e^{int^A}$$

10.10 The matrix-vector weak form - I

The last integral we consider is the following:

$$\int_{\partial\Omega_{\bar{t}_i}^e} w_i^h \bar{t}_i \, dS$$

Remember that there is no summation over the index i implied in this case.

We can expand the component of the weighting function by using our basis functions.

$$\int_{\partial\Omega_{\bar{t}_i}^e} w_i^h \bar{t}_i \, dS = \int_{\partial\Omega_{\bar{t}_i}^e} \left(\sum_A N^A c_{i_e}^A \right) \bar{t}_i \, dS$$

Again we can take the sum and the nodal degree of freedom out of the integral.

$$\sum_A c_{i_e}^A \int_{\partial\Omega_{\bar{t}_i}^e} N^A \bar{t}_i \, dS$$

The next step is to do a change of variables and turn this into an integral over a a surface of our bi-unit parent domain. Over $\partial\Omega_{\bar{t}_i}^e$, we may be need to construct local coordinates so we can get the correct mapping. We will call these $\widetilde{x}_1$ and $\widetilde{x}_{2s}$. The local coordinate Jacobian is by definition

$$\underline{J}_s = \frac{d\underline{\widetilde{x}}}{d\underline{\xi}}$$

Remember that $\underline{\xi}$ is only two of the three possible coordinate axes in the parent domain. The choice of which $\underline{\xi}$ component to use is our choice.

Since it is likely that not all faces of the node will lie on the Neumann boundary, we will specify the index of nodes which lie on the Neumann boundary by the set $\mathcal{A}_N$. The summation will also take place over this set of indices.

$$\sum_{A \in \mathcal{A}_N} c_{i_e}^A \int_{\partial\Omega_{\bar{t}_i}^\xi} N^A \bar{t}_i \det\left[\underline{J}_s\right] dS^\xi$$

Depending on the choice of the face in the parent domain, dS^ξ will differ.

Let's denote the result of our integral here as $\overline{F}_{i_e}^{\bar{t}^A}$,

$$\sum_{A \in \mathcal{A}_N} c_{i_e}^A \overline{F}_{i_e}^{\bar{t}^A}$$

We can define a new quantity so we can get rid of the restriction over the node indices. Let $F_{i_e}^{\bar{t}^A}$ be

$$F_{i_e}^{\bar{t}^A} = \begin{cases} \overline{F}_{i_e}^{\bar{t}^A} & \text{if } A \in \mathcal{A}_N \\ 0 & \text{otherwise} \end{cases}$$

At this point, our finite dimensional weak form is

$$\sum_e \sum_{A,B} \underline{c}_e^{A\intercal} \underline{K}_e^{AB} \underline{d}_e^B = \sum_e \sum_A \underline{c}_e^{A\intercal} \underline{F}_e^{int^A} + \sum_{i=1}^{N_{sd}} \sum_{e \in E_i} \sum_A c_{i_e}^A \bar{F}_{i_e}^{\bar{t}^A}$$

10.11 The matrix-vector weak form - II

Each node in an element e is associated with degree of freedom vectors $\underline{c}_e^A$ and $\underline{d}_e^B$.

Let's define a vector $\underline{c}_e^\intercal$ and $\underline{d}_e$ which are vectors containing all the degrees of freedom for Ω^e

$$\underline{c}_e^\intercal = \begin{bmatrix} \underline{c}_e^1 & \underline{c}_e^2 & \dots & \underline{c}_e^{N_{ne}} \end{bmatrix} \qquad \underline{d}_e = \begin{bmatrix} \underline{d}_e^1 \\ \underline{d}_e^2 \\ \vdots \\ \underline{d}_e^{N_{ne}} \end{bmatrix}$$

Its clear that

$$\underline{c}_e^\intercal, \underline{d}_e^\intercal \in \mathbb{R}^{N_{ne} \times N_{sd}}$$

Defining these vectors will allow us to get rid of the summation over A and B. This will also increase the dimensionality of the $\underline{K}_e$ matrix and it will no longer depend on A and B. Since we are also getting rid of the summation over A in first term on the right hand side of the summation symbol, $\underline{F}$ will not depend on A anymore.

So the finite dimensional matrix vector weak form is

$$\sum_e \underline{c}_e^\intercal \underline{K}_e \underline{d}_e = \sum_e \underline{c}_e^\intercal \underline{F}_e^{int} + \sum_{i=1}^{N_{sd}} \sum_{e \in E_i} \sum_A c_{i_e}^A \bar{F}_{i_e}^{\bar{t}^A}$$

Since $\underline{c}_e^\intercal$ is a $1 \times \left(N_{ne} \times N_{sd}\right)$ - 1 by 24 row vector - and $\underline{d}_e$ is a $\left(N_{ne} \times N_{sd}\right) \times 1$ column vector, $\underline{K}_e$ must have dimensions $\left(N_{ne} \times N_{sd}\right)^2$. Similarly, $\underline{F}_e^{int}$ must be a $\left(N_{ne} \times N_{sd}\right) \times 1$ column vector.

The matrix $\underline{\underline{K}}_e$ can be thought of a 8×8 matrix as made of 3×3 blocks, $\underline{\underline{K}}_e^{AB}$ for $A, B = 1, \ldots, N_{n_e}$:

$$\underline{\underline{K}}_e = \begin{bmatrix} \underline{\underline{K}}_e^{11} & \cdots & \underline{\underline{K}}_e^{1N_{ne}} \\ \vdots & & \\ \underline{\underline{K}}_e^{N_{ne}1} & \cdots & \underline{\underline{K}}_e^{N_{ne}N_{ne}} \end{bmatrix}$$

We can look as $\underline{F}_e^{int}$ as a $N_{n_e} \times 1$ column vector made of 3×1 vectors

$$\underline{F}_e^{int} = \begin{bmatrix} \underline{F}_e^{int^1} \\ \vdots \\ \underline{F}_e^{int^{N_{ne}}} \end{bmatrix}$$

10.12 Assembly of the global matrix-vector equations - I

Taking off from the matrix vector equations we wrote last time, and the next step will be to perform finite element assembly.

To do so, let's first define the global $\underline{c}$ and $\underline{\overline{d}}$ vectors which will contain every entry of each nodal degree of freedom for each element. In particular, $\underline{c}$ contains only the degrees of freedom entries that is not associated with nodes on the Dirichlet boundary (they are by definition zero so they are discarded). $\underline{\overline{d}}$ will have N_D more entries than $\underline{c}$, where N_D is the number of degree of freedom vector entries associated Dirichlet boundary conditions.

In general, N_D need not be a a multiple spatial dimension, N_{sd}. This is because we can apply Dirichlet conditions to some particular coordinate

direction and not on others.

$$\underline{c} = \begin{bmatrix} c_1^1 \\ c_2^1 \\ c_3^1 \\ \vdots \\ c_1^A \\ c_2^A \\ c_3^A \\ \vdots \end{bmatrix} \qquad \overline{\underline{d}} = \begin{bmatrix} d_1^1 \\ d_2^1 \\ d_3^1 \\ \vdots \\ d_1^A \\ d_2^A \\ d_3^A \\ \vdots \\ d_1^{N_{ne}} \\ d_2^{N_{ne}} \\ d_3^{N_{ne}} \end{bmatrix}$$

In writing $\underline{c}$ we are assuming that global node 1 is not on the Dirichlet boundary.

Also recall that d_i^A entries that correspond to the Dirichlet condition will be taken out of $\overline{\underline{d}}$. This will reduce the dimensionality of the d vector.

So let's rewrite

$$\sum_e \underline{c}_e^\intercal \underline{K}_e \underline{d}_e = \underline{c}^\intercal \overline{\underline{K}} \overline{\underline{d}},$$

where $\overline{\underline{K}}$ is obtained from assembling each element stiffness matrix.

$$\overline{\underline{K}} = \mathop{\mathbf{A}}_{e=1}^{N_{el}} \underline{K}_e$$

$\overline{\underline{K}}$ has dimensions $(N_n \times N_{sd} - N_D) \times (N_n \times N_{sd})$.

For figure 17, we know that the element stiffness matrix for the two elements will take the following form:

$$\underline{K}_{e_1} = \begin{bmatrix} \underline{K}_{e_1}^{11} & \cdots & \underline{K}_{e_1}^{18} \\ \vdots & \ddots & \vdots \\ \underline{K}_{e_1}^{81} & \cdots & \underline{K}_{e_1}^{88} \end{bmatrix} \qquad \underline{K}_{e_2} = \begin{bmatrix} \underline{K}_{e_2}^{11} & \cdots & \underline{K}_{e_2}^{18} \\ \vdots & \ddots & \vdots \\ \underline{K}_{e_2}^{81} & \cdots & \underline{K}_{e_2}^{88} \end{bmatrix}$$

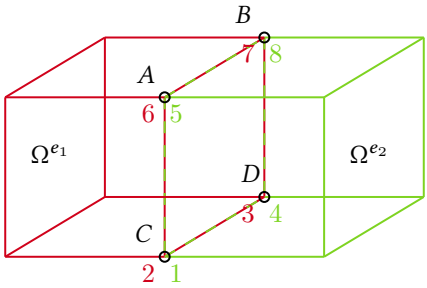

Figure 17: Consider the assembly for a pair of hexahedral elements Ω^{e_1} and Ω^{e_2}. The red and green numbers are the local node numbering for the corresponding element. A, B, C, D are the global node numbering for the 4 nodes shared between Ω^{e_1} and Ω^{e_2}.

10.13 Assembly of the global matrix-vector equations - II

Continuing from the previous subsection, we can zoom into a portion of $\overline{\underline{K}}$ matrix with row numbers A, B, C, D counting from top to bottom, and columns numbers A, B, C, D from left to right.

$$\overline{\underline{K}} = \begin{bmatrix} \underline{K}_{e_1}^{66} + \underline{K}_{e_2}^{55} & & & \underline{K}_{e_1}^{63} + \underline{K}_{e_2}^{54} \\ & \underline{K}_{e_1}^{77} + \underline{K}_{e_2}^{88} & & \\ & & & \\ & & \underline{K}_{e_1}^{32} + \underline{K}_{e_2}^{41} & \end{bmatrix}$$

We wrote in a few of the entries. The superscript numbering on each of the contributions at positions $\overline{K}_{ij}$ correspond to the local numbering of the global nodes i and j. For example, at row A and column D of the matrix, the superscript on the e_1 contribution is 63, since global node A is local node 6 and global D is local node 3 in e_1, as we defined in figure 17.

As we know, there will be more contributions at each node and column from elements who also share node A, B, C, D.

Similarly, we can see how Ω^{e_1} and Ω^{e_2} contribute to global forcing vector, $\underline{F}^{int}$. The global forcing vector is obtained from assembling all the contribu-

tions from each element

$$\sum_e \underline{c}_e^{\intercal} \underline{F}_e^{int} = \underline{c}^T \underline{F}^{int}$$

Let's zoom into rows A, B, C, D for the forcing vector:

$$\underline{F}^{int} = \begin{bmatrix} \underline{F}_{e_1}^{int^6} + \underline{F}_{e_2}^{int^5} \\ \underline{F}_{e_1}^{int^7} + \underline{F}_{e_2}^{int^8} \\ \underline{F}_{e_1}^{int^2} + \underline{F}_{e_2}^{int^1} \\ \underline{F}_{e_1}^{int^3} + \underline{F}_{e_2}^{int^4} \end{bmatrix}$$

Again, there will likely be more contributions in each row from other elements that also share global nodes A, B, C, D.

10.13.1 Coding assignment 3 - II

Recall that we described the element stiffness matrices as being $N_{n_e} \times N_{n_e}$ (8×8), where each entry is a 3 by 3 matrix. In the actual implementation, $\underline{K}_e$ is a 24×24 matrix. Each entry in this matrix is $K_{ik_e}^{AB}$, where $i, k = 1, \ldots, N_{sd}$.

$$K_{ik_e}^{AB} = \int_{\Omega^\xi} N_{,j}^A \mathbb{C}_{ijkl} N_{,l}^B \det \left[\underline{J}(\underline{\xi}) \right] \, dV^\xi$$

Given A, B, i, k the element of the matrix is accessed as follows:

```
Klocal[3*A + i][3*B + k]
```

10.13.2 Dirichlet boundary conditions - I

There is one term we have yet to tackle:

$$\sum_{i=1}^{N_{sd}} \sum_{e \in E_i} \sum_A c_{i_e}^A \bar{F}_{i_e}^{t^A}$$

This is the traction/Neumann boundary condition term.

We state that this term will show up as

$$\sum_{i=1}^{N_{sd}} \sum_{e \in E_i} \sum_{A} c_{i_e}^{A} \bar{F}_{i_e}^{A} = \underline{c}^{\intercal} \underline{\bar{F}}$$

To write out what $\underline{c}^{\intercal} \underline{\bar{F}}$ is, we will consider each of the three possible Neumann boundaries (one for each spatial dimension) one at a time.

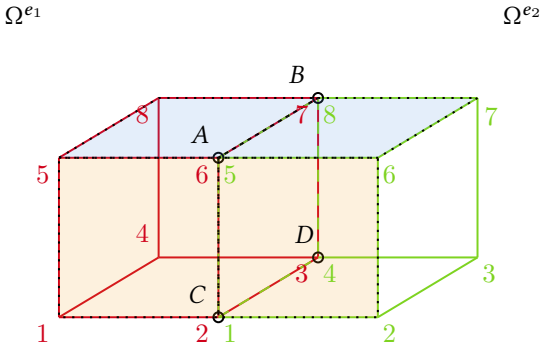

Figure 18: Recall Ω^{e_1}, Ω^{e_2}. The two faces of the elements shaded in orange represent that they they coincide with $\partial\Omega_{\bar{t}_1}$. The two blue faces represent that they coincide with $\partial\Omega_{\bar{t}_2}$.

Following figure 18, we will write the entries in the $\underline{c}^{\intercal}$ and $\underline{\bar{F}}$ corresponding to nodes A, B, C, D.

Each node will be associated with three columns in the $\underline{c}^{\intercal}$ row vector, corresponding to three weighting function degrees of freedom - given that none of the dof vector component correspond to a Dirichlet boundary condition. Likewise, each node will be associated with three rows in the $\underline{\bar{F}}$ column vector. A row associated with a node will be zero if the that node does not lie on any of $\partial\Omega_{\bar{t}_i}$, for $i = 1, 2, 3$.

Only showing the contributions from global nodes A, B, C, D:

$$\underline{c}^{\intercal} = \begin{bmatrix} \dots & c_1^A\, c_2^A\, c_3^A & \dots & c_1^B\, c_2^B\, c_3^B & \dots & c_1^C\, c_2^C\, c_3^C & \dots & c_1^D\, c_2^D\, c_3^D & \dots \end{bmatrix}$$

$$\underline{F}^{\bar{t}} = \begin{bmatrix} \vdots \\ \underline{F}^{\bar{t}^6}_{1_{e_1}} + \underline{F}^{\bar{t}^5}_{1_{e_2}} \\ \underline{F}^{\bar{t}^6}_{2_{e_1}} + \underline{F}^{\bar{t}^5}_{2_{e_2}} \\ 0 \\ \vdots \\ 0 \\ \underline{F}^{\bar{t}^7}_{2_{e_1}} + \underline{F}^{\bar{t}^8}_{2_{e_2}} \\ 0 \\ \vdots \\ \underline{F}^{\bar{t}^2}_{1_{e_1}} + \underline{F}^{\bar{t}^1}_{1_{e_2}} \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

10.14 Dirichlet boundary conditions - II

In the previous subsections, we took care of converting the sum over elements in the matrix vector weak form. The only thing left to do is to account for the Dirichlet boundary conditions in the global matrix vector weak form.

$$\underline{c}^\intercal \overline{\underline{K}\underline{d}} = \underline{c}^\intercal \underline{F}^{int} + \underline{c}^\intercal \underline{F}^{\bar{t}}$$

The dimensions of $\overline{\underline{K}}$ is $(N_n \times N_{sd} - N_d) \times (N_n \times N_{sd})$. N_n is the number of nodes in the problem, N_{sd} is the number of spatial dimensions, and N_d is the total number of weighting function degree of freedom vector components that are zero due to the Dirichlet boundary condition.

Suppose that d_1^A, d_2^B are known Dirichlet boundary conditions. We know when we carry out the matrix vector product, columns $N_{sd}\times A+1$ and $N_{sd}\times B+2$ will be known.

We can move the columns and the known degrees of freedom to the right hand side. This reduces the dimensionality of $\underline{\overline{K}}$ and $\underline{\overline{d}}$. $\underline{\overline{K}}$ is now a $(N_n \times N_{sd} - N_d)^2$ square matrix, which we will denote without the overline. $\underline{\overline{d}}$ has N_d fewer entries, and we will also write it without the overline.

$$\underline{c}^\intercal \underline{K}\underline{d} = \underline{c}^\intercal \underbrace{\left(\underline{F}^{int} + \underline{F}^{\bar{t}} - d_1^A \underline{\overline{K}}^{N_{sd}\times A+1} - d_2^B \underline{\overline{K}}^{N_{sd}\times B+2}\right)}_{\underline{F}}$$

$$\implies \underline{c}^\intercal \left(\underline{K}\underline{d} - \underline{F}\right) = 0, \ \forall \ \underline{c} \in \mathbb{R}^{N_n \times N_{sd} - N_D}$$

Since the weak form must hold for all weighting function in $\mathcal{V}^h$ this is enforced in the fact that the equation must also hold for all $\underline{c} \in \mathbb{R}^{N_n \times N_{sd} - N_D}$. This allows us to write

$$\underline{K}\underline{d} = \underline{F},$$

which we can use to solve for $\underline{d}$.

There exists a solution $\underline{d}$ when $\underline{K}$ is invertible. $\underline{K}$ is invertible when it is a positive definite matrix. The fact that $\underline{K}$ is positive definite comes from the $\underline{\mathbb{C}}$ is positive definite, and that the "Dirichlet boundary conditions eliminate rigid body motion".

10.14.1 Unit 10 Quiz

1. Consider the constitutive relations for linearized elasticity for an isotropic material and suppose that the Poisson ratio = 0 for this material. Which of the following is true:

    - There is no coupling between the stress and strain in mutually orthogonal directions

2. Question 2 In steady state linearized elasticity, you cannot specify the u_1 displacement component as a Dirichlet boundary condition, while also specifying the T_2 and T_3 components of the traction as Neumann conditions on that same boundary.

- False

3. Consider two neighboring hexahedral elements in a finite element implementation of steady state linearized elasticity, and suppose that they share a common face that does not lie on any of the domain boundaries, nor intersect any of them. When you assemble the global stiffness matrix, $\underline{K}$ for this problem, how many entries in this matrix will have contributions from both elements?

   - 12

4. Question 4 In linearized elasticity all components of the stress can be specified as a Neumann condition on some boundary.

   - False

11 Linear, parabolic partial differential equations for a scalar unknown in three dimensions (Unsteady heat conduction and mass diffusion)

11.1 The strong form

In this new unit, we will move to parabolic PDEs for a scalar variable in three dimensions. The problems we will consider are unsteady are unsteady (time dependent) mass and heat conduction.

The additional component of our problem is that every point on the domain will have an additional first order time dependent term. (The nature of heat conduction and mass diffusion is first order.)

In three dimensions, we can think of a small volume element at some location $\underline{x}$ with some flux flowing in and out of it specified by the net flux vector $\underline{j}$, as well as some internal heating specified by $\underline{f}$. The temperature at $\underline{x}$ will have a non-zero time derivative.

Here is the strong form. Given u_g, j_n, f, the constitutive relation

$$j = -\kappa_{ij} u_{,j},$$

and a constant ρ, find $u(\underline{x}, t)$ such that

$$\rho \frac{\partial u}{\partial t} = -j_{i,i} + f$$

holds in $\Omega \times [0, T]$ (a combination of the spatial domain and a time $t \in [0, T]$), with the boundary conditions

$$\begin{cases} u = u_g & \text{on } \partial\Omega_u \\ -j_i n_i = j_n & \text{on } \partial\Omega_j \end{cases}$$

The statement of the problem is not complete. Since this is an time-dependent problem, we also need to provide initial conditions. The initial condition is specified as

$$u(\underline{x}, 0) = u_0(\underline{x})$$

The new coefficient, ρ, is specific heat per unit volume. The heat equation arises from the first law of thermodynamics.

If we are considering the case of mass diffusion, then $\rho = 1$.

11.2 The weak form, and finite-dimensional weak form - I

Recall the strong form of our problem.

Strong form of linear parabolic PDE for a scalar variable: Given u_g, j_n, f, the constitutive relation

$$j_i = -\kappa_{ij} u_{,j},$$

and the specific heat ρ, find $u(\underline{x}, t)$ such that the following PDE holds on $\Omega \times [0, T]$

$$\rho \frac{\partial u}{\partial t} = -j_{i,i} + f,$$

with the boundary conditions

$$\begin{cases} u = u_g & \text{on } \partial\Omega_u \\ -j_i n_i = j_n & \text{on } \partial\Omega_j \end{cases}$$

and the single initial condition (arising due to our first order time dependence)

$$u(\underline{x}, 0) = u_0(\underline{x})$$

We will arrive at the weak form of the equation by multiplying the strong form by a weighting function, and integrating over the domain.

Consider $w \in \mathcal{V} = \{w | w = 0 \text{ on } \partial\Omega_u\}$:

$$\int_\Omega w \rho \frac{\partial u}{\partial t} \, dV = \int_\Omega -w j_{i,i} \, dV + \int_\Omega w f \, dV$$

Apply integration by parts to the term in the middle:

$$\int_\Omega w\rho\frac{\partial u}{\partial t}\,dV = \int_\Omega w_{,i}j_i\,dV + \int_\Omega wf\,dV - \int_{\partial\Omega} w\underbrace{j_i n_i}_{=-j_n}\,dS$$

The right most surface integral is obtained from applying the divergence theorem. We also see that the dot product between $\underline{j}$ and $\underline{n}$ is the heat in flux vector, $-j_n$.

The surface integral can be broken up into integration over the Neumann and Dirichlet boundaries. By definition of w, integration over the Dirichlet boundary will be zero.

$$\underbrace{\int_{\partial\Omega_u} wj_i n_i\,dS}_{=0} + \int_{\partial\Omega_j} wj_i n_i\,dS$$

Let's also invoke the constitutive relation, $j_i = -\kappa_{ij}u_{,j}$, giving

$$\int_\Omega w\rho\frac{\partial u}{\partial t}\,dV + \int_\Omega w_{,i}\kappa_{ij}u_{,j}\,dV = \int_\Omega wf\,dV + \int_{\partial\Omega_j} wj_n\,dS$$

Before we finish off, let's write out the finite dimensional weak form, realizing that any attempts to solve the infinite dimensional weak form is no easier than solving the strong form.

Finite-dimensional weak form of a 3D linear parabolic PDE for a scalar variable: Our goal is to find $u^h \in \mathcal{S}^h \subset S, \mathcal{S}^h = \{u^h \in H^1(\Omega) | u^h = u_g \text{ on } \partial\Omega_u\}$ such that for call $w^h \in \mathcal{V}^h \subset \mathcal{V}, \mathcal{V}^h = \{w^h \in H^1(\Omega) | w^h = 0 \text{ on } \partial\Omega_u\}$, the following holds

$$\int_\Omega w^h\rho\frac{\partial u^h}{\partial t}\,dV + \int_\Omega w^h_{,i}\kappa_{ij}u^h_{,j}\,dV = \int_\Omega w^h f\,dV + \int_{\partial\Omega_j} w^h j_n\,dS$$

11.3 The weak form, and finite-dimensional weak form - II

With the finite dimensional weak form, we then partition our domain into subdomains. On each of the element subdomains, the trial solution now depends on time. What we will do is to allow time dependence in the nodal degrees of freedoms:

$$u_e^h(\underline{x}, t) = \sum_A^{N_{ne}} N^A(\underline{x}(\underline{\xi}))d_e^A(t)$$

By expanding the trial solution as a sum of basis functions, we are discretizing the domain. We have not discretized time. For this reason, our formulation of is also call a semi-discrete finite element formulation.

For the weighting function, we see from the finite dimensional weak form that there are no integration over time, or derivatives in time. So the weighting function do not require any time dependence.

$$w_e^h = \sum_A^{N_{ne}} N^A(\underline{x}(\underline{\xi}))c_e^A$$

With the procedure we have developed in the past weeks, we can say immediately say that the following holds:

$$\int_\Omega w_{,i}^h \kappa_{ij} u_{,j}^h \, dV = \underline{c}^\intercal \underline{\underline{K}} \underline{d}(t)$$
$$\int_\Omega w^h f \, dV + \int_{\partial\Omega_j} w^h j_n \, dS = \underline{c}^\intercal \underline{F}$$

We can have time dependence in the Dirichlet boundary condition, $u_g = u_g(t)$, heat influx, $j_n = j_n(t)$, and $f = f(\underline{x}, t)$. So, $\underline{F}$ can equal to $\underline{F}(t)$.

11.4 Basis functions, and the matrix-vector weak form - I

The new term involving the time derivative is the one we need to worry about.

$$\int_\Omega w^h \rho \frac{\partial u^h}{\partial t}\, dV$$

the time derivative on u over each element is

$$\left[\frac{\partial u^h}{\partial t}\right]_e = \sum_A N^A \dot{d}_e^A,$$

where the dot over d_e^A denotes the time derivative.

Let's write the integral over the domain as a sum over elements:

$$\int_\Omega w^h \rho \frac{\partial u^h}{\partial t}\, dV = \sum_e \int_{\Omega^e} \left[\sum_A N^A c_e^A\right] \rho \left[\sum_B N^B \dot{d}_e^B\right] dV$$

Let's manipulate the terms a little:

$$\sum_e \sum_{A,B} c_e^A \underbrace{\left(\int_{\Omega^e} N^A \rho N^B\, dV\right)}_{M_e^{AB}} \dot{d}_e^B$$

The integral, which we denote as M_e^{AB} is a term whose combination of components we have not encountered.

Over the element, we can assemble the product into matrix vector form. Again, we will let

$$\underline{c}_e^\intercal = \left[c_e^1 \dots c_e^{N_{ne}}\right]$$

and define the vector of time dependent degree of freedoms.

$$\underline{\dot{d}}_e^\intercal = \left[\dot{d}_e^1 \dots \dot{d}_e^{N_{ne}}\right]$$

Defining the element level vectors lets us get rid of the summation over A and B.

$$\sum_e \sum_{A,B} c_e^A \underbrace{\left(\int_{\Omega^e} N^A \rho N^B \, dV \right)}_{M_e^{AB}} \dot{d}_e^B = \sum_e \underline{c}_e^\intercal \overline{\underline{M}}_e \dot{\overline{\underline{d}}}_e$$

The matrix $\overline{\underline{M}}_e$ is often called the element mass matrix. This name is attributed to any matrix that is formed by integrating the product of basis functions (no derivatives) over an element. We also added bars over the mass matrix and trial solution degrees of freedom in anticipation that the dimensions of these objects will be reduced if the element lies on a Dirichlet boundary.

For a general element Ω^e, such that $\partial\Omega^e \cap \partial\Omega_u = \emptyset$, $\underline{c}_e^\intercal$ will have dimensions $1 \times N_{n_e}$, and $\overline{\underline{M}}_e$ will be a square matrix of dimensions $N_{n_e} \times N_{n_e}$.

- The element mass matrix will also be symmetric, since $N^A N^B = N^B N^A$.
- One can also show that the element mass matrix is a positive definite matrix:

$$\underline{c}_e^\intercal \overline{\underline{M}}_e \underline{c}_e \geq 0 \; \forall \; \underline{c}_e \in \mathbb{R}^{N_{n_e}}$$

- The element mass matrix is also called the consistent element mass matrix.
- A lumped element mass matrix is formed by summing the columns on every row, and putting the sum on the diagonal, while setting other entries to be zero.

$$\widetilde{\underline{M}}_e^{AB} = \begin{bmatrix} \sum_B M_e^{1B} & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & \sum_B M_e^{N_{n_e} N_{n_e}} \end{bmatrix}$$

In another manner, we can define the lumped matrix as

$$\widetilde{\underline{M}}_e = \begin{cases} \sum_{c=1}^{N_{n_e}} M_e^{AC} & A = B \\ 0 & A \neq B \end{cases}$$

11.5 Basis functions, and the matrix-vector weak form - II

In the last section, we concluded that

$$\int_\Omega w^h \rho \frac{\partial u^h}{\partial t}\, dV = \sum_e \underline{c}_e^\intercal \overline{\underline{M}}_e \overline{\underline{\dot{d}}}_e$$

The global degree of freedom vectors are:

$$\underline{c}^\intercal = \begin{bmatrix} c^1 & c^2 & \dots & c^{N_{n_e}-N_D} \end{bmatrix}$$
$$\overline{\underline{\dot{d}}}^\intercal = \begin{bmatrix} \overline{\dot{d}}^1 & \overline{\dot{d}}^2 & \dots & \overline{\dot{d}}^{N_{n_e}-N_D} \end{bmatrix}$$

The size of the $\underline{c}$ vector is $(N_{n_e}-N_D)\times 1$ since we remove the weighting function degrees of freedom associated with nodes that are on the Dirichlet boundary.

The global mass matrix is

$$\overline{\underline{M}} = \mathop{\mathbin{\mathop{\Large A}}}_{e=1}^{N_{el}} \overline{\underline{M}}_e$$

11.6 Dirichlet boundary conditions; the final matrix-vector equations

Recall our finite dimensional weak form:

$$\underbrace{\int_\Omega w^h \rho \frac{\partial u^h}{\partial t}\, dV}_{\underline{c}^\intercal \overline{\underline{M}}\overline{\underline{\dot{d}}}} + \underbrace{\int_\Omega w^h_{,i} \kappa_{ij} u^h_{,j}\, dV}_{\underline{c}^\intercal \underline{K}\underline{d}} = \underbrace{\int_\Omega w^h f\, dV + \int_{\partial\Omega_j} w^h j_n\, dS}_{\underline{c}^\intercal \underline{F}}$$

If were working with a steady state problem, the first term on the left involving the mass matrix will not arise. Say that we have accounted for the Dirichlet boundary conditions on the other terms, we will look at what to do with the Dirichlet boundary condition on the new time-dependent term.

Let's write out $\underline{c}^\intercal \underline{\overline{M}\dot{\overline{d}}}$:

$$\begin{bmatrix} c^1 \dots c^{N_{ne}-N_D} \end{bmatrix} \begin{bmatrix} \left\{\underline{\overline{M}}^A\right\} & \left\{\underline{\overline{M}}^B\right\} \end{bmatrix} \begin{bmatrix} \dot{\overline{d}}^1 \\ \vdots \\ \dot{\overline{d}}^A \\ \vdots \\ \dot{\overline{d}}^B \\ \vdots \\ \dot{\overline{d}}^{N_{ne}} \end{bmatrix}$$

Suppose that the degrees of freedom $\dot{\overline{d}}^A$ and $\dot{\overline{d}}^B$ are given by the Dirichlet boundary condition; $\left\{\underline{\overline{M}}^A\right\}$ represents the column A of the global mass matrix, which is multiplied with $\dot{\overline{d}}^A$ degree of freedom.

We will remove columns A and B from the global mass matrix and the corresponding time dependent trial solution degree of freedom from $\underline{\dot{\overline{d}}_e}$, and move them to the right hand side:

$$\underline{c}^\intercal \underline{M\dot{d}} + \underline{c}^\intercal \underline{Kd} = \underline{c}^\intercal \underbrace{\left(\underline{F} - \dot{d}^A \underline{\overline{M}}^A - \dot{d}^B \underline{\overline{M}}^B\right)}_{\underline{F}}$$

Let's rename the sum of the terms within the right hand side bracket to be $\underline{F}$, the updated global forcing vector with time dependent degree of freedoms.

Rearranging,

$$\underline{c}^\intercal \left(\underline{M\dot{d}} + \underline{Kd} - \underline{F}\right) = 0 \;\forall\; \underline{c} \in \mathbb{R}^{N_{ne}-N_D} \implies \underline{M\dot{d}} + \underline{Kd} = \underline{F}$$

This is the equation for our semi-discrete matrix vector problem. It is semi-discrete because we have only discretized spatially.

In the coming section, we will look at how to take care of the time dependent terms.

11.7 Time discretization; the Euler family - I

We came to the following matrix vector equation in our last subsection.

$$\underline{M}\dot{\underline{d}} + \underline{K}\underline{d} = \underline{F}$$

We realize that our boundary conditions have already been accounted for.

We can then think of the equation as a first order ODE for the vector unknown $\underline{d}$, where $\underline{d} \in \mathbb{R}^{N_{df}}$, where df is an integer representing the number of degrees of freedom we are solving for. First order ODEs requires initial conditions.

In our statement of the strong form, the initial condition was:

$$u(\underline{x}, 0) = u_0(\underline{x})$$

This translates into our matrix-vector weakform as

$$\underline{d}(0) = \underline{d}_0 = \begin{bmatrix} u_0(\underline{x}^A) \\ \vdots \\ u_0(\underline{x}^{N_{df}}) \end{bmatrix}$$

is a column vector of u evaluated at $t = 0$ at each nodal point.

Our global mass matrix was made of consistent element mass matrices. We mentioned that it was also possible to use a lumped element mass matrix. We can define globally lumped mass matrix as a matrix the diagonal entry on row A is equal to sum of all entries in row A of the global consistent mass matrix. In other words,

$$M_l^{AB} = \begin{cases} \sum_C M^{AC} & A = B \\ 0 & A \neq B \end{cases}$$

The subscript l denotes "lumped".

We can look at the integration algorithms for our problem, which depends on how we choose to perform the time discretization.

In this series of lectures, we will use a finite difference method in time. Space-time Galerkin finite element methods do exist. They carry out integration over Ω and $[0,T]$. The accuracy with respect to time of these methods are of higher order than what we will be developing here.

The first step is the divide the time interval $[0,T]$ into N sub-intervals: $[t_0,t_1],[t_1,t_2],\ldots,[t_{N-1},t_N]$, such that $t_0=0$, $t_N=T$.

Consider an interval $[t_n,t_{n+1}]$, $n\in[0,N-1]$. We will perform time stepping: "knowing the solution at t_n, find the solution at t_{n+1}".

We define $d(t_n)$ as the time-exact solution at $t=t_n$. (If we were able to integrate our ODE exactly, then this is the solution we will get at t_n.)

We define d_n as the algorithmic solution we obtain by applying a method to integrate the time-discretized ODE.

We will go from the time-exact ODE:

$$\underline{M}\dot{\underline{d}}(t_n)+\underline{K}\underline{d}(t_n)=\underline{F}(t_n)$$

to the time discrete ODE:

$$\underline{M}\underline{v}_n+\underline{K}\underline{d}_n=\underline{F}_n,$$

where $\underline{v}_n$ stands for "velocity", used in the sense that this is time-discretized approximation the time-rate of change of $\underline{d}(t_n)$ degrees of freedom.

11.8 Time discretization; the Euler family - II

Consider the time-discretized ODE at time t_{n+1}:

$$\underline{M}\underline{v}_{n+1}+\underline{K}\underline{d}_{n+1}=\underline{F}_{n+1},$$

The integration algorithm we will invoke are sometimes called the "Euler family" for first order ODEs. For a general ODE

$$\dot{y}=f(y),$$

the algorithm proposed by the Euler family is

$$\frac{y_{n+1} - y_n}{\Delta t} = f(y_{n+\alpha}),$$

where

$$\Delta t = t_{n+1} - t_n$$

is the time step, and α is a real number in the closed interval $[0, 1]$.

When

$\alpha = 0$	Forward Euler
$\alpha = 1$	Backward Euler
$\alpha = 1/2$	Midpoint rule; Crank-Nicholson method

We can design methods to work for any $\alpha \in [0, 1]$.

We define $y_{n+\alpha}$ as

$$y_{n+\alpha} = \alpha y_{n+1} + (1 - \alpha) y_n$$

In the Euler family algorithms, we approximate every time derivative as a linearly varying quantity over all time interval. The choice of α is where we are able to take liberty. Different choices of α gives different properties for our method.

11.9 The v-form and d-form

Let's apply the idea of the Euler family in our time-discretized ODE. We convert from

$$\underline{M}\underline{v}_{n+1} + \underline{K}\underline{d}_{n+1} = \underline{F}_{n+1},$$

To the following form:

$$\begin{aligned} \underline{d}_{n+1} &= \underline{d}_n + \Delta t \underline{v}_{n+\alpha} \\ \underline{v}_{n+\alpha} &= \alpha \underline{v}_{n+1} + (1 - \alpha) \underline{v}_n, \end{aligned}$$

given $\underline{d}_0$.

There are two approaches to solving this time discretized ODE. First of which is called the "v-method". To begin, we rewrite $\underline{d}_{n+1}$:

$$\begin{aligned}\underline{d}_{n+1} &= \underline{d}_n + \Delta t\left(\alpha \underline{v}_{n+1} + (1-\alpha)\underline{v}_n\right) \\ &= \underbrace{\underline{d}_n + (1-\alpha)\Delta t \underline{v}_n}_{\widetilde{\underline{d}}_{n+1}} + \alpha \Delta t \underline{v}_{n+1} \\ &= \widetilde{\underline{d}}_{n+1} + \alpha \Delta t \underline{v}_{n+1}\end{aligned}$$

The first two terms in the rewritten equation only depends on n. For now, let's denote this term as $\widetilde{\underline{d}}_{n+1}$, for reasons that will become clear.

We are getting into the realm of predictor-corrector methods. The first term $\widetilde{\underline{d}}_{n+1}$ is called the predictor. and the last term $\alpha \Delta t \underline{v}_{n+1}$ is the corrector. ("$\widetilde{\underline{d}}_{n+1}$ is our predictor, it is our guess for $\underline{d}_{n+1}$, which requires the correction term $\alpha \Delta t \underline{v}_{n+1}$".)

The v-method involves rewriting our time-discretized ODE in terms of $\underline{v}$ only.

$$\begin{aligned}&\underline{M}\underline{v}_{n+1} + \underline{K}\widetilde{\underline{d}}_{n+1} + \alpha \Delta t \underline{K}\underline{v}_{n+1} = \underline{F}_{n+1} \\ &\implies \left(\underline{M} + \alpha \Delta t \underline{K}\right)\underline{v}_{n+1} = \underline{F}_{n+1} - \underline{K}\widetilde{\underline{d}}_{n+1} \\ &\implies \underline{v}_{n+1} = \left(\underline{M} + \alpha \Delta t \underline{K}\right)^{-1}\left(\underline{F}_{n+1} - \underline{K}\widetilde{\underline{d}}_{n+1}\right)\end{aligned}$$

Some remarks:

1. If the global lumped mass matrix used, and $\alpha = 0$, then

$$\underline{v}_{n+1} = \underline{M}_l^{-1}\left(\underline{F}_{n+1} - \underline{K}\widetilde{\underline{d}}_{n+1}\right).$$

   This is also called an explicit method. (The lumped mass matrix is a diagonal matrix. Recall that if the diagonal entries of a diagonal matrix are non-zero, then the inverse is a matrix with the original entries raised to the power of -1.)

2. If $\alpha \neq 0$, this is also known as the implicit method.

We come our second method: d-method. Our goal is to eliminate $\underline{v}$ from our equation. Let's isolate $\underline{v}_{n+1}$ from our predictor-corrector form.

$$\underline{d}_{n+1} = \widetilde{\underline{d}}_{n+1} + \alpha\Delta t\underline{v}_{n+1} \implies \underline{v}_{n+1} = \frac{\underline{d}_{n+1} - \widetilde{\underline{d}}_{n+1}}{\alpha\Delta t} \text{ for } \alpha \neq 0$$

We substitute this into our time-discretized ODE:

$$\begin{aligned}
&\underline{M}\underline{v}_{n+1} + \underline{K}\widetilde{\underline{d}}_{n+1} + \alpha\Delta t\underline{K}\underline{v}_{n+1} = \underline{F}_{n+1} \\
&\implies \underline{M}\left(\frac{\underline{d}_{n+1} - \widetilde{\underline{d}}_{n+1}}{\alpha\Delta t}\right) + \underline{K}\underline{d}_{n+1} = \underline{F}_{n+1} \\
&\implies \left(\underline{M} + \alpha\Delta t\underline{K}\right)\underline{d}_{n+1} = \alpha\Delta t\underline{F}_{n+1} - \underline{M}\widetilde{\underline{d}}_{n+1} \\
&\implies \underline{d}_{n+1} = \left(\underline{M} + \alpha\Delta t\underline{K}\right)^{-1}\left(\alpha\Delta t\underline{F}_{n+1} - \underline{M}\widetilde{\underline{d}}_{n+1}\right)
\end{aligned}$$

Use a lumped mass matrix in this case would ease the set up of our right hand side terms.

11.10 Analysis of the integration algorithms for first order, parabolic equations; modal decomposition - I

We will apply modal decomposition to analyze our solution.

To ease our analysis, we will rewrite our equations:

$$\underline{M}\left(\frac{\underline{d}_{n+1} - \underline{d}_n}{\Delta t}\right) + \underline{K}\underline{d}_{n+\alpha} = \underline{F}_{n+\alpha}$$

In the equation above, we directly applied a finite difference approximation to the time-exact time derivative of $\underline{d}$. The Euler family comes in when we say all other terms are being evaluated at $n + \alpha$.

We hope to gain insight to the stability and consistency of the time-integration algorithm. We will consider the homogeneous version of the ODE. ("If there were no forcing, how does time-discretization and the Euler family of algorithms effect the evolution of the problem?")

So our ODE with it's initial condition is:

$$\underline{\underline{M}}\left(\frac{\underline{d}_{n+1}-\underline{d}_n}{\Delta t}\right)+\underline{\underline{K}}\underline{d}_{n+\alpha}=\underline{0}$$

given $\underline{d}_0$.

In order to perform the modal decomposition, we will invoke a related, generalized eigenvalue problem:

$$\underline{\underline{K}}\underline{\phi}=\lambda\underline{\underline{M}}\underline{\phi},\underline{\phi}\in\mathbb{R}^{n_{df}}$$

We recover the standard eigenvalue problem if $\underline{\underline{M}}$ is the identity matrix.

Let λ_m, $m=1,\ldots,n_{df}$, be an eigenvalue, with the corresponding eigenvector $\underline{\phi}_m$. These eigenvalue eigenvector pairs then satisfy

$$\underline{\underline{K}}\underline{\phi}_m=\lambda_m\underline{\underline{M}}\underline{\phi}_m$$

We can show that it is possible to construct a set of vectors $\{\underline{\psi}_m|m=1,\ldots,n_{df}\}$ from the set $\{\underline{\phi}_m|m=1,\ldots,n_{df}\}$ which are orthonormal with respect to $\underline{\underline{M}}$ (or "M-orthonormal").

$$\underline{\psi}_m\cdot\underline{\underline{M}}\underline{\psi}_k=\delta_{mk}$$

11.11 Analysis of the integration algorithms for first order, parabolic equations; modal decomposition - II

To perform the orthnormalization process, we begin with our set of linearly independent eigenvectors $\{\underline{\phi}_m|m=1,\ldots,n_{df}\}$. We know that this set of eigenvectors are linearly independent since both $\underline{\underline{K}}$ and $\underline{\underline{M}}$ are positive definite (this condition guarantees the existence of linearly independent eigenvectors).

The actual algorithm used is the Gram-Schmidt process.

With the M-orthonormal set, $\{\underline{\psi}_m|m=1,\ldots,n_{df}\}$, constructed, we can move on to understand how we can perform our analysis.

A property of the ψ_m vectors are

$$\begin{aligned}\underline{\psi}_m \cdot (\underline{K}\underline{\psi}_k) &= \lambda_k \underline{\psi}_m \cdot (\underline{M}\underline{\psi}_k) \\ &= \lambda_k \delta_{mk}\end{aligned}$$

There are no summation implied over the second index k.

The set $\left\{\underline{\psi}_m\right\}$ are linearly independent, and forms a basis in $\mathbb{R}^{n_{df}}$. Then any vector $\underline{d}$ can be written as a linear combination of vectors in the basis:

$$\underline{d} = \sum_{m=1}^{n_{df}} d_m \underline{\psi}_m$$

To get the coefficients d_m,

$$\begin{aligned}\underline{\psi}_k \cdot \left(\underline{M}\underline{d}\right) &= \underline{\psi}_k \cdot \sum_{m=1}^{n_{df}} d_m \underline{M}\underline{\psi}_m \\ &= \sum_{m=1}^{n_{df}} d_m \underbrace{\underline{\psi}_k \cdot \left(\underline{M}\underline{\psi}_m\right)}_{\delta_{km}} \\ &= d_k\end{aligned}$$

So any coefficient d_k is uniquely given by

$$d_k = \underline{\psi}_k \cdot \left(\underline{M}\underline{d}\right)$$

Expanding some vector $\underline{d}$ by the M-orthonormal set of basis vectors is what we mean by the modal decomposition.

$$\underline{d} = \sum_{m=1}^{n_{df}} d_m \underline{\psi}_m$$

- Each $\underline{\psi}_m$ is called a mode.
- Each coefficient d_m is called a modal coefficient of $\underline{d}$

11.12 Modal decomposition and modal equations - I

We recognized in the last section that there exists a generalized eigenvalue problem that we can identity which provides the basis for the modal analysis of the linear parabolic system we will carry out in this section.

We will perform a modal decomposition of the homogeneous version of the time-exact ODE, given the initial condition $\underline{d}(0) = \underline{d}_0$:

$$\underline{M}\dot{\underline{d}} + \underline{K}\underline{d} = \underline{0}$$

We already know the modal decomposition of $\underline{d}$. Different from our notation in previous subsection, we will change the subscripts in d_m and $\underline{\psi}_m$ to superscripts:

$$\underline{d} = \sum_{m=1}^{n_{df}} d^m(t)\underline{\psi}^m$$

The $\underline{d}$ vector as we have defined it depends on time. We placed the time dependence on the coefficients d^m as opposed to the basis vectors $\underline{\psi}$ since the matrices involved generalized eigenvalue problem which gave rise to these basis vectors do not depend on time (since we are looking at a linear parabolic system - they can have dependence, but our problem would be non-linear).

Taking a single time derivative gives:

$$\dot{\underline{d}} = \sum_{m=1}^{n_{df}} \dot{d}^m(t)\underline{\psi}^m$$

Substituting into our ODE:

$$\underline{M}\sum_{m=1}^{n_{df}} \dot{d}^m(t)\underline{\psi}^m + \underline{K}\sum_{m=1}^{n_{df}} d^m(t)\underline{\psi}^m = \underline{0}$$

By linearity,

$$\sum_{m=1}^{n_{df}} \dot{d}^m \underline{M}\underline{\psi}^m + \sum_{m=1}^{n_{df}} d^m \underline{K}\underline{\psi}^m = \underline{0}$$

We dot the entire left hand side with the an eigenvector:

$$\underline{\psi}^L \cdot \left[\sum_{m=1}^{n_{df}} \dot{d}^m \underline{M} \underline{\psi}^m \right] + \underline{\psi}^L \cdot \left[\sum_{m=1}^{n_{df}} d^m \underline{K} \underline{\psi}^m \right] = 0$$

The zero on the right is now the scalar zero.

11.13 Modal decomposition and modal equations - II

We continue to carry out our derivation from last the last subsection.

$$\underline{\psi}^L \cdot \left[\sum_{m=1}^{n_{df}} \dot{d}^m \underline{M} \underline{\psi}^m \right] + \underline{\psi}^L \cdot \left[\sum_{m=1}^{n_{df}} d^m \underline{K} \underline{\psi}^m \right] = 0$$

Let's move the dot products into the sum,

$$\left[\sum_{m=1}^{n_{df}} \dot{d}^m \underline{\psi}^L \cdot \underline{M} \underline{\psi}^m \right] + \left[\sum_{m=1}^{n_{df}} d^m \underline{\psi}^L \cdot \underline{K} \underline{\psi}^m \right] = 0$$

Using our definition of the set of $\left\{ \underline{\psi}^m \right\}$ vectors, we know

$$\left[\sum_{m=1}^{n_{df}} \dot{d}^m \delta_{Lm} \right] + \left[\sum_{m=1}^{n_{df}} d^m \lambda^m \delta_{Lm} \right] = 0$$

Accounting for the sum over the index m, we get

$$\dot{d}^L + \lambda^L d^L = 0 \qquad \forall L = 1, \ldots, N_{df}$$

This is called a single degree of freedom modal equation.

Now we will do the same for the time-discrete, homogeneous ODE.

$$\underline{M} \left(\frac{\underline{d}_{n+1} - \underline{d}_n}{\Delta t} \right) + \underline{K} \underline{d}_{n+\alpha} = \underline{0}$$

Let's multiply through my Δt, and expand $\underline{d}_{n+\alpha}$ term.

$$\underline{M} \left(\underline{d}_{n+1} - \underline{d}_n \right) + \Delta t \underline{K} \left(\alpha \underline{d}_{n+1} + (1 - \alpha) \underline{d}_n \right) = \underline{0}$$

For each $\underline{d}_n$ vector, we substitute the modal decomposition of $\underline{d}$:

$$\underline{d}_{n+1} \sum_m d_{n+1}^m \underline{\psi}^m$$

Make the substitution and factor the matrices that multiply $\underline{d}_n$ and $\underline{d}_{n+1}$,

$$\left(\underline{M} + \alpha \Delta t \underline{K}\right) \sum_m d_{n+1}^m \underline{\psi}^m - \left(\underline{M} - (1-\alpha)\Delta t \underline{K}\right) \sum_m \underline{d}_n^m \underline{\psi}^m = \underline{0}$$

Let's dot both sides of the equation with $\underline{\psi}^L$ to convert a vector equation into a scalar equation.

$$\underline{\psi}^L \cdot \left(\underline{M} + \alpha \Delta t \underline{K}\right) \sum_m d_{n+1}^m \underline{\psi}^m - \underline{\psi}^L \cdot \left(\underline{M} - (1-\alpha)\Delta t \underline{K}\right) \sum_m \underline{d}_n^m \underline{\psi}^m = 0$$

Invoking linearity, we get

$$\sum_m d_{n+1}^m \left(\underline{\psi}^L \cdot \underline{M}\underline{\psi}^m + \alpha \Delta t \underline{\psi}^L \cdot \underline{K}\underline{\psi}^m\right) - \sum_m d_n^m \left(\underline{\psi}^L \cdot \underline{M}\underline{\psi}^m - (1-\alpha)\Delta t \underline{\psi}^L \cdot \underline{K}\underline{\psi}^m\right) = 0$$

Using the orthogonality of $\underline{\psi}^m$ vectors,

$$\sum_m d_{n+1}^m \left(\delta_{Lm} + \alpha \Delta t \lambda^m \delta_{Lm}\right) - \sum_m d_n^m \left(\delta_{Lm} - (1-\alpha)\Delta t \lambda^m \delta_{Lm}\right) = 0$$

The summation over the index m and with the Kronecker delta gets us

$$d_{n+1}^L \left(1 + \alpha \Delta t \lambda^L\right) - d_n^L \left(1 - (1-\alpha)\Delta t \lambda^L\right) = 0 \qquad \forall L = 1, \ldots, N_{df}$$

This is the single degree of freedom model equation for the time-discrete problem.

11.14 Modal equations and stability of the time-exact single degree of freedom systems - I

Recall the single degree of freedom modal equations we derived.

For the time exact case:

$$\dot{d}^L + \lambda^L d^L = 0 \qquad \forall L = 1, \ldots, N_{df}$$

This is incomplete without initial conditions.

$$\begin{aligned} d^L(0) &= \underline{\psi}^L \cdot \underline{M}\underline{d}(0) \\ &= d_0^L \end{aligned}$$

For the time-discrete case:

$$d_{n+1}^L \left(1 + \alpha \Delta t \lambda^L\right) - d_n^L \left(1 - (1-\alpha)\Delta t \lambda^L\right) = 0 \qquad \forall L = 1, \ldots, N_{df}$$

Where the initial condition, d_0^L is given.

Recall that to arrive at our single degree of freedom modal equation for the time discrete case, we discretized the time exact matrix vector equations, and applied modal decomposition.

But we could have also started with the time exact single degree of freedom modal equation, and applied time discretization.

This comes from the fact that modal decomposition and time discretization are "essentially linear operations, and so they commute".

Not all linear operations commute, but in our case they do.

11.15 Modal equations and stability of the time-exact single degree of freedom systems - II

We want to get a sense of the stability of our equations. We first look to understand the stability of the time exact case.

The equations we have derived holds for every mode, $L = 1, \ldots, N_{df}$. So we will drop the use of the index L. in our equation. But we will add a superscript h to λ This is to denote that λ^h corresponding to a mode is obtained after discretization.

$$\dot{d} + \lambda^h d = 0 \qquad\qquad d(0) = d_0$$

We can directly write the solution to this ODE.

$$d(t) = d_0 \exp\left(-\lambda^h t\right)$$

We also know that $\lambda^h \leq 0$. This comes from the fact that $\underline{M}$ is positive definite, and $\underline{K}$ is positive semi-definite.

From this additional information, we see that $d(t_{n+1}) \leq d(t_n)$, since we have an exponential decay. The modal coefficients are monotonically decreasing. Alternatively, we can look at the ratio of $d(t_{n+1})$ and $d(t_n)$. Provided that $d(t_n) \neq 0$, we have

$$\frac{d(t_{n+1})}{d(t_n)} \leq 1$$

Recall that our model equations come from applying the modal decomposition to a homogeneous matrix vector equation equation: there is no forcing.

We have chosen to look at the homogeneous case to reveal this fundamental characteristic of our equation.

Let's look at our time-discrete modal equation.

$$d_{n+1}\left(1 + \alpha\Delta t\lambda^h\right) - d_n\left(1 - (1-\alpha)\Delta t\lambda^h\right) = 0$$

Rearranging

$$\frac{d_{n+1}}{d_n} = \underbrace{\frac{1 - (1-\alpha)\Delta t\lambda^h}{1 + \alpha\Delta t\lambda^h}}_{A\text{: amplification factor}}$$

We see that the right hand side is a term that reflects how our modal coefficients are scaled for every time step. Its clear that

1. A depends on Δt, how big or small our time step is will affect our solution
2. A depends on α, the choice of integration algorithms matter
3. A depends on λ^h, which is not an obvious result. The stability of our solution also depends the spatial discretization

To reflect the monotonically decaying behaviour as we saw in the time exact case, we define the stability criterion to be

$$|A| \leq 1$$

We have restricted our selves further by saying that the absolute value of A must be less than 1. We did this our time discrete solution can go negative.

11.15.1 Modal coefficient vs Nodal value

Consider a 1 dimensional heat conduction problem. We expect a positive initial conditions would cause the temperature of other nodes to rise as the temperature equilibriates. While the nodal values of initial condition nodes decrease, the nodal values of other nodes increase. Recall the the collection of nodal values (nodal degree of freedom), is contained in the vector $\underline{d}$.

In a modal decomposition, we write $\underline{d}$ as a linear combination of different mode shapes, represented by vectors $\underline{\psi}$ which belong to $\mathbb{R}^{N_{df}}$. We know that the modal coefficients decay in time, at rate given by a corresponding eigenvalue λ^h.

11.16 Stability of the time-discrete single degree of freedom systems

Recall the definition of our amplification factor.

$$\frac{d_{n+1}}{d_n} = \frac{1-(1-\alpha)\Delta t\lambda^h}{1+\alpha\Delta t\lambda^h} = A$$

We know that $\alpha \in [0,1]$, $\Delta t > 0$, and $\lambda^h \geq 0$. Our stability criterion is that the magnitude of the amplification factor is less than or equal to 1 for for every time step, or

$$-1 \leq A \leq 1$$

This is an "linear stability criterion".

- The ODE we are solving is a linear ODE
- For linear problems, we can establish the stability requirement by limiting the solution to stay constant/decay from one step to another other
- The physics of non-linear problems can lead to solutions growing under certain regimes, so this is not the correct condition to use

We can also write our criterion as

$$-1\left(1+\alpha\Delta t\lambda^h\right) \leq 1-(1-\alpha)\Delta t\lambda^h \leq 1+\alpha\Delta t\lambda^h$$

Let's first consider the right hand side inequality

$$1 - (1-\alpha)\Delta t\lambda^h \leq 1 + \alpha\Delta t\lambda^h$$
$$-(1-\alpha)\Delta t\lambda^h \leq \alpha\Delta t\lambda^h$$

Rearranging

$$0 \leq (\alpha + 1 - \alpha)\Delta t\lambda^h$$
$$0 \leq \Delta t\lambda^h$$

We see that the right hand side inequality requires that the product of the time step and the eigenvalue be greater than or equal to zero. This is always true by construction, and does not depend on α, which determines our time integration algorithm.

Let's consider the left hand side inequality

$$-1\left(1 + \alpha\Delta t\lambda^h\right) \leq 1 - (1-\alpha)\Delta t\lambda^h$$

Rearranging

$$\begin{aligned}
0 &\leq 1 - (1-\alpha)\Delta t\lambda^h + \left(1 + \alpha\Delta t\lambda^h\right) \\
&\implies 0 \leq 1 - (1-\alpha)\Delta t\lambda^h + (1 + \alpha\Delta t\lambda^h) \\
&\implies 0 \leq 2 - (1-\alpha)\Delta t\lambda^h + \alpha\Delta t\lambda^h \\
&\implies 0 \leq 2 + (2\alpha - 1)\Delta t\lambda^h \\
&\implies (1-2\alpha)\Delta t\lambda^h \leq 2
\end{aligned}$$

Consider

1. The case when $\alpha \geq 1/2$. Then our inequality becomes

   $$(1-2\alpha)\Delta t\lambda^h \leq 0$$

   This means that

   $$2 \geq (1-2\alpha)\Delta t\lambda^h,$$

   which holds for all $\Delta t > 0$. This is an unconditional stability. No matter Δt, we will always get a stable solution, though be it an inaccurate one. Recall that $\alpha = 1/2$ and $\alpha = 1$ are well known as the Crank-Nicolson method and the backwards Euler's method.

2. Let's look at the case when $\alpha \in [0, 0.5)$. In this case,

$$(1 - 2\alpha) > 0 \qquad \forall \alpha \in [0, 0.5)$$

This allows us to divide both sides by $(1 - 2\alpha)\lambda^h$.

$$\Delta t \leq \frac{2}{(1 - 2\alpha)\lambda^h}$$

This is a conditional stability. Our time step cannot be too large, otherwise the solution blows up.

λ^h is inversely proportional to the granularity of our spatial discretization. (This was briefly mentioned in the video lectures. Not in enough detail for me to understand it.)

With the semi-discrete method, as we make our mesh finer (increasing our temporal discretization), we require a smaller time step when we are working with a conditionally stable method. Such as the forward Euler's method.

11.17 Behaviour of higher-order modes; consistency - I

We will summarize the results we found for stability. When considering the parameter α, we found that

$$\alpha \begin{cases} \geq 1/2 & \text{unconditionally stable, any } \Delta t > 0 \\ < 1/2 & \text{conditionally stable, } \Delta t \leq \frac{2}{(1-2\alpha)\lambda^h} \end{cases}$$

Since λ^h is different for different modes, Δt must satisfy the criterion for all modes for conditional stability. This implies that

$$\Delta t \leq \frac{2}{(1 - 2\alpha)\lambda^h_{max}}$$

We are not considering the absolute value of λ^h since all λ^h are non-negative by construction.

Without proof, we also stated that λ^h is proportional to h^{-2}. This this case, $\Delta t \approx h^2$. This tells us that as we refine our finite element mesh, our time step needs to get smaller.

Recall the definition of the amplification factor in the time-discrete problem:

$$A = \frac{d_{n+1}}{d_n} = \frac{1-(1-\alpha)\Delta t\lambda^h}{1+\alpha\Delta t\lambda^h} \implies d_{n+1} = Ad_n$$

We want to use this fact to look at the behaviour of high order modes.

A mode is of "high order" when it has a large eigenvalue. For the system, it turns out that these are the modes with higher spatial frequencies. (A larger mode numbering m and l do not necessarily mean that the mode is high order.)

In the time exact problem, recall our solution to single degree of freedom modal equation

$$d(t) = d_0 \exp\left(-\lambda^h t\right)$$

We can interpret the time-exact amplification factor as

$$A_{ex} = \exp\left(\lambda^h \Delta t\right)$$

Let's inspect the limit as $\Delta t\lambda^h$ of A for the time discrete case.

$$\lim_{\Delta t\lambda^h\to\infty} \frac{1-(1-\alpha)\Delta t\lambda^h}{1+\alpha\Delta t\lambda^h} = \lim_{\Delta t\lambda^h\to\infty} \frac{\frac{1}{\Delta t\lambda^h}-(1-\alpha)}{\frac{1}{\Delta t\lambda^h}+\alpha}$$

For higher order modes, the backward Euler's method behaves most like the time-exact amplification factor. We say that the backward Euler's method tends to dissipate/attenuate away the higher order modes. This is called "numerical dissipation".

11.18 Behaviour of higher-order modes; consistency - II

Since the backward Euler's method results in an algorithmic amplification factor that is most similar to the time exact amplification factor, it is often preferred over other time integration schemes.

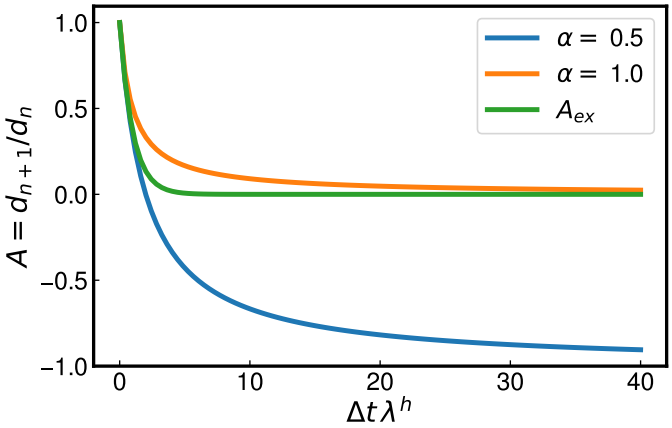

Figure 19: Different behaviours of the amplification for both the time-exact and time-discrete case. We see that for $\Delta t\lambda^h \to \infty$, A_{ex} tends to zero exponentially. For $\alpha = 1$ (backward Euler's method), A approaches 0 for large $\Delta t\lambda^h$. For $\alpha = 0.5$, A approaches -1. For $\alpha = 1$, A diverges to $-\infty$.

It is no surprise that the amplification factor when using forward Euler's method is unbounded for large $\Delta t\lambda^h$ since we already know that Euler family methods for $\alpha < 1/2$ are only conditionally stable.

The fact that for $\alpha = 1$, $A \to -1$ for higher order modes tells us that

$$d_{n+1} = -d_n$$

and we can expect oscillatory behaviour in the nodal coefficients d.

We state that it is possible to time-average the model coefficients to eliminate oscillatory behaviour in nodal coefficients of higher order modes. How this is done in practice is not discussed.

Our discussion on stability of the homogeneous problem prepares us to talk about convergence. But before this, we need to talk about the consistency of our finite dimensional solution.

Let's recall our single degree of freedom model equation for the homoge-

neous time-discrete problem:

$$d_{n+1}^L \left(1 + \alpha \Delta t \lambda^L\right) - d_n^L \left(1 - (1-\alpha)\Delta t \lambda^L\right) = 0 \qquad \forall L = 1, \ldots, N_{df}$$

Omitting the numbering L and adding a superscript h to λ:

$$d_{n+1} \left(1 + \alpha \Delta t \lambda^h\right) - d_n \left(1 - (1-\alpha)\Delta t \lambda^h\right) = 0$$

Now we will return the forcing term to the right hand side.

$$d_{n+1} \left(1 + \alpha \Delta t \lambda^h\right) - d_n \left(1 - (1-\alpha)\Delta t \lambda^h\right) = \Delta t F_{n+\alpha}$$

where

$$F_{n+\alpha}^L = \underline{\psi}^L \cdot \underline{F}_{n+\alpha}$$

Rearranging:

$$\begin{aligned} 0 &= d_{n+1} \left(1 + \alpha \Delta t \lambda^h\right) - d_n \left(1 - (1-\alpha)\Delta t \lambda^h\right) - \Delta t F_{n+\alpha} \\ &= d_{n+1} - d_n A - \underbrace{\frac{\Delta t}{\left(1 + \alpha \Delta t \lambda^h\right)} F_{n+\alpha}}_{L_n} \end{aligned}$$

What happens if we were to substitute in the time exact equation?

Let's denote $d(t_{n+1})$ to be the time-exact modal coefficient at t_{n+1} corresponding to λ^h.

$$d(t_{n+1}) - d(t_n) A - L_n$$

The question is whether the above equation equals zero. In general, this equation does not equal zero.

$$d(t_{n+1}) - d(t_n) A - L_n = \Delta t \tau(t_n)$$

In general it equals Δt times some quantity τ that depends on t_n. (The Δt is on the right hand side since the way we arrived at this equation involves multiplying through by Δt.) This is a consistency condition.

We define a method to be a consistent method when

$$\tau(t_n) \leq c \Delta t^k, k > 0,$$

where c is a constant. When this condition is true, in the limit as $\Delta t \to 0$, $\Delta t \tau(t_n) \to 0$: the time-discrete equation admits the exact solution in the limit as $\Delta t \to 0$.

The power k is called the order of accuracy. It turns out

$$k = \begin{cases} 2 & \alpha = 1/2 \\ 1 & \text{otherwise} \end{cases}$$

11.19 Convergence - I

We will first define the error for the finite element solution, before moving onto speaking about convergence.

We define the error to be

$$\underline{d}_{n+1} - \underline{d}(t_{n+1}) = \underline{e}_{n+1} \in \mathbb{R}^{N_{df}}$$

We can perform a modal decomposition on $\underline{e}$:

$$\underline{e}_{n+1} = \sum_m e^m_{n+1} \underline{\psi}^m$$

We state the convergence criterion to be

$$\lim_{n+1 \to \infty} \underline{e}_{n+1} \cdot \underline{\underline{M}} \underline{e}_{n+1} = 0$$

Since $\underline{\underline{M}}$ is positive definite, $\underline{e}_{n+1} \cdot \underline{\underline{M}} \underline{e}_{n+1} = 0$ iff $\underline{e}_{n+1}$ is 0 by definition.

Let's convert this condition to a condition on the modal coefficients.

$$\begin{aligned}
\underline{e}_{n+1} \cdot \underline{\underline{M}}\underline{e}_{n+1} &= \left(\sum_m e^m_{n+1} \underline{\psi}^m \right) \cdot \underline{\underline{M}} \left(\sum_L e^L_{n+1} \underline{\psi}^L \right) \\
&= \sum_{m,L} e^m_{n+1} \left(\underline{\psi}^m \cdot \underline{\underline{M}} \underline{\psi}^L \right) e^L_{n+1} \\
&= \sum_{m,L} e^m_{n+1} \delta_{mL} e^L_{n+1} \\
&= \sum_m e^m_{n+1} e^m_{n+1} \\
&= \sum_m \left(e^m_{n+1} \right)^2
\end{aligned}$$

We see that as $\underline{e}_{n+1} \cdot \underline{\underline{M}}\underline{e}_{n+1}$ tends to zero, the modal coefficients must tend to zero.

So

$$\lim_{n+1 \to \infty} \underline{e}_{n+1} \cdot \underline{\underline{M}}\underline{e}_{n+1} = 0$$

holds iff

$$\lim_{n+1 \to \infty} e^m_{n+1} = 0$$

If we the the difference of the following equations:

$$\begin{aligned}
d_{n+1} - d_n A - L_n &= 0 \\
d(t_{n+1}) - d(t_n) A - L_n &= \Delta t \tau(t_n)
\end{aligned}$$

We get

$$e_{n+1} - e_n A = -\Delta t \tau(t_n)$$

which is a recursive equation. We can find a formula for e_{n+1} involving only e_0:

$$e_{n+1} = A^{n+1} e_0 - \sum_{i=0}^{n} A^i \Delta t \tau(t_{n-i})$$

e_0 is zero as the initial condition for is taken to be error-free.

11.20 Convergence - II

Continuing from our previous section, we have

$$e_{n+1} = -\sum_{i=0}^{n} A^i \Delta t \tau(t_{n-i})$$

We can see that

$$|e_{n+1}| = \left|\sum_{i=0}^{n} A^i \Delta t \tau(t_{n-i})\right| \leq \sum_{i=0}^{n} \left|A^i \Delta t \tau(t_{n-i})\right|$$

By the triangle inequality. (The statement that the absolute of the sum is bounded from above by the sum of the absolute values.)

Then, we can invoke the Cauchy-Schwarz inequality, to say that the absolute value of a product is bounded from above by the product absolute values.

$$\sum_{i=0}^{n} \left|A^i \Delta t \tau(t_{n-i})\right| \leq \sum_{i=0}^{n} \left|A^i\right| |\Delta t \tau(t_{n-i})|$$

Now we can invoke stability. We know that for a stable method, $A \leq 1$. So we can get a further upper bound:

$$\sum_{i=0}^{n} \left|A^i\right| |\Delta t \tau(t_{n-i})| \leq \sum_{i=0}^{n} |\Delta t \tau(t_{n-i})|$$

Next, we invoke the consistency property, and that we know $\Delta t > 0$.

$$\sum_{i=0}^{n} |\Delta t \tau(t_{n-i})| \leq \Delta t \sum_{i=0}^{n} \left|c\Delta t^k\right| \leq \Delta t(n+1) \left|c\Delta t^k\right| = t_{n+1} \left|c\Delta t^k\right|$$

From the consistency property, we know that $k > 0$. For this reason, as $\Delta t \to 0$, $t_{n+1} \left|c\Delta t^k\right| \to 0$.

And

$$\lim_{\Delta t \to 0} |e_{n+1}| \leq 0$$

we indeed have convergence. This is an example of the Lax Theorem: consistency and stability implies convergence.

We have completed our study of the finite element method to solve linear parabolic problems.

In our formulation, we used a standard spatial discretization, but used the finite difference method for time discretization.

Since the problem is first order in time, we looked at the Euler family of algorithms. We analyzed the stability, look at higher order modes, consistency, and how these things leads us to convergence.

α	Name	Stability	k	Higher order modes
0	Forward Euler	Conditional	1	$A \to -\infty$
1/2	Midpoint rule	Unconditional	2	$A \to -1$
1	Backward Euler	Unconditional	1	$A \to 0$

11.20.1 Coding assignment

To time-step our solution using the v-method

1. Compute $\widetilde{\underline{d}}$

$$\widetilde{\underline{d}} = \underline{d}_n + (1-\alpha)\Delta t \underline{v}_n$$

2. With $\widetilde{\underline{d}}$, we want to find $\underline{v}_{n+1}$ by inverting a system of equations

$$\underline{v}_{n+1} = \left(\underline{M} + \alpha \Delta t \underline{K}\right)^{-1} \left(\underline{F}_{n+1} - \underline{K}\widetilde{\underline{d}}_{n+1}\right)$$

3. With $\underline{v}_{n+1}$, we can compute $\underline{d}_{n+1}$

$$\underline{d}_{n+1} = \tilde{\underline{d}} + \alpha \Delta t \underline{v}_{n+1}$$

12 Linear, hyperbolic partial differential equations for a vector unknown in three dimensions (Linear elastodynamics)

12.1 The strong and weak forms

We now move on to study linear hyperbolic PDEs in vector unknowns. The canonical example is linear elastodynamics in 3-D.

In our previous study of linear elliptic PDEs for vector unknowns, we considered the example of linearized elasticity at steady state.

We looked at how our body of interest would deform, but we did not consider dynamic effects. These effect lead to mechanical wave propagation. From this, we can also look at body travelling through space, and while deforming at the same time.

Over a body of interest, represented by Ω, we decompose its surface, $\partial\Omega$ into a Dirichlet and Neumann subsets.

$$\partial\Omega = \partial\Omega_{u_i} \cup \partial\Omega_{\bar{t}_i}, i = 1, 2, 3,$$

where

$\partial\Omega_i$	Dirichlet
$\partial\Omega_{\bar{t}_i}$	Neumann

and the sets are disjoint ($\partial\Omega_{u_i} \cap \partial\Omega_{\bar{t}_i} = \emptyset$).

Given $u_i^g, \bar{t}_i$ (denotes traction), f_i, ρ, and initial conditions u_{i_0}, v_{i_0} ($i = 1, 2, 3$, these are components of vectors), and the constitutive relations:

$$\sigma_{ij} = \mathbb{C}_{ijkl}\epsilon_{kl}$$
$$\epsilon_{kl} = \frac{1}{2}\left(\frac{\partial u_k}{\partial x_l} + \frac{\partial u_l}{\partial x_k}\right)$$

find $u_i(\underline{x}, t)$ (the displacement field) such that the following holds over $\Omega \times [0, T]$ (the spatial and time domain):

$$\rho \frac{\partial^2 u_i}{\partial t^2} = \sigma_{ij,j} + f_i$$

with the boundary conditions:

$$u_i(\underline{x}, t) = u_i^g(\underline{x}, t) \qquad \forall \underline{x} \in \partial\Omega_{u_i}$$

We are allowing the Dirichlet boundary conditions to also have time dependence.

$$\sigma_{ij} n_j = \bar{t}_i(\underline{x}, t) \qquad \forall \underline{x} \in \partial\Omega_{\bar{t}_i}$$

Since our PDE is second order in time, we need two initial conditions:

$$u_i(\underline{x}, 0) = \underline{u}_{i_0}(\underline{x}) \qquad \forall \underline{x} \in \Omega$$

and

$$\dot{u}_i(\underline{x}, 0) = v_{i_0}(\underline{x}) \qquad \forall \underline{x} \in \Omega$$

These two initial conditions specifies the initial displacement of every point on body, as well as the initial velocity of every point on the body.

Let's convert our problem to its infinite dimensional weak form. Given all the previous data, we want to find $u_i \in \mathcal{S} = \mathcal{S} = \{u_i | u_i = u_i^g \text{ on } \partial\Omega_{u_i}\}$, $i = 1, 2, 3$, such that for all $w_i \in \mathcal{V} = \{w_i | w_i = 0 \text{ on } \partial\Omega_{u_i}\}$, $i = 1, 2, 3$ the following holds

$$\int_\Omega w_i \rho \frac{\partial^2 u_i}{\partial t^2}\, dV + \int_\Omega w_{i,j}\sigma_{ij}\, dV = \int_\Omega w_i f_i\, dV + \sum_{i=1}^{3} \int_{\partial\Omega_{\bar{t}_i}} w_i \bar{t}_i\, dS$$

This equation differs from our weak form for steady state linear elasticity in

that there is an addition of the first term on the left. This term comes form multiplying the left hand side fo the strong form with w_i, and integrating over the domain. No integration by parts was required.

12.2 The finite-dimensional and matrix-vector weak forms - I

In the last section, we stated the infinite dimension weak form. Let's now state finite dimensional weak form.

We want to find $u_i^h \in \mathcal{S}^h \subset \mathcal{S}$, where $\mathcal{S}^h = \{u_i^h \in H^1(\Omega) | u_i^h = u_i^g \text{ on } \partial\Omega_{u_i}\}$, such that for all $w_i^h \in \mathcal{V}^h \subset \mathcal{V}$, where $\mathcal{V}^h = \{w_i^h \in H^1(\Omega) | w_i^h = 0 \text{ on } \partial\Omega_{u_i}\}$, the following integral equations hold

$$\int_\Omega w_i^h \rho \frac{\partial^2 u_i^h}{\partial t^2}\, dV + \int_\Omega w_{i,j}^h \sigma_{ij}^h\, dV = \int_\Omega w_i^h f_i\, dV + \sum_{i=1}^{3} \int_{\partial\Omega_{\bar{t}_i}} w_i^h \bar{t}_i\, dS$$

This is the finite dimensional weak form.

If we substitute the constitutive relations for σ_{ij}^h,

$$\int_\Omega w_{i,j}^h \sigma_{ij}^h\, dV = \int_\Omega w_{i,j}^h \mathbb{C}_{ijkl} \epsilon_{lk}^h\, dV$$

From our study of linearized elasticity, we found that this term contributed $\underline{c}^\intercal \underline{\underline{K}} \underline{d}$.

We found that the right hand terms together contribute $\underline{c}^\intercal \underline{F}$.

For the first term on the left hand side, we can reference our study of the linear parabolic problem. We state that that the contribution of this term to the matrix vector equation is of the form

$$\underline{c}^\intercal \underline{\underline{M}} \ddot{\underline{d}}$$

Remember that we are solving for a vector unknown. So $\ddot{\underline{d}}$ will have more entries than there are nodes, since each node has three scalar degrees of freedom.

There is a detail in the form of the mass matrix, $\underline{M}$. Recall that

$$\int_\Omega w_i^h \rho \frac{\partial^2 u_i^h}{\partial t^2}\, dV = \sum_e \int_{\Omega^e} w_i^h \rho \frac{\partial^2 u_i^h}{\partial t^2}\, dV$$

12.3 The finite-dimensional and matrix-vector weak forms - II

Let's look at the element integral.

$$\int_{\Omega^e} w_i^h \rho \frac{\partial^2 u_i^h}{\partial t^2}\, dV = \sum_{A,B} c_{e_i}^A \left(\int_{\Omega_e} \rho N^A N^B\, dV \right) \ddot{d}_{e_i}^B$$

where $c_{e_i}^A$ is the i^{th} weighting function degree of freedom, at node A of the element e. We also imply a summation over index i for the spatial dimensions.

Since we have a summation over i, we can equivalently write our sum as

$$\sum_{A,B} c_{e_i}^A \left(\int_{\Omega_e} \rho N^A N^B \delta_{ij}\, dV \right) \ddot{d}_{e_j}^B$$

We added an additional Kronecker delta function in the integrand, and we have a summation over i and j.

We can now write out our element matrix vector equations. We define the vector $\underline{c}_e^\intercal$ to be

$$\underline{c}_e^\intercal = \begin{bmatrix} \underline{c}_e^{1^\intercal} & \underline{c}_e^{2^\intercal} & \dots & \underline{c}_e^{N_{ne}^\intercal} \end{bmatrix}$$

where for example $\underline{c}_e^{1^\intercal}$ is a row vector whose entries are the three scalar degrees of freedom for node 1 and element e.

We define $\underline{\ddot{d}}_e^\intercal$ to be

$$\underline{\ddot{d}}_e^\intercal = \begin{bmatrix} \underline{\ddot{d}}_e^{1^\intercal} & \underline{\ddot{d}}_e^{2^\intercal} & \dots & \underline{\ddot{d}}_e^{N_{ne}^\intercal} \end{bmatrix}$$

where for example $\underline{\ddot{d}}_e^{1^\intercal}$ is a row vector containing the second time derivative of three scalar degrees of freedom for the trial solution at node 1 for element e.

From our study of the parabolic problem, we found that the summation over indices A and B gave rise to a $N_{n_e} \times N_{n_e}$ element mass matrix.

This time, we also have a summation over i and j, for $i, j = 1, 2, 3$ (or $1, \dots N_{sd}$ in general). This means that the entry AB of the consistent element mass matrix is given by

$$\underline{\underline{M}}_e^{AB} = \int_{\Omega_e} \rho N^A N^B \, dV \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The element mass matrix has dimensions $[N_{n_e} \times N_{sd}]^2$.

In summary, we have

$$\sum_{A,B} c_{e_i}^A \left(\int_{\Omega_e} \rho N^A N^B \delta_{ij} \, dV \right) \ddot{d}_{e_j}^B = \underline{c}_e^\intercal \underline{\underline{M}}_e \underline{\ddot{d}}$$

After the assembly process, our finite element equations are

$$\underline{\underline{M}} \underline{\ddot{d}} + \underline{\underline{K}} \underline{d} = \underline{F}$$

Unlike the steady state problem, there is an addition term $\underline{\underline{M}} \underline{\ddot{d}}$. We also have two initial conditions:

$$\underline{d}(0) = \underline{d}_0$$
$$\underline{\dot{d}}(0) = \underline{v}_0 = \begin{bmatrix} v_0(\underline{x}^A) \\ \vdots \end{bmatrix}$$

12.4 The time-discretized equations

Recall our finite element equations with its two initial conditions.

In addition to the mass and stiffness matrix we see, it is popular to include a "damping matrix", to account for "structural damping". This manifests from the roots of the finite element method in structural mechanics.

We define the damping matrix to be

$$\underline{C} = a\underline{M} + b\underline{K}$$

where $a, b \in \mathbb{R}$. This model of damping was known as "Rayleigh damping".

With the damping matrix, our matrix vector problem becomes

$$\underline{M}\ddot{\underline{d}} + \underline{C}\dot{\underline{d}} + \underline{K}\underline{d} = \underline{F}$$

with two initial conditions.

"The notion of damping is seen in the addition of the single time derivative of $\underline{d}$. The introduction of the this extra term suggests the effect of viscosity."

Now we consider the time discretization using a finite difference method. We discretize the time interval $[0, T] = [t_0 = 0, t_1] \cup [t_0, t_1] \cup \ldots [t_{N-1}, t_N = T]$ into N sub-intervals.

We define $\underline{d}_n$ to be the time-discrete approximation to the displacement field $\underline{d}(t_n)$. Let's define $\underline{v}_n$ to be the approximation to $\dot{\underline{d}}(t_n)$. Similarly, $\underline{a}_n$ be the approximation to $\ddot{\underline{v}}$.

$$\underline{M}\underline{a}_{n+1} + \underline{C}\underline{v}_{n+1} + \underline{K}\underline{d}_{n+1} = \underline{F}_{n+1}$$

The initial conditions in the time-discrete case are that $\underline{d}_0$ and $\underline{v}_0$ are known.

As the Euler family of algorithms were used to time integrate first order ODEs, we have the Newmark family of algorithms for second order ODEs.

While the Euler family involved a single parameter, α, the Newmark family involves two parameters:

$$\gamma \in [0, 1]$$
$$2\beta \in [0, 1]$$

The Newmark family works by

$$\underline{d}_{n+1} = \underline{d}_n + \Delta t \underline{v}_n + \frac{\Delta t^2}{2}\left((1 - 2\beta)\underline{a}_n + 2\beta\underline{a}_{n+1}\right)$$

Where

$$\underline{v}_{n+1} = \underline{v}_n + \Delta t\left((1 - \gamma)\underline{a}_n + \gamma\underline{a}_{n+1}\right)$$

To find the solution, we will use what is known as the a-method (a stands for acceleration).

The predictors for the a-method is

$$\underline{\widetilde{d}}_{n+1} = \underline{d}_n + \Delta t \underline{v}_n + \frac{\Delta t^2}{2}(1-2\beta)\underline{a}_n$$
$$\underline{\widetilde{v}}_{n+1} = \underline{v}_n + \Delta t(1-\gamma)\underline{a}_n$$

We see that the predictors are simply the the Newmark family equation for $\underline{d}_{n+1}$ and $\underline{v}_{n+1}$ with the $\underline{a}_{n+1}$ term discarded. This is reasonable. These are the quantities we can compute at time n.

The correctors are the terms (marked with a brace in the following equations) such that the sum of predictor and corrector recovers $\underline{d}_{n+1}$ or $\underline{v}_{n+1}$ as defined by the Newmark family.

$$\underline{d}_{n+1} = \underline{\widetilde{d}}_{n+1} + \underbrace{\Delta t^2 \beta \underline{a}_{n+1}}$$
$$\underline{v}_{n+1} = \underline{\widetilde{v}}_{n+1} + \underbrace{\Delta t \gamma \underline{a}_{n+1}}$$

Let's substitute these equations into our time-discrete matrix vector equations:

$$\underline{M}\underline{a}_{n+1} + \underline{C}\left(\underline{\widetilde{v}}_{n+1} + \Delta t \gamma \underline{a}_{n+1}\right) + \underline{K}\left(\underline{\widetilde{d}}_{n+1} + \Delta t^2 \beta \underline{a}_{n+1}\right) = \underline{F}_{n+1}$$

This equation is now in terms of the predictors, and $\underline{a}_{n+1}$.

Let's factor out $\underline{a}_{n+1}$. Since we assume the predictors to be known at any step n, we can move them to the right hand side.

$$\left[\underline{M} + \underline{C}\Delta t\gamma + \underline{K}\Delta t^2 \beta\right]\underline{a}_{n+1} = \underline{F}_{n+1} - \underline{K}\underline{\widetilde{d}}_{n+1} - \underline{C}\underline{\widetilde{v}}_{n+1}$$

In this form, we can solve for $\underline{a}_{n+1}$ for each time step n. Using a, we can compute $\underline{d}_{n+1}$ and $\underline{v}_{n+1}$.

To start up the algorithm, we need to know $\underline{a}_0$. Using our initial conditions, we solve for $\underline{a}_0$ using time-discrete matrix vector equation

$$\underline{M}\underline{a}_0 = \underline{F}_0 - \underline{C}\underline{v}_0 - \underline{K}\underline{d}_0$$

12.5 Stability - I

Our analysis is based on a suitable chosen eigenvalue problem. Using this eigenvalue problem, we construct a decomposition of our ODE in modes.

The generalized eigenvalue problem we invoke is

$$\omega^2 \underline{M}\underline{\psi} = \underline{K}\underline{\psi}$$

In this context, ω represents a natural frequency of oscillation.

The eigenvectors $\underline{\psi}^L$, $L = 1, \ldots, N_{df}$ are M-orthonormal. From this, we can expand any vectors in $\mathbb{R}^{N_{df}}$ that show up in the problem in terms of the modes $\underline{\psi}$

$$\underline{d} = \sum_{L=1}^{N_{df}} d^L \underline{\psi}^L$$

We can apply modal decomposition to vectors in the time exact matrix vector equation:

$$\underline{M} \sum_{L=1}^{N_{df}} \ddot{d}^L \underline{\psi}^L + \underline{C} \sum_{L=1}^{N_{df}} \dot{d}^L \underline{\psi}^L + \underline{K} \sum_{L=1}^{N_{df}} d^L \underline{\psi}^L = \underline{F}$$

$$\sum_{L=1}^{N_{df}} \underline{M} \ddot{d}^L \underline{\psi}^L + \sum_{L=1}^{N_{df}} \left(a\underline{M} + b\underline{K}\right) \dot{d}^L \underline{\psi}^L + \sum_{L=1}^{N_{df}} d^L \underline{K}\underline{\psi}^L = \underline{F}$$

$$\underline{\psi}^m \cdot \sum_{L=1}^{N_{df}} \underline{M} \ddot{d}^L \underline{\psi}^L + \underline{\psi}^m \cdot \sum_{L=1}^{N_{df}} \left(a\underline{M} + b\underline{K}\right) \dot{d}^L \underline{\psi}^L + \underline{\psi}^m \cdot \sum_{L=1}^{N_{df}} d^L \underline{K}\underline{\psi}^L = \underline{\psi}^m \cdot \underline{F}$$

$$\sum_{L=1}^{N_{df}} \ddot{d}^L \delta_{mL} + \sum_{L=1}^{N_{df}} \left(a\delta_{mL} + b\omega^{L^2} \delta_{mL}\right) \dot{d}^L + \sum_{L=1}^{N_{df}} d^L \omega^{L^2} \delta_{mL} = \underline{\psi}^m \cdot \underline{F}$$

$$\ddot{d}^m + \left(a + b\omega^{m^2}\right) \dot{d}^m + d^m \omega^{m^2} = \underline{\psi}^m \cdot \underline{F}$$

Let's omit the modal indices on d and ω. We also add a superscript h to ω remind us that they depend on $\underline{M}$ and $\underline{K}$ which comes from spatial

discretization.

$$\ddot{d} + \left(a + b\omega^{h^2}\right)\dot{d} + d\omega^{h^2} = \underline{\psi} \cdot \underline{F}$$

We define the modal damping ratio to be

$$\xi_m^h = \frac{a}{\omega^{m^h}} + b\omega^{m^h}$$

Omitting the modal index m,

$$\ddot{d} + \xi^h \omega^h \dot{d} + d\omega^{h^2} = \underline{\psi} \cdot \underline{F}$$

This is our single degree of freedom modal equation. Equivalently, we can also say that our equation is

$$\ddot{d} + 2\xi^h \omega^h \dot{d} + d\omega^{h^2} = \underline{\psi} \cdot \underline{F}$$

since the constants a and b in ξ^h can be varied.

Consider the homogeneous version of our ODE.

$$\ddot{d} + 2\xi^h \omega^h \dot{d} + d\omega^{h^2} = 0$$

Let's rewrite it as two first order ODEs. We define

$$\underline{y} = \begin{bmatrix} d \\ \dot{d} \end{bmatrix}$$

If we define $v = \dot{d}$, and $u = d$, then

$$\frac{d}{dt}\underline{y} = \begin{bmatrix} \dot{u} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \omega^{h^2} & -2\xi^h \omega^h \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \omega^{h^2} & -2\xi^h \omega^h \end{bmatrix} \underline{y}$$

Similarly, the time discretized homogeneous problem in modal form is:

$$a_{n+1} + 2\xi^h \omega^h v_{n+1} + \omega^h d_{n+1} = 0$$

where a_{n+1}, v_{n+1}, and d_{n+1} are the modal coefficients for a particular mode.

12.6 Stability - II

For the time-discretized homogeneous problem, we can similarly perform a reduction to two first order time discretized ODEs. We will not present the details here.

$$\underline{y}_{n+1} = \underline{A}\underline{y}_n + \underline{L}_n$$

where

$$\underline{y}_n = \begin{bmatrix} d_n \\ v_n \end{bmatrix}$$

and $\underline{A}$ is the "amplification matrix". The matrices $\underline{A}$ and $\underline{L}$ involves the Newmark family parameters, γ and 2β.

When we are concerned with the homogeneous problem (as we are when we look at stability), the $\underline{L}_n$ term is not present. In this case,

$$\underline{y}_n = \underline{A}^n \underline{y}_0$$

We now give a summary of stability results. Remember the following conditions needs to hold for all modes.

When

$$2\beta \geq \gamma \geq 1/2$$

we have unconditional stability.

When

$$\begin{cases} \gamma \geq 1/2 \\ 0 \leq 2\beta \leq \gamma \\ \omega^h \Delta t \leq \Omega_{\text{critical}} \end{cases}$$

we have conditional stability. Here, Ω_{critical} is defined to be

$$\Omega_{\text{critical}} = \frac{\xi^h (\gamma - 1/2) + \left(\gamma/2 - \beta + \xi^{h^2} (\gamma - 1/2)^2\right)^{1/2}}{\gamma/2 - \beta}$$

We have damping when $\xi^h > 0$. So its clear that the effect of damping is to increase the quantity Ω_{critical} ("called the critical frequency").

We can also look at the case when $\xi^h = 0$ (no damping). We have the undamped critical frequency to be

$$\Omega^u_{\text{critical}} = \left(\frac{\gamma}{2} - \beta\right)^{-1/2}$$

This is a lower bound to the critical frequency:

$$\Omega^u_{\text{critical}} \leq \Omega_{\text{critical}}$$

We see $\Omega^u_{\text{critical}}$ is a more stringent condition on $\omega^h \Delta t$.

Table 2: Summary of some Newmark family methods. k is the order of accuracy.

Method	Type	β	γ	$\Omega^u_{\text{critical}}$	k
Trapezoidal rule	Implicit	1/4	1/2	Unconditional	2
Linear acceleration	Implicit	1/6	1/2	$2\sqrt{3}$	2
Average acceleration	Implicit	1/12	1/2	$\sqrt{6}$	2
Central difference	Explicit	0	1/2	2	2

12.7 Behaviour of higher-order modes

We will sketch out some parts of the analysis that lead to the summary of stability results for the Newmark family of algorithms. We will not work them out in complete detail.

This stability analysis is based on an eigenvalue analysis of the amplification matrix.

We define the spectral radius of the amplification matrix to be $r(\underline{A})$:

$$r(\underline{A}) = \max_i \left|\lambda_i(\underline{A})\right|$$

(The spectral radius is the largest eigenvalue of $\underline{A}$.)

Let λ_i for $i = 1, 2, \ldots, n$ be eigenvalues of a matrix $\underline{A} \in \mathbb{C}^{n \times n}$, the spectral radius of $\underline{A}$ is $\rho(\underline{A})$, the largest absolute value of all λ_i.
For Hermitian matrices, the spectral radius of a matrix is equal to its operator norm.

Accounting for the case that $\underline{A}$ has complex eigenvalues,

$$r(\underline{A}) = \max_i \sqrt{\lambda_i(\underline{A})\overline{\lambda}_i(\underline{A})}$$

We propose that, when λ_i are distinct, (implying the eigenvectors are also linearly independent), $r(\underline{A}) \leq 1$. And when λ_i are the same ($i = 1, 2$ since $\underline{A}$ is a 2×2 matrix; the eigenvectors are linearly dependent), $r(\underline{A}) < 1$.

Its clear that, for large powers of $\underline{A}$, $\underline{A}$ stays bounded when $r(\underline{A}) \leq 1$, and λ_i are distinct. We can see that this is true when we diagonalize $\underline{A}$ to compute its powers.

But, when λ_i are the same, and the eigenvectors are linearly dependent, $\underline{A}$ cannot be diagonalized, and

$$\underline{A}^n = \underline{Q} \begin{bmatrix} \lambda^n & n\lambda^{n-1} \\ 0 & \lambda^n \end{bmatrix} \underline{Q}^{-1}$$

the off-diagonal entries become unbounded.

We define the stability condition to be

$$r(\underline{A}) \leq 1$$

The next step in the analysis is to solve for λ_i. From linear algebra, we know that the eigenvalues are the roots to the characteristic equation, given by

$$\lambda^2 - 2A_1\lambda + A_2 = 0$$

where

$$A_1 = \frac{1}{2}\mathrm{tr}(\underline{A})$$
$$A_2 = \det(\underline{A})$$

And the roots are given by

$$\lambda_1, \lambda_2 = A_1 \pm \sqrt{A_1^2 - A_2}$$

Since the spectral radius depends on λ_i, and we wrote λ_i in terms of A_1 and A_2, we can state our stability condition in terms of A_1 and A_2.

Without proof, we state them here

$$\text{stable if } \begin{cases} -\frac{(A_2+1)}{2} \leq A_1 \leq \frac{A_2+1}{2} & A_1 < 1 \\ -1 < A_1 < 1 & A_2 = 1 \end{cases}$$

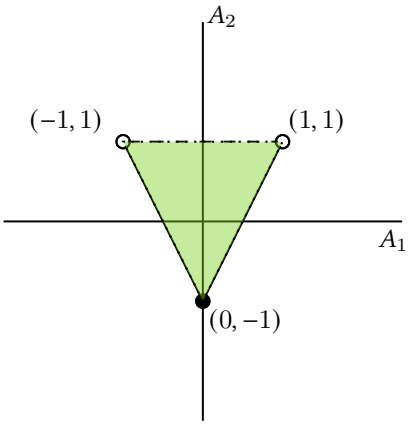

Figure 20: The region of convergence given in A_1, A_2 space is shaded in green.

12.8 Convergence

In the previous section, we illustrated how we can approach the stability analysis for the modal equations for linear elastodynamics.

We will look at the effect of the amplification factor to high order modes.

Without proof, we state that the higher order modes are non-decaying if the eigenvalues of $\underline{A}$ are all real.

To get complex eigenvalues - and decaying higher order modes - we require

$$A_1^2 - A_2 < 0$$

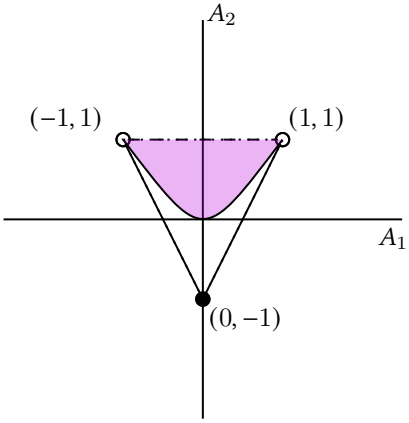

Figure 21: The region of stability and decay of higher order modes on the A_1, A_2 plane.

Without proof, we give the following relation that β, γ must satisfy for the damping of higher order modes.

$$\beta \geq \frac{1}{4}\left(\gamma + \frac{1}{2}\right)^2$$

Analogous to our plot of the amplification factor versus $\Delta\lambda^h$ in our study of the parabolic ODE, here we have a plot of the spectral radius to $\Delta t 2\pi\omega^2$.

Now we will look at consistency (to build up to convergence). Let's bring back the matrix $\underline{L}_n$ into our equation.

$$\underline{y}_{n+1} = \underline{A}\underline{y}_n + \underline{L}_n$$

Let's denote $\underline{y}(t_n)$ as the time exact solution to the modal equation at time t_n. Substituting:

$$\underline{y}(t_{n+1}) = \underline{A}\underline{y}(t_n) + \underline{L}_n$$

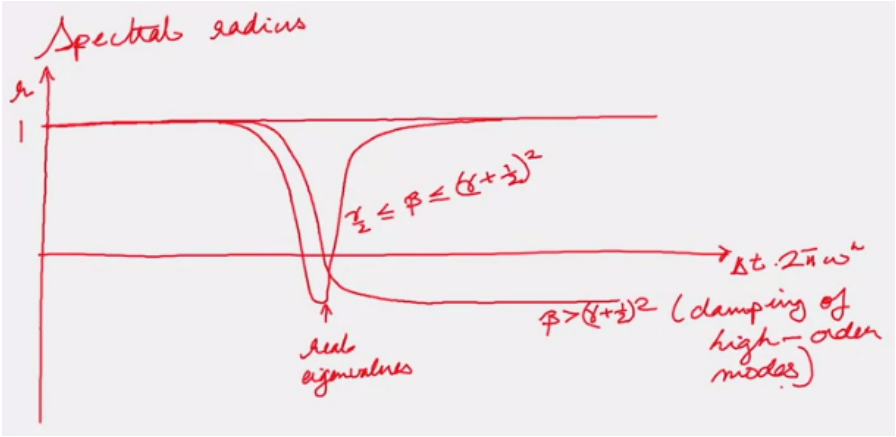

Figure 22: If β satisfies the requirement for unconditional stability, but remains less than $(\gamma + 1/2)^2/4$, we get a behaviour that the spectral radius initially drops, then rises back to one. (**Note**: there are errors in the inequalities for β in this plot.) In the minima, we get real eigenvalues. When β satisfies the condition for damping of higher order modes, we see that the spectral radius starts from 1, decays, then asymptotically approaches some other value.

But this equality does not generally hold. To make this equality hold, we need to add a term $\Delta t \underline{\tau}(t_n)$.

$$\underline{y}(t_{n+1}) = \underline{\underline{A}}\underline{y}(t_n) + \underline{L}_n + \Delta t \underline{\tau}(t_n)$$

Consistency requires

$$\underline{\tau}(t) = \underline{c}\Delta t^k = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \Delta t^k \qquad\qquad c_1, c_2 \in \mathbb{R},$$

and k is the order of accuracy.

To look at consistency, we need to look at the error. We define

$$\underline{e}_{n+1} = \underline{\underline{A}}^{n+1}\underline{e}_0 - \sum_{i=0}^{n} \Delta t \underline{\underline{A}}^i \underline{\tau}(t_n)$$

(Recall from the parabolic problem that $\underline{e}_0 = 0$.) We can show that as Δt tends to 0, the error also ends to zero.