ELECTRONIC NOTES

CEE598 - Finite Elements for Engineers: Module 1

S. D. Rajan
Department of Civil Engineering
Arizona State University
Tempe, AZ 85287-5306

Initial Version: Fall 1998
Last Revision: Fall 2013

Finite Elements for Engineers

School of Sustainable Engineering and the Built Environment
Arizona State University
Tempe, AZ 85287-5306
Phone 480.965.1712 • Fax 480.965.0557
e-mail s.rajan@asu.edu

Table of Contents

Introduction

"Change is inevitable, growth is optional." Anon.

This course is a part of the graduate program in the Civil, Environmental and Sustainable Engineering and the Aerospace and Mechanical Engineering programs at Arizona State University.

Who should take this course?

Finite elements has become the defacto industry standard for solving multi-disciplinary engineering problems that can be described by equations of calculus. Applications cut across several industries by virtue of the applications – solid mechanics (civil, aerospace, automotive, mechanical, biomedical, electronic), fluid mechanics (geotechnical, aerospace, electronic, environmental, hydraulics, biomedical, chemical), heat transfer (automotive, aerospace, electronic, chemical), acoustics (automotive, mechanical, aerospace), electromagnetics (electronic, aerospace) and many, many more.

Course Objectives

- To understand the basic ideas behind the finite element method.
- To understand and apply the Galerkin's Method in solving (multi-disciplinary) one-dimensional engineering problems.

Prerequisites

- Mathematics – Linear Algebra, Numerical Analysis, Partial and/or ordinary differential equations.
- Knowledge of undergraduate core material from (at least two of) solid mechanics, fluid mechanics, heat transfer and electromagnetics. Knowledge in one area should be at an advanced level (aligned with the undergraduate major).
- Knowledge of a high level programming language and use of computer-based tools.

Instructor-Student Interaction

To successfully meet the course objectives it is necessary that the students avail themselves of all the resources – discussion forums, e-mail, chat rooms, libraries. Keep the instructor and teaching assistant informed of all your concerns. The web pages connected with this course will contain instructions on how to communicate with the instructor regarding the questions you may have or turning in the assignments etc.

Computer Programs and Computer Aids

There is one computer program **1DBVP©** that is available electronically. This is a program that works under Windows as any Windows GUI application.

The usage of other computer programs (including commercial programs that you may have access to) that are more sophisticated and practical will follow in Modules 2 and 3.

Notation

"As complexity rises, precise statements lose meaning, and meaningful statements lose precision."
Lotfi Zadeh

Vectors

$\mathbf{a}_{n\times 1}$ column vector with n rows

a_i element i of vector $\mathbf{a}$

$\mathbf{b}_{1\times m}$ row vector with m columns

Matrices

$\mathbf{A}_{m\times n}$ matrix with m rows and n columns

A_{ij} element row i and column j of matrix

Others

y' Derivative of y (or, $\frac{dy}{dx}$)

L (Units of) length
F (Units of) force
M (Units of) mass
t (Units of) time
T (Units of) temperature
E (Units of) energy

Greek Alphabets

Lower Case	Upper Case	English Word	Lower Case	Upper Case	English Word	Lower Case	Upper Case	English Word
α	A	alpha	β	B	Beta	γ	Γ	Gamma
δ	Δ	delta	ε	E	epsilon	ζ	Z	zeta
η	H	eta	θ	Θ	theta	ι	I	lota
κ	K	kappa	λ	Λ	lambda	μ	M	mu
ν	N	nu	ξ	Ξ	xi	o	O	omicron
π	Π	pi	ρ	P	rho	σ	Σ	sigma
τ	T	tau	υ	Υ	upsilon	ϕ	Φ	phi
χ	X	chi	ψ	Ψ	psi	ω	Ω	omega

Lesson Plan

"Problems cannot be solved by the same level of thinking that created them." A. Einstein.

Module 1 is divided into four topics. Each topic has several lessons designed to focus on the critical issues. With each lesson there is a set of objectives. There are several review problems at the end of every topic. Solutions to most problems are also provided. Note that the set of problems represents the minimal set needed to understand the material. You should solve more problems from some of the referenced texts.

Topic 1 covers most of the prerequisites. Hence I have not listed the books to refer to nor added my comments. Your undergraduate books or similar books are adequate. The review problems should be very handy and bring you up to speed.

Topic 1: Review of Background Material.

Lesson 1: Solid Mechanics.
Lesson 2: Heat Transfer.
Lesson 3: Fluid Mechanics.
Lesson 4: Electrostatics & Electromagnetics
Lesson 5: Linear Algebra and Numerical Analysis.
Lesson 6: Differential Equations and Calculus
Review Exercises

Topic 2 is an introduction to finite elements. I have provided almost all of the material – text, examples and exercises.

Topic 2: The Six Major Steps.

Lesson 1: Overview of Finite Elements.
Lesson 2: The Six Steps.
Review Exercises

Topics 3 and 4 form the backbone of Module 1. The solution technique that is very commonly used, the Galerkin's Method, forms Topic 3. You should spend an adequate amount of time here to understand the basics.

Topic 3: The Galerkin's Method.

Lesson 1: Method of Weighted Residuals.
Lesson 2: Applying the Galerkin's Method.
Review Exercises

In Topic 4 we will look at the engineering problems described as one-dimensional boundary-value (BVP) problems. These problems will be solved using the Galerkin-Based Finite Element Method.

Topic 4: One-Dimensional Problems.

Lesson 1: The Element Concept.
Lesson 2: Solid Mechanics Examples.
Lesson 3: Heat Transfer and Fluid Flow Examples.
Lesson 4: Higher-Order Elements.
Lesson 5: Mesh Refinement and Convergence.
Review Exercises

Topic 1: Review

"I'll play with it first and tell you what it is later." Miles Davis.

Lesson 1: Solid Mechanics

Objectives: Knowledge of the following ideas and topics is crucial to understanding the subsequent lessons. The major objectives are listed below.

- To recognize and know the properties of simple structural systems
  Truss
  Beam
  Frame

- To understand the concepts associated with
  Equilibrium
  Stress and Strain
  Strain-Displacement Relations
  Stress-Strain Relations
  Simple theories of failure
  Plane Elasticity – Plane stress and strain

- To be able to compute
  Strain energy
  Work potential
  Total Potential Energy

Lesson 2: Heat Transfer

Objectives: Knowledge of the following ideas and topics is crucial to understanding the subsequent lessons. The major objectives are listed below.

- To understand the concepts associated with
  Conservation of energy
  Newton's Law of Cooling
  Fourier's Law
  Conduction, convection and radiation

- To recognize and know the properties of simple heat transfer problems involving
  Steady-state conduction
  Steady-state conduction and convection

- To be able to compute
  Thermal energy

Lesson 3: Fluid Mechanics

Objectives: Knowledge of the following ideas and topics is crucial to understanding the subsequent lessons. The major objectives are listed below.

- To understand the concepts associated with
  Mass balance
  Energy balance
  Momentum balance
  Pipe flow

- To recognize and know the properties of simple fluid mechanics problems involving
  Inviscid compressible and incompressible flow
  Viscous flow (newtonian/non-newtonian, laminar/transition/turbulent, compressible/incompressible)
  Flow through porous media

Lesson 4: Electromagnetics & Electrostatics

Objectives: Knowledge of the following ideas and topics is crucial to understanding the subsequent lessons. The major objectives are listed below.

- To understand the concepts associated with
  Electrostatic Fields
  Magnetostatic Fields

- To recognize and know the numerical solution to
  Maxwell's Equations
  Wave Equations

Lesson 5: Linear Algebra & Numerical Analysis

Objectives: Knowledge of the following ideas and topics is crucial to understanding the subsequent lessons. The major objectives are listed below.

- To understand the concepts associated with
  Vectors
  Matrices
  Polynomial approximation and interpolation

- To recognize and know the numerical solution to
  Linear algebraic equations
  Linear eigenvalue problem
  Numerical integration
  Numerical differentiation

- To be able to write a program in a high-level language to generate
  A library of matrix operations (addition, transpose, multiplication etc.)
  The solution to a set of linear algebraic equations

Lesson 6: Differential Equations and Calculus

Objectives: Knowledge of the following ideas and topics is crucial to understanding the subsequent lessons. The major objectives are listed below.

- To understand the concepts associated with
  Ordinary Differential Equations
  Partial Differential Equation
  Divergence Theorem

- To recognize and know the properties of
  Initial Value Problems
  Boundary Value Problems
  Elliptic and Parabolic Partial Differential Equations

- To be able to
  Integrate by parts
  Derive classical solution to simple ODEs
  Derive classical solution to simple PDEs

Review Exercises

SOLID MECHANICS

Problem T1L1-1

In a plane strain problem for a material with modulus of elasticity $E = 20(10^6)\ psi$ and Poisson's ratio $\nu = 0.3$, the state of stress at a point is given by $\{\sigma_x = 20\,ksi, \sigma_y = 0, \tau_{xy} = -7\,ksi\}$. Compute (a) σ_z, and (b) ε_x.

Problem T1L1-2

In a two-dimensional (plane stress) body, the $x-y$ displacements designated as (u,v) are given as (in mm)

$$u(x,y) = x^2 - 3y^2 + 8y$$
$$v(x,y) = 3x + 6y - 8xy$$

Determine the state of strain at the point $(2,-4)$. Using the material properties from Problem 1, determine the state of stress at the point.

Problem T1L1-3

In a solid body, the state of stress at a point is given by $\{\sigma_x, \sigma_y, \sigma_z, \tau_{xy}, \tau_{yz}, \tau_{zx}\} = \{10,-5,15,20,-10,5\}$MPa. Find the state of stress on a plane whose normal has the direction cosines $\left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}\right)$. Find the principal stresses at the point and the orientations of the principal planes.

Problem T1L1-4

Consider the slender rod as shown in the Fig. T1LR-1. The rod has a circular cross-section of radius $0.5\,\text{cm}$. The strain at any point is given as $\varepsilon_x = 2x^2 - 8x$. What is the strain energy in the bar if the bar is made of aluminum?

Fig. T1LR-1

Problem T1L1-5

Consider the slender rod as shown in Fig. T1LR-2. The rod has a circular cross-section of radius 0.5 cm and is made of aluminum. What is the potential energy of the bar?

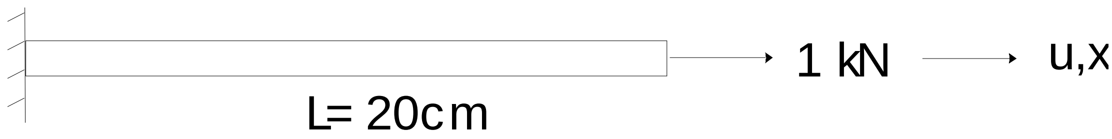


Fig. T1LR-2

Problem T1L1-6

A simply supported beam is shown in the Fig. T1LR-3. Compute the largest transverse displacement and rotation in the beam. EI is a constant.

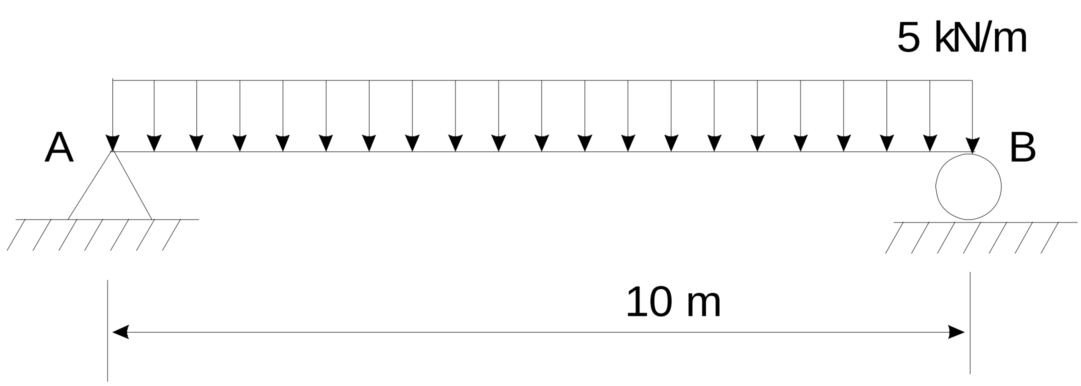


Fig. T1LR-3

HEAT TRANSFER

Problem T1L2-1

The inside and outside surfaces of a window glass are at $40^\circ C$ and $10^\circ C$, respectively. The glass has a thermal conductivity of $0.8\dfrac{W}{m\cdot^\circ C}$. The glass is 75 cm by 50 cm and 1.25 cm thick. Determine the heat loss through the glass over a period of 45 minutes.

Problem T1L2-2

An insulating board is to be used to limit the heat loss to $50\dfrac{W}{m^2}$ for a temperature difference of $125^\circ C$ across it. The thermal conductivity of the material is $0.04\dfrac{W}{m\cdot^\circ C}$. Determine the thickness of the insulating board.

Problem T1L2-3

The inside surface of an insulation layer is maintained at $T_1 = 200^\circ C$. The outside surface is dissipating heat by convection into air at $T_\infty = 20^\circ C$. The insulating layer has a thickness of $5cm$ and thermal conductivity of $1.5\dfrac{W}{m\cdot^\circ C}$. What is the minimum value of the heat transfer coefficient at the outside surface, if the temperature T_2 at the outside surface should not exceed $100^\circ C$.

Problem T1L2-4

Determine the interface temperature T_1 and the surface temperature T_3 of the composite wall shown in the Fig. T1LR-4. $k_1 = 0.1\dfrac{W}{m\cdot^\circ C}$, $k_2 = 1.0\dfrac{W}{m\cdot^\circ C}$ and $k_3 = 2.0\dfrac{W}{m\cdot^\circ C}$.

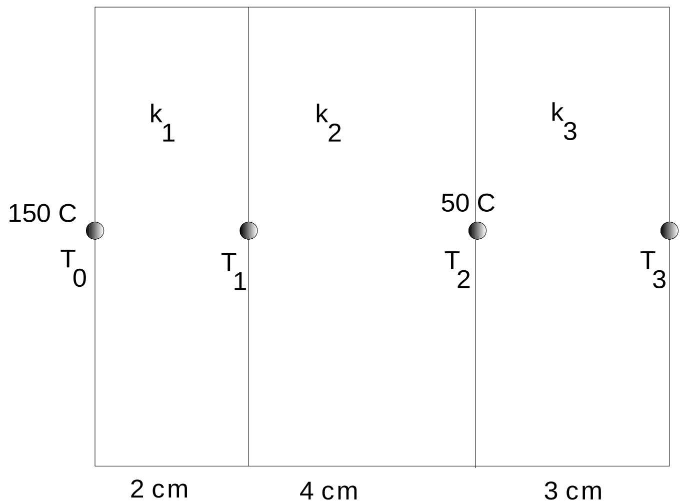


Fig. T1LR-4

FLUID MECHANICS

Problem T1L3-1

A pipeline with a 28 cm inside diameter is carrying liquid at a flow rate of $0.03\frac{m^3}{s}$. A reducer is placed in the line, and the outlet diameter is 15 cm. Determine the velocity at the beginning and end of the reducer.

Problem T1L3-2

For the piping system shown below, the distance from A to B is 4500 ft, and the main line is made of 10-nominal, schedule 40 wrought iron pipe. The attached loop is 8-nominal, schedule 40 wrought iron pipe. The flow Q_1 is $0.3\frac{ft^3}{s}$ of water. Determine the flow rate in both branches.

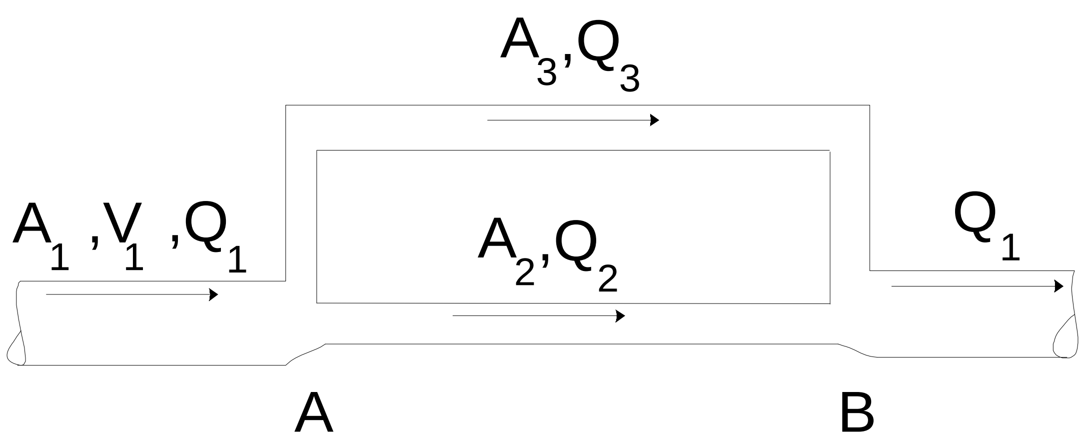


Fig. T1LR-5

LINEAR ALGEBRA & NUMERICAL ANALYSIS

Problem T1L5-1

Define the following terms.
(a) Symmetric matrix
(b) Singular matrix
(c) Rank of a matrix and rank deficiency
(d) Positive definite matrix

Problem T1L5-2

Given the matrix

$$\mathbf{A} = \begin{bmatrix} 4 & -5 & 6 \\ -5 & 3 & 8 \\ 6 & 8 & 10 \end{bmatrix}$$

Compute its determinant. Is the matrix positive definite?

Problem T1L5-3

Solve the set of linear equations given below.

$$\begin{bmatrix} 7 & -3 & 1 \\ -3 & 8 & 2 \\ 1 & 2 & 10 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} 4 \\ 0 \\ -2 \end{Bmatrix}$$

Problem T1L5-4

Numerically integrate the following expressions and compare the answers with the analytical solution.

(a) $\int_3^5 (x^3 - 12x^2 - 3x + 10)\, dx$

(b) $\int_{-1}^{1} \frac{2x-1}{x^2 - 6x + 13}\, dx$

Problem T1L5-5

Numerically differentiate the following expressions and compare the answers with the analytical solution.

(a) $f(x) = 12x^3 + 8x - 10$ at $x = -3$.

(b) $f(x)=\dfrac{3x-4}{4x+10}$ at $x=1$.

Problem T1L5-6

Given a set of four points $(1,8)$, $(2,12)$, $(3,10)$, and $(4,8)$, do a least-squares fit of a quadratic polynomial.

DIFFERENTIAL EQUATIONS & CALCULUS

Problem T1L6-1

Differentiate between an ordinary differential equation (ODE) and a partial differential equation (PDE)?

The boundary conditions to a differential equation can be classified as – essential (Dirichlet), natural (Neumann) or mixed (Robin). What are the traits of each of these types of boundary conditions?

What is meant by a problem being well-posed?

Problem T1L6-2

Explain the terms - Initial Value Problems, Boundary Value Problems, Elliptic and Parabolic Partial Differential Equations.

Problem T1L6-3

Integrate the following expressions

(a) $\int x^2 \cos x \, dx$

(b) $\int x e^{-x} dx$

Problem T1L6-4

Solve the differential equations given below.

(a) $x^2 \dfrac{dy}{dx} + 2xy - x + 1 = 0 \qquad y(1) = 0$

(b) $\dfrac{d}{dx}\left((x+1) \dfrac{dy(x)}{dx} \right) = 0 \qquad 1 < x < 2$

$$y(x=1) = 1 \qquad \left(-(x+1)\frac{dy}{dx} \right)_{x=2} = 1$$

Topic 2: The Six Major Steps

"A common mistake that people make when trying to design something completely foolproof is to underestimate the ingenuity of complete fools." Douglas Adams

Lesson 1: Overview of Finite Elements

Objectives: In this lesson we will look at the "what and when" of finite element analysis. The major objectives are listed below.

- To understand what is meant by "finite element method".
- To understand when the finite element method can be used.

Suggested Reading

Reference	*Pages*
T1	1-2
T2	4-7
T3	3-32

What is the Finite Element Method?

Definition

The Finite Element Method (FEM) is a numerical technique to obtain approximate solutions to a wide variety of engineering problems where the variables are related by means of algebraic, differential and integral equations.

Ponder over the keywords that are underlined.

Approach

Briefly, the engineering problem that is described by governing differential or integral equations is transformed to a set of algebraic equations or eigenequations. These equations are solved numerically.

Flavors of FEM

FEM is not one technique or methodology. There are several different approaches to formulating the problem solution. Three approaches are presented below.

Direct Stiffness Method: Historically this was the first approach used in the area of structural mechanics. We will see the details of this method in the next lesson.

Variational Approach: The original problem is converted to a functional[1] that is extremized. The extremum gives the solution. We will look at this approach very briefly with respect to solid mechanics problems.

Weighted Residuals Approach: An approximate solution is assumed symbolically. The error in the approximate solution is weighted over the domain of the problem and minimized. We will use this approach as the major approach throughout the course.

Let us revisit the definition. What are the characteristics of the governing equations and the solution? The characteristics will determine whether we have a

- Time-independent problem – static (solid mechanics), steady-state (heat transfer)
- Time-dependent problem – dynamic and wave propagation (solid mechanics), diffusion (heat transfer)
- Linear problem – Solution obtained in one pass since the problem parameters are constants
- Nonlinear problem - Solution obtained in several passes since the problem parameters are not constants but interdependent
- Eigenproblem – modal analysis (solid mechanics), acoustics, heat transfer

There are further differences in problem formulation – *p*-finite elements, stress hybrid finite elements (solid mechanics), edge elements (electromagnetics).

---

[1] A functional is a function of a function.

Usage

There are three major stages in any finite element solution.

Pre-Processing: The physical system is converted to a finite element model. This is the most labor intensive and time-consuming step.

Solution: Once the pre-processor has been used to build the model, the finite element analysis can take place. This is the most compute intensive step.

Post-Processing: A post-processor is a program or part of a program that is used to examine the results from the FE analysis. Post-processing is usually graphical but can also be query or report based.

Applications of the Finite Element Method

Finite elements have been used to solve problems in a variety of industries such as

Automotive – stress analysis, heat transfer, crashworthiness.
Civil structures – stress analysis, structural dynamics.
Mechanical – stress analysis, dynamics.
Aerospace – structural analysis, fracture mechanics, fluid flow.
Biomedical – stress analysis, fluid flow.
Electromagnetics – parasitic, harmonic.
Semiconductor (Electronic) Packaging – stress analysis, contact, impact.

Commercial Codes

Some of the commercial programs that are finite-element based are listed below.

ABAQUS: http://www.hks.com/
ADINA: http://www.adina.com/
ALGOR: http://www.algor.com/
ANSYS: http://www.ansys.com/
COSMOS: http://www.cosmosm.com/
MSC-NASTRAN: http://www.msc.com/

Specialized Preprocesors

Parametric Technology: http://www.ptc.com/
PATRAN: http://www.msc.com
Solidworks: http://www.solidworks.com/
Unigraphics: http://www.ugs.com

Finite Element Terminology (the basics)

Fig. T2L1-1 shows a model of a roof truss that is commonly used in residential buildings. The truss members are usually made of wood. Note that the model contains numbers (1 through 14) that denote locations of certain key points on the truss.

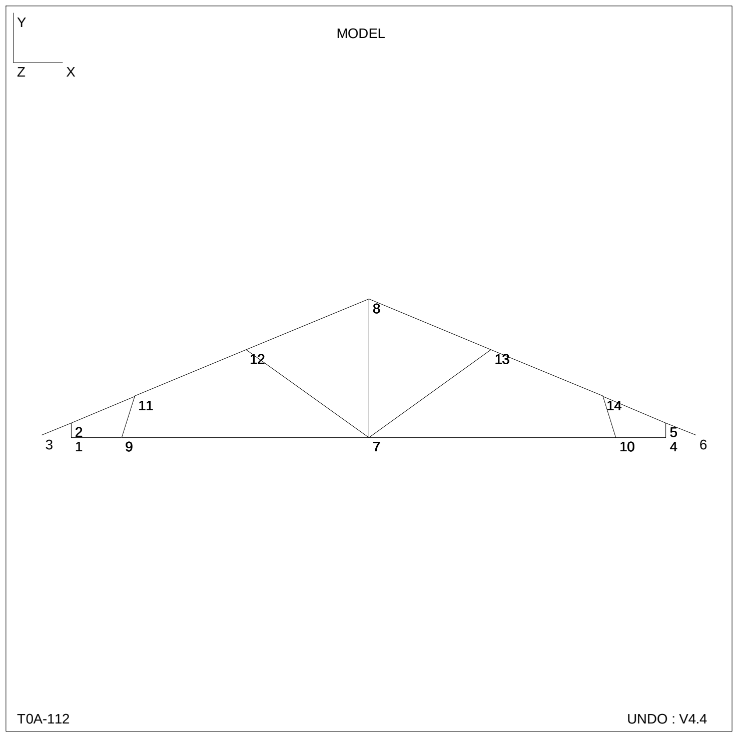


Fig. T2L1-1 Model of roof truss

Fig. T2L1-2 shows the finite element model of a barricade that is used to test air bag explosive charges. The barricade is made of two chambers - the supply chamber and the main chamber. The top of the chambers is vented (opening as shown in the figure). The isometric view shows a collection of triangles used to model the walls.

Using these two examples, we will look at some of the commonly used terms with finite element models.

Elements

Elements can be thought of as the basic building blocks. The collection of the **finite** number elements forms the finite element mesh. In Fig. T2L1-1, the elements are the truss members between two key points. In Fig. T2L1-2, the triangles are the elements. The finite element mesh[2] is an approximation to the physical problem.

[2] The collection of finite elements forms the finite element mesh.

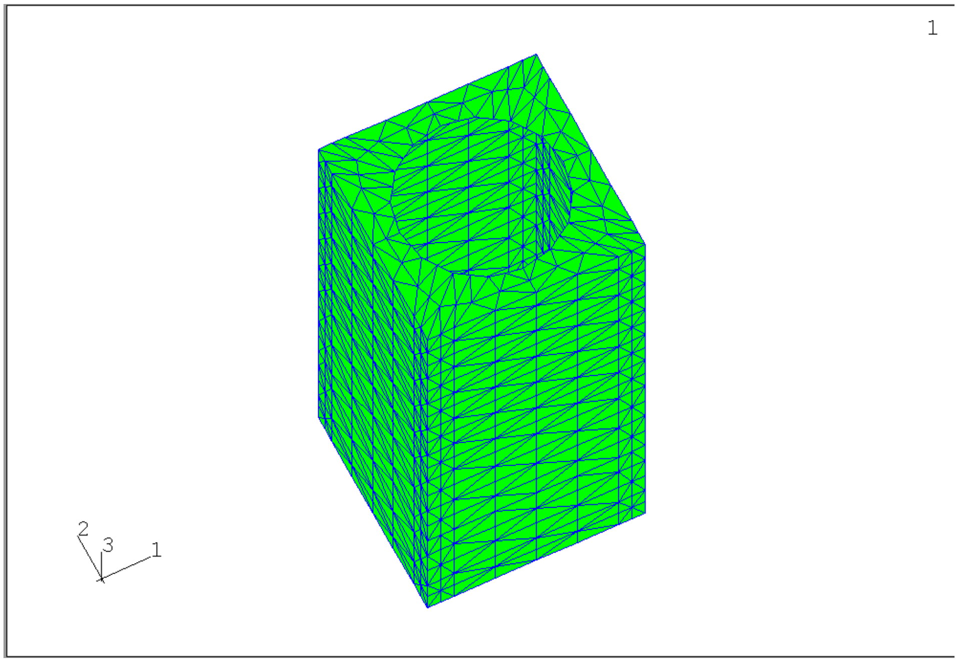


Fig. T2L1-2 Finite element model of a barricade

Nodes
Nodes are the key points. Nodes are used to define elements. In Fig. T2L1-1, two nodes are needed to define an element – a straight line. In Fig. T2L1-2, three nodes are needed to define an element – the triangle.

Element Properties
Since the elements have a physical sense, they are described in terms of element properties - geometry, dimensions and material. To describe the members in the truss, we need to know the shape of the members (typically rectangular), the dimensions (sides of the rectangle) and the material properties (properties of the wood). To describe the triangles, we need to know the *thickness* of the triangle (the thickness of the walls etc. of the chambers) and the material properties of the material making the chamber walls or sides.

Loads
In both the above examples, loads are applied to the structure. The *responses* that we are looking for are quantities such as displacements, stresses, strains etc. that result from the loads applied on the structure.

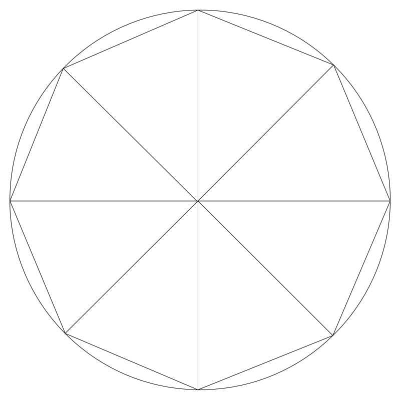
Boundary Conditions
Finally, in order to compute the response quantities, we need to know how the truss or the barricade is supported. Since the supports are on the boundary of the structure, they are called

boundary conditions. As we will see later, there are other examples of boundary conditions apart from supports.

An Illustrative Example T2L1-1

Problem Statement
Consider the following problem. We need to estimate the area of a circle of radius R. We will assume that we do **not** know the exact answer. However, we know some of the rudiments of trigonometry and geometry.

Solution
Step 1: We will assume that we know the formula for the area of a triangle. We will split the circle into a collection of triangles (that can be used as an approximation to the circle) as shown as Mesh A. Note that the mesh of triangles will provide a lower bound to the exact solution.



Mesh A
Fig. T2L1-3

We can see the concept of elements (and nodes) with this example. The continuous region (circle) is represented in each mesh by a finite number of triangles. Each triangle is a finite element. Note that in the mesh every triangle is the same as all the others. When we have this situation, the mesh is known as a uniform mesh. While it is not always necessary to work with a uniform mesh, it is appropriate for this problem.

Step 2: For a typical triangle that subtends an angle θ at the center of the circle as shown below

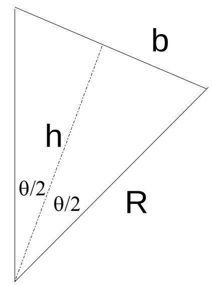


Fig. T2L1-4

$$b = R\sin(\theta/2) \tag{T2L1-1}$$
$$h = R\cos(\theta/2) \tag{T2L1-2}$$

If there are n triangles in the mesh, $\theta = \dfrac{2\pi}{n}$. Hence, the area of one triangle is

$$a_e = \frac{R^2}{2}\sin\left(\frac{2\pi}{n}\right) \tag{T2L1-3}$$

Step 3: Now the estimate the area of the circle we use the obvious concept – the sum of the areas of all the triangles is the area of the circle. In other words

$$A_A^{(n)} = \sum_{e=1}^{n} a_e = \frac{nR^2}{2}\sin\left(\frac{2\pi}{n}\right) \tag{T2L1-4}$$

Step 4: Given the numerical value of R, we can estimate the area of the circle by substituting the numerical values in the above equation. It is apparent that the accuracy is a function of n. In other words, increasing n decreases the error. However note that one has to contend with the numerical problems by increasing n.

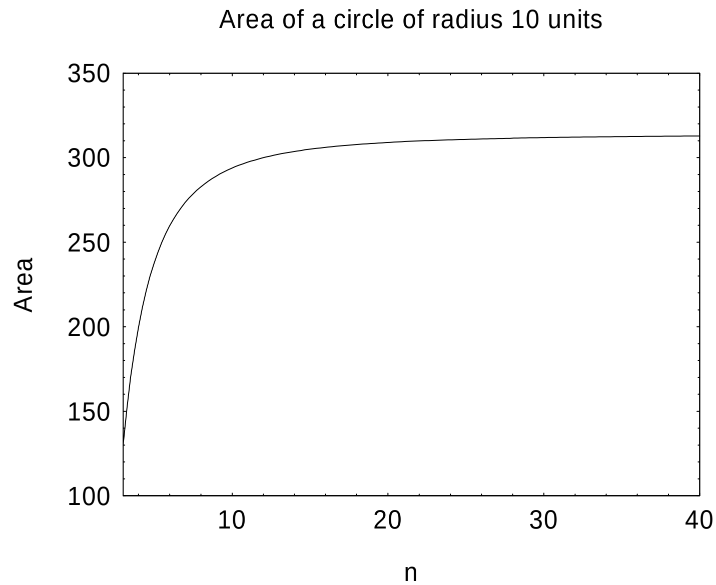
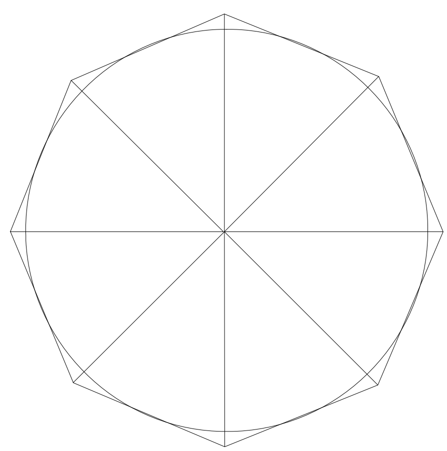


Fig. T2L1-5

Now, consider the collection of triangles that can be used as an approximation to the circle as shown as Mesh B. The mesh of triangles will provide an upper bound to the exact solution.



Mesh B

Fig. T2L1-6

Conclusions

Let's summarize the steps. First, we split the circle (problem *domain*) into simple shape whose properties are known. The process of splitting the problem domain into elements (or, creating the finite element mesh) is known as *discretization*. Second, we computed the properties of a typical element. Third, we estimated the area of the circle used the concept of summing the areas of the individual triangles. This is known as *assembly* where the properties of the elements are assembled to form the properties of the system. In this example, the property of the system was a scalar (a single unknown). Last, the equation was solved numerically to obtain the solution.

In the next lesson we will see these steps in greater detail.

When should finite elements be used?

Finite elements while extremely powerful should be used with caution. FEM should **NOT** be used as a black box.

Let us look at some of the sources of error in a finite element solution.

Translating the Physical System to the FE Model
There are inherent approximations when a physical system is transformed into a finite element model. The approximations include not only those involving the geometry, boundary conditions, and like, but also the approximations inherent in the FE formulation (arising out of the assumptions).

Errors in Parameter Values
The material properties, the loads etc. are best estimates when input into a model.

Numerical Errors
The solution is obtained numerically and is susceptible to numerical errors – truncation and round-off.

Of these three, the first two play a very major role for most of the problems.

Finally, note that there are some classes of problems that can be solved more appropriately by other techniques such as finite differences, finite strips, boundary elements, etc.

Lesson 2: The Six Steps

Objectives: In this lesson we will look at the six steps in a typical finite element solution.

- To understand the basics of the direct stiffness method.
- To understand the concepts associated with primary unknowns, elements, nodes, element equations, assembly, boundary conditions and solution.
- To be able to apply the method to solving a variety of one-dimensional problems.

The Six Steps

We will look at the six major steps in a typical finite element problem by using examples from different engineering areas. These examples will deal with discrete systems. These are systems similar to the roof truss example shown earlier. The elements are essentially one-dimensional. Their geometry and properties are described by a single spatial variable x. Refer to the Illustrative Example T2L1-1 from the previous lesson to reflect on some of the terminology.

Step 1: Discretization

The process of breaking the system or structure into elements is quite simple (unlike creating the triangles in the barricade example) for systems involving discrete elements. The discrete elements are the finite elements.

Step 2: Element Equations

To develop the equations that capture the behavior of a typical element, we need to know the physics of the problems. In the next few sections we will illustrate how this is done for discrete problems in a variety of engineering problems.

We will look at some engineering applications in the areas of solid mechanics, heat transfer, fluid flow and electrical networks. The motivation is to show the similarities between these different specialty areas and point out that the differences are purely based on the physics of the problem.

Linear Springs (Primary unknowns: Displacements)

The physical problem is one involving a collection of springs that are acted upon by external forces. The intent is to calculate the deformation and force in every spring. We will use the following terminology - the *primary unknown* is displacement (this is what we wish to compute first) and the force in the spring is a *derived variable* (one that can be computed provided we have the values of the primary variables). Consider the free-body diagram of a typical linear spring i as shown in the figure below.

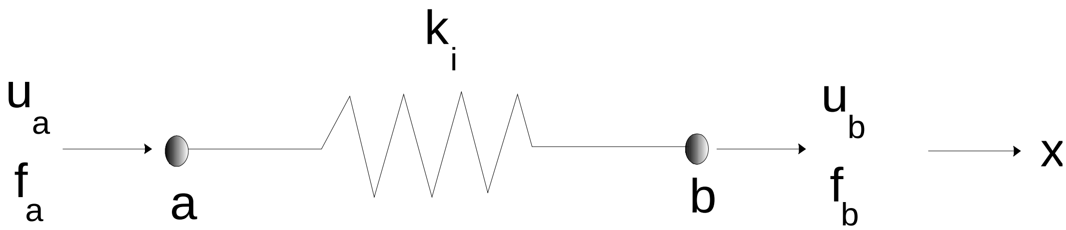
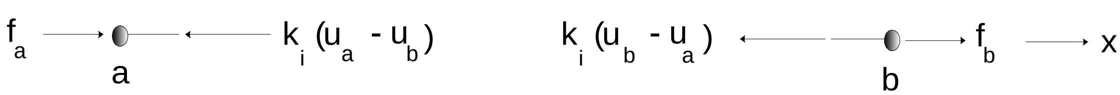


Fig. T2L2-1

The spring element is defined in terms of two nodes that are labeled a and b. u_a represents the displacement of node a and f_a is the force acting at node a. Note that the displacements and forces are all shown acting in the positive coordinate direction. In a general derivation we always assume that all the unknown quantities are positive (the solution to an actual problem will tell us the sign associated with a specific unknown). The spring constant of the element is k_i. In any linear spring

$$u = \frac{f}{k} \tag{T2L2-1}$$

Now let us look at the free-body diagrams of the two nodes.



Using the concept of equilibrium

$$f_a = k_i(u_a - u_b) = k_i u_a - k_i u_b \tag{T2L2-2}$$

$$f_b = k_i(u_b - u_a) = k_i u_b - k_i u_a \tag{T2L2-3}$$

These two equations can be written in the form of matrix equations as

$$k_i \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{Bmatrix} u_a \\ u_b \end{Bmatrix} = \begin{Bmatrix} f_a \\ f_b \end{Bmatrix} \tag{T2L2-4}$$

or, $$\mathbf{k}_{2\times2}\mathbf{u}_{2\times1} = \mathbf{f}_{2\times1} \tag{T2L2-5}$$

where $\mathbf{k}_{2\times2}$ is the element stiffness matrix
$\mathbf{u}_{2\times1}$ is the element nodal displacements vector
$\mathbf{f}_{2\times1}$ is the element nodal forces vector

These are the element equations that capture the concepts of equilibrium-compatibility for a typical spring.

1D Heat Flow (Primary unknowns: Temperature)

The physical problem is one involving a collection of thermal elements that are essentially in a one-dimensional heat flow. The intent is to calculate the temperature distribution and the heat flux. Consider the diagram of a typical thermal element i as

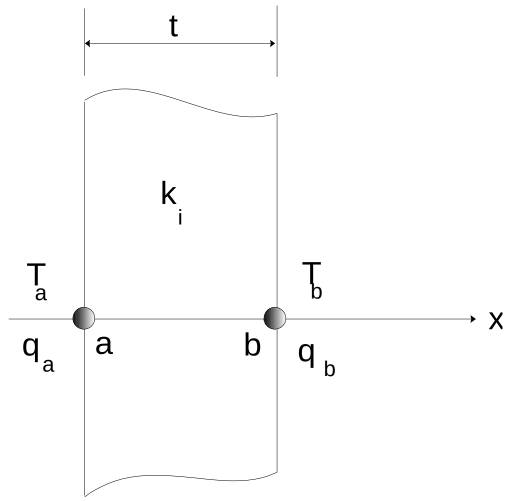


Fig. T2L2-2

Using Fourier's Law

$$q = -kA\frac{dT}{dx} \tag{T2L2-6}$$

where q is the heat flux or energy flow in the x direction(positive if entering the element), k is the thermal conductivity that is assumed to be a constant, T is the temperature and A is the area. With reference to the typical element, at the two nodes we have

$$q_a = k_i A\frac{(T_a - T_b)}{t} \tag{T2L2-7}$$

$$q_b = k_i A\frac{(T_b - T_a)}{t} \tag{T2L2-8}$$

These equations can be written in the matrix form as

$$\frac{k_i A}{t}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{Bmatrix} T_a \\ T_b \end{Bmatrix} = \begin{Bmatrix} q_a \\ q_b \end{Bmatrix} \tag{T2L2-9}$$

or, $$\mathbf{k}_{2\times2}\mathbf{T}_{2\times1} = \mathbf{q}_{2\times1} \tag{T2L2-10}$$

where $\mathbf{k}_{2\times2}$ is the element thermal conductivity matrix

$\mathbf{T}_{2\times1}$ is the element nodal temperature vector

$\mathbf{q}_{2\times1}$ is the element nodal flux vector

These are the element equations that capture the concepts of conservation of energy or energy balance for a typical thermal element.

Pipe Flow (Primary unknowns: Pressure)

The physical problem is one involving a collection of pipes. The intent is to calculate the pressure (at the ends of the pipe) and the flow. Let us assume that the pipes are circular and the flow is fully developed laminar flow. Consider a typical fluid pipe element i as

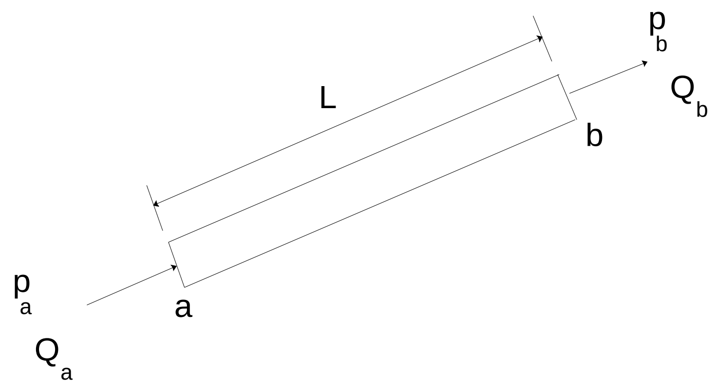


Fig. T2L2-3

Using Darcy's Law, the pressure drop between the ends of a pipe

$$p_1 - p_2 = \frac{128QL\mu}{\pi D^4} \quad \text{(T2L2-11)}$$

where L is the length of the pipe, Q is the flow (positive if entering the element), μ is the dynamic viscosity and D is the pipe diameter. The flows entering the ends of the typical pipe are given as

$$Q_a = \frac{\pi D^4}{128L\mu}(p_a - p_b) \quad \text{(T2L2-12)}$$

$$Q_b = \frac{\pi D^4}{128L\mu}(p_b - p_a) \quad \text{(T2L2-13)}$$

These equations can be written in the matrix form as

$$\frac{\pi D^4}{128L\mu}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{Bmatrix} p_a \\ p_b \end{Bmatrix} = \begin{Bmatrix} Q_a \\ Q_b \end{Bmatrix} \quad \text{(T2L2-14)}$$

or, $$\mathbf{k}_{2\times2}\mathbf{p}_{2\times1} = \mathbf{Q}_{2\times1} \quad \text{(T2L2-15)}$$

where $\mathbf{k}_{2\times2}$ is the element fluidity matrix

$\mathbf{p}_{2\times1}$ is the element nodal pressure vector

$\mathbf{Q}_{2\times1}$ is the element nodal flow vector

These are the element equations that capture the concepts of conservation of mass or mass balance for a typical fluid pipe element.

Electrical Network (Primary unknowns: Voltage)

The physical problem is one involving an electrical network. The intent is to calculate the voltage (at the ends of the resistors or elements). Consider a typical element i as

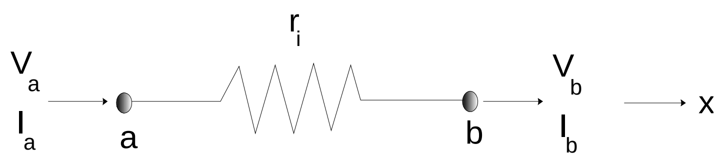


Fig. T2L2-4

Using Ohms's Law, the voltage drop between the ends of an element

$$V_1 - V_2 = rI \tag{T2L2-16}$$

where r is the resistance, I is the current (positive if entering the element), and V is the voltage. The current entering the ends of the typical element are given as

$$I_a = \frac{1}{r_i}(V_a - V_b) \tag{T2L2-17}$$

$$I_b = \frac{1}{r_i}(V_b - V_a) \tag{T2L2-18}$$

These equations can be written in the matrix form as

$$\frac{1}{r_i}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{Bmatrix} V_a \\ V_b \end{Bmatrix} = \begin{Bmatrix} I_a \\ I_b \end{Bmatrix} \tag{T2L2-19}$$

or, $$\mathbf{k}_{2\times2}\mathbf{V}_{2\times1} = \mathbf{I}_{2\times1} \tag{T2L2-20}$$

where $\mathbf{k}_{2\times2}$ is the element resistivity matrix

$\mathbf{V}_{2\times1}$ is the element nodal voltage vector

$\mathbf{I}_{2\times1}$ is the element nodal current vector

These are the element equations that capture the concepts of electrical flow and continuity in a typical resistor.

Step 3: Assembly
The obvious question is "What do we do with the element equations?" The answer is provided with respect to an example shown below.

Illustrative Example T2L2-1

The figure shows a collection of 4 springs that is loaded with two concentrated forces.

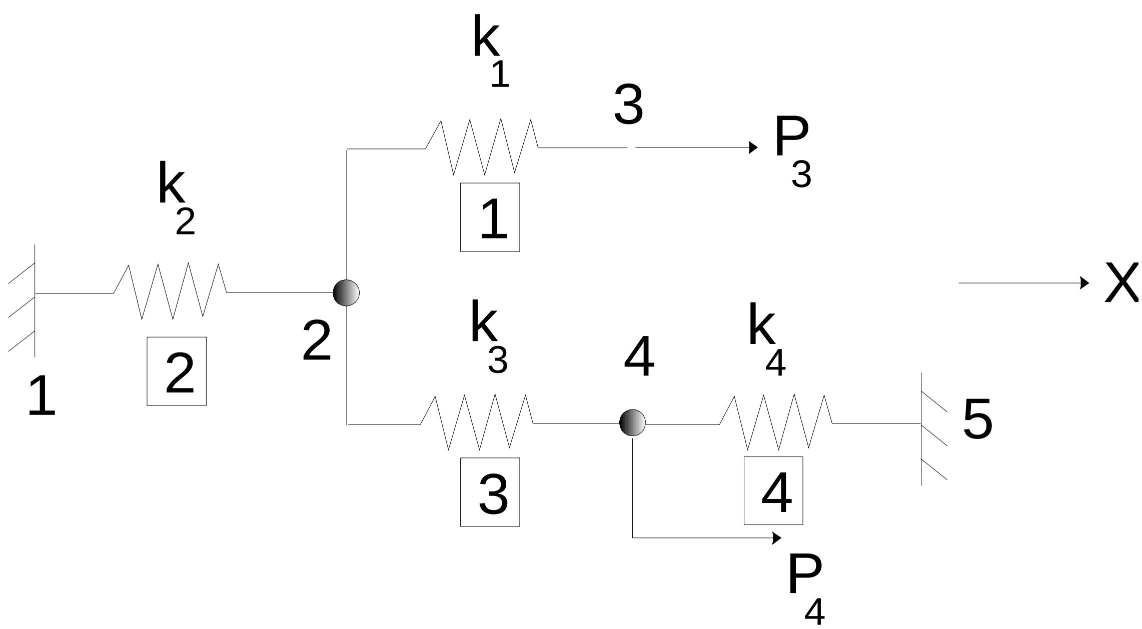


Fig. T2L2-5

The first task is to number the nodes. We have labeled the nodes starting at 1, consecutively all the way to 5. Designating a node as 1 or 2 … is arbitrary. Next, the elements are labeled starting at 1 consecutively all the way to 4. Hence, the finite element mesh has 5 nodes and 4 elements. Each node has one unknown variable associated with it – the displacement at the node. In FE terminology, we state that there is one degree of freedom (DOF) per node. The problem is described in terms of a total of 5 DOF - $\{D_1, D_2, D_3, D_4, D_5\}$. What we have done here is what needs to be done in Step 1 with respect to discretizing a problem.

Let's go ahead and implement Step 2 for this problem.

Element 1: $a=2$ and $b=3$

$$k_1\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{Bmatrix} D_2 \\ D_3 \end{Bmatrix} = \begin{Bmatrix} f_2^1 \\ f_3^1 \end{Bmatrix} \qquad \text{(T2L2-21)}$$

Element 2: $a = 1$ and $b = 2$

$$k_2\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{Bmatrix} D_1 \\ D_2 \end{Bmatrix} = \begin{Bmatrix} f_1^2 \\ f_2^2 \end{Bmatrix} \tag{T2L2-22}$$

Element 3: $a = 2$ and $b = 4$

$$k_3\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{Bmatrix} D_2 \\ D_4 \end{Bmatrix} = \begin{Bmatrix} f_2^3 \\ f_4^3 \end{Bmatrix} \tag{T2L2-23}$$

Element 4: $a = 4$ and $b = 5$

$$k_4\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{Bmatrix} D_4 \\ D_5 \end{Bmatrix} = \begin{Bmatrix} f_4^4 \\ f_5^4 \end{Bmatrix} \tag{T2L2-24}$$

A slightly different notation is used above with respect to applying Eqn. (T2L2-4). The element nodal displacements are mapped from the generic u_a and u_b to the appropriate global designation D_n. The forces acting at the ends of the element are denoted f_i^j meaning that the force is acting at end i of element j (the superscript j is necessary to differentiate the contributions made by different elements that share a node).

The assembly process is now to take these 8 equilibrium-compatibility equations and <u>construct 5 equilibrium-compatibility equations for the entire system</u>. The system equations look like

$$\begin{bmatrix} K_{11} & K_{12} & K_{13} & K_{14} & K_{15} \\ K_{21} & K_{22} & K_{23} & K_{24} & K_{25} \\ K_{31} & K_{32} & K_{33} & K_{34} & K_{35} \\ K_{41} & K_{42} & K_{43} & K_{44} & K_{45} \\ K_{51} & K_{52} & K_{53} & K_{54} & K_{55} \end{bmatrix}\begin{Bmatrix} D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \end{Bmatrix} = \begin{Bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \\ F_5 \end{Bmatrix} \tag{T2L2-25a}$$

or, $$\mathbf{K}_{5\times5}\mathbf{D}_{5\times1} = \mathbf{F}_{5\times1} \tag{T2L2-25b}$$

where $\mathbf{K}$ is the global (or, structural, or, system) stiffness matrix, $\mathbf{D}$ and $\mathbf{F}$ are the vector of global nodal displacements and global nodal forces respectively.

But how do we construct Eqns. (T2L2-25)? Now back to the element equations. The first equation for element 1 is the equilibrium-compatibility of element 1 along D_2. Similarly, the second equation for element 2 and the first equation for element 3 are the equilibrium-compatibility along D_2. Hence these three equations combine to generate the second global equation. In a similar manner we can collect the appropriate equations for the other four DOFs.

However, one recommended way to assemble is to go through sequentially one element at a time and update Eqn. (T2L2-25). Hence after using the Eqns. (T2L2-21) for element 1 we have

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & k_1 & -k_1 & 0 & 0 \\ 0 & -k_1 & k_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{Bmatrix} D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \end{Bmatrix} = \begin{Bmatrix} 0 \\ f_2^1 \\ f_3^1 \\ 0 \\ 0 \end{Bmatrix} \tag{T2L2-26}$$

After using Eqns. (T2L2-22) for element 2 we have

$$\begin{bmatrix} k_2 & -k_2 & 0 & 0 & 0 \\ -k_2 & k_1+k_2 & -k_1 & 0 & 0 \\ 0 & -k_1 & k_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{Bmatrix} D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \end{Bmatrix} = \begin{Bmatrix} f_1^2 \\ f_2^1 + f_2^2 \\ f_3^1 \\ 0 \\ 0 \end{Bmatrix} \tag{T2L2-27}$$

After using Eqns. (T2L2-23) for element 3 we have

$$\begin{bmatrix} k_2 & -k_2 & 0 & 0 & 0 \\ -k_2 & k_1+k_2+k_3 & -k_1 & -k_3 & 0 \\ 0 & -k_1 & k_1 & 0 & 0 \\ 0 & -k_3 & 0 & k_3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{Bmatrix} D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \end{Bmatrix} = \begin{Bmatrix} f_1^2 \\ f_2^1 + f_2^2 + f_2^3 \\ f_3^1 \\ f_4^3 \\ 0 \end{Bmatrix} \tag{T2L2-28}$$

Finally, after all the four elements have been assembled

$$\begin{bmatrix} k_2 & -k_2 & 0 & 0 & 0 \\ -k_2 & k_1+k_2+k_3 & -k_1 & -k_3 & 0 \\ 0 & -k_1 & k_1 & 0 & 0 \\ 0 & -k_3 & 0 & k_3+k_4 & -k_4 \\ 0 & 0 & 0 & -k_4 & k_4 \end{bmatrix} \begin{Bmatrix} D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \end{Bmatrix} = \begin{Bmatrix} f_1^2 \\ f_2^1 + f_2^2 + f_2^3 \\ f_3^1 \\ f_4^3 + f_4^4 \\ f_5^4 \end{Bmatrix} \tag{T2L2-29}$$

Now comparing the system load vector to the Fig. T2L2-5, $f_2^1 + f_2^2 + f_2^3 = 0$, $f_3^1 = P_3$ and $f_4^3 + f_4^4 = P_4$. F_1 and F_5 are not known and are the support reactions that develop at the two supports. Interestingly enough $D_1 = 0$ and $D_5 = 0$, a fact that we will use in the next step.

Hence the equations reduce to

$$\begin{bmatrix} k_2 & -k_2 & 0 & 0 & 0 \\ -k_2 & k_1+k_2+k_3 & -k_1 & -k_3 & 0 \\ 0 & -k_1 & k_1 & 0 & 0 \\ 0 & -k_3 & 0 & k_3+k_4 & -k_4 \\ 0 & 0 & 0 & -k_4 & k_4 \end{bmatrix} \begin{Bmatrix} D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \end{Bmatrix} = \begin{Bmatrix} F_1 \\ 0 \\ P_3 \\ P_4 \\ F_5 \end{Bmatrix} \tag{T2L2-30}$$

The rest of the problem will be solved using numerical data. Let

$$k_1 = 10\frac{lb}{in} = k_4 \text{ , } k_3 = 15\frac{lb}{in} \text{ , } k_2 = 20\frac{lb}{in} \text{ , } F_3 = 100\,lb \text{ , } F_4 = 50\,lb$$

Step 4: Imposition of Boundary Conditions

Eqns. (T2L2-29) cannot be solved for two reasons – first we do not know the RHS vector completely and second, the $\mathbf{K}$ matrix is singular! The boundary conditions for the given problem refer to the essential boundary conditions (EBC), i.e. $D_1 = 0$ and $D_5 = 0$. Imposing the EBCs will make it possible to solve the system equations.

Since $D_1 = 0$, the first equation is not necessary to solve for the other unknowns. Similarly, since $D_5 = 0$, the last equation is not necessary to solve for the other unknowns. These two conditions can be simply imposed by modifying the first and the last equations **and** the columns corresponding to D_1 and D_5, i.e. the first and the last *columns*[3]. Hence,

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & k_1+k_2+k_3 & -k_1 & -k_3 & 0 \\ 0 & -k_1 & k_1 & 0 & 0 \\ 0 & -k_3 & 0 & k_3+k_4 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{Bmatrix} D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ P_3 \\ P_4 \\ 0 \end{Bmatrix} \tag{T2L2-31a}$$

or, $$\overline{\mathbf{K}}_{5\times5}\mathbf{D}_{5\times1} = \overline{\mathbf{F}}_{5\times1} \tag{T2L2-31b}$$

where the parameters are the modified global stiffness, global displacement and modified global load vector respectively.

[3] The LHS is $\mathbf{KD}$ and involves matrix multiplication.

This method of imposing the EBC is known as the direct or elimination approach. Examine the above equations – the first and the last equations are decoupled from the rest and solving them yields $D_1 = 0$ and $D_5 = 0$ which is what we set out to do.

Now substituting the numerical values, we have

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 45 & -10 & -15 & 0 \\ 0 & -10 & 10 & 0 & 0 \\ 0 & -15 & 0 & 25 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{Bmatrix} D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 100 \\ 50 \\ 0 \end{Bmatrix} \tag{T2L2-32}$$

Step 5: Solution of Primary Unknowns

The system equations can be solved using a variety of techniques. Some of the more important techniques used in FE programs are

(a) Direct Method: Gaussian Elimination Technique (examples include LDL^T or Cholesky Decomposition for symmetric, positive definite coefficient matrix), and

(b) Iterative Method: Preconditioned Conjugate Gradient Method.

While the theory is somewhat standard for these techniques, there exist tens of ways of implementing the solution techniques.

Solving Eqns. (T2L2-32), we have $\{D_1, D_2, D_3, D_4, D_5\} = \{0,5,15,5,0\}\ in$.

Step 6: Obtaining Derived Variables

Using the nodal displacements we can now compute the forces in the springs and the support reactions.

Using Eqns. (T2L2-4) for each element (units are in and lb).

Element 1:

$$10\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{Bmatrix} 5 \\ 15 \end{Bmatrix} = \begin{Bmatrix} -100 \\ 100 \end{Bmatrix} = \begin{Bmatrix} f_a \\ f_b \end{Bmatrix} \tag{T2L2-33}$$

Element 2:

$$20\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{Bmatrix} 0 \\ 5 \end{Bmatrix} = \begin{Bmatrix} -100 \\ 100 \end{Bmatrix} = \begin{Bmatrix} f_a \\ f_b \end{Bmatrix} \tag{T2L2-34}$$

Element 3:

$$15\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{Bmatrix} 5 \\ 5 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix} = \begin{Bmatrix} f_a \\ f_b \end{Bmatrix} \tag{T2L2-35}$$

Element 4:

$$10\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{Bmatrix} 5 \\ 0 \end{Bmatrix} = \begin{Bmatrix} 50 \\ -50 \end{Bmatrix} = \begin{Bmatrix} f_a \\ f_b \end{Bmatrix} \tag{T2L2-36}$$

Let's interpret one of these results by selecting Element 2. The FBD for the element is

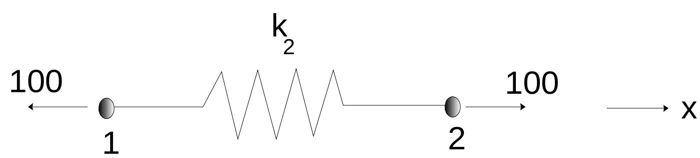


Fig. T2L2-6

indicating that the spring is in equilibrium (and in tension). Draw the FBDs for all the other elements.

Finally, let's draw the FBD for nodes 1 and 2.

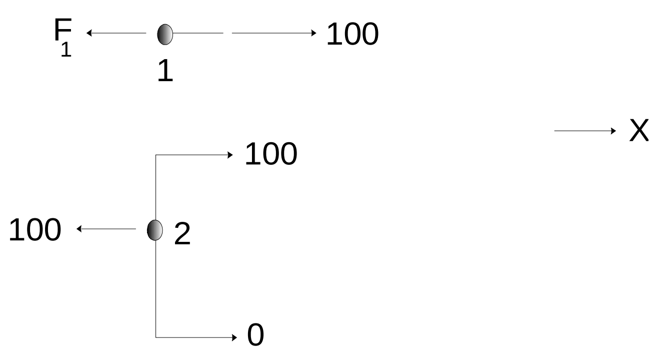


Fig. T2L2-7

Again note that each node should be in equilibrium. Using the FBD of node 1

$$\overset{\rightarrow +}{\sum} F_x = 0 = -F_1 + 100 \Rightarrow F_1 = 100$$

and the reaction at node 1 (a support) is $100\,lb$. An examination of Node 2 FBD shows that it is in equilibrium.

Some Important Observations

- The element stiffness matrix $\mathbf{k}$ is symmetric and singular.
- The global stiffness matrix $\mathbf{K}$ is symmetric, banded and singular.
- The modified global stiffness matrix $\overline{\mathbf{K}}$ is symmetric, banded and nonsingular (positive definite).
- It will be necessary to look at the imposition of the EBC again in order to handle situations where the EBCs are non-homogeneous.

Consider the set of equations shown below.

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} b_1 \\ b_2 \\ b_3 \end{Bmatrix} \tag{T2L2-38}$$

We will first rewrite the equations as

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 + A_{13}x_3 &= b_1 \\ A_{21}x_1 + A_{22}x_2 + A_{23}x_3 &= b_2 \\ A_{31}x_1 + A_{32}x_2 + A_{33}x_3 &= b_3 \end{aligned} \tag{T2L2-39}$$

Let the EBC be such that $x_2 = c$, a known constant. Imposing the EBC in the first and last equations, we have

$$\begin{aligned} A_{11}x_1 + \qquad A_{13}x_3 &= b_1 - A_{12}x_2 = b_1 - A_{12}c \\ A_{31}x_1 + \qquad A_{33}x_3 &= b_3 - A_{32}x_2 = b_3 - A_{32}c \end{aligned}$$

Since $x_2 = c$, the second equation can be modified to

$$(0)x_1 + (1)x_2 + (0)x_3 = c$$

Collecting the three modified equations, we now have

$$\begin{bmatrix} A_{11} & 0 & A_{13} \\ 0 & 1 & 0 \\ A_{31} & 0 & A_{33} \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} b_1 - A_{12}c \\ c \\ b_3 - A_{32}c \end{Bmatrix} \tag{T2L2-40}$$

which can be solved for the unknown vector $\mathbf{x}$. Note that c can be any known number including **zero**.

Summary

The six major steps in a typical finite element solution are as follows.

Step 1: Discretization
The problem domain is discretized into a collection of simple shapes (or, elements).

Step 2: Element Equations
Knowing the physics of the problem once can develop the element equations for a typical element. These equations are symbolic, and can be evaluated numerically by substituting the numerical values for the symbols (problem parameters). The primary unknowns are evaluated at the nodes of the elements.

Step 3: Assembly
The element equations for each element in the FE mesh needs to be assembled into a set of global equations that represent the properties of the entire system.

Step 4: Imposition of Boundary Conditions
The global equations cannot be solved unless the boundary conditions are imposed. The boundary conditions reflect which of the primary unknowns have known values.

Step 5: Solution of Primary Unknowns
The modified global equations from Step 4 are solved for the primary unknowns. One has to be careful about issues like condition number of the stiffness matrix and round-off errors.

Step 6: Obtaining Derived Variables
The evaluated values of the primary unknowns are used to obtain derived values using the relationships between the primary unknowns and the derived variables.

Finally, one must use the computed answers to check whether the physical solution makes sense. In the example, we checked for the equilibrium of the nodes and elements.

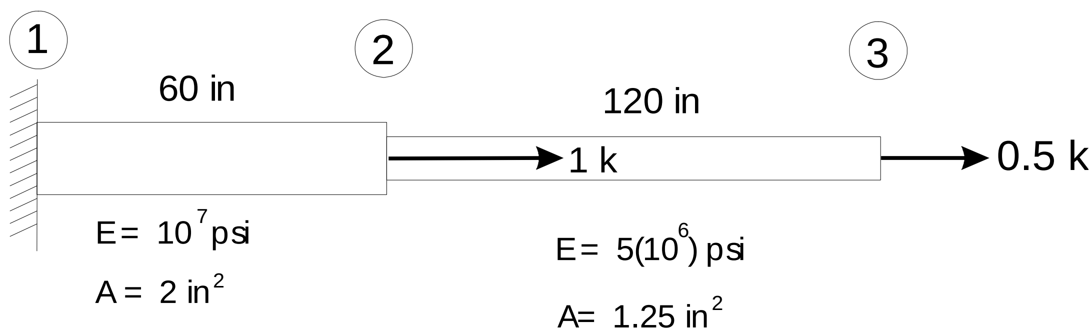
Review Exercises

Problem T2L1-1

Estimate the area of the circle of unit radius using Mesh A and Mesh B. Compare the two solutions.

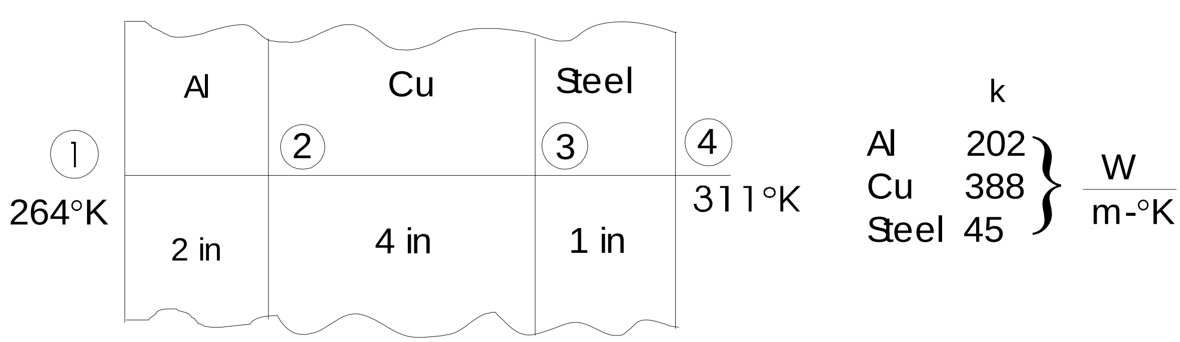
Problem T2L2-1

A uniform bar is made up of two sections is loaded by concentrated nodal forces as shown. Use a two-element finite element model to compute the nodal displacements, element forces and stresses. Hint: The stiffness k of a bar of length L, uniform cross-sectional area A and modulus of elasticity E is $k = \dfrac{AE}{L}$.



Problem T2L2-2

A composite wall consists of layers of aluminum, copper, and steel. The surface temperatures are $264^\circ K$ and $311^\circ K$. Use a three-element model to determine the interfacial temperatures and the heat flow per unit area.

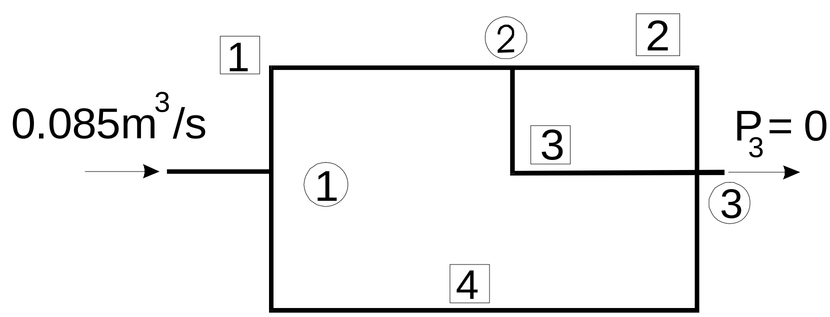


Problem T2L2-3

Laminar incompressible flow occurs in the branched network of circular pipes. If $0.085 \frac{m^3}{s}$ fluid enters and leaves the piping network, use a three-node, four-element model to compute the fluid nodal pressures and the volume flow rate in each pipe.

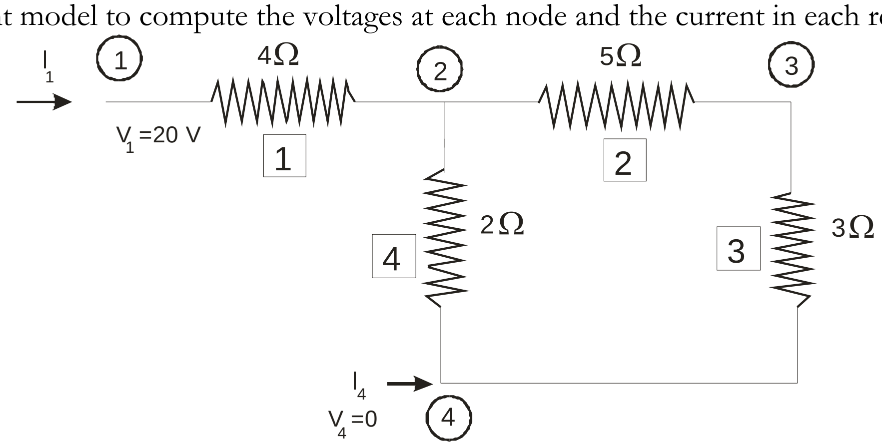
Pipe	Diameter (cm)	Length (m)
1	7.6	30.5
2	5.1	45.8
3	5.1	45.8
4	10.2	91.5

The dynamic viscosity is $0.96(10^{-3})\frac{N \cdot s}{m^2}$.



Problem T2L2-4

The voltages at the output terminals of the DC circuit have the values shown. Use a four-node finite element model to compute the voltages at each node and the current in each resistor.



Topic 3: The Galerkin's Method

"Engineering problems are under-defined, there are many solutions, good, bad and indifferent. The art is to arrive at a good solution. This is a creative activity, involving imagination, intuition and deliberate choice." Ove Arup

Lesson 1: Method of Weighted Residuals.

Objectives: In this lesson we will look at Method of Weighted Residuals to solve one-dimensional differential equations and the Galerkin's method in particular.

- To understand the Method of Weighted Residuals.
- To understand the Galerkin's Method and concepts associated with trial solution, optimizing criteria, accuracy.

Background

One-dimensional boundary-value problems have several applications some of which we saw in the context of the direct stiffness method in the previous topic. We will list some additional problems below.

Solid Mechanics	Transverse deflection of a cable
Hydrodynamics	One-dimensional flow in an inviscid, incompressible fluid
Magnetostatics	One-dimensional magnetic potential distribution
Heat Conduction	One-dimensional heat flow in a solid medium
Electrostatics	One-dimensional electric potential distribution

Consider a one-dimensional boundary value problem (also known as equilibrium problem) given by

$$\frac{d^2y}{dx^2} + P(x)\frac{dy}{dx} + Q(x)y = F(x) \qquad \text{(T3L1-1)}$$

If P and Q are constants, then

$$\frac{d^2y}{dx^2} + P\frac{dy}{dx} + Qy = F(x) \qquad \text{(T3L1-2)}$$

The total solution is

$$y(x) = Ae^{\lambda_1 x} + Be^{\lambda_2 x} + C_0 F(x) + C_1 F'(x) + ... \qquad \text{(T3L1-3)}$$

where the constants of integration A and B can be found by substituting the two boundary conditions into Eqn. (T3L1-2). Note that we need two boundary conditions for the problem to be well-posed.

The boundary conditions are of three types.
(a) The function y may be specified. This is known as the Dirichlet or Essential boundary condition.
(b) The derivative y' may be specified. This is known as the Neumann or Natural boundary condition.
(c) The function y and the derivative y' may be specified. This is known as the Mixed or Robin boundary condition.

In the FE solution, an approximate or trial solution $\tilde{y}(x)$ is constructed and solved for[4]. The FE approach has three distinct operations.

[4] Those familiar with the finite difference solution will note that there are two approaches to solving the boundary-value ODE – the shooting (initial-value) method and the equilibrium (boundary-value) method.

(a) A trial solution $\tilde{y}(x)$ is constructed.

(b) An optimizing criterion is applied to $\tilde{y}(x)$.

(c) An estimation of the accuracy of $\tilde{y}(x)$ is made.

Trial Solution

The trial solution is constructed with a finite number of terms as

$$\tilde{y}(x;a) = \phi_0(x) + a_1\phi_1(x) + a_2\phi_2(x) + ... + a_n\phi_n(x) \qquad \text{(T3L1-4)}$$

where $\phi_i(x)$ are known trial or basis functions and the coefficients a_i are undetermined parameters known as degrees of freedom (DOF). The purpose of the trial function $\phi_0(x)$ is to satisfy some or all of the boundary conditions. The most common form of trial solutions is to use polynomials. We will see more about this later.

Optimizing Criterion

The optimizing criterion is used to generate the appropriate equations so that we can solve for the numerical values of the coefficients a_i. As you can guess, the optimizing criterion is not unique and different approaches define what is meant by the "best possible approximation" to the exact solution. The two most common forms are

(a) The Method of Weighted Residuals (MWR) – applicable when the problem is described by differential equations, and

(b) The Ritz Variational Method (RVM) - applicable when the problem is described by integral (or, variational) equations.

In this lesson, the focus is on the former method. In the MWR, the criteria minimize an expression of error in the differential equation. In the RVM, an attempt is made to extremize (typically, minimize) a physical quantity. We will see this approach in the second module of this course.

Accuracy Estimate

Since the FE solution is usually approximate and since for practical problems it is virtually impossible to generate or estimate the exact solution, we need a measure to tell us how close the FE solution is to the exact solution. We will illustrate the ideas and estimates later.

Method of Weighted Residuals

There are at least four different optimizing criteria and we will see them in this lesson. Consider the differential equation (T3L1-1) rewritten as

$$\frac{d^2 y}{dx^2} + P(x)\frac{dy}{dx} + Q(x)y - F(x) = 0 \qquad \text{(T3L1-5a)}$$

or, $$\Im(y) - F(x) = 0 \qquad \text{(T3L1-5b)}$$

Substituting the trial solution, we have, in general

$$R(x;a) = \Im(\tilde{y}) - F(x) \neq 0 \qquad \text{(T3L1-6)}$$

$R(x;a)$ is known as the residual or error in the solution. In the MWR, the optimizing criterion is to find the numerical values for a_i which will make $R(x;a)$ as close to zero as possible for all values of x throughout the domain of the problem. Once a specific criterion is applied, a set of algebraic equations is produced. As you can see, the process is to transform the original (linear) ODE to a set of linear algebraic equations.

Collocation Method

In this method, for each of the parameter a_i a point x_i is chosen and the residual is set to zero at that point.

$$R(x_i;a) = 0 \qquad i = 1,..,n \qquad \text{(T3L1-7)}$$

The points x_i are known as the collocation points. Note that we are setting the error in the residual, not the error in the solution, to zero.

Subdomain Method

In this method, for each of the parameter a_i an interval Δx_i is chosen and the average of the residual is set to zero.

$$\frac{1}{\Delta x_i}\int_{\Delta x_i} R(x;a)\, dx = 0 \qquad i = 1,..,n \qquad \text{(T3L1-8)}$$

The intervals Δx_i are called the subdomains.

Least-squares Method

In this method, with respect to each a_i we minimize the integral over the entire domain, the square of the residual.

$$\frac{\partial}{\partial a_i}\int_{\Omega} R^2(x;a)\,dx = 0 \qquad i = 1,..,n \tag{T3L1-9}$$

The Galerkin's Method

In this method, for each a_i we require that a weighted average of the residual over the entire domain be zero. The weighting functions are the trial functions $\phi_i(x)$ associated with each a_i.

$$\int_{\Omega} R(x;a)\phi_i(x)dx = 0 \qquad i = 1,..,n \tag{T3L1-10}$$

The natural question is "Which of these techniques is the most appropriate?" A detailed answer is outside the scope of this course. However, experience has shown that the Galerkin's Method is the most suitable for the type of finite element applications to be seen in this course.

In the next lesson, we will apply the Galerkin's Method in the classical sense. In the next topic we will overcome the drawbacks of the classical approach and use the finite element concepts.

Lesson 2: Galerkin's Method

Objectives: In this lesson we will look at applying the Galerkin's Method (in the theoretical or classical sense) and understanding its traits.

- To apply the Galerkin's Method.
- To understand the concepts of trial solution, accuracy, classical solution, numerical solution and convergence.

An Illustrative Example T3L2-1

Let us look at an example to see how the Galerkin's Method works. Consider the problem

DE: $$\frac{d}{dx}\left(x\frac{dy(x)}{dx}\right)=\frac{2}{x^2} \qquad 1 \le x \le 2 \qquad \text{(T3L2-1)}$$

BC: $$y(x=1)=2 \qquad \text{(T3L2-2)}$$

$$\left(-x\frac{dy}{dx}\right)_{x=2}=\frac{1}{2} \qquad \text{(T3L2-3)}$$

Classical (or, Theoretical) Solution

Let us assume the trial solution as

$$\tilde{y}=a_1+a_2x+a_3x^2+a_4x^3 \qquad \text{(T3L2-4a)}$$

Hence,

$$\frac{d\tilde{y}}{dx}=a_2+2a_3x+3a_4x^2 \qquad \text{(T3L2-4b)}$$

In order to satisfy the BC, we need

$$\tilde{y}(x=1)=2=a_1+a_2(1)+a_3(1)^2+a_4(1)^3$$

or, $$a_1+a_2+a_3+a_4=2 \qquad \text{(T3L2-5)}$$

And,

$$\left(-x\frac{d\tilde{y}}{dx}\right)_{x=2}=\frac{1}{2}=-2a_2-8a_3-24a_4$$

or, $$a_2+4a_3+12a_4=-\frac{1}{4} \qquad \text{(T3L2-6)}$$

Eqns. (T3L2-5) and (T3L2-6) are known as constraint equations since the 4 a_i's in Eqn. (T3L2-4a) are no longer independent. The two constraint equations render only 2 out of the 4 a_i's as independent parameters (or, DOF). Using the two constraint equations to write a_1 and a_2 in terms of the other two, we have

$$a_1=2-a_2-a_3-a_4 \qquad \text{(T3L2-7)}$$

$$a_2=-\frac{1}{4}-4a_3-12a_4 \qquad \text{(T3L2-8)}$$

Substituting these two equations in (T3L2-4a)

$$\tilde{y}=2-\frac{1}{4}(x-1)+a_3(x-1)(x-3)+a_4(x-1)(x^2+x-11) \qquad \text{(T3L2-9)}$$

This equation is now in the familiar form

$$\tilde{y}(x;a) = \phi_0(x) + a_1\phi_1(x) + a_2\phi_2(x) \qquad \text{(T3L2-10)}$$

where $$\phi_0(x) = 2 - \frac{1}{4}(x-1)$$

$$\phi_1(x) = (x-1)(x-3)$$

$$\phi_2(x) = (x-1)(x^2+x-11)$$

and we will assume for the sake of convenience that a_1 and a_2 in (T3L2-10) are the same as a_3 and a_4 in (T3L2-9).

We will now write the residual as

$$R(x;a) = \frac{d}{dx}\left(x\frac{d\,\tilde{y}(x)}{dx} \right) - \frac{2}{x^2} \neq 0 \qquad \text{(T3L2-11)}$$

Substituting (T3L2-10) in the above equation, we have

$$R(x;a) = -\frac{1}{4} + 4(x-1)a_1 + 3(3x^2-4)a_2 - \frac{2}{x^2} \qquad \text{(T3L2-12)}$$

Now using the Galerkin's Method (Eqn. (T3L1-10))

$$\int_1^2 \left(-\frac{1}{4} + 4(x-1)a_1 + 3(3x^2-4)a_2 - \frac{2}{x^2} \right)(x-1)(x-3)\,dx = 0$$

$$\int_1^2 \left(-\frac{1}{4} + 4(x-1)a_1 + 3(3x^2-4)a_2 - \frac{2}{x^2} \right)(x-1)(x^2+x-11)\,dx = 0 \qquad \text{(T3L2-13a)}$$

Integrating yields two equations

$$\begin{bmatrix} \frac{5}{3} & \frac{41}{5} \\ \frac{41}{5} & \frac{81}{2} \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} -\frac{29}{6} + 8\ln 2 \\ -\frac{211}{16} + 24\ln 2 \end{Bmatrix} \qquad \text{(T3L2-13b)}$$

And solving

$$a_1 = 2.138 \text{ and } a_2 = -0.348$$

Hence, substituting in Eqn. (T3L2-10) yields

$$\tilde{y}(x;a) = 2 - \frac{1}{4}(x-1) + 2.138(x-1)(x-3) - 0.348(x-1)(x^2+x-11)$$

$$= -0.348x^3 + 2.138x^2 - 4.629x + 4.839$$

(T3L2-14)

There is an important concept associated with a derived term - *flux* that we have not seen before and should have. Flux is defined as

$$\tau(x) = -x\frac{dy}{dx} \qquad \text{(T3L2-15)}$$

Flux has a physical interpretation and we will discuss the concepts in the next topic. Substituting in Eqn. (T3L2-15), we have

$$\begin{aligned}\tilde{\tau}(x;a) &= \frac{1}{2}+\frac{1}{4}(x-2)-4.276x(x-2)+1.043x(x-2)(x+2) \\ &= 1.043x^3-4.276x^2+4.629x\end{aligned} \tag{T3L2-16}$$

Finally, note that this solution is called the theoretical (or, classical) solution because of the manner in which the two boundary conditions were imposed. In this methodology, the BCs were imposed on the trial solution itself so that the trial solution would satisfy the BCs exactly. However, this in no way implies that the solution is correct in the interior of the problem domain.

Some Important Observations

- Why did we start with a trial solution that is a cubic polynomial? Since there are two BCs that must be imposed, the *lowest* order polynomial that we could have used would have been a quadratic polynomial. Imposing the BCs then would have left one free parameter (or, one DOF). The choice of using a cubic polynomial was an arbitrary choice and we were left with two free parameters.
- The classical way of applying the BCs can become extremely cumbersome with more complex problems.
- How do we know whether the solution is good? One way to answer this question is to look at the concept of *convergence*. The exact solution to the problem is

$$y(x) = \frac{2}{x} + \frac{1}{2}\ln x \tag{T3L2-17a}$$

and

$$\tau(x) = \frac{2}{x} - \frac{1}{2} \tag{T3L2-17b}$$

Convergence can be checked by starting with a low-order polynomial and increasing the order of the polynomial gradually. The different trial functions should converge to a solution. If we had started with a quadratic polynomial, the Galerkin's solution would have been

$$\begin{aligned}\tilde{y}(x;a) &= 2 - \frac{1}{4}(x-1) + 0.427(x-1)(x-3) \\ &= 0.427x^2 - 1.958x + 3.531\end{aligned} \tag{T3L2-18}$$

and

$$\tilde{\tau} = \frac{1}{2} + \frac{1}{4}(x-2) - 0.854x(x-2) = -0.854x^2 + 1.958x \tag{T3L2-19}$$

In the previous section, with the trial solution as a cubic polynomial the solution was

$$\begin{aligned}\tilde{y}(x;a) &= 2 - \frac{1}{4}(x-1) + 2.138(x-1)(x-3) - 0.348(x-1)(x^2+x-11) \\ &= -0.348x^3 + 2.138x^2 - 4.629x + 4.839\end{aligned}$$

and

$$\begin{aligned}\tilde{\tau}(x;a) &= \frac{1}{2} + \frac{1}{4}(x-2) - 4.276x(x-2) + 1.043x(x-2)(x+2) \\ &= 1.043x^3 - 4.276x^2 + 4.629x\end{aligned}$$

What if we had started with a quartic polynomial? The solution is then

$$\begin{aligned}\tilde{y}(x;a) = 2 &- \frac{1}{4}(x-1) + 3.3725(x-1)(x-3) - 0.8881(x-1)(x^2+x-11) + \\ &+ 0.0864(x-1)(x^3+x^2+x-31) \\ &= 0.0864x^4 - 0.888x^3 + 3.3725x^2 - 5.848x + 5.277\end{aligned} \tag{T3L2-20}$$

and

$$\tilde{\tau}(x;a) = \frac{1}{2} + \frac{1}{4}(x-2) - 6.745x(x-2) + 2.664x(x-2)(x+2) - 0.346x(x-2)(x^2+2x+4)$$
$$= -0.346x^4 + 2.664x^3 - 6.745x^2 + 5.848x$$

(T3L2-20)

We will now plot the three trial functions and the exact solution both for the function and the flux.

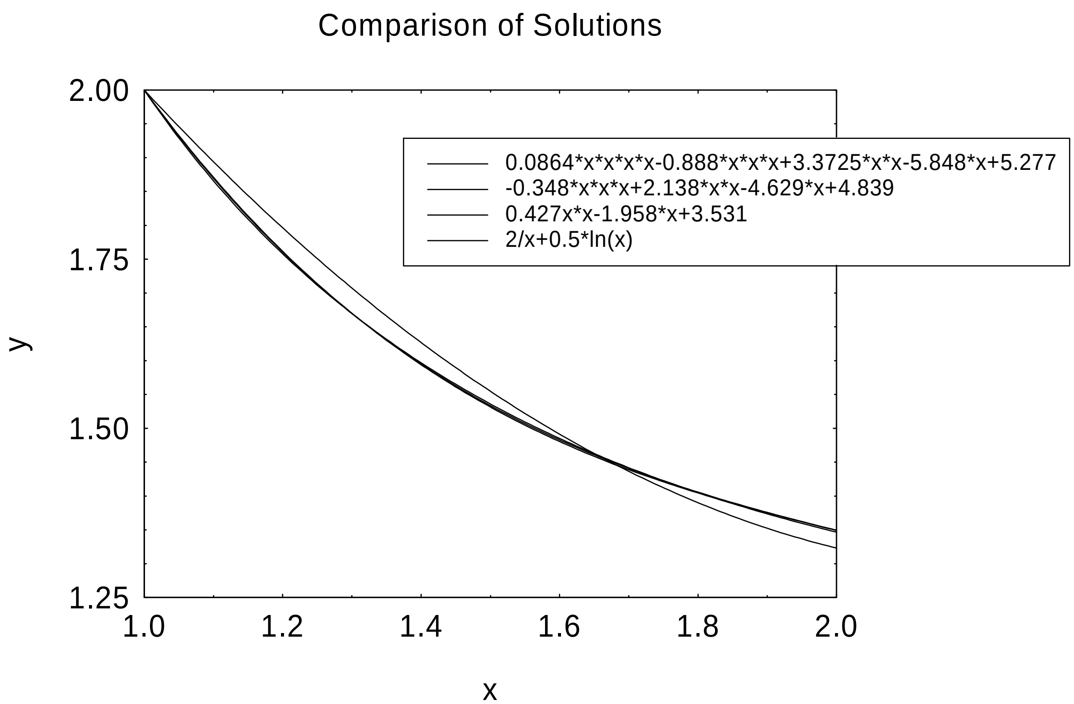


Fig. T3L2-1 Comparison of the function

While some difference can be noted between the quadratic and the exact solution, there is very little difference between the cubic, quartic and the exact solutions. Note that all the solutions have the same function value at the left boundary point $x = 1$.

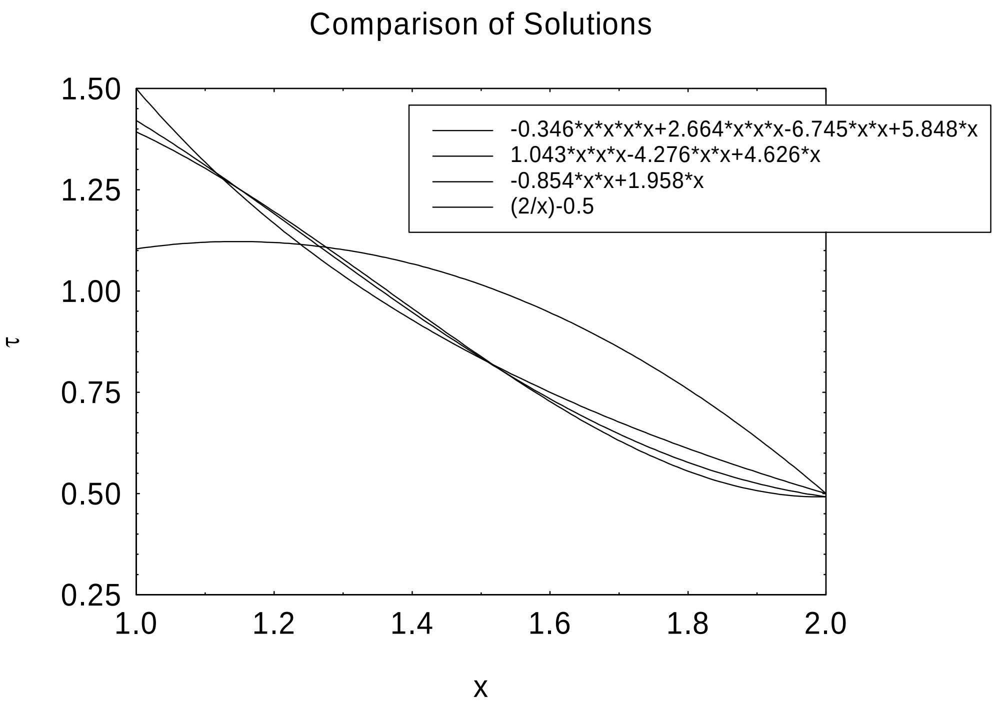


Fig. T3L2-2 Comparison of the flux

The differences between the exact solution and the three Galerkin's solutions with respect to the flux is a different matter altogether. The error in the flux from the quadratic solution is large. However, the error decreases with the cubic and the quartic functions. Note that all the solutions have the same flux value at the right boundary point $x = 2$. The message here is quite clear – (a) small errors in the function do not translate to small errors in the flux that involve the derivatives of the function, and (b) increasing order (polynomials) trial functions yield better and converging solutions[5].

[5] Later we will see that the trial functions must have certain properties for this to be true. If we keep on increasing order of the trial solution will we converge to the exact solution for this problem?

Summary

The residual function is the function corresponding to the original differential equation (with all the terms on the LHS) in which an approximate solution is substituted. It measures how close the approximation is to satisfying the DE but does not tell us how close the approximation is to the exact solution. The Method of Weighted Residuals converts the original DE into a set of algebraic equations that are much easier to solve. There are different approaches used in terms of weighting the residual. Of all the techniques discussed, the Galerkin's Method is by far the most suitable for the type of problems encountered in engineering analysis.

In the lessons in this topic, we saw how to assume the approximate solution called the trial solution and use the Galerkin's Method to generate the algebraic equations. The trial solution is usually a polynomial because of the properties that they possess (continuous, differential, easy to handle etc.). We saw how to enforce the boundary conditions on the trial solution. We also looked at the *flux* term. Finally, we also saw how to obtain more accurate solutions by increasing the order of the polynomial in the trial solution.

There are two problems with this classical (or, theoretical approach). First, the trial solution is valid for the entire problem domain. Hence it is not possible to accurately model problems in which there are known discontinuities in the solution and the flux. Second, the manner in which the boundary conditions are enforced can become cumbersome for more complex problems. In this next topic, we will see how to overcome both these drawbacks.

Review Exercises

Problem T3L2-1

Consider the differential equation

$$\frac{d^2y}{dx^2} + \frac{y}{4} = 0 \qquad\qquad 0 < x < \pi$$

with $y(0) = 1$ and $y(\pi) = 0$. Find the solution using the Galerkin's Method by assuming the trial solution as $y(x;a) = a_1 + a_2 x + a_3 x^2 + a_4 x^3$. Compare with the exact solution.

Problem T3L2-2

Consider the differential equation

$$\frac{d}{dx}\left(x^2 \frac{dy}{dx}\right) = \frac{1}{12}\left(-30x^4 + 204x^3 - 351x^2 + 110x\right) \qquad\qquad 0 < x < 4$$

with $y(0) = 1$ and $y(4) = 0$. Find the solution using the Galerkin's Method by assuming the trial solution as (i) a quadratic polynomial, and (ii) a cubic polynomial. Compare to the exact solution.

Problem T3L2-3

Consider the differential equation

$$\frac{d}{dx}\left((x+1)\frac{dy(x)}{dx}\right) = 0 \qquad\qquad 1 < x < 2$$

with $\qquad y(x=1) = 1 \qquad\qquad \left(-(x+1)\frac{dy}{dx}\right)_{x=2} = 1$

(a) Using the Galerkin's method with a quadratic polynomial for the trial solution obtain an approximate solution for both the function and the flux.
(b) Obtain a second approximate solution using a cubic polynomial for the trial solution.
(c) Compare the two solutions with each other and the exact solution.

Topic 4: 1D Problems

""Mach 2 travel feels no different." a passenger commented on an early Concorde flight. "Yes," Sir George replied. "That was the difficult bit."- Sir George Edwards, co-director of Concorde development Quoted in Kenneth Owen, "Concorde, New Shape in the Sky"

Lesson 1: The Element Concept

Objectives: In this lesson we will look at Galerkin's Method to solve one-dimensional boundary-value problems.

- To understand the 1D Boundary Value Problems (BVP).
- To apply the Galerkin's Method to solving the 1D BVP.

Integration by parts

Evaluate $I = \int f(x)g(x)dx$.

Define $u = f(x)$

$dv = g(x)dx$

Compute $du = f'(x)\,dx$

$v = \int g(x)\,dx$

Then $I = \int u(x)dv = u(x)v(x) - \int v(x)\,du$

Overcoming Drawbacks

In the previous topic we looked at applying the Galerkin's Method in the classical (or, theoretical) sense. To make the approach general and useful we need to deviate to a new format as shown below.

Step 1: Let the trial solution be assumed as (Eqn. (T3L1-4))

$$\tilde{y}(x;a) = \varphi_0(x) + a_1\varphi_1(x) + a_2\varphi_2(x) + ... + a_n\varphi_n(x) = \varphi_0(x) + \sum_{j=1}^{n} a_j\varphi_j(x) \qquad \text{(T4L1-1)}$$

We will stick with the example from the previous topic. Using the definition of the residual, we have

$$R(x;a) = \frac{d}{dx}\left(x \frac{d\,\tilde{y}(x)}{dx} \right) - \frac{2}{x^2} \qquad \text{(T4L1-2)}$$

Since we have n parameter trial function, we need n residual equations (see Eqn. (T3L1-10))

$$\int_{x_a}^{x_b} R(x;a)\phi_i(x)dx = 0 \qquad i = 1,...,n \qquad \text{(T4L1-3)}$$

Notice that we have changed the limits of integration (instead of using 1 and 2 as the limits) just to make the derivation general and more useful. Substituting, we have

$$\int_{x_a}^{x_b} \left[\frac{d}{dx}\left(x \frac{d\,\tilde{y}(x)}{dx} \right) - \frac{2}{x^2} \right] \phi_i(x)dx = 0 \qquad i = 1,...,n \qquad \text{(T4L1-4)}$$

Step 2: Integrate by parts the highest derivative term in the residual equations (which would be first term containing the second-order derivative). Rearranging the terms, we have

$$\int_{x_a}^{x_b} x \frac{d\,\tilde{y}}{dx}\frac{d\phi_i}{dx}dx = -\int_{x_a}^{x_b} \frac{2}{x^2}\phi_i dx - \left[\left(-x\frac{d\,\tilde{y}}{dx} \right) \phi_i \right]_{x_a}^{x_b} \qquad i = 1,...,n \qquad \text{(T4L1-5)}$$

Three observations here – (a) the highest order derivative in Eqn. (T4L1-5) is now lower (only first order derivatives compared to second-order in Eqn. (T4L1-4)), and (b) the "stiffness" term is on the LHS and the "loading" terms are on the RHS, and (c) the loading term contains the boundary flux term.

Step 3: Substituting Eqn. (T4L1-1) in Eqn. (T4L1-5) and noting that

$$\frac{d\,\tilde{y}(x;a)}{dx} = \frac{d\phi_0(x)}{dx} + \sum_{i=1}^{n} a_i \frac{d\phi_i(x)}{dx} \qquad \text{(T4L1-6)}$$

we have

$$\sum_{j=1}^{n}\left(\int_{x_a}^{x_b}\frac{d\phi_i}{dx}x\frac{d\phi_j}{dx}dx\right)a_j=-\int_{x_a}^{x_b}\frac{2}{x^2}\phi_i\,dx-\left[\left(-x\frac{d\tilde{y}}{dx}\right)\phi_i\right]_{x_a}^{z_b}$$

$$-\int_{x_a}^{x_b}\frac{d\phi_i}{dx}x\frac{d\phi_0}{dx}dx \qquad i=1,..,n \tag{T4L1-7}$$

Let

$$K_{ij}=\int_{x_a}^{x_b}\frac{d\phi_i}{dx}x\frac{d\phi_j}{dx}dx$$

and $$F_i=-\int_{x_a}^{x_b}\frac{2}{x^2}\phi_i\,dx-\left[\left(-x\frac{d\tilde{y}}{dx}\right)\phi_i\right]-\int_{x_a}^{x_b}\frac{d\phi_i}{dx}x\frac{d\phi_0}{dx}dx \tag{T4L1-8}$$

Then we can rewrite Eqn. (T4L1-7) in the matrix notation as

$$\begin{bmatrix} K_{11} & K_{12} & . & . & K_{1n} \\ K_{21} & K_{22} & . & . & K_{2n} \\ . & . & . & . & . \\ . & . & . & . & . \\ K_{n1} & K_{n2} & . & . & K_{nn} \end{bmatrix}\begin{Bmatrix} a_1 \\ a_2 \\ . \\ . \\ a_n \end{Bmatrix}=\begin{Bmatrix} F_1 \\ F_2 \\ . \\ . \\ F_n \end{Bmatrix} \tag{T4L1-9}$$

or,

$$\mathbf{K}_{n\times n}\mathbf{a}_{n\times 1}=\mathbf{F}_{n\times 1} \tag{T4L1-10}$$

The above equations are remarkably similar to Eqns. (T2L2-25a) encountered in the Direct Stiffness Method. The stiffness matrix $\mathbf{K}$ is symmetric since

$$K_{ji}=\int_{x_a}^{x_b}\frac{d\phi_j}{dx}x\frac{d\phi_i}{dx}dx=K_{ij} \tag{T4L1-8b}$$

Step 4: Use a specific form of the trial solution

In the previous topic we used a polynomial as the trial solution. Polynomials are easy to deal with and we will stick with that approach here. Let us assume the solution as

$$\tilde{y}=a_1+a_2x+a_3x^2=\sum_{j=1}^{n}a_j\phi_j(x) \tag{T4L1-11}$$

implying that

$$\phi_1(x)=1 \qquad \phi_2(x)=x \qquad \phi_3(x)=x^2 \tag{T4L1-12}$$

The only difference between Eqn. (T4L1-1) and the above equation is the ϕ_0 term. In this modified approach we will be applying the BCs numerically and hence there is no need for the ϕ_0 term. Now we are ready to compute the terms in Eqn. (T4L1-8). Using (T4L1-12)

$$\frac{d\phi_1}{dx}=0 \qquad \frac{d\phi_2}{dx}=1 \qquad \frac{d\phi_3}{dx}=2x \tag{T4L1-13}$$

We will compute a typical term in the stiffness matrix

$$K_{23}=\int_{x_a}^{x_b}(1)(x)(2x)dx=\frac{2}{3}\left(x_b^3-x_a^3\right) \tag{T4L1-14}$$

and a force term (in two parts)

$$F_2^{\text{int}}=-\int_{x_a}^{x_b}\frac{2}{x^2}\,x\,dx=-2\ln\frac{x_b}{x_a} \tag{T4L1-15}$$

and

$$F_2^{bnd}=\left(-x\frac{d\tilde{y}}{dx}\right)_{x_a}x_a-\left(-x\frac{d\tilde{y}}{dx}\right)_{x_b}x_b \tag{T4L1-16}$$

where F_2^{int} is the interior load term and F_2^{bnd} is the boundary load term. Similarly, the flux can be expressed as

$$\tilde{\tau}=-x\frac{d\tilde{y}}{dx}=-a_2x-2a_3x^2 \tag{T4L1-17}$$

Step 5: Generate the algebraic equations as

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{2}\left(x_b^2-x_a^2\right) & \frac{2}{3}\left(x_b^3-x_a^3\right) \\ 0 & \frac{2}{3}\left(x_b^3-x_a^3\right) & \left(x_b^4-x_a^4\right) \end{bmatrix}\begin{Bmatrix} a_1 \\ a_2 \\ a_3 \end{Bmatrix}=\begin{Bmatrix} 2\left(\frac{1}{x_b}-\frac{1}{x_a}\right) \\ -2\ln\frac{x_b}{x_a} \\ -2(x_b-x_a) \end{Bmatrix}+\begin{Bmatrix} \tilde{\tau}|_{x_a}-\tilde{\tau}|_{x_b} \\ \tilde{\tau}|_{x_a}\;x_a-\tilde{\tau}|_{x_b}\;x_b \\ \tilde{\tau}|_{x_a}\;x_a^2-\tilde{\tau}|_{x_b}\;x_b^2 \end{Bmatrix} \tag{T4L1-18}$$

We can now proceed further by substituting the problem data.

Step 6: Substituting the numerical data.

$x_a=1 \qquad\qquad x_b=2$

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{3}{2} & \frac{14}{3} \\ 0 & \frac{14}{3} & 15 \end{bmatrix}\begin{Bmatrix} a_1 \\ a_2 \\ a_3 \end{Bmatrix}=\begin{Bmatrix} -1 \\ -2\ln 2 \\ -2 \end{Bmatrix}+\begin{Bmatrix} \tilde{\tau}|_{x=1}-\tilde{\tau}|_{x=2} \\ \tilde{\tau}|_{x=1}-\tilde{\tau}|_{x=2}\;(2) \\ \tilde{\tau}|_{x=1}-\tilde{\tau}|_{x=2}\;(4) \end{Bmatrix} \tag{T4L1-19}$$

Step 7: Applying the BCs

First we will apply the natural boundary condition $\left(-x\frac{dy}{dx}\right)_{x=2} = \frac{1}{2}$

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{3}{2} & \frac{14}{3} \\ 0 & \frac{14}{3} & 15 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \end{Bmatrix} = \begin{Bmatrix} -1 \\ -2\ln 2 \\ -2 \end{Bmatrix} + \begin{Bmatrix} \tilde{\tau}\,|_{x=1} - \frac{1}{2} \\ \tilde{\tau}\,|_{x=1} - 1 \\ \tilde{\tau}\,|_{x=1} - 2 \end{Bmatrix} \tag{T4L1-20}$$

Since the NBC is applied to the system equations it does not guarantee that the final solution will satisfy the NBC. In the classical approach, we enforced the NBC up front on the trial solution itself so that the final solution did satisfy the NBC.

Applying the EBC $y(x=1)=2$ is different than what we saw in the Direct Stiffness Method (Topic 2). This is because there is no equation directly associated with $y(x)$. Imposing the EBC on the trial solution itself (Eqn. (T4L1-11))

$$a_1 + a_2 + a_3 = 2 \tag{T4L1-21}$$

We can write a_3 in terms of the other two parameters as

$$a_3 = 2 - a_1 - a_2 \tag{T4L1-22}$$

Eliminating a_3 from Eqns. (T4L1-20) using the above equation, we have

$$\begin{bmatrix} 0 & 0 \\ -\frac{14}{3} & \frac{3}{2}-\frac{14}{3} \\ -15 & \frac{14}{3}-15 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} \tilde{\tau}\,|_{x=1} - \frac{3}{2} \\ \tilde{\tau}\,|_{x=1} - 2\ln 2 - \frac{31}{3} \\ \tilde{\tau}\,|_{x=1} - 34 \end{Bmatrix} \tag{T4L1-23}$$

Eliminating the last equation (from above, Eqn. (1)-Eqn. (3), Eqn.(2)-Eqn. (3))

$$\begin{bmatrix} 15 & \frac{31}{3} \\ \frac{31}{3} & \frac{43}{6} \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} \frac{65}{2} \\ \frac{71}{3} - 2\ln 2 \end{Bmatrix} \tag{T4L1-24}$$

The final equations are symmetric, do not contain any boundary terms and can be solved numerically.

Step 8: Solving the system equations

$$a_1 = 3.719 \qquad a_2 = -2.254 \tag{T4L1-25}$$

Substituting in Eqn. (T4L1-22),

$$a_3 = 0.535. \tag{T4L1-26}$$

Hence

$$\tilde{y} = 3.719 - 2.254x + 0.535x^2 \qquad \text{(T4L1-27)}$$

and

$$\tilde{\tau} = 2.254x - 1.070x^2 \qquad \text{(T4L1-28)}$$

While we have overcome some obstacles with this procedure, the process has still a major drawback. We again revisit the issue of imposing the boundary conditions.

The Element Concept (Finally!)

In the preceding section, the trial solution was assumed to be applicable for the entire problem domain. We will depart from this approach and go through the process of discretization (or, creating a mesh with elements). This will enable us to develop a very general methodology making it convenient to do a variety of things including applying the boundary conditions.

One-Element Solution

Step 4: We will once again stick with the same problem as the last section. Let us assume that the domain is discretized into a single element. But just as in the Direct Stiffness Approach, let us now assume a typical element as shown below.

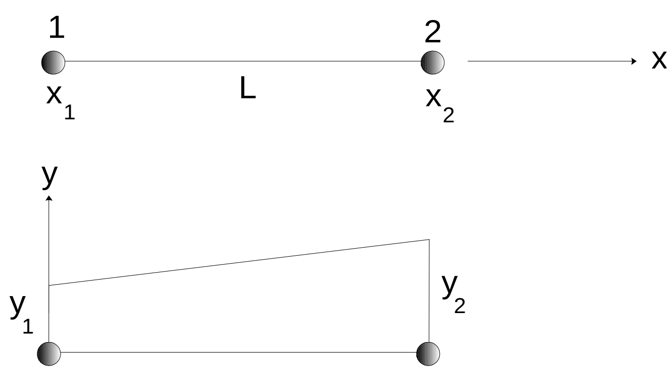


Fig. T4L1-1

The element is described by two nodes that are labeled 1 and 2 and are located at x_1 and x_2 respectively, with the length of the element as L. Let us also assume that the solution varies linearly over the element and is y_1 at node 1 and y_2 at node 2 noting that at this stage these values (called *nodal values*) are unknowns. We can describe the variation of the solution over the element as a linear **interpolation** using the two nodal values. In other words

$$\tilde{y}(x) = a_1 + a_2 x = \phi_1(x) y_1 + \phi_2(x) y_2 \qquad \text{(T4L1-29)}$$

This is indeed the trial solution that we have used as the starting point in the preceding sections. Using the end conditions

$$y(x = x_1) = y_1 \qquad y(x = x_2) = y_2 \qquad \text{(T4L1-30)}$$

and substituting in the first part of Eqn. (T4L1-29) we obtain

$$y_1 = a_1 + a_2 x_1$$

$$y_2 = a_1 + a_2 x_2 \qquad \text{(T4L1-31)}$$

Solving for the a_i's, we have

$$a_1 = \frac{x_2 y_1 - x_1 y_2}{x_2 - x_1} \qquad a_2 = \frac{y_2 - y_1}{x_2 - x_1} \qquad \text{(T4L1-32)}$$

Therefore,

$$\tilde{y}(x) = \frac{x_2 y_1 - x_1 y_2}{x_2 - x_1} + \frac{y_2 - y_1}{x_2 - x_1} x \qquad \text{(T4L1-33)}$$

Rearranging

$$\tilde{y}(x) = \frac{x_2 - x}{x_2 - x_1} y_1 + \frac{x - x_1}{x_2 - x_1} y_2 = \frac{x_2 - x}{L} y_1 + \frac{x - x_1}{L} y_2 \qquad \text{(T4L1-34)}$$

The trial functions

$$\phi_1 = \frac{x_2 - x}{L} \qquad \phi_2 = \frac{x - x_1}{L} \qquad \text{(T4L1-35)}$$

are known as the shape functions. They have special properties as shown below.

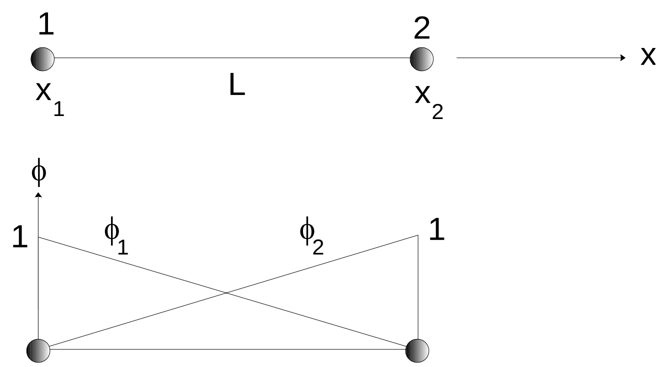


Fig. T4L1-2

Going back to Step 4 in the preceding section, we have no $\phi_0(x)$[6] and the ϕ_i's are given by Eqn. (T4L1-35). Hence for a typical element

$$\begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} \begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} = \begin{Bmatrix} F_1 \\ F_2 \end{Bmatrix} \qquad \text{(T4L1-36)}$$

Let us evaluate couple of typical terms.

$$k_{12} = \int_{x_1}^{x_2} \frac{d\phi_1}{dx} x \frac{d\phi_2}{dx} dx = \int_{x_1}^{x_2} \left(\frac{-1}{L}\right)(x)\left(\frac{1}{L}\right) dx = -\frac{1}{2}\frac{x_1 + x_2}{L} \qquad \text{(T4L1-37)}$$

$$F_1^{\text{int}} = -\int_{x_1}^{x_2} \frac{2}{x^2} \phi_1 \, dx = -\frac{2}{x_1} + \frac{2}{x_2 - x_1} \ln \frac{x_2}{x_1} \qquad \text{(T4L1-38)}$$

---

[6] It is not necessary to have this term since the BCs will be imposed numerically.

$$F_2^{bnd} = -\left(-x \frac{d\tilde{y}}{dx} \right)_{x_b} \qquad \text{(T4L1-39)}$$

Step 5: With all the other terms evaluated similarly,

$$\frac{1}{2L}\begin{bmatrix} (x_1+x_2) & -(x_1+x_2) \\ -(x_1+x_2) & (x_1+x_2) \end{bmatrix}\begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} = \begin{Bmatrix} -\dfrac{2}{x_1}+\dfrac{2}{L}\ln\dfrac{x_2}{x_1} \\ \dfrac{2}{x_2}-\dfrac{2}{L}\ln\dfrac{x_2}{x_1} \end{Bmatrix} + \begin{Bmatrix} \tilde{\tau}\,|_{x=1} \\ -\tilde{\tau}\,|_{x=2} \end{Bmatrix} \qquad \text{(T4L1-40)}$$

These are the element equations similar to Eqn. (T2L2-4) etc. The flux expression is of the form

$$\tilde{\tau} = -x\frac{d\tilde{y}}{dx} = \frac{x}{x_2 - x_1}(y_1 - y_2) \qquad \text{(T4L1-41)}$$

Step 6: Substituting the numerical values

$$x_1 = 1 \qquad\qquad x_2 = 2$$

we have

$$\frac{1}{2}\begin{bmatrix} 3 & -3 \\ -3 & 3 \end{bmatrix}\begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} = \begin{Bmatrix} -2+2\ln 2 \\ 1-2\ln 2 \end{Bmatrix} + \begin{Bmatrix} \tilde{\tau}\,|_{x=1} \\ -\tilde{\tau}\,|_{x=2} \end{Bmatrix} \qquad \text{(T4L1-42)}$$

Step 7: Applying the boundary conditions

First we will apply the natural boundary condition $\tau_{x=2} = \dfrac{1}{2}$. This results in

$$\frac{1}{2}\begin{bmatrix} 3 & -3 \\ -3 & 3 \end{bmatrix}\begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} = \begin{Bmatrix} -2+2\ln 2 \\ 1-2\ln 2 \end{Bmatrix} + \begin{Bmatrix} \tilde{\tau}\,|_{x=1} \\ -\dfrac{1}{2} \end{Bmatrix} \qquad \text{(T4L1-43)}$$

Now applying the EBC $y(x=1) = y_1 = 2$ is done in a manner described in the section containing Eqn. (T2L2-39). The equations reduce to

$$\begin{bmatrix} 1 & 0 \\ 0 & \dfrac{3}{2} \end{bmatrix}\begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} = \begin{Bmatrix} 2 \\ \dfrac{7}{2}-2\ln 2 \end{Bmatrix} \qquad \text{(T4L1-43)}$$

Step 8: Solving

$$y_1 = 2 \qquad\qquad y_2 = 1.409 \qquad \text{(T4L1-44)}$$

Hence, the approximate solution over the element is found by substituting these values in Eqns. (T4L1-34) and (T4L1-41) yielding

$$\tilde{y} = 2.591 - 0.591x \qquad \text{(T4L1-45)}$$

and

$$\tilde{\tau} = 0.591x \qquad \text{(T4L1-46)}$$

Comparing this solution to the exact solution shows that the results are quite in error. This is because of the linear trial solution that was assumed and because of the fact that we used only one element.

Two-Element Solution

The finite element mesh for the two-element solution is shown below.

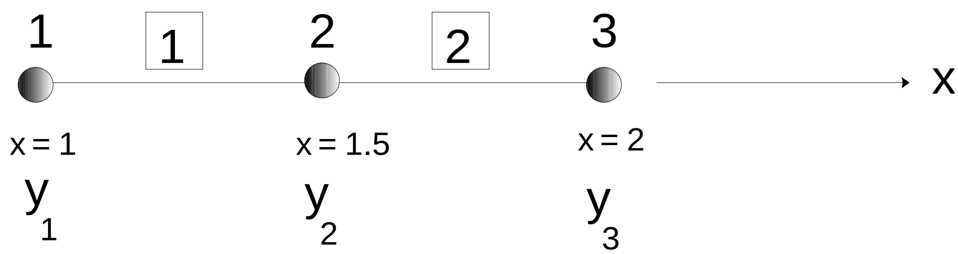


Fig. T4L1-3

This is a uniform mesh since both the elements are geometrically identical. The unknowns at the three nodes are labeled y_1, y_2, y_3. Just as in the Direct Stiffness Method, we will generate the element equations.

Element 1: $x_1 = 1$ and $x_2 = 1.5$

$$\frac{1}{2(0.5)}\begin{bmatrix} (1+1.5) & -(1+1.5) \\ -(1+1.5) & (1+1.5) \end{bmatrix}\begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} = \begin{Bmatrix} -\frac{2}{1}+\frac{2}{0.5}\ln\frac{1.5}{1} \\ \frac{2}{1.5}-\frac{2}{0.5}\ln\frac{1.5}{1} \end{Bmatrix} + \begin{Bmatrix} \left(\tilde{\tau}\,|_{x=1}\right)_1 \\ \left(-\tilde{\tau}\,|_{x=1.5}\right)_1 \end{Bmatrix} \qquad \text{(T4L1-47)}$$

Element 2: $x_1 = 1.5$ and $x_2 = 2$

$$\frac{1}{2(0.5)}\begin{bmatrix} (1.5+2) & -(1.5+2) \\ -(1.5+2) & (1.5+2) \end{bmatrix}\begin{Bmatrix} y_2 \\ y_3 \end{Bmatrix} = \begin{Bmatrix} -\frac{2}{1.5}+\frac{2}{0.5}\ln\frac{2}{1.5} \\ \frac{2}{2}-\frac{2}{0.5}\ln\frac{2}{1.5} \end{Bmatrix} + \begin{Bmatrix} \left(\tilde{\tau}\,|_{x=1.5}\right)_2 \\ \left(-\tilde{\tau}\,|_{x=2}\right)_2 \end{Bmatrix} \qquad \text{(T4L1-48)}$$

We need to expand the notation to differentiate between the two elements, i.e. $\left(\tilde{\tau}\,|_{x=1}\right)_1$ represents the flux at $x = 1$ for element 1. Assembling the two equations to create the system equations, we obtain

$$\begin{bmatrix} 2.5 & -2.5 & 0 \\ -2.5 & 6.0 & -3.5 \\ 0 & -3.5 & 3.5 \end{bmatrix} \begin{Bmatrix} y_1 \\ y_2 \\ y_3 \end{Bmatrix} = \begin{Bmatrix} -2+4\ln\dfrac{3}{2} \\ 4\ln\dfrac{8}{9} \\ 1-4\ln\dfrac{4}{3} \end{Bmatrix} + \begin{Bmatrix} \left(\tilde{\tau}\,|_{x=1}\right)_1 \\ 0 \\ \left(-\tilde{\tau}\,|_{x=2}\right)_2 \end{Bmatrix} \tag{T4L1-49}$$

Note that the second term in the boundary load vector is zero. This is because when we assemble the system equations we assume that

$$\left(\tilde{\tau}\,|_{x=1.5}\right)_1 = \left(\tilde{\tau}\,|_{x=1.5}\right)_2 \tag{T4L1-50}$$

In other words the flux is assumed to be continuous across the two elements. As before, since this constraint is not enforced at the system level (but a mere substitution is made for the flux on the RHS), the final solution will **not** show this flux continuity across the elements[7].

Now we need to impose the boundary conditions. Imposing the NBC first, we have

$$\begin{bmatrix} 2.5 & -2.5 & 0 \\ -2.5 & 6.0 & -3.5 \\ 0 & -3.5 & 3.5 \end{bmatrix} \begin{Bmatrix} y_1 \\ y_2 \\ y_3 \end{Bmatrix} = \begin{Bmatrix} -2+4\ln\dfrac{3}{2} \\ 4\ln\dfrac{8}{9} \\ 1-4\ln\dfrac{4}{3} \end{Bmatrix} + \begin{Bmatrix} \tilde{\tau}\,|_{x=1} \\ 0 \\ -\dfrac{1}{2} \end{Bmatrix} \tag{T4L1-51}$$

Now imposing the EBC yields

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 6.0 & -3.5 \\ 0 & -3.5 & 3.5 \end{bmatrix} \begin{Bmatrix} y_1 \\ y_2 \\ y_3 \end{Bmatrix} = \begin{Bmatrix} 2 \\ 4\ln\dfrac{8}{9}+5 \\ \dfrac{1}{2}-4\ln\dfrac{4}{3} \end{Bmatrix} \tag{T4L1-52}$$

Solving the equations yields

$$y_1 = 2 \qquad y_2 = 1.551 \qquad y_3 = 1.365 \tag{T4L1-53}$$

Hence, for element 1 (using Eqns. (T4L1-34) and (T4L1-41))

$$\left(\tilde{y}(x)\right)_1 = 2\left(\frac{1.5-x}{0.5}\right) + 1.551\left(\frac{x-1}{0.5}\right) = -0.898x + 2.898 \tag{T4L1-54}$$

$$\left(\tilde{\tau}(x)\right)_1 = 0.898x \tag{T4L1-55}$$

[7] This is similar to the imposition of the natural boundary condition at the system level.

and for element 2

$$\left(\tilde{y}(x)\right)_2 = 1.551\left(\frac{2.0-x}{0.5}\right) + 1.365\left(\frac{x-1.5}{0.5}\right) = -0.372x + 2.109 \qquad \text{(T4L1-54)}$$

$$\left(\tilde{\tau}(x)\right)_2 = 0.372x \qquad \text{(T4L1-55)}$$

Let's compare the one-element and the two-element solutions.

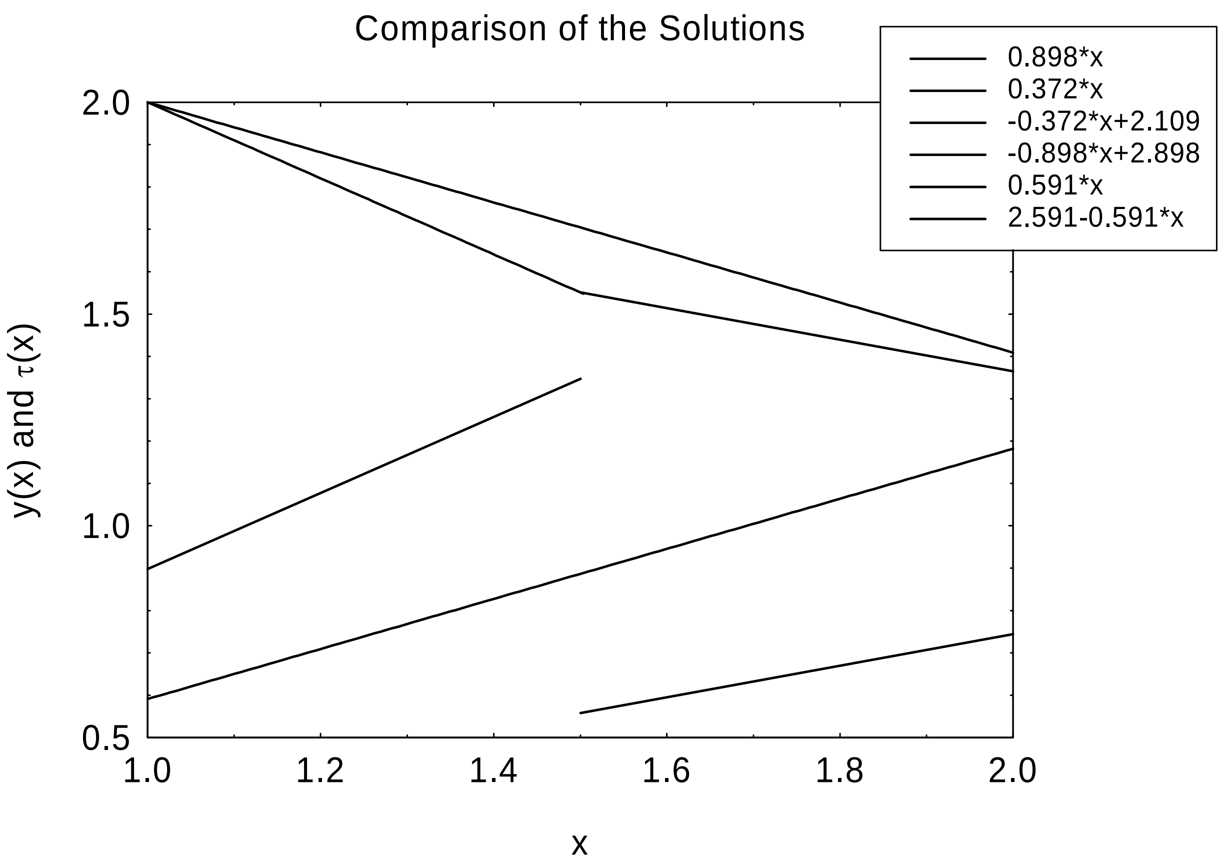


Fig. T4L1-4

(1) Both the solutions satisfy the EBC at $x=1$. They should, as we have imposed the EBC on the final system equations.

(2) There is an improvement in the two-element solution with respect to $y(x)$ (An error of 4.6% and 1.3% for the one-element and two-element solutions, respectively at $x=2$).

(3) The error in the one-element flux is large throughout the domain. The error in the flux from the two-element solution is much less (An error of 136% and 49% for the one-element and two-element solutions, respectively at $x=2$).

(4) While the function is continuous throughout the domain, the flux is discontinuous for the two-element model at the element interface at $x=1.5$.

Review of the Steps

Step 1: Assume the trial solution of the form

$$\tilde{y}(x;a) = a_0 + a_1 x + ... + a_n x^n = \sum_{i=1}^{n} y_i \phi_i(x)$$

where n is equal to the number of DOF in the element, y_i are the nodal values and $\phi_i(x)$ are the shape functions. Construct the residual and substitute the trial solution in the residual equations. Note that there are as many residual equations as there are DOF in the element.

Step 2: Integrate by parts the highest derivative term. Integrating by parts will not only lower the highest derivative term but also generate a boundary (flux) term.

Step 3: Rewrite the equations so that the stiffness related terms are on the left and the load terms (interior and boundary) are on the right.

Step 4: To generate the equations completely we need to assume the exact form of the trial solution, i.e. fix the value of n. This will enable us to generate the terms in the stiffness matrix and the load vector. We can now generate the element equations in the form

$$\mathbf{k}_{n\times n}\mathbf{u}_{n\times 1} = \mathbf{f}_{n\times 1}$$

We also need to generate the expression for the flux.

Step 5: Using the problem data, we can generate the element equations for all the elements in the model. These equations are then assembled into the system or global equations of the form

$$\mathbf{K}_{N\times N}\mathbf{U}_{N\times 1} = \mathbf{F}_{N\times 1}$$

for a model with N system DOF.

Step 6: Using the problem data, we impose the NBCs first. Then we impose the EBCs resulting in equations of the form

$$\mathbf{K}_{N\times N}\mathbf{U}_{N\times 1} = \mathbf{F}_{N\times 1}$$

Step 7: Solve the system equations for the primary nodal unknowns $\mathbf{U}$.

Step 8: Using the primary unknowns we can compute the flux in each element.

Review Exercises

In the following problems the quadratic element is to be used. To obtain the shape functions for the quadratic element refer to Lesson 4.

Problem T4L1-1

Consider the differential equation

$$\frac{d}{dx}\left((x+1)\frac{dy(x)}{dx}\right)=0 \qquad 1<x<2$$

with $y(x=1)=1$ $\qquad \left(-(x+1)\frac{dy}{dx}\right)_{x=2}=1$

(a) Using the element concept, use the linear polynomial as the trial solution. Apply the steps outlined in this section to obtain an approximate solution for both the function and the flux.
(b) Repeat the problem but now use a quadratic polynomial. Compare the two solutions.
(c) Are these solutions different than those obtained in Problem T3L2-3?

Problem T4L1-2

Consider the differential equation

$$\frac{d}{dx}\left(x^2\frac{dy}{dx}\right)=\frac{1}{12}\left(-30x^4+204x^3-351x^2+110x\right) \qquad 0<x<4$$

with $y(0)=1$ and $y(4)=0$. Using the element concept, use a quadratic polynomial as the trial solution. Obtain an approximate solution for both the function and the flux. Compare this to the solution obtained for Problem T3L2-2. *(Hint: While a linear element goes with a linear polynomial and requires an element with 2 nodes, a quadratic element requires an element with 3 nodes. See Lesson 4 for details.)*

One-Dimensional Boundary Value Problem

The one-dimensional BVP is described by (with the positive x direction left to right)

DE: $$-\frac{d}{dx}\left(\alpha(x)\frac{dy(x)}{dx}\right)+\beta(x)y(x)=f(x) \qquad x_a < x < x_b \qquad \text{(T4L1-56)}$$

BCs: At $x=x_a$ either $y=y_a$ or $\tau=c_a y+d_a$

At $x=x_b$ either $y=y_b$ or $\tau=c_b y+d_b$ (T4L1-57)

where $\alpha(x)$, $\beta(x)$, and $f(x)$ are known functions, y_a, y_b, c_a, c_b, d_a and d_b are constants, and $\tau=-\alpha\dfrac{dy}{dx}$ is the flux. The BCs are EBC, NBC or mixed. Note that the mixed BC reduces to a NBC by setting $c_a=0$ and $c_b=0$. We will look at specific engineering problems that are governed by this differential equation in the next lesson.

We will use the Galerkin's Method to generate the element equations. The following steps pertain to a typical element in the mesh.

Step 1: Assume the trial solution as

$$\tilde{y}(x;a)=\sum_{j=1}^{n} y_j\phi_j(x) \qquad \text{(T4L1-58)}$$

As in the previous section, we do not have the $\phi_0(x)$ term since the boundary conditions will be imposed numerically. We will drop the ~(tilde) notation for convenience sake. Substituting in the residual equations and integrating over the domain Ω of the element, we have

$$\int_\Omega \left[-\frac{d}{dx}\left(\alpha(x)\frac{dy(x)}{dx}\right)+\beta(x)y(x)-f(x)\right]\phi_i(x)\,dx=0 \qquad i=1,2,..,n \qquad \text{(T4L1-59)}$$

Step 2: Integrating by parts the highest-order derivative, we have

$$\int_\Omega \left[\alpha(x)\frac{dy}{dx}\frac{d\phi_i}{dx}+\beta(x)y(x)\phi_i\right]dx=\int_\Omega f(x)\phi_i dx-\left[\tau\,\phi_i\right]^\Gamma \qquad i=1,2,..,n \qquad \text{(T4L1-60)}$$

where Γ represents the boundary of the element.

Step 3: Using Eqn. (T4L1-58) in (T4L1-60)

$$\sum_{j=1}^{n}\left[\int_\Omega \frac{d\phi_i}{dx}\alpha(x)\frac{d\phi_j}{dx}dx+\int_\Omega \phi_i(x)\beta(x)\phi_j(x)\,dx\right]y_j=$$

$$\int_\Omega f(x)\phi_i(x)\,dx-\left[\tau\phi_i\right]^\Gamma \qquad i=1,2.,,n \qquad \text{(T4L1-61)}$$

Step 4: Let us use the linear interpolation two-node element from the earlier section $(n = 2)$

$$\phi_1 = \frac{x_2 - x}{L} \qquad \phi_2 = \frac{x - x_1}{L} \tag{T4L1-62}$$

Substituting Eqn. (T4L1-62) in Eqn. (T4L1-61), we have the element equations as

$$\begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} \begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} = \begin{Bmatrix} F_1^{\text{int}} \\ F_2^{\text{int}} \end{Bmatrix} + \begin{Bmatrix} F_1^{bnd} \\ F_2^{bnd} \end{Bmatrix} \tag{T4L1-63}$$

Let us look at a typical stiffness term first.

$$k_{11} = \int_{x_1}^{x_2} \left(-\frac{1}{x_2 - x_1} \right) \alpha(x) \left(-\frac{1}{x_2 - x_1} \right) dx + \int_{x_1}^{x_2} \left(\frac{x_2 - x}{x_2 - x_1} \right) \beta(x) \left(\frac{x_2 - x}{x_2 - x_1} \right) dx \tag{T4L1-64}$$

To evaluate the above equation we need to know $\alpha(x)$ and $\beta(x)$. The functions can then be substituted and the integral evaluated. Instead, we will assume that we know the numerical value of these two functions at the centroid of the element, i.e. at $x = \frac{x_1 + x_2}{2}$. For this element in which the solution is assumed to vary linearly, this is the most accurate point to evaluate the integral. Let us denote the centroidal values (constants) as $\bar{\alpha}$ and $\bar{\beta}$. Now, integrating

$$k_{11} = \frac{\bar{\alpha}}{L} + \frac{\bar{\beta} L}{3} \tag{T4L1-65}$$

We can use a similar strategy with the load terms. When all the terms are evaluated we have,

$$\left[\begin{array}{c|c} \frac{\bar{\alpha}}{L} + \frac{\bar{\beta} L}{3} & -\frac{\bar{\alpha}}{L} + \frac{\bar{\beta} L}{6} \\ \hline -\frac{\bar{\alpha}}{L} + \frac{\bar{\beta} L}{6} & \frac{\bar{\alpha}}{L} + \frac{\bar{\beta} L}{3} \end{array} \right] \left\{ \begin{array}{c} y_1 \\ \hline y_2 \end{array} \right\} = \left\{ \begin{array}{c} \frac{\bar{f} L}{2} \\ \hline \frac{\bar{f} L}{2} \end{array} \right\} - \left\{ \begin{array}{c} \left[\tau \phi_1 \right]^{\Gamma} \\ \hline \left[\tau \phi_2 \right]^{\Gamma} \end{array} \right\} \tag{T4L1-66}$$

The term that requires special treatment here is the last term. In general

$$\left[\tau \phi_i \right]^{\Gamma} = \left[\tau \phi_i \right]_{x_2} - \left[\tau \phi_i \right]_{x_1} \tag{T4L1-67}$$

From Eqn. (T4L1-57), we have

$$\tau = (cy + d) \text{ for } x = x_a \text{ and } \tau = (cy + d) \text{ for } x = x_b .$$

Substituting this in Eqn. (T4L1-67) for $i = 1$

$$\left[\tau\varphi_1\right]^{\Gamma} = \left[\tau\varphi_1\right]_{x_2} - \left[\tau\varphi_1\right]_{x_1} = 0 - \left[\tau\right]_{x_1} = -\left(cy+d\right)_{x_1} = -c_1 y_1 - d_1 \quad \text{(T4L1-68)}$$

Similarly, for $i = 2$

$$\left[\tau\varphi_2\right]^{\Gamma} = \left[\tau\varphi_2\right]_{x_2} - \left[\tau\varphi_2\right]_{x_1} = \left[\tau\right]_{x_2} - 0 = \left(cy+d\right)_{x_2} = c_2 y_2 + d_2 \quad \text{(T4L1-69)}$$

Note that the unknown y appears in the two equations and must be moved to the LHS. Therefore the modified element equations are

$$\left[\begin{bmatrix} \frac{\bar{\alpha}}{L} + \frac{\bar{\beta}L}{3} & -\frac{\bar{\alpha}}{L} + \frac{\bar{\beta}L}{6} \\ -\frac{\bar{\alpha}}{L} + \frac{\bar{\beta}L}{6} & \frac{\bar{\alpha}}{L} + \frac{\bar{\beta}L}{3} \end{bmatrix} - c_1 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + c_2 \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\right] \begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} =$$

$$\begin{Bmatrix} \frac{\bar{f}L}{2} \\ \frac{\bar{f}L}{2} \end{Bmatrix} + \begin{Bmatrix} d_1 \\ -d_2 \end{Bmatrix} \quad \text{(T4L1-70)}$$

The c_1 term exists provided $x_1 = x_a$. Similarly, c_2 term exists provided $x_2 = x_b$. The flux at the center of the element is given by

$$\tau = -\alpha \frac{dy}{dx} = -\frac{\bar{\alpha}}{L}\left(y_2 - y_1\right) \quad \text{(T4L1-71)}$$

The element equations are ready and we now need to look at engineering problems that are described as 1D BVP to illustrate these steps and the rest of the solution.

Finally a warning – boundary conditions can be dangerous to your health! Applying the BCs incorrectly is one of the most common forms of errors.

Lesson 2: Solid Mechanics Examples

Objectives: In this lesson we will look at Galerkin's Method to solve one-dimensional solid mechanics problems.

- To understand the relationship between some of the solid mechanics problems and the general 1D BVP discussed in the previous lesson.
- To solve an example so as to understand the process.

Axial Deformation of an Elastic Rod

Consider the elastic rod as shown in figure below. The axial displacement is denoted $u = u(x)$ and the axial loading (body force) on the rod is $w = w(x)$. A sample set of boundary conditions is shown where $u = 0$ (EBC) on the left end and the force F (NBC) on the right end is zero.

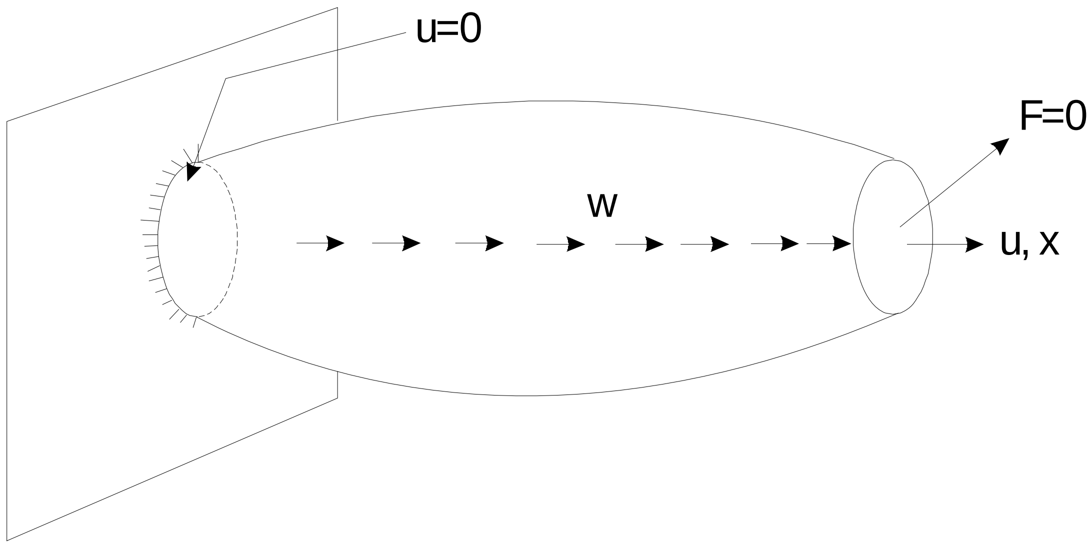


Fig. T4L2-1

The governing differential equation is given as

$$-\frac{d}{dx}\left(A(x)E(x)\frac{du(x)}{dx} \right) = w(x)A(x) \qquad \text{(T4L2-1)}$$

with the boundary conditions as

At $x = x_a$, $u(x = x_a) = u_a$ or NBC (T4L2-2)

At $x = x_b$, $u(x = x_b) = u_b$ or NBC (T4L2-3)

The NBC is of the form

$$\overline{X} = n_x \sigma_x \qquad \text{(T4L2-3a)}$$

where n_x is the direction cosine of the outward normal, $\overline{X}$ is the force per unit area and is positive if it acts in the positive x direction. Let us look at some possibilities with respect to the boundary conditions.

Rod has a known displacement u_a *(incl. zero) at the left end*

$$u = u_a$$

There is a concentrated force F_a *applied at the left end in the positive x direction* $(n_x = -1)$

$$F_a = -AE\left.\frac{du}{dx}\right|_{x=x_a} \quad \text{or} \quad \sigma_a = -E\left.\frac{du}{dx}\right|_{x=x_a}$$

There is a concentrated force F_b *applied at the right end in the positive x direction* $(n_x = 1)$

$$F_b = AE\left.\frac{du}{dx}\right|_{x=x_b} \quad \text{or} \quad \sigma_b = E\left.\frac{du}{dx}\right|_{x=x_b}$$

Using the process discussed in the previous lesson, Eqn. (T4L1-70) reduces to

$$\frac{\overline{AE}}{L}\left[\begin{array}{c|c} 1 & -1 \\ \hline -1 & 1 \end{array}\right]\left\{\begin{array}{c} u_1 \\ \hline u_2 \end{array}\right\} = \left\{\begin{array}{c} \dfrac{\overline{wAL}}{2} \\ \hline \dfrac{\overline{wAL}}{2} \end{array}\right\} + \left\{\begin{array}{c} F_1 \\ \hline F_2 \end{array}\right\} \qquad \text{(T4L2-4)}$$

$$\mathbf{k}_{2\times 2}\mathbf{u}_{2\times 1} = \mathbf{f}_{2\times 1} \qquad \text{(T4L2-5)}$$

A dimensional analysis will show that - $\overline{A} \equiv \mathrm{L}^2$, $\overline{E} \equiv \dfrac{\mathrm{F}}{\mathrm{L}^2}$, $u \equiv \mathrm{L}$ and $\overline{wA} \equiv \dfrac{\mathrm{F}}{\mathrm{L}}$. The LHS and the RHS have units of F. While we will look formally at systems modeled with discrete structural elements (truss and frame) in the second module, it should be noted that the stiffness matrix (in the local coordinate system aligned with the axis of the element) is the stiffness matrix for a truss element provided

(a) the end of the element are pin connections,

(b) the cross-sectional area and the modulus of elasticity are constant within the element, and

(c) $f(x) = 0$ since no element loads are allowed on a truss member.

An Illustrative Example T4L2-1

Figure shows a bar made of a material with modulus of elasticity of 200 GPa. Compute the nodal displacements, element stresses and support reactions.

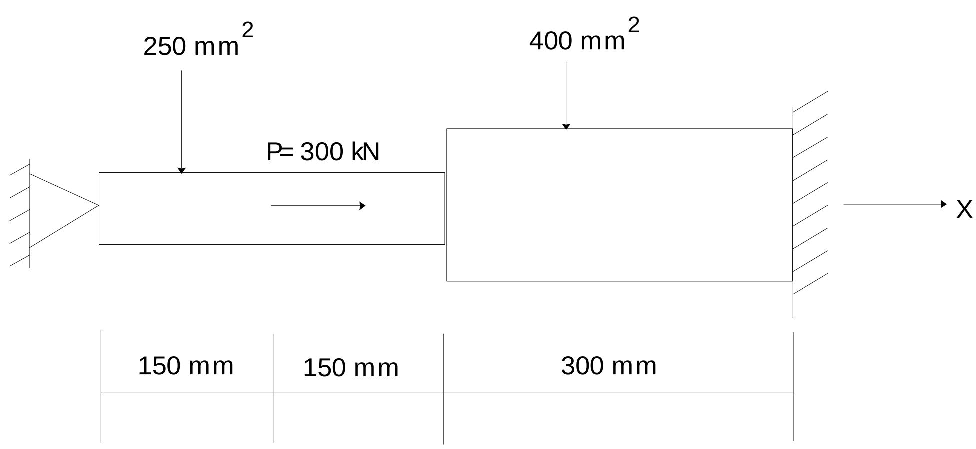


Fig. T4L2-2a

Solution: Let us use m and N as the problem units. Let us use a three-element model placing a node where the concentrated force acts[8]. The FE model is shown below. We have EBCs at the left and the right ends.

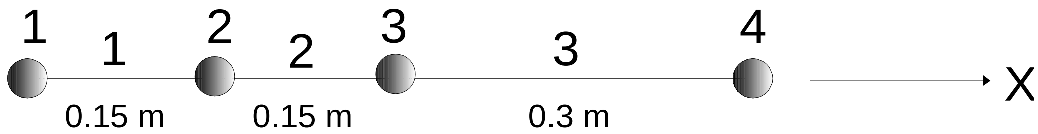


Fig. T4L2-2b

Element 1: $E = 200(10^9)\dfrac{N}{m^2}$, $A = 250(10^{-6})\,m^2$, $L = 0.15m$, $w = 0$. Hence the element equations are

$$\begin{bmatrix} 3.333(10^8) & -3.333(10^8) \\ -3.333(10^8) & 3.333(10^8) \end{bmatrix} \begin{Bmatrix} U_1 \\ U_2 \end{Bmatrix} = \begin{Bmatrix} F_1^1 \\ -F_2^1 \end{Bmatrix}$$

Element 2: $E = 200(10^9)\dfrac{N}{m^2}$, $A = 250(10^{-6})\,m^2$, $L = 0.15m$, $w = 0$. Hence the element equations are

$$\begin{bmatrix} 3.333(10^8) & -3.333(10^8) \\ -3.333(10^8) & 3.333(10^8) \end{bmatrix} \begin{Bmatrix} U_2 \\ U_3 \end{Bmatrix} = \begin{Bmatrix} F_1^2 \\ -F_2^2 \end{Bmatrix}$$

Element 3: $E = 200(10^9)\dfrac{N}{m^2}$, $A = 400(10^{-6})\,m^2$, $L = 0.30m$, $w = 0$. Hence the element equations are

$$\begin{bmatrix} 2.667(10^8) & -2.667(10^8) \\ -2.667(10^8) & 2.667(10^8) \end{bmatrix} \begin{Bmatrix} U_3 \\ U_4 \end{Bmatrix} = \begin{Bmatrix} F_1^3 \\ -F_2^3 \end{Bmatrix}$$

Assembling the equations, we have

---

[8] This is not necessary since we can use a Dirac Delta function $w(x)$ to define the concentrated force and then compute the equivalent forces acting at the nodes of the element.

$$10^8 \begin{bmatrix} 3.333 & -3.333 & 0 & 0 \\ -3.333 & 6.667 & -3.333 & 0 \\ 0 & -3.333 & 6 & -2.667 \\ 0 & 0 & -2.667 & 2.667 \end{bmatrix} \begin{Bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{Bmatrix} = \begin{Bmatrix} F_1^1 \\ 300(10^3) \\ 0 \\ F_2^3 \end{Bmatrix}$$

Imposing the boundary conditions, we have

$$10^8 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 6.667 & -3.333 & 0 \\ 0 & -3.333 & 6 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{Bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 300(10^3) \\ 0 \\ 0 \end{Bmatrix}$$

Solving, the nodal displacements are

$$\{U_1, U_2, U_3, U_4\} = \{0, 6.23, 3.46, 0\}(10^{-4})\, m$$

The force in each element is computed as follows. Note that flux is negative of the member force.

Element 1

$$F_1 = \frac{A_1 E_1}{L_1}(U_2 - U_1) = \frac{5(10^7)}{0.15}(6.23)(10^{-4}) = 2.077(10^5)N \quad \text{(Tension)}$$

This should also be equal and opposite to the support reaction at the left end, i.e. $R_{left} = 2.077(10^5)\, N \leftarrow$.

Element 2

$$F_2 = \frac{A_2 E_2}{L_2}(U_3 - U_2) = \frac{5(10^7)}{0.15}(3.46 - 6.23)(10^{-4}) = -9.2333(10^4)N \quad \text{(Compression)}$$

Element 3

$$F_3 = \frac{A_3 E_3}{L_3}(U_4 - U_3) = \frac{8(10^7)}{0.30}(0 - 3.46)(10^{-4}) = -9.2333(10^4)N \text{ (Compression)}$$

This should also be equal and opposite to the support reaction at the right end, i.e. $R_{right} = 9.2333(10^4)\,N \leftarrow$.

We can now examine the results. The displacements satisfy the EBC at the left and right ends. The force equilibrium is satisfied since the sum of the two support reactions is equal to the applied force. Also the forces in elements 2 and 3 are equal.

Thermal Loads Example T4L2-2

Consider a bar of length L, constant cross-section A, modulus of elasticity E and coefficient of thermal expansion, α. Let the temperature change in the bar be ΔT. The initial strain is given by

$$\varepsilon_0 = \alpha \Delta T \qquad \text{(ET4L2-2-1)}$$

The equivalent nodal loads due to the temperature change in the element for the 1D-C^0 linear element is given by

$$\mathbf{f}_{2\times1} = \begin{Bmatrix} f_1 \\ f_2 \end{Bmatrix} = AE\alpha \Delta T \begin{Bmatrix} -1 \\ 1 \end{Bmatrix} \qquad \text{(ET4L2-2-2)}$$

The resulting load vector is shown in Fig. ET4L2-2-1.

Fig. ET4L2-2-1

Example (a)
Fig. ET4L2-2-2 shows a bar made of steel. The length of the bar is 2 m. The cross-sectional area is $0.001m^2$. The bar is initially at $25^\circ C$. The temperature is increased to $100^\circ C$. Compute the stress in the bar.

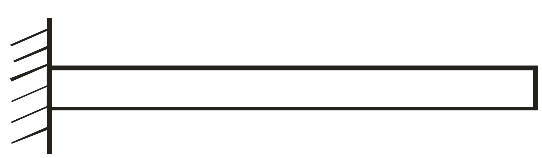


Fig. ET4L2-2-2

Solution: We will solve this problem using a one element mesh with the linear element. The element equations are as follows.

$$\frac{(0.001)\left(200\times10^{9}\right)}{2}\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}\begin{Bmatrix}D_1\\ D_2\end{Bmatrix}=(0.001)\left(200\times10^{9}\right)\left(11.7\times10^{-6}\right)(75)\begin{Bmatrix}-1\\ 1\end{Bmatrix}$$

or
$$\begin{bmatrix}10^{8} & -10^{8}\\ -10^{8} & 10^{8}\end{bmatrix}\begin{Bmatrix}D_1\\ D_2\end{Bmatrix}=\begin{Bmatrix}-175500\\ 175500\end{Bmatrix}$$

Since these are the system equations, applying the essential BC at the left end, we have

$$10^{8}D_2=175500\Rightarrow D_2=0.001755\,\text{m}$$

This indicates that the bar is expanding, as it should! Now

$$\varepsilon=\frac{D_2-D_1}{L}=\frac{0.001755}{2}=0.0008775$$
$$\varepsilon_0=\left(11.7\times10^{-6}\right)(75)=0.0008775$$
$$\sigma=E\left(\varepsilon-\varepsilon_0\right)=200\times10^{9}\left(0.0008775-0.0008775\right)=0$$

Since the bar is unstrained (free to expand), there is no stress in the bar.

Example (b)

Fig. ET4L2-2-3 shows a bar made of steel. The length of the bar is 2 m. The cross-sectional area is $0.001m^2$. The bar is initially at $25^\circ C$. The temperature is increased to $100^\circ C$. Compute the stress in the bar.

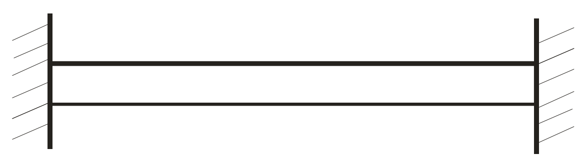


Fig. ET4L2-2-3

Solution: We will solve this problem using a two element mesh with the linear elements. The element equations are as follows.

Element 1

$$\frac{(0.001)\left(200\times10^{9}\right)}{1}\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}\begin{Bmatrix}D_1\\ D_2\end{Bmatrix}=(0.001)\left(200\times10^{9}\right)\left(11.7\times10^{-6}\right)(75)\begin{Bmatrix}-1\\ 1\end{Bmatrix}$$

or
$$\begin{bmatrix}(2)10^{8} & -(2)10^{8}\\ -(2)10^{8} & (2)10^{8}\end{bmatrix}\begin{Bmatrix}D_1\\ D_2\end{Bmatrix}=\begin{Bmatrix}-175500\\ 175500\end{Bmatrix}$$

Element 2

$$\frac{(0.001)(200\times10^9)}{1}\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}\begin{Bmatrix}D_2\\ D_3\end{Bmatrix}=(0.001)(200\times10^9)(11.7\times10^{-6})(75)\begin{Bmatrix}-1\\ 1\end{Bmatrix}$$

or $$\begin{bmatrix}(2)10^8 & -(2)10^8\\ -(2)10^8 & (2)10^8\end{bmatrix}\begin{Bmatrix}D_2\\ D_3\end{Bmatrix}=\begin{Bmatrix}-175500\\ 175500\end{Bmatrix}$$

Assembling the equations, applying the essential BC at the left and right ends, we have

$$(4)10^8 D_2 = 0 \Rightarrow D_2 = 0$$

This indicates that the bar is neither expanding nor contracting, as it should!

Element 1

$$\varepsilon=\frac{D_2-D_1}{L}=0$$
$$\varepsilon_0=(11.7\times10^{-6})(75)=0.0008775$$
$$\sigma=E(\varepsilon-\varepsilon_0)=200\times10^9(0-0.0008775)=-1.755\times10^8\,Pa\Rightarrow\sigma=175MPa\,(C)$$
$$F=\sigma A=(-1.755\times10^8)(0.001)=-175500\,N$$

Element 2

$$\varepsilon=\frac{D_3-D_2}{L}=0$$
$$\varepsilon_0=(11.7\times10^{-6})(75)=0.0008775$$
$$\sigma=E(\varepsilon-\varepsilon_0)=200\times10^9(0-0.0008775)=-1.755\times10^8\,Pa\Rightarrow\sigma=175MPa\,(C)$$
$$F=\sigma A=(-1.755\times10^8)(0.001)=-175500\,N$$

Since the bar is completely restrained (prevented from expanding), the bar is in compression as illustrated by the stress in both bars. The FBDs of the elements and node 2 are shown in Fig. ET4L2-2-4.

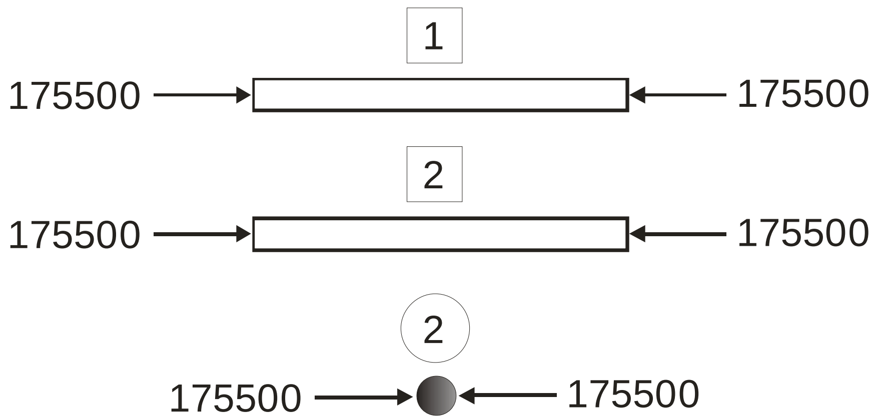


Fig. ET4L2-2-4 Element and nodal FBDs

Example (c)
Fig. ET4L2-2-5 shows a bar made of steel and aluminum of equal lengths. The total length of the bar is 2 m. The cross-sectional area is $0.001 m^2$. The entire bar is initially at $25^\circ C$. The temperature is uniformly increased to $100^\circ C$. Compute the stress distribution in the bar.

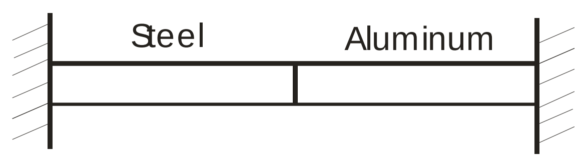


Fig. ET4L2-2-5

Solution: We will solve this problem using a two element mesh with the linear elements. The element equations are as follows.

Element 1 (Steel)

$$\frac{(0.001)\left(200\times10^9\right)}{1}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{Bmatrix} D_1 \\ D_2 \end{Bmatrix} = (0.001)\left(200\times10^9\right)\left(11.7\times10^{-6}\right)(75)\begin{Bmatrix} -1 \\ 1 \end{Bmatrix}$$

or
$$\begin{bmatrix} (2)10^8 & -(2)10^8 \\ -(2)10^8 & (2)10^8 \end{bmatrix}\begin{Bmatrix} D_1 \\ D_2 \end{Bmatrix} = \begin{Bmatrix} -175500 \\ 175500 \end{Bmatrix}$$

Element 2 (Aluminum)

$$\frac{(0.001)\left(70\times10^9\right)}{1}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{Bmatrix} D_2 \\ D_3 \end{Bmatrix} = (0.001)\left(70\times10^9\right)\left(23.0\times10^{-6}\right)(75)\begin{Bmatrix} -1 \\ 1 \end{Bmatrix}$$

or
$$\begin{bmatrix} (0.7)10^8 & -(0.7)10^8 \\ -(0.7)10^8 & (0.7)10^8 \end{bmatrix}\begin{Bmatrix} D_2 \\ D_3 \end{Bmatrix} = \begin{Bmatrix} -120750 \\ 120750 \end{Bmatrix}$$

Assembling the equations, applying the essential BC at the left and right ends, we have

$$(2.7)10^8 D_2 = 175500 - 120750 = 54750 \Rightarrow D_2 = 0.000202778\,m$$

This indicates that the steel part of the bar is expanding and the aluminum part of the bar is contracting.

Element 1 (Steel)

$$\varepsilon = \frac{D_2 - D_1}{L} = 0.000202778$$

$$\varepsilon_0 = \left(11.7\times10^{-6}\right)\left(75\right) = 0.0008775$$

$$\sigma = E\left(\varepsilon - \varepsilon_0\right) = 200\times10^9\left(0.000202778 - 0.0008775\right) = -1.349\times10^8\,Pa \Rightarrow \sigma = 135MPa\,(C)$$

$$F = \sigma A = \left(-1.349\times10^8\right)\left(0.001\right) = -134900\,N$$

Element 2 (Aluminum)

$$\varepsilon = \frac{D_3 - D_2}{L} = -0.000202778$$

$$\varepsilon_0 = \left(23.0\times10^{-6}\right)\left(75\right) = 0.001725$$

$$\sigma = E\left(\varepsilon - \varepsilon_0\right) = 70\times10^9\left(-0.000202778 - 0.001725\right) = -1.349\times10^8\,Pa \Rightarrow \sigma = 135MPa\,(C)$$

$$F = \sigma A = \left(-1.349\times10^8\right)\left(0.001\right) = -134900\,N$$

Since the bar is completely restrained (prevented from expanding), the bar is in compression as illustrated by the stress in both bars. The FBDs of the elements and node 2 are shown in Fig. ET4L2-2-6.

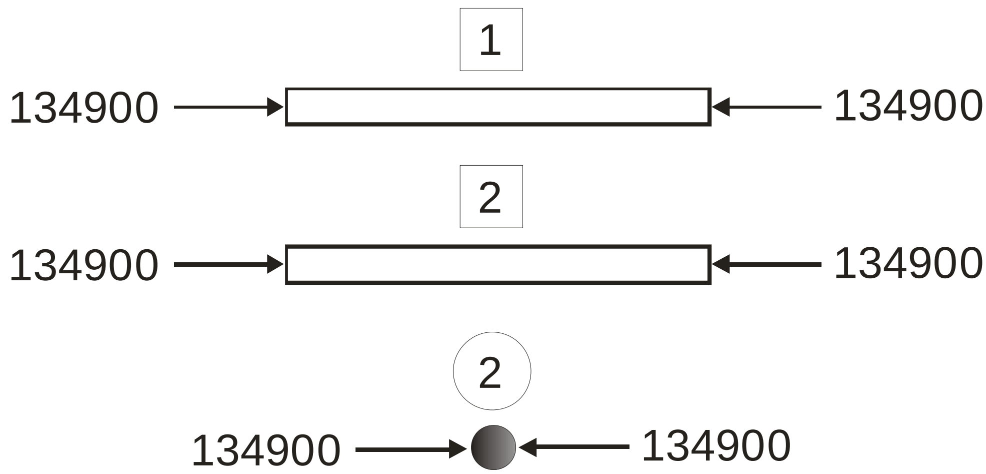


Fig. ET4L2-2-6 Element and nodal FBDs

Note that the support reaction at the left end is $134900N\,(\rightarrow)$ and the support reaction at the right end is $134900N\,(\leftarrow)$.

Lesson 3: Heat Transfer and Electrostatic Examples

Objectives: In this lesson we will look at Galerkin's Method to solve one-dimensional heat transfer and fluid flow problems.

- To understand the relationship between some of the heat transfer and fluid flow problems and the general 1D BVP discussed in the previous lesson.
- To solve an example so as to understand the process.

One-Dimensional Heat Conduction and Convection Problem

Consider a long, thin rod as shown in figure below. The temperature $T = T(x)$, $q = q(x)$ is the heat flux, $k = k(x)$ is the thermal conductivity, $Q = Q(x)$ is the interior volume heat source, $A = A(x)$ is the cross-sectional area, $h = h(x)$ is the convective heat transfer coefficient, $l = l(x)$ is the circumference, T_∞ is the ambient temperature. A sample set of boundary conditions is shown with an NBC on the left end and EBC on the right end. We could also have a case involving a mixed BC on the left or right ends.

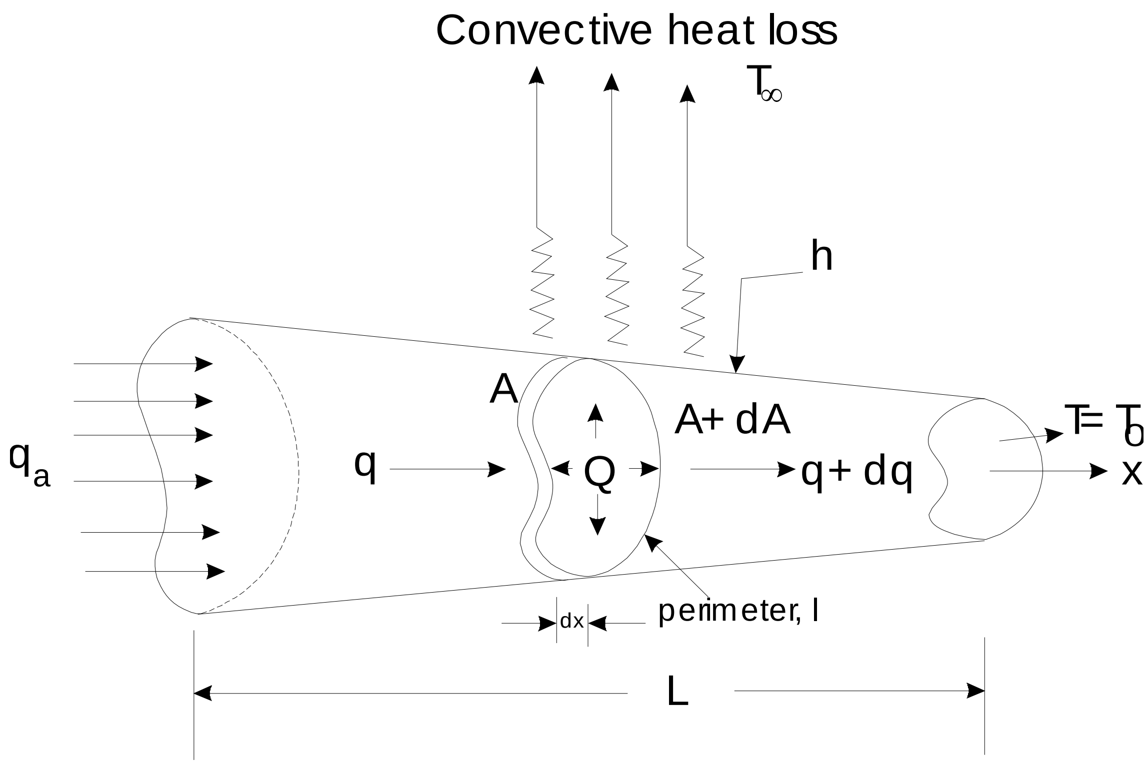


Fig. T4L3-1

The governing differential equation is given as

$$-\frac{d}{dx}\left(A(x)k(x)\frac{dT(x)}{dx}\right) + h(x)l(x)T(x) = Q(x)A(x) + h(x)l(x)T_\infty \qquad \text{(T4L3-1a)}$$

or, for a cylindrical rod with a constant cross-sectional area

$$-\frac{d}{dx}\left(k(x)\frac{dT(x)}{dx}\right) + \frac{hl}{A}T(x) = Q(x) + \frac{hl}{A}T_\infty \qquad \text{(T4L3-1a)}$$

with the boundary conditions as

At $x = x_a$, $T = T_a$ or $-q = c_a$ or $-q = h_a(T - T_a^\infty)$ (T4L3-2a)

At $x = x_b$, $T = T_b$ or $q = c_b$ or $q = h_b(T - T_b^\infty)$ (T4L3-2b)

Looking ahead to two and three-dimensional problems, these boundary conditions are special cases of the general form (for specified heat flow)

$$q_x n_x + q_y n_y + q_z n_z = -q_S \quad \text{(T4L3-3a)}$$

if heat q_S is flowing **into** the surface S, and (n_x, n_y, n_z) are the direction cosines of the outward normal from the surface. Similarly, for free convection from surface S, we have

$$q_x n_x + q_y n_y + q_z n_z = h(T_S - T_\infty) \quad \text{(T4L3-3b)}$$

Possible boundary conditions at the left end $(n_x = -1)$

Left end is at a known temperature, $T = T_a$

$$T = T_a$$

Heat (q_a) *is flowing into the left end*

$$q = q_a$$

Left end is insulated

$$q = 0$$

Free convection is taking place at the left end (ambient temperature is T_∞^a *and the convective coef is* h_a, $T > T_\infty^a$*)*

$$-q = h_a T - h_a T_a^\infty \qquad \text{or} \quad q = -h_a T + h_a T_a^\infty \quad \text{(T3L3-3c)}$$

Possible boundary conditions at the right end $(n_x = 1)$

Right end is at a known temperature, $T = T_b$

$$T = T_b$$

Heat (q_b) *is flowing out of the right end*

$$q = q_b$$

Right end is insulated

$$q = 0$$

Free convection is taking place at the right end (ambient temperature is T_∞^b *and the convective coef is* h_b, $T > T_\infty^b$*)*

$$q = h_b T - h_b T_b^\infty \quad \text{(T3L3-3d)}$$

Comparing these equations to the general form of the 1D BVP[9],

$$y(x) = T(x) \qquad \alpha(x) = k(x) \qquad \beta(x) = \frac{hl}{A} \quad \text{(T4L3-4a)}$$

$$f(x) \equiv Q(x) + \frac{hl}{A} T_\infty \qquad \tau = -k\frac{dT}{dx} = q \quad \text{(T4L3-4b)}$$

Compare $\quad \tau = c_a y + d_a \quad$ and $\quad q = -h_a T + h_a T_a^\infty$

---

[9] The sign convention is as follows – energy (or, heat flow) into the surface or boundary is positive.

and $\tau = c_b y + d_b$ and $q = h_b T - h_b T_b^\infty$, and we have

$$c_a = -h_a = c_1 \quad c_b = h_b = c_2 \quad d_a = h_a T_a^\infty = d_1 \quad d_b = -h_b T_b^\infty = d_2 \tag{T4L3-4c}$$

Using the natural and mixed boundary conditions as discussed above, and substituting the above in to Eqn. (T4L1-70), we have

$$\left[\left[\begin{array}{c|c} \dfrac{\bar{k}}{L} + \dfrac{\overline{hl}L}{3A} & -\dfrac{\bar{k}}{L} + \dfrac{\overline{hl}L}{6A} \\ \hline -\dfrac{\bar{k}}{L} + \dfrac{\overline{hl}L}{6A} & \dfrac{\bar{k}}{L} + \dfrac{\overline{hl}L}{3A} \end{array}\right] + h_1 \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & 0 \end{array}\right] + h_2 \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & 1 \end{array}\right]\right] \begin{Bmatrix} T_1 \\ T_2 \end{Bmatrix} =$$

$$\frac{L}{2} \left\{\begin{array}{c} \bar{Q} + \dfrac{\overline{hl}}{A} T_\infty \\ \hline \bar{Q} + \dfrac{\overline{hl}}{A} T_\infty \end{array}\right\} + \left\{\begin{array}{c} q_1 \\ \hline -q_2 \end{array}\right\} + \left\{\begin{array}{c} h_1 T_\infty^1 \\ \hline h_2 T_\infty^2 \end{array}\right\} \tag{T4L3-5a}$$

$$\mathbf{k}_{2\times2}\mathbf{u}_{2\times1} = \mathbf{f}_{2\times1} \tag{T4L3-5b}$$

Note that on the LHS and the RHS some of the components will be zero depending on whether the boundary condition is NBC or mixed. A dimensional analysis will show that - $T \equiv \mathrm{T}$, $\bar{A} \equiv \mathrm{L}^2$, $\bar{k} \equiv \dfrac{\mathrm{E}}{\mathrm{tLT}}$, $\bar{h} \equiv \dfrac{\mathrm{E}}{\mathrm{tL^2T}}$, $l = \mathrm{L}$, $\bar{Q} \equiv \dfrac{\mathrm{E}}{\mathrm{tL^3}}$, $c \equiv \dfrac{\mathrm{E}}{\mathrm{tL^2}}$, the LHS and the RHS have the units $\dfrac{\mathrm{E}}{\mathrm{tL^2}}$.

Illustrative Example T4L3-1

The figure shows a composite wall made of three materials. The outer temperature is $20^\circ C$. Convection heat transfer place on the inner surface of the wall with $T_\infty = 800^\circ C$ and $h = 25 \dfrac{W}{m^2 \cdot^\circ C}$. The thermal conductivities are $k_1 = 20 \dfrac{W}{m \cdot^\circ C}, k_2 = 30 \dfrac{W}{m \cdot^\circ C}, k_3 = 50 \dfrac{W}{m \cdot^\circ C}$. We need to determine the temperature distribution in the wall.

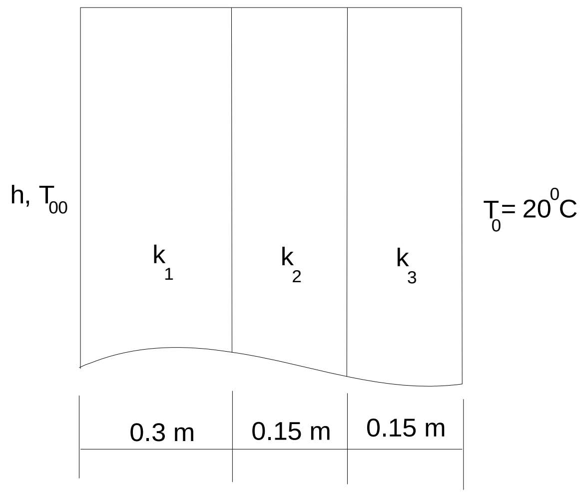


Fig. T4L3-2a

Solution: We will use a three-element model as shown. Note the following:

(a) This is a problem where the convective heat exchange (gain) takes place from the left end. The mixed boundary condition can be expressed as $q_x n_x = -q = -h_a\left(T_a^\infty - T_a\right)$. The leading minus sign on the right is because heat is flowing into the surface. Or, $q = -h_a T_a + h_a T_a^\infty$ (same as before).

(b) There is no convective heat exchange from the top and bottom of the wall that are assumed to be very tall compared to the thickness of the wall. Hence, $h = 0 = l$. We will assume a unit area for all computations, i.e $A = 1m^2$. There is no internal heat generation, i.e. $\overline{Q} = 0$.

(c) The right end is at a specified temperature (EBC). For the sake of convenience, we will not include the boundary flux load terms by assuming inter-element flux continuity.

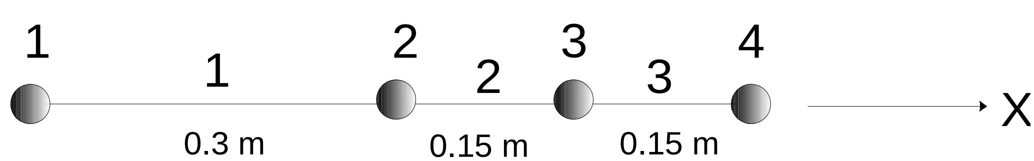


Fig. T4L3-2b

Element 1: $\overline{k} = 20\dfrac{W}{m \cdot {}^\circ C}$, $L = 0.3m$, $A = 1m^2$, $h = 0$, $l = 0$, $h_1 = 25\dfrac{W}{m^2 \cdot {}^\circ C}$, $T_1^\infty = 800^\circ C$, $h_2 = 0$, $\overline{Q} = 0$, $c_1 = 0$ and $c_2 = 0$. The element equations are as follows.

$$\left[\begin{array}{c|c} \frac{20}{0.3}+25 & -\frac{20}{0.3} \\ \hline -\frac{20}{0.3} & \frac{20}{0.3} \end{array}\right]\left\{\begin{array}{c} T_1 \\ \hline T_2 \end{array}\right\}=\left\{\begin{array}{c} 25(800) \\ \hline 0 \end{array}\right\}$$

Element 2: $\bar{k}=30\frac{W}{m\cdot^\circ C}$, $L=0.15m$, $A=1m^2$, $h=0$, $l=0$, $h_1=0$, $h_2=0$, $\overline{Q}=0$, $c_1=0$ and $c_2=0$. The element equations are as follows.

$$\left[\begin{array}{c|c} \frac{30}{0.15} & -\frac{30}{0.15} \\ \hline -\frac{30}{0.15} & \frac{30}{0.15} \end{array}\right]\left\{\begin{array}{c} T_2 \\ \hline T_3 \end{array}\right\}=\left\{\begin{array}{c} 0 \\ \hline 0 \end{array}\right\}$$

Element 3: $\bar{k}=50\frac{W}{m\cdot^\circ C}$, $L=0.15m$, $A=1m^2$, $h=0$, $l=0$, $h_1=0$, $h_2=0$, $\overline{Q}=0$, $c_1=0$ and $c_2=0$. The element equations are as follows.

$$\left[\begin{array}{c|c} \frac{50}{0.15} & -\frac{50}{0.15} \\ \hline -\frac{50}{0.15} & \frac{50}{0.15} \end{array}\right]\left\{\begin{array}{c} T_3 \\ \hline T_4 \end{array}\right\}=\left\{\begin{array}{c} 0 \\ \hline 0 \end{array}\right\}$$

Assembling the equations

$$\begin{bmatrix} 91.6667 & -66.6667 & 0 & 0 \\ -66.6667 & 266.6667 & -200.0 & 0 \\ 0 & -200.0 & 533.3333 & -333.3333 \\ 0 & 0 & -333.3333 & 333.3333 \end{bmatrix}\begin{Bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \end{Bmatrix}=\begin{Bmatrix} 20{,}000 \\ 0 \\ 0 \\ 0 \end{Bmatrix}$$

There is no natural boundary condition (the mixed BC was taken care of when the element equations were generated). To enforce the EBC $T_4=20$, we use the elimination approach. The modified equations are

$$\begin{bmatrix} 91.6667 & -66.6667 & 0 & 0 \\ -66.6667 & 266.6667 & -200.0 & 0 \\ 0 & -200.0 & 533.3333 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{Bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \end{Bmatrix} = \begin{Bmatrix} 20,000 \\ 0 \\ 6666.6667 \\ 20 \end{Bmatrix}$$

Solving the equations we have (note decreasing temperature from node 1 to 4),

$$\{T_1, T_2, T_3, T_4\} = \{304.76, 119.05, 57.14, 20\}^\circ C$$

We can now compute the flux in each element.

Element 1:

$$\tau^1 = -\frac{k_1}{L_1}(T_2 - T_1) = -\frac{20}{0.3}(119.05 - 304.76) = 12380.95\frac{W}{m^2}$$

Element 2:

$$\tau^2 = -\frac{k_2}{L_2}(T_3 - T_2) = -\frac{30}{0.15}(57.14 - 119.05) = 12380.95\frac{W}{m^2}$$

Element 3:

$$\tau^3 = -\frac{k_3}{L_3}(T_4 - T_3) = -\frac{50}{0.15}(20.0 - 57.1) = 12380.95\frac{W}{m^2}$$

We will now examine the results. The solution satisfies the EBC at the right end. The flux is constant throughout the domain as it should be.

Another Illustrative Example T4L3-2

Fig. T4L3-3 shows a circular cross-section pin fin. It has a diameter of 0.3125" and a length of 5". At the root, the temperature is $T_0 = 150^\circ F$. The ambient temperature is $T_\infty = 80^\circ F$, the convective coefficient is $h = 6\dfrac{BTU}{h \cdot ft^2 \cdot^\circ F}$, and the thermal conductivity is $k = 24.8\dfrac{BTU}{h \cdot ft \cdot^\circ F}$. Determine the temperature distribution in the fin.

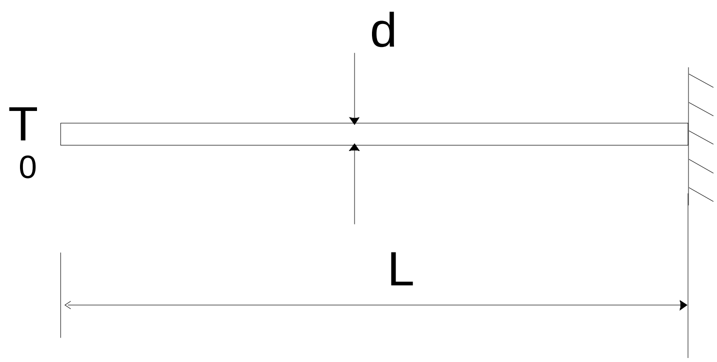


Fig. T4L3-3

Solution: We will use a two-element model to solve the problem. The FE model is shown below.

Fig. T4L3-4

Note the following - (a) In this problem the left end is tied to an EBC. (b) There is no convective heat exchange from the right end (NBC with $q=0$). (c) There is no internal heat generation, i.e. $\overline{Q}=0$. We will select ft as the units for length.

Element 1: $\bar{k} = 24.8\dfrac{BTU}{h \cdot ft \cdot^\circ F}$, $L = 0.208\,ft$, $A = \pi\dfrac{d^2}{4} = 5.326(10^{-4})\,ft^2$, $T_\infty = 80^\circ F$,

$h = 6\dfrac{BTU}{h \cdot ft^2 \cdot^\circ F}$, $l = \pi d = 0.0818\,ft$, $h_1 = 0$, $h_2 = 0$, $\overline{Q} = 0$, $c_1 = 0$ and $c_2 = 0$. The element equations are as follows.

$$\left[\begin{array}{c|c} \dfrac{24.8}{0.208} + \dfrac{6(0.0818)(0.208)}{3(5.326e-4)} & -\dfrac{24.8}{0.208} + \dfrac{6(0.0818)(0.208)}{6(5.326e-4)} \\ \hline -\dfrac{24.8}{0.208} + \dfrac{6(0.0818)(0.208)}{6(5.326e-4)} & \dfrac{24.8}{0.208} + \dfrac{6(0.0818)(0.208)}{3(5.326e-4)} \end{array}\right] \begin{Bmatrix} T_1 \\ \hline T_2 \end{Bmatrix} = \frac{0.208}{2}\begin{Bmatrix} \dfrac{6(0.0818)}{5.326e-4}(80) \\ \hline \dfrac{6(0.0818)}{5.326e-4}(80) \end{Bmatrix}$$

or,
$$\left[\begin{array}{c|c} 183.12 & -87.285 \\ \hline -87.285 & 183.12 \end{array}\right] \begin{Bmatrix} T_1 \\ \hline T_2 \end{Bmatrix} = \begin{Bmatrix} 7667 \\ \hline 7667 \end{Bmatrix}$$

Element 2: Same as element 1 except we use T_2 and T_3.

Assembling, we have

$$\begin{bmatrix} 183.12 & -87.285 & 0 \\ -87.285 & 366.25 & -87.285 \\ 0 & -87.285 & 183.12 \end{bmatrix} \begin{Bmatrix} T_1 \\ T_2 \\ T_3 \end{Bmatrix} = \begin{Bmatrix} 7667 \\ 15334 \\ 7667 \end{Bmatrix}$$

The NBC term is zero. To apply the EBC we use the elimination technique resulting in

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 366.245 & -87.285 \\ 0 & -87.285 & 183.12 \end{bmatrix} \begin{Bmatrix} T_1 \\ T_2 \\ T_3 \end{Bmatrix} = \begin{Bmatrix} 150 \\ 28427 \\ 7667 \end{Bmatrix}$$

Solving, we have

$$\{T_1, T_2, T_3\} = \{150, 98.82, 88.97\}^\circ F$$

The temperature decreases from left to right as expected. The element flux is computed as follows.

Element 1:

$$\tau^1 = -\frac{k_1}{L_1}(T_2 - T_1) = -\frac{24.8}{0.208}(98.82 - 150) = 6102 \frac{BTU}{h \cdot ft^2}$$

Element 2:

$$\tau^2 = -\frac{k_2}{L_2}(T_3 - T_2) = -\frac{24.8}{0.208}(88.97 - 98.82) = 1174 \frac{BTU}{h \cdot ft^2}$$

Let us carry out a few checks. The solution satisfies the EBC of $150^\circ F$ at $x = 0$. The flux at the right end must be zero. However, with this crude two-element model the error is large. To obtain a better accuracy we must refine the mesh, i.e. add more elements to the mesh. We will look at this aspect in the next two sections.

Electrostatic Field

Fig. T4L3-4(a) and (b) show an essentially one-dimensional electrostatic field.

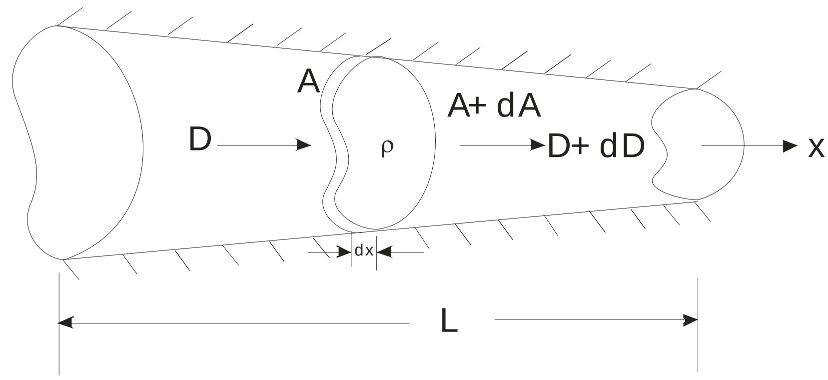


Fig. T4L3-4 (a) A dielectric rod with its lateral surface electrically insulated

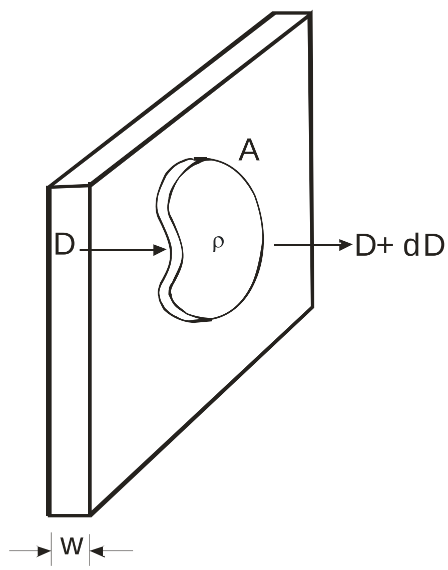


Fig. T4L3-4 (b) A wall or panel of dielectric material

The conservation of electric charge as per Gauss' Law states that

$$DA + \rho A dx = \left(D + dD\right)\left(A + dA\right) \qquad \text{(T4L3-6)}$$

where

$D(x)$ displacement of electric charge, Q/L^2

$\rho(x)$ charge density, Q/L^3

Neglecting second-order terms we have

$$\frac{d}{dx}\big(D(x)A(x)\big) = \rho(x)A(x) \tag{T4L3-7}$$

The constitutive equation is given as

$$D(x) = \varepsilon(x)E(x) \tag{T4L3-8}$$

where

$E(x)$ electrostatic field, V/L

$\varepsilon(x)$ permittivity, C/L

ε is a material property that measures how easily charged particles are displaced in a dielectric. It is related to another property, the dielectric constant K, by the relation $K = \varepsilon/\varepsilon_0$ where ε_0 is the permittivity of a vacuum. In addition we also need the relation

$$E(x) = -\frac{d\Phi(x)}{dx} \tag{T4L3-9}$$

where $\Phi(x)$ is the electrostatic potential (units: V). Eqn. (T4L3-9) is the corollary of Coulomb's Law which is a physical law describing the nature of electrostatic forces between charged particles. Combining Eqn. (T4L3-8) and (T4L3-9) yields the constitutive equation

$$D(x) = -\varepsilon(x)\frac{d\Phi(x)}{dx} \tag{T4L3-10}$$

Substituting Eqn. (T3L3-10) into (T4L3-7) yields the differential equation

$$-\frac{d}{dx}\left(\varepsilon(x)A(x)\frac{d\Phi(x)}{dx}\right) = \rho(x)A(x) \tag{T4L3-11}$$

The potential $\Phi(x)$ is the unknown function. The displacement $D(x)$ plays the role of a flux. The boundary conditions are either essential ($\Phi(x)$ is known) or natural ($D(x)$ is known).

Lesson 4: Higher-Order Elements

Objectives: In this lesson we will look at the characteristics of elements that use higher-order interpolation.

- To understand the properties of higher-order elements.
- To derive and use in an example a higher-order element.

Higher-Order Interpolation

So far we have seen the element concept illustrated using linear interpolation on an element described by two nodes. We also saw in the classical Galerkin's solution methodology that superior results were obtained using higher order interpolation. Can we not tie this concept to the element concept? The answer is a resounding "Yes".

Fig. T4L4-1 shows the linear element (a) on which the solution is defined by a linear polynomial, (b) is described by two nodes with nodal values y_1 and y_2, and (c) the resulting shape functions that use the nodal values to interpolate the solution.

$$\tilde{y}(x) = a_1 + a_2 x = \phi_1(x) y_1 + \phi_2(x) y_2 \tag{T4L1-29}$$

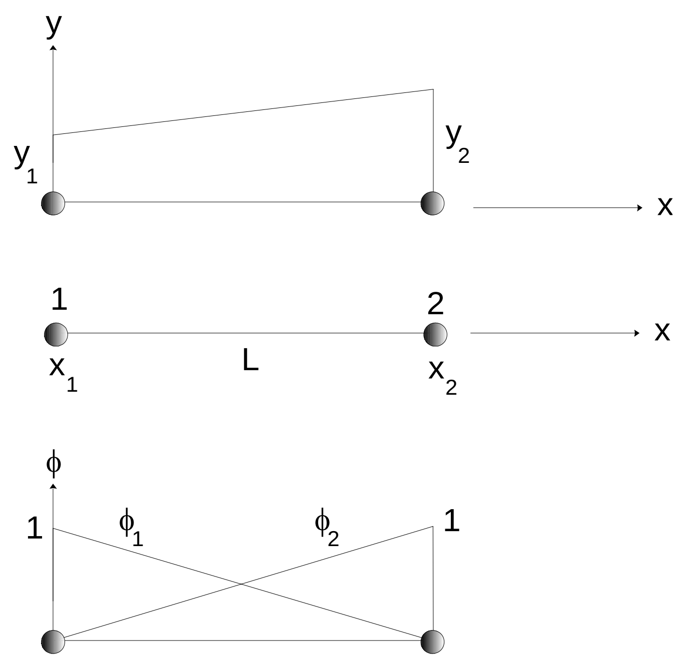


Fig. T4L4-1

Fig. T4L4-2 shows the quadratic element (a) on which the solution is defined by a quadratic polynomial, (b) is described by three nodes with nodal values y_1, y_2 and y_3, and (c) the resulting shape functions that use the nodal values to interpolate the solution.

$$\tilde{y}(x) = a_1 + a_2 x + a_3 x^2 = \phi_1(x) y_1 + \phi_2(x) y_2 + \phi_3(x) y_3 \tag{T4L4-1}$$

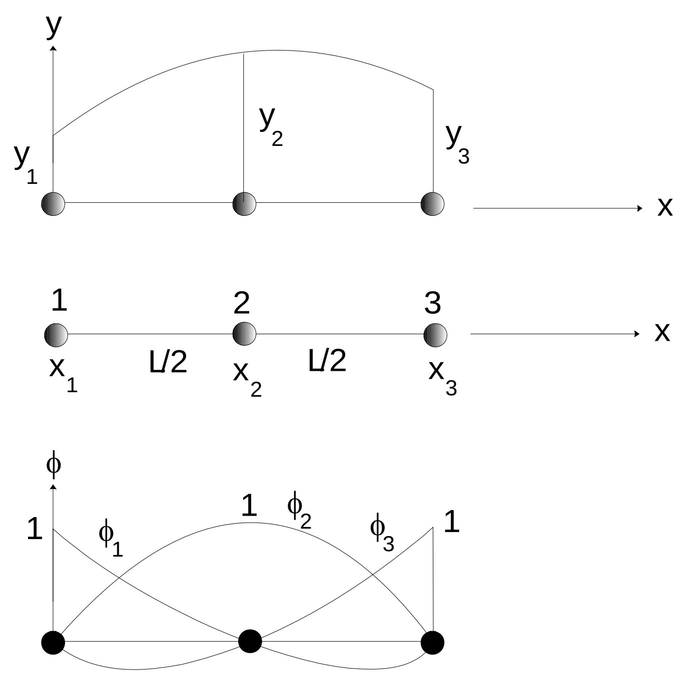


Fig. T4L4-2

A few comments are in order.

(a) For a one-dimensional element, we need to have nodes at the ends of the elements so that one element can be tied to the element next to it. So we need a minimum of two nodes.

(b) When we have one degree of freedom per node, the number of nodes in the element is equal to the number of coefficients in the trial solution. In other words, we need a total of two nodes for the linear element and a total of three nodes for the quadratic element. Only then will we have sufficient equations to write the coefficients of the polynomials in terms of the nodal values.

(c) The number of shape functions is also equal to the number of nodal values. We should now look at the properties of the shape functions. First, the trial solution must be complete – a quadratic polynomial must have the constant, linear **and** quadratic terms (e.g. having the constant, quadratic and cubic terms is not allowed). This is done to ensure that the element can assume all possible solution modes. Second, the shape functions satisfy the following conditions

$$\phi_i(x_j) = \delta_{ij} \tag{T4L4-2}$$

where δ_{ij} is the Kronecker's Delta. In other words, the shape function will have a unit value at the node it is associate with and zeros at the other nodes (see the plots shown above). This also ensures that the shape functions are linearly independent.

Now on with the quadratic element. Using Eqn. (T4L4-1) and the nodal conditions, we have

$$\begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} y_1 \\ y_2 \\ y_3 \end{Bmatrix} \tag{T4L4-3}$$

These equations are similar to Eqn. (T4L1-31). Solving for the a_i's and collecting like terms, we have

$$\tilde{y}(x) = \frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)} y_1 + \frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)} y_2 + \frac{(x-x_1)(x-x_2)}{(x_3-x_1)(x_3-x_2)} y_3 \tag{T4L4-4}$$

Hence,

$$\phi_1(x) = \frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)} \tag{T4L4-5a}$$

$$\phi_2(x) = \frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)} \tag{T4L4-5b}$$

$$\phi_3(x) = \frac{(x-x_1)(x-x_2)}{(x_3-x_1)(x_3-x_2)} \tag{T4L4-5c}$$

There is an easier way to compute the shape functions that we shall see in the next module. We still have two questions to answer before we can generate the element equations for the quadratic element. First, "Where is node 2 located within the element?" Again, a detailed answer will be generated in Module 2. For the time being let us assume that it is located at the center of the element. Hence,

$$(x_2 - x_1) = (x_3 - x_2) = \frac{L}{2} \qquad (x_3 - x_1) = L \tag{T4L4-6}$$

Second, "How do we handle the $\alpha(x)$, $\beta(x)$ and $f(x)$ terms?" We will develop a sophisticated technique in Module 2. For the time being let us assume that they are constants within the element. Hence,

$$\left[\begin{bmatrix} \frac{7\alpha}{3L} & -\frac{8\alpha}{3L} & \frac{\alpha}{3L} \\ -\frac{8\alpha}{3L} & \frac{16\alpha}{3L} & -\frac{8\alpha}{3L} \\ \frac{\alpha}{3L} & -\frac{8\alpha}{3L} & \frac{7\alpha}{3L} \end{bmatrix} + \begin{bmatrix} \frac{4\beta L}{30} & \frac{2\beta L}{30} & -\frac{\beta L}{30} \\ \frac{2\beta L}{30} & \frac{16\beta L}{30} & \frac{2\beta L}{30} \\ -\frac{\beta L}{30} & \frac{2\beta L}{30} & \frac{4\beta L}{30} \end{bmatrix} - g_1 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right.$$

$$\left. + h_3 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right] \begin{Bmatrix} y_1 \\ y_2 \\ y_3 \end{Bmatrix} = \begin{Bmatrix} \frac{fL}{6} \\ \frac{4fL}{6} \\ \frac{fL}{6} \end{Bmatrix} + \begin{Bmatrix} c_1 \\ 0 \\ -c_3 \end{Bmatrix} \qquad \text{(T4L4-7)}$$

The flux in the element is given by

$$\tilde{\tau} = -\alpha(x)\frac{d\tilde{y}}{dx} = -\alpha\left(\frac{2(2x - x_2 - x_3)}{L^2} y_1 + \frac{(-4)(2x - x_1 - x_3)}{L^2} y_2 + \frac{2(2x - x_1 - x_2)}{L^2} y_3\right)$$

$$= -\frac{\alpha}{L^2}\left[x(4y_1 - 8y_2 + 4y_3) + x_1(4y_2 - 2y_3) - 2x_2(y_1 + y_3) - 2x_3(y_1 - 2y_2)\right] \qquad \text{(T4L4-8)}$$

In a similar manner we can generate higher order elements involving cubic, quartic, quintic etc. trial functions with four, five, six etc. nodes per element. The manner in which we identify these elements is usually as follows. These elements are designated as 1D-C^m *interpolation order* element where C^m denotes that the problem variable and its derivatives up to order *m* are continuous across element boundaries. In the element formulation that we have discussed so far, only the problem variable is continuous across the element boundaries. Hence the elements are designated 1D-C^0 linear element, or 1D-C^0 quadratic element etc.

Illustrative Example T4L4-1

Let us resolve the last problem from the previous section. This time we will use the quadratic element. Let us assume that the number of nodes is the same. The FE model is shown in Fig. T4L4-3.

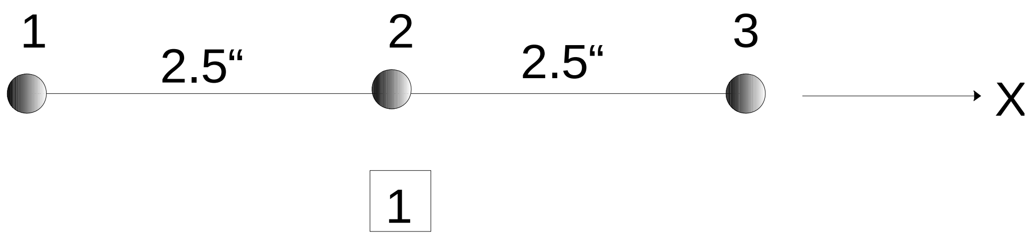


Fig. T4L4-3

Element 1: $\bar{k} = 24.8 \dfrac{BTU}{h \cdot ft \cdot^\circ F}$, $L = 0.416\,ft$, $A = \pi \dfrac{d^2}{4} = 5.326(10^{-4})\,ft^2$, $T_\infty = 80^\circ F$,

$h = 6 \dfrac{BTU}{h \cdot ft^2 \cdot^\circ F}$, $l = \pi d = 0.0818\,ft$, $h_1 = 0$, $h_2 = 0$, $\overline{Q} = 0$, $c_1 = 0$ and $c_2 = 0$. The element equations are as follows ($\alpha = k$, $\beta = \dfrac{hl}{A}$, $f = \dfrac{hlT_\infty}{A}$)

$$\left(\begin{bmatrix} \dfrac{7(24.8)}{3(0.416)} & -\dfrac{8(24.8)}{3(0.416)} & \dfrac{(24.8)}{3(0.416)} \\ -\dfrac{8(24.8)}{3(0.416)} & \dfrac{16(24.8)}{3(0.416)} & -\dfrac{8(24.8)}{3(0.416)} \\ \dfrac{(24.8)}{3(0.416)} & -\dfrac{8(24.8)}{3(0.416)} & \dfrac{7(24.8)}{3(0.416)} \end{bmatrix} + \right.$$

$$\left. \begin{bmatrix} \dfrac{4(921.52)(0.416)}{30} & \dfrac{2(921.52)(0.416)}{30} & -\dfrac{(921.52)(0.416)}{30} \\ \dfrac{2(921.52)(0.416)}{30} & \dfrac{16(921.52)(0.416)}{30} & \dfrac{2(921.52)(0.416)}{30} \\ -\dfrac{(921.52)(0.416)}{30} & \dfrac{2(921.52)(0.416)}{30} & \dfrac{4(921.52)(0.416)}{30} \end{bmatrix} \right) \begin{Bmatrix} T_1 \\ T_2 \\ T_3 \end{Bmatrix} =$$

$$\begin{Bmatrix} \dfrac{(73721.4)(0.416)}{6} \\ \dfrac{4(73721.4)(0.416)}{6} \\ \dfrac{(73721.4)(0.416)}{6} \end{Bmatrix}$$

Or,
$$\begin{bmatrix} 181.7 & -116.38 & -1.4256 \\ -116.38 & 488.33 & -116.38 \\ -1.4256 & -116.38 & 181.7 \end{bmatrix} \begin{Bmatrix} T_1 \\ T_2 \\ T_3 \end{Bmatrix} = \begin{Bmatrix} 5111.3 \\ 20445 \\ 5111.3 \end{Bmatrix}$$

Applying the boundary conditions (EBC for T_1), we have

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 488.33 & -116.38 \\ 0 & -116.38 & 181.7 \end{bmatrix} \begin{Bmatrix} T_1 \\ T_2 \\ T_3 \end{Bmatrix} = \begin{Bmatrix} 150 \\ 37902 \\ 5325.2 \end{Bmatrix}$$

Solving,

$$\{T_1, T_2, T_3\} = \{150, 99.84, 93.26\}^\circ F$$

The flux in the element is computed using Eqn (T4L4-8) and are given by

$$\tau(x = 0.0879') = 6382 \frac{BTU}{h \cdot ft^2}$$

$$\tau(x = 0.3281') = 383.1 \frac{BTU}{h \cdot ft^2}$$

The reason for selecting the two locations will be explained in the next module. Bear in mind that the flux distribution in this element is, as Eqn. (T4L4-8) shows, linear unlike the linear element in which the flux is a constant. Let's compare the two results.

	Temperature ($^\circ F$)		Flux $\left(\dfrac{BTU}{h \cdot ft^2}\right)$	
Node	Linear Elements	Quadratic Element	Linear Elements	Quadratic Element
1	150	150		
2	98.92	99.84	6102	6382
3	88.97	93.26	1174	383.1

The nodal temperatures are close but the flux values are quite different.

Lesson 5: Mesh Refinement and Convergence

Objectives: In this lesson we will look at the approach to applying the FE method to solve a problem.

- To understand the concepts associated with mesh refinement and convergence.
- To resolve one or more previous examples with different meshes and use the concept of convergence to obtain the solution.

Mesh Refinement

In the previous lessons we learnt two important facts. First, increasing the number of degrees of freedom increases the accuracy of the solution. In other words, if we keep on increasing the number of elements (and nodes) in the FE model, we should obtain better solutions that converge to the exact result. This is known as *h-convergence*. The *h* notation refers to the size of the elements. Second, increasing the order of the polynomial in the trial solution also increases the accuracy of the solution. In other words, if we keep on increasing the element order (and hold the number of elements constant), we should obtain better solutions that converge to the exact result. This is known as *p-convergence*. The *p* notation refers to the polynomial order. We could also combine the two and obtain what is known as *hp-convergence*.

The advantages of low-order finite elements are (a) the size of the element matrices is small, (b) the computational ease with which the elements can be generated, and (c) the size of the hand-band width of the structural stiffness matrix is small. The major disadvantage is that the convergence is slow. The comments pertaining to higher-order elements are just the opposite.

The FE mesh need not be uniform nor contain just one type of element. It may be advantageous to use the lower-order elements where the solution does not change rapidly (flux has a low value) and use the higher-order elements in regions where higher accuracy is required. By the same token, the mesh can be finer where higher accuracy is required.

The biggest disadvantage of any FE solution is the computational expense in solving the system equations. Hand calculations (with help from equation solvers in calculators) are tedious if the number of unknowns is greater than about 10. However, the finite element method is a numerical technique that is ideal for computer-based solutions. As we mentioned in the first topic, the amount of human time taken to create the data and examine the results is far greater than the time taken to solve most FE problems.

We will illustrate these ideas using the results from the computer program. If you wish you may jump to the section "Using the 1DBVP© Program" to familiarize yourself with the program before reading the next section.

An Illustrative Example T4L5-1

Let solve Example T4L3-2 using the concepts discussed earlier. We will monitor three response quantities – temperature at the right end, the flux at the left end that is the highest flux in the model and the flux at the right end that should converge to zero. The results are summarized below.

Model ID	Number of elements	Element Type	Temperature at right end	Flux at right end	Flux at left end
1	2	Linear	89	1 174	6 102
2	4	- do -	90.5	542	7 804
3	8	- do -	90.9	266	8 970
4	16	- do -	91	132	9 664
5	32	- do -	91	66	10 044
6	64	- do -	91	33	10 244
7	128	- do -	91	17	10 346
1	2	Quadratic	91.1	417	8 039
2	4	- do -	91	220	9 137
3	8	- do -	91	111	9 767
4	16	- do -	91	55	10 102
5	32	- do -	91	28	10 275
6	64	- do -	91	14	10 362

The temperature at the left end converges very rapidly to $91^\circ F$. The flux at the right end appears to converge linearly. The flux at the left end converges much more slowly for the linear element compared to the quadratic element.

Figs. T4L5-1 and T4L5-2 show the plots for the flux at the left and right ends for the linear and quadratic elements.

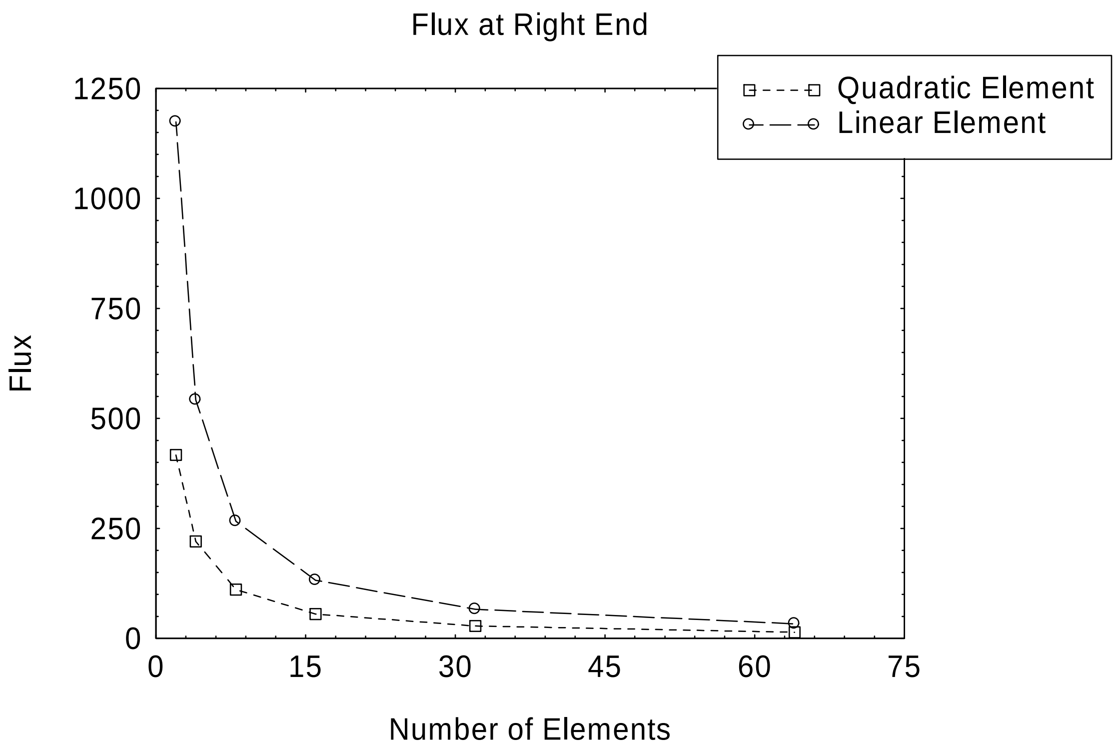


Fig. T4L5-1

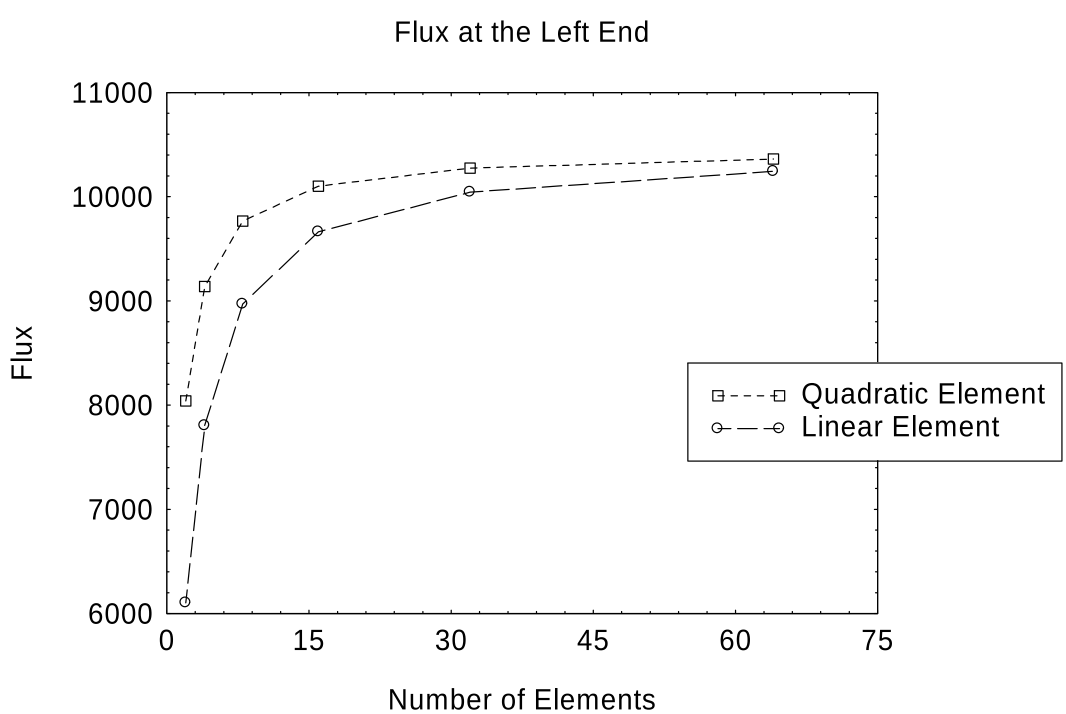


Fig. T4L5-2

Summary

We achieved two primary goals with the lessons in this topic. First, we generalized the Galerkin's Method by tying it to the element concept. This made it easier to generate the trial solutions that are valid over an arbitrary problem subdomain (the element!). The side effect is that we can handle problems with known discontinuities in the solution, e.g. the flux. Second, we looked at the manner in which the boundary conditions could be applied more easily. While the problem's essential boundary conditions were satisfied exactly, the higher-order boundary conditions were satisfied only in the limit.

The derivation of the element equations is perhaps the most important step when implementing the FE method. We saw how we could generate a family of trial solutions and the associated elements. There are two important issues that we will deal with in the next module that will make it easier to generate the element equations – a more rational way to generate the shape functions, and the isoparametric formulation.

It should be pointed out that seemingly different engineering problems are all governed by the same differential equation. However, one still needs to contend with the physics behind the parameters and the equations.

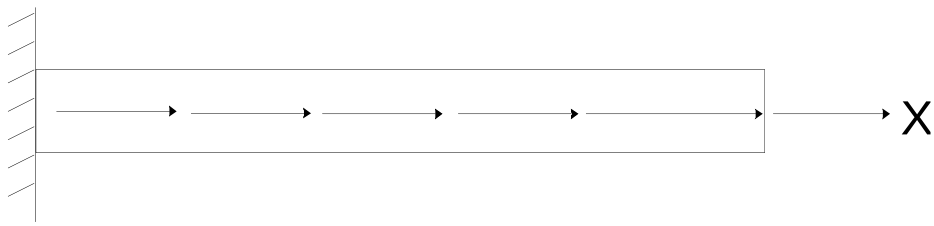
Better solutions are obtained by (a) increasing the number of elements in the mesh or, (b) by increasing the order of the element (trial solution), or (c) both. With the easy availability of FE computer programs, it is now possible to obtain accurate solutions in a relatively short period of time. The 1DBVP© program illustrates (in its own manner) the aspects associated with pre-processing, solution and post-processing.

Review Exercises

In all the problems below, generate all the steps by hand except solve the systems equations using a programmable calculator or a computer program. Finally check your answers using the 1DBVP© program. Explain the differences in answers, if any.

Problem T4L2-1

Consider the $4''$ bar shown below that is loaded by an axial surface traction given by the function $f(x) = x^2\ lb/in$.



The bar properties are as follows - $E = 30(10^6)\ psi$ and $A = 2\ in^2$. Determine the stresses in the bar using

(a) One linear element.
(b) Two linear elements.
(c) Four linear elements.

Comment on the results.

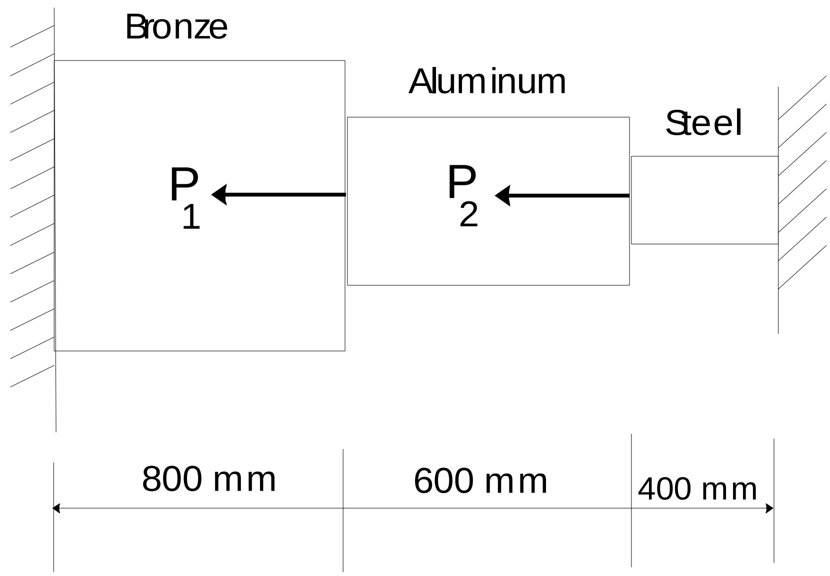
Problem T4L2-2

Resolve Problem T4L2-1 but use quadratic elements instead.

Problem T4L2-3

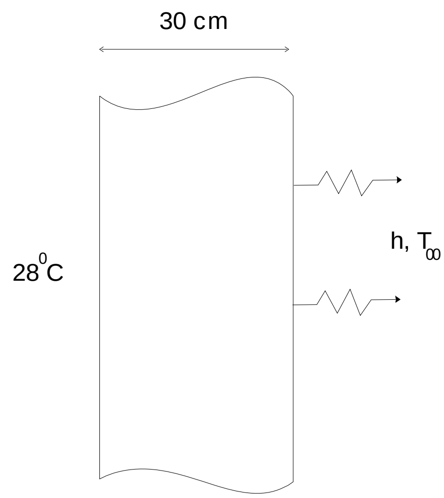
The structure shown in the figure below is subjected to an increase in temperature $\Delta T = 80^\circ C$. In addition the loads are given as follows: $P_1 = 60\ kN$, $P_2 = 75\ kN$. Determine the displacements, stresses, and support reactions using linear elements. The material properties are as follows.

Material	$A\,(mm^2)$	$E\ (GPa)$	$\alpha(mm/mm\cdot^\circ C)$
Bronze	2400	83	18.9e-6
Aluminum	1200	70	23e-6
Steel	600	200	11.7e-6



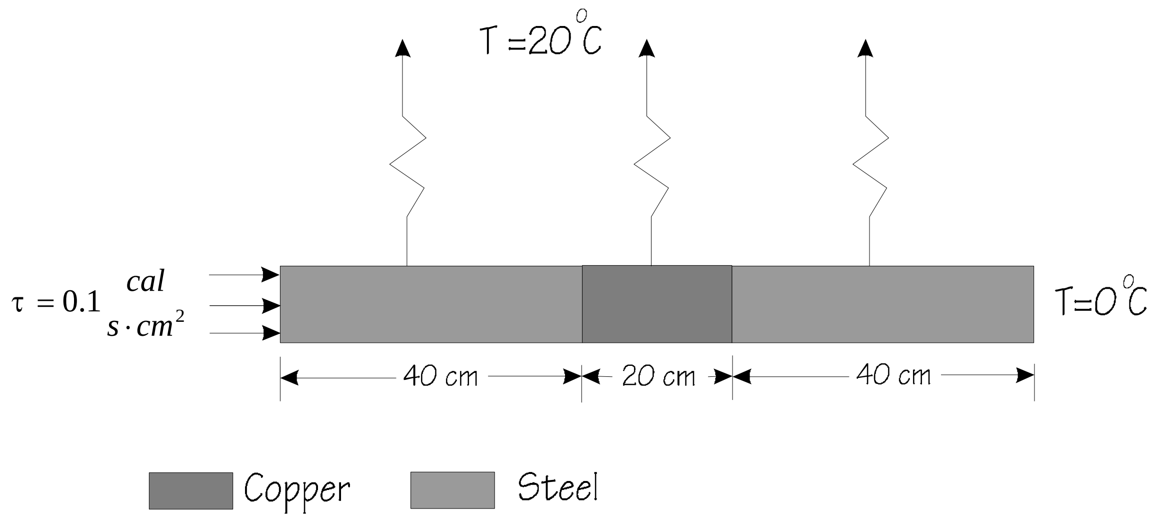
Problem T4L3-1

Consider a brick wall of thickness $L = 30cm$, $k = 0.7 W/m\cdot^\circ C$. The inner surface is at $28^\circ C$ and the outer surface is exposed to cold air at $T_\infty = -15^\circ C$. The heat transfer coefficient associated with the outside surface is $h = 40 W/m^2\cdot^\circ C$. Determine the steady-state temperature distribution within the wall and also the heat flux through the wall. Assume a one-dimensional heat flow. Use a two linear-element model.



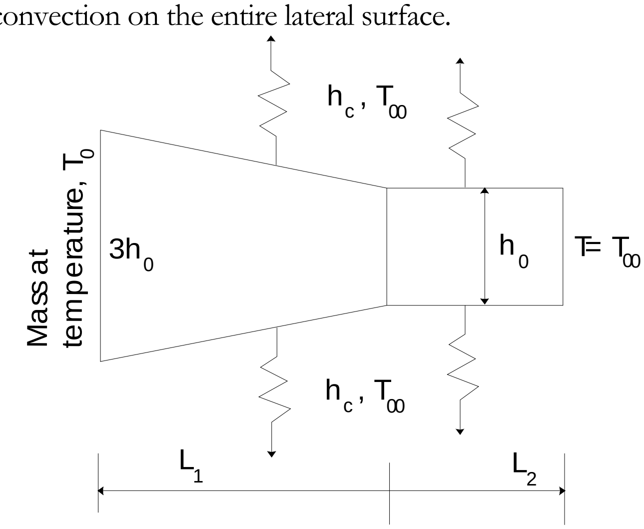
Problem T4L3-2

Figure shows a thin, cylindrical rod, 1m long, composed of two different materials – the two end sections each 40 cm long, are made of steel $\left(k = 0.12 \dfrac{cal - cm}{s - cm^2 - ^\circ C} \right)$, and the center section is made of copper $\left(k = 0.92 \dfrac{cal - cm}{s - cm^2 - ^\circ C} \right)$ is 20 cm long. The cross section is circular, with a radius of 2 cm. Heat is flowing into the left at a steady rate of $0.1 cal/s - cm^2$. The temperature of the right end is maintained at a constant $0^\circ C$. The rod is in contact with air at an ambient temperature of $20^\circ C$ so there is free convection from the lateral surface. The convective coefficient is given as $1.5 \times 10^{-4} \dfrac{cal}{s - cm^2 - ^\circ C}$. Determine the temperature and flux distribution in the rod using (a) only linear elements and (b) only quadratic elements.



Problem T4L3-3

A variable area rectangular cross-section fin transmits heat away from a mass as indicated in the figure. The thickness of the fin in the direction perpendicular to the paper is ten times that shown in the plane of the paper. There is convection on the entire lateral surface.



With $h_0 = 5\ cm$, $L_1 = L_2 = 10\ cm$, $T_0 = 400^\circ C$, $T_\infty = 100^\circ C$, $h_c = 10^{-3}\ W/mm^2 \cdot^\circ C$ and $k = 0.30 W/mm^2 \cdot^\circ C$, find the temperature and flux distribution.

Problem T4L3-4

Consider a slab of thickness 0.1 m with a thermal conductivity $k = 40W/\left(m\cdot^{\circ}C\right)$ in which energy is generated at a constant rate of $10^6\,W/m^3$. The boundary surface at $x = 0$ is insulated, and the one at $x = 0.1m$ is subjected to convection with a heat transfer coefficient of $200W/\left(m^2\cdot^{\circ}C\right)$ into an ambient at a temperature of $150^{\circ}C$. Find the temperature and flux distribution.

Problem T4L3-5

When radial heat flow takes place in cylindrical bodies, the differential equation is different than what we saw in 1D-BVP problem. Consider steady-state radial heat conduction in a long solid cylinder of radius $r = b$, in which energy is generated at a rate of $Q(r)W/m^3$. The temperature distribution $T(r)$ in the cylinder is governed by the following heat conduction equation

$$-\frac{1}{r}\frac{d}{dr}\left[kr\frac{dT(r)}{dr}\right] = Q(r) \qquad a < r < b$$

with the appropriate boundary conditions at $r = a$ and $r = b$. Develop the appropriate element equations for solving this problem using the Galerkin's Method for the (a) 1D-C^0 linear element and (b) 1D-C^0 quadratic element.

Problem T4L3-6

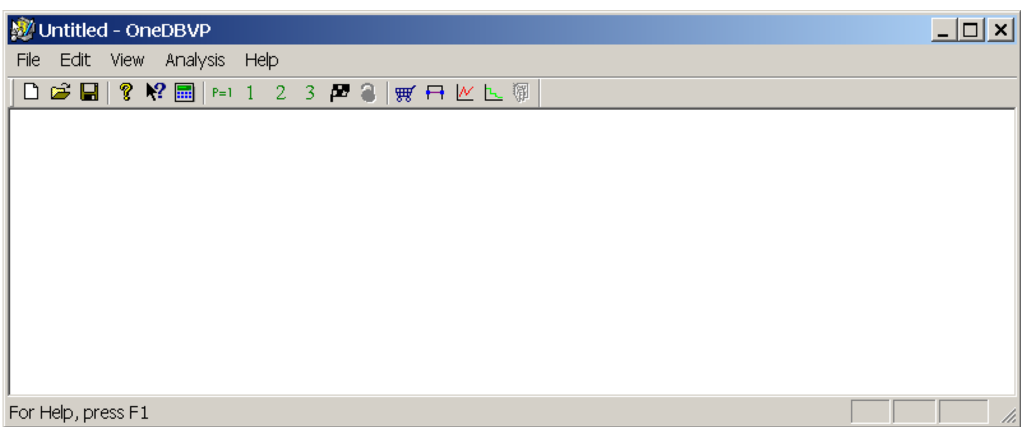
A 10-cm diameter solid chrome-nickel rod with thermal conductivity of $k = 20W/\left(m\cdot^{\circ}C\right)$ is heated electrically by the passage of an electric current which generates energy within the rod at a uniform rate of $10^7\,W/m^3$. The surface of the rod is subjected to convection with a heat transfer coefficient $h = 200W/\left(m\cdot^{\circ}C\right)$ into an ambient at $30^{\circ}C$. Find the temperature distribution in the rod.

Using the OneDBVP© Program

The OneDBVP program is designed to solve the problem described by equations (T4L1-56) and (T4L1-57). It can be used on computer systems running any version of the Windows OS (98, NT, 2000 and XP). Install the program as you would any other Microsoft Windows program. A few things to note before we get started with the program.

(a) The first step is to get the problem data ready. **Note that the program solves the problem as described by Eqns. (T4L1-56) and (T4L1-57)**. In other words, it does not specifically solve a solid mechanics problem or a heat transfer problem etc. You must relate the parameters from the problem to the general one-dimensional BVP as we have done in Lessons 2-4.
(b) The program does not assume any units. The problem units must be consistent.
(c) Finally we will discuss the program terminology. The FE mesh is described in terms of segments. A segment contains one or more elements that have exactly the same properties. Only the nodes that make up the elements have different locations and numbers. The problem domain is divided into one or more segments. The program numbers the nodes and elements (starting at 1) starting with the leftmost segment. Note that you create the segments; 1DBVP program creates the elements, nodes and loads.

Once you launch the program, you should see the following window.



Using the program is as easy as 1-2-3. Press F1 to invoke the on-line help. Go through the on-line help on understand how to use the program. The $\alpha(x)$, $\beta(x)$ and $f(x)$ terms are assumed to be described by a quadratic polynomial of the form ax^2+bx+c and the input to the program are the a,b,c values. The type of boundary condition is recognized by the program as being an EBC, a NBC or a mixed BC. The first two need an additional real input. The mixed BC needs two values - the g/h and c values. Finally, an interior concentrated flux if one exists, is input separately .

Problem	Example T4L2-1	Example T4L3-1	Example T4L3-2
Units	N, m, C, s	W, m, C, s	BTU, ft, F, s
# of Segments	3	3	1
Segment 1			
Element Order	1	1	1
# of elements	1	1	2
Left End Coor.	0	0	0
Right End Coor.	0.15	0.3	0.416
$\alpha(x)$	0, 0, 5e7	0, 0, 20	0, 0, 24.8
$\beta(x)$	0, 0, 0	0, 0, 0	0, 0, 921.517
$f(x)$	0, 0, 0	0, 0, 0	0, 0, 73721.367
Segment 2			
Element Order	1	1	
# of elements	1	1	
Left End Coor.	0.15	0.3	
Right End Coor.	0.30	0.45	
$\alpha(x)$	0, 0, 5e7	0, 0, 30	
$\beta(x)$	0, 0, 0	0, 0, 0	
$f(x)$	0, 0, 0	0, 0, 0	
Segment 3			
Element Order	1	1	
# of elements	1	1	
Left End Coor.	0.30	0.45	
Right End Coor.	0.60	0.6	
$\alpha(x)$	0, 0, 8e7	0, 0, 50	
$\beta(x)$	0, 0, 0	0, 0, 0	
$f(x)$	0, 0, 0	0, 0, 0	
Left End BC	EBC	Mixed	EBC
Left End BC Value(s)	0.0	-25, 20000	150
Right End BC	EBC	EBC	NBC
Right End BC Value(s)	0.0	20	0
Concentrated Flux Data			
Location, Value	0.15, 300e3		

Program Limitations

There are no limitations except those imposed by the operating system.

Troubleshooting

The program checks the input for a variety of errors. Most of these deal with the validity of the input data. The error messages are shown on the screen. Apart from the input errors the most common error message is "Ill-posed problem…." This usually is an indication that the boundary conditions have not been imposed correctly.

Solving Linear Algebraic Equations

A class of problems in finite element analysis involves the solution of the algebraic equations

$$\mathbf{K}_{n\times n}\mathbf{D}_{n\times m} = \mathbf{F}_{n\times m} \tag{1}$$

where $\mathbf{K}_{n\times n}$ is the system matrix, $\mathbf{D}_{n\times m}$ is the matrix of primary unknowns and $\mathbf{F}_{n\times m}$ is the matrix of right-hand side (RHS) vectors. The number of unknowns is n and the number of RHS vectors is m. Typically, $n >> m$. In the context of finite element analysis, $\mathbf{K}_{n\times n}$ is mostly symmetric and positive definite. We will assume that $\mathbf{K}_{n\times n}$ is symmetric and positive definite

Equation solvers can be categorized as being either direct or iterative. Direct solvers are those that solve Eqns. (1) in a non-iterative fashion. The procedure involves one pass or two passes through the n equations. On the other hand, iterative solvers transform the problem into an equivalent problem and the solution is obtained iteratively.

There are at least four major issues in solving Eqn. (1).

(a) How much of storage space will be used?
(b) How can numerically accurate solutions be obtained?
(c) How much time will be taken to obtain the solution?
(d) How much of additional effort is needed if an additional solution (to a new right-hand side vector) vector is to be generated?

There are other issues such as parallelizing and vectorizing the solution procedure on specialized hardware and software systems but they are beyond the scope of the discussions here.

Storage Schemes

The structural stiffness matrix is usually sparse. Typically (especially for larger problems), the nonzero entries make up a few percent (1-10%) of the entire matrix. Researchers have devised several storage schemes to minimize the amount of storage space needed by recognizing that $\mathbf{K}$ is sparse and that the locations of the nonzero entries are known once the structural model is defined.

Banded and Skyline: The $\mathbf{K}$ matrix is symmetric and banded. A banded matrix can be though of as a special case of a sparse matrix. Consider the planar truss shown below (Fig. 1).

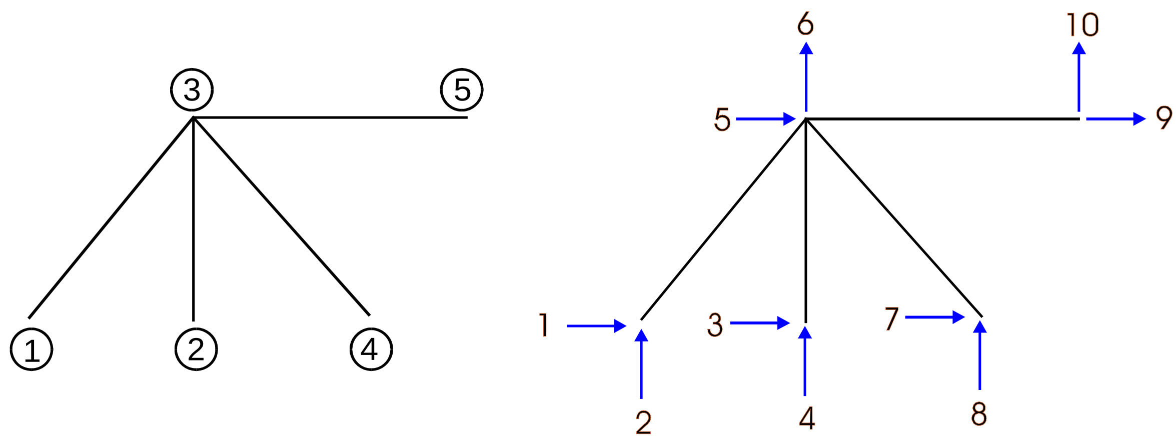


Fig. 1 A planar truss shows the node numbers and the global degrees-of-freedom at the nodes

There are 10 degrees-of-freedom in the system. If we go through the process of constructing the element equations and forming the system equations (symbolically not numerically), we can generate the map for the **K** matrix. Fig. 2 shows only the nonzero entries in the upper triangular portion of the matrix.

$$\begin{bmatrix} K_{11} & K_{12} & & & K_{15} & K_{16} & & & & \\ & K_{22} & & & K_{25} & K_{26} & & & & \\ & & K_{33} & K_{34} & K_{35} & K_{36} & & & & \\ & & & K_{44} & K_{45} & K_{46} & & & & \\ & & & & K_{55} & K_{56} & K_{57} & K_{58} & K_{59} & K_{5,10} \\ & & & & & K_{66} & K_{67} & K_{68} & K_{69} & K_{6,10} \\ & & & & & & K_{77} & K_{78} & & \\ & & & & & & & K_{88} & & \\ & & & & & & & & K_{99} & K_{9,10} \\ & & & & & & & & & K_{10,10} \end{bmatrix}$$

Fig. 2 Upper triangular portion of the system stiffness matrix

As can be deduced from Fig. 2, the HBW of the given matrix is 6. We can store this matrix as a rectangular matrix as follows.

$$\mathbf{K}_{10\times 6}^{banded} = \begin{bmatrix} K_{11} & K_{12} & 0 & 0 & K_{15} & K_{16} \\ K_{22} & 0 & 0 & K_{25} & K_{26} & 0 \\ K_{33} & K_{34} & K_{35} & K_{36} & 0 & 0 \\ K_{44} & K_{45} & K_{46} & 0 & 0 & 0 \\ K_{55} & K_{56} & K_{57} & K_{58} & K_{59} & K_{5,10} \\ K_{66} & K_{67} & K_{68} & K_{69} & K_{6,10} & 0 \\ K_{77} & K_{78} & 0 & 0 & 0 & 0 \\ K_{88} & 0 & 0 & 0 & 0 & 0 \\ K_{99} & K_{9,10} & 0 & 0 & 0 & 0 \\ K_{10,10} & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Fig. 3 Stiffness matrix stored as a banded matrix

The relationship (mapping) between the elements in the original $\mathbf{K}$ in Fig. 2 and the banded form $\mathbf{K}^{banded}$ in Fig. 3 can be derived simply as follows.

$$K_{i,j} = 0 \text{ if } \left(j - i + 1\right) > HBW \tag{2}$$

$$K_{i,j} = 0 \text{ if } \left(j < i\right) \tag{3}$$

else $$K_{i,j} \Rightarrow K_{i,j-i+1}^{banded} \tag{4}$$

As we can see in Fig. 4, a good number of the elements are zero.

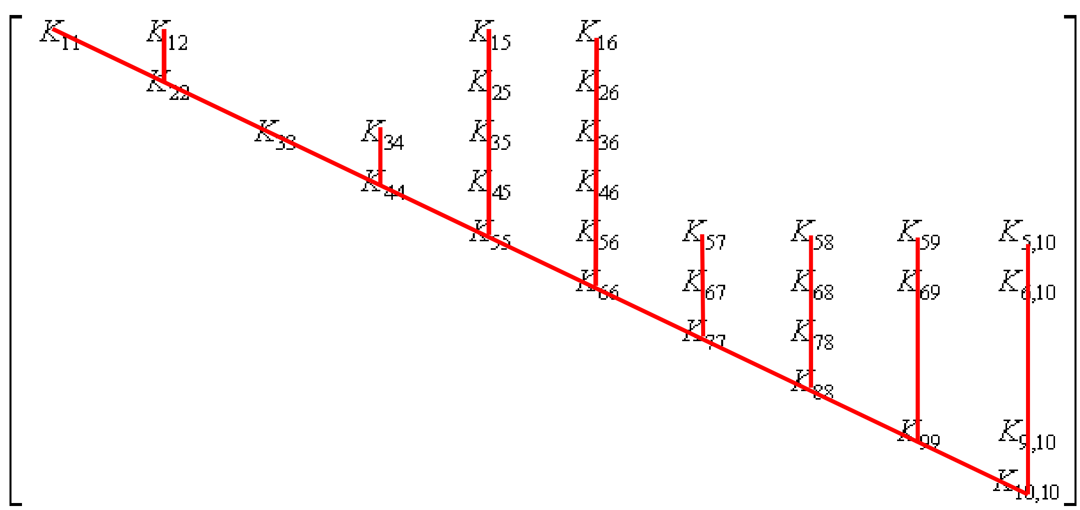


Fig. 4 The "skyline" profile

We can improve the storage scheme by storing only the elements that lie within the skyline of the matrix as shown in Fig. 4. The stiffness matrix is stored in a vector as follows.

$$\mathbf{K}_{35\times1}^{skyline} = \left\{K_{11}, K_{22}, K_{12}, K_{33}, K_{44}, K_{34}, ..., K_{10,10}, K_{9,10}, K_{8,10}, K_{7,10}, K_{6,10}, K_{5,10}\right\} \tag{5}$$

Note that each column is stored starting with the diagonal element of that column followed by all the other elements in that column until the last nonzero entry in that column. To facilitate mapping the original elements of the stiffness matrix, an additional indexing vector, $\mathbf{D}_{n+1}^{loc}$ is created that has $(n+1)$ elements. These elements store the location of the diagonal element (of each column) with the last element storing the (last element+1) in the stiffness matrix. In other words, the last element contains one more than the total number of entries in the skyline profile. Going back to the current example, we have the following $\mathbf{D}_{n+1}^{loc}$ vector.

$$\mathbf{D}_{11}^{loc} = \left\{1,2,4,5,7,12,18,21,25,30,36\right\} \tag{6}$$

The relationship (mapping) between the elements in the original $\mathbf{K}$ in Fig. 1 and the skyline form $\mathbf{K}^{skyline}$ in Fig. 4 can be derived as follows.

$$K_{i,j} = 0 \text{ if } \left(j-i\right) > \left(D_{j+1}^{loc} - D_j^{loc}\right) \tag{7}$$

$$K_{i,j} = 0 \text{ if } \left(j < i\right) \tag{8}$$

$$K_{i,j} \Rightarrow l = D_j^{loc} + j - i \Rightarrow K_l^{skyline} \tag{9}$$

For example, to locate K_{68} we note that (a) $i=6, j=8$, (b) $D_8^{loc} = 21$, (c) $l = 21+8-6 = 23$. Hence, $K_{68} \Rightarrow K_{23}^{skyline}$.

Finally, let us compare the storage requirements of the three schemes – full, banded, and skyline, assuming that the stiffness matrix is stored in the double precision format, and that two integer words make up a single double precision word $(q = hbw, m = \mathbf{D}_{n+1}^{loc} - 1)$.

Storage Scheme	**What is to be stored?**	**Equivalent integer words**
Full	$\mathbf{K}_{n\times n}$	$2n^2$
Banded	$\mathbf{K}_{n\times hbw}^{banded}$	$2nq$
Skyline	$\mathbf{K}_m^{skyline}, \mathbf{D}_{n+1}^{loc}$	$2m+n+1$

Using our example, we have the three values in the last column as $200, 120$ and 81 integer words – significant savings with increasing sophistication of the storage scheme.

Sparse: It is not evident with our simple example that the stiffness matrix is sparse. The % sparsity of the stiffness matrix is $\frac{35}{100}\times 100 = 35\%$. As the size of the problem increases, in most instances, the sparsity of the stiffness matrix increases (or the number of nonzero terms decreases). Consider the stiffness matrix shown in Fig. 4. In one of the sparse storage schemes, the stiffness matrix is stored in a vector rowwise with only the non-zero entries being stored[10] in $\mathbf{K}_{m\times 1}^{sparse}$. Using our previous example, we have the following.

$$\mathbf{K}_{31\times 1}^{sparse} = \left\{K_{11}, K_{12}, K_{15}, K_{16}, K_{22}, K_{25}, K_{26}, \ldots, K_{88}, K_{99}, K_{9,10}, K_{10,10}\right\} \tag{10}$$

To access the entries in the matrix, two additional (indexing) vectors are needed. The first, $\mathbf{C}_{m\times 1}$ is used to store the column numbers of the nonzero entries. Again, we have with our example

$$\mathbf{C}_{31\times 1} = \left\{1,2,5,6,2,5,6,3,4,5,6,\ldots,8,9,10,10\right\} \tag{11}$$

The second, $\mathbf{R}_{(n+1)\times 1}$, stores the starting location of each row. Again, we have with our example

$$\mathbf{R}_{11\times 1} = \left\{1,5,8,12,15,21,26,28,29,31,32\right\} \tag{12}$$

Discussion of sparse equation solvers is outside the scope of this document.

Offline: There are special situations when the solution procedure operates on matrices that are written on and retrieved from computer hard disk. This is typically done when the size of the stiffness matrix and other associated vectors/matrices in any storage format is several times the size of the computer's random access memory (RAM). Discussion of offline equation solvers is outside the scope of this document.

DIRECT SOLVERS

2.1 Gaussian Elimination

There are two phases to the solution – forward elimination and backward substitution. In the forward elimination phase, the basic idea is to take the set of equations from the original form

[10] It should be noted that zero entries within the skyline profile may become nonzero during the solution phase.

$$\begin{bmatrix} K_{11} & K_{12} & K_{13} & K_{1i} & K_{1n} \\ K_{21} & K_{22} & K_{23} & K_{2i} & K_{2n} \\ K_{31} & K_{32} & K_{33} & K_{3i} & K_{3n} \\ & & & & \\ K_{i1} & K_{i2} & K_{i3} & K_{ii} & K_{in} \\ & & & & \\ K_{n1} & K_{n2} & K_{n3} & K_{ni} & K_{nn} \end{bmatrix} \begin{Bmatrix} D_1 \\ D_2 \\ D_3 \\ \\ D_i \\ \\ D_n \end{Bmatrix} = \begin{Bmatrix} F_1 \\ F_2 \\ F_3 \\ \\ F_i \\ \\ F_n \end{Bmatrix} \tag{13}$$

to an upper triangular matrix of the form

$$\begin{bmatrix} K_{11} & K_{12} & K_{13} & K_{1i} & K_{1n} \\ 0 & K_{22}^{(1)} & K_{23}^{(1)} & K_{2i}^{(1)} & K_{2n}^{(1)} \\ 0 & 0 & K_{33}^{(2)} & K_{3i}^{(2)} & K_{3n}^{(2)} \\ & & & & \\ 0 & 0 & 0 & K_{ii}^{(i-1)} & K_{in}^{(i-1)} \\ & & & & \\ 0 & 0 & 0 & 0 & K_{nn}^{(n-1)} \end{bmatrix} \begin{Bmatrix} D_1 \\ D_2 \\ D_3 \\ \\ D_i \\ \\ D_n \end{Bmatrix} = \begin{Bmatrix} F_1 \\ F_2^{(1)} \\ F_3^{(2)} \\ \\ F_i^{(i-1)} \\ \\ F_n^{(n)} \end{Bmatrix} \tag{14}$$

With each step through the equations one unknown is eliminated until only D_n remains as the unknown in the equation. In step $k\left(k=1,2,...,n-1\right)$

$$K_{ij}^{(k)} = K_{ij}^{(k-1)} - \frac{K_{ik}^{(k-1)}}{K_{kk}^{(k-1)}} K_{kj}^{(k-1)} \quad i,j = k+1,...,n \tag{15}$$

$$F_i^{(k)} = F_i^{(k-1)} - \frac{K_{ik}^{(k-1)}}{K_{kk}^{(k-1)}} F_k^{(k-1)} \quad i = k+1,...,n \tag{16}$$

If $K_{kk}^{(k-1)} \le \varepsilon$, where ε is a small positive constant, then the system of equations is linearly dependent.

In the backward substitution phase (dropping the superscript)

$$D_n = \frac{F_n}{K_{nn}} \tag{17}$$

and $$D_i = \frac{F_i - \sum_{j=i+1}^{n} K_{ij} D_j}{K_{ii}} \quad i = n-1, n-2, ..., 1 \tag{18}$$

Algorithm

1: Forward Elimination. Loop through rows, $k = 1, \ldots, n-1$.

2: Check if $\left|K_{kk}\right| < \varepsilon$. If yes, stop.

3: Loop through columns, $i = k+1, \ldots, n$.

4: Compute constant, $c = \dfrac{K_{ik}}{K_{kk}}$.

5: Loop through $j = k+1, \ldots, n$.

6: Set $K_{ij} = K_{ij} - cK_{kj}$.

7: End loop j.

8: Set $F_i = F_i - cF_k$.

9: End loop i.

10: End loop k.

11: Backward substitution. Set $D_n = F_n / K_{nn}$.

12: Loop through all rows, $i = n-1, \ldots, 1$.

13: Compute $sum = \sum_{j=i+1}^{n} K_{ij} D_j$.

14: Compute $D_i = \dfrac{F_i - sum}{K_{ii}}$.

15: End loop i.

Example 1
Solve the following set of equations using Gaussian Elimination method.

$$\begin{bmatrix} 10 & -5 & 2 \\ 3 & 20 & 5 \\ -2 & 7 & 15 \end{bmatrix} \begin{Bmatrix} D_1 \\ D_2 \\ D_3 \end{Bmatrix} = \begin{Bmatrix} 6 \\ 58 \\ 57 \end{Bmatrix}$$

Solution

Forward Substitution

Noting that $n = 3$, the successive snapshots as we go through the algorithm are as follows.

$$k = 1 \Rightarrow \begin{bmatrix} 10 & -5 & 2 \\ & 21.5 & 4.4 \\ & 6 & 15.4 \end{bmatrix} \begin{Bmatrix} D_1 \\ D_2 \\ D_3 \end{Bmatrix} = \begin{Bmatrix} 6 \\ 56.2 \\ 58.2 \end{Bmatrix}$$

$$k = 2 \Rightarrow \begin{bmatrix} 10 & -5 & 2 \\ & 21.5 & 4.4 \\ & & 14.172 \end{bmatrix} \begin{Bmatrix} D_1 \\ D_2 \\ D_3 \end{Bmatrix} = \begin{Bmatrix} 6 \\ 56.2 \\ 42.5163 \end{Bmatrix}$$

Backward Substitution

$$D_3 = \frac{42.5163}{14.172} = 3$$

$$i = 2 \Rightarrow D_2 = \frac{56.2 - 4.4(3)}{21.5} = 2$$

$$i = 1 \Rightarrow D_1 = \frac{6 - (-4)}{10} = 1$$

Hence the solution is $\mathbf{D}_{3\times1} = \begin{Bmatrix} 1 \\ 2 \\ 3 \end{Bmatrix}$.

2.2 Cholesky Decomposition or $\mathbf{L\hat{D}L^T}$ Factorization (or, Decomposition)

The matrix $\mathbf{A}$ in $\mathbf{Ax}=\mathbf{b}$ can be factored as $\mathbf{A}=\mathbf{LU}$ where $\mathbf{L}$ is a lower triangular matrix with 1's on the diagonals and $\mathbf{U}$ is an upper triangular matrix. The proof can be found in a book on linear algebra or numerical analysis. When $\mathbf{A}$ is symmetric and positive definite as is $\mathbf{K}$ in Eqn. (1), the matrix can be factored as $\mathbf{K}=\mathbf{L\hat{D}L^T}$ where $\mathbf{L}$ is a lower triangular matrix with 1's on the diagonals and $\mathbf{\hat{D}}$ is a diagonal matrix with positive entries. The $\mathbf{L\hat{D}L^T}$ decomposition is a variation of Cholesky Decomposition, and provides a very effective solution to the system equilibrium equations.

The solution proceeds as follows. Starting with Eqns. (1) we first factor $\mathbf{K}=\mathbf{L\hat{D}L^T}$ by finding the $\mathbf{L}$ and $\mathbf{\hat{D}}$ matrices.

$$\begin{bmatrix} K_{11} & K_{12} & . & K_{1n} \\ K_{12} & K_{22} & . & K_{2n} \\ . & . & . & . \\ K_{1n} & K_{2n} & . & K_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & . & 0 \\ L_{21} & 1 & . & 0 \\ . & . & . & . \\ L_{n1} & L_{n2} & . & 1 \end{bmatrix} \begin{bmatrix} \hat{D}_1 & 0 & . & 0 \\ 0 & \hat{D}_2 & . & 0 \\ . & . & . & . \\ 0 & 0 & . & \hat{D}_n \end{bmatrix} \begin{bmatrix} 1 & L_{21} & . & L_{n1} \\ 0 & 1 & . & L_{n2} \\ . & . & . & . \\ 0 & 0 & . & 1 \end{bmatrix}$$

$$= \left[\begin{array}{c|c|c|c|c} \hat{D}_1 & \hat{D}_1 L_{21} & \hat{D}_1 L_{31} & . & \hat{D}_1 L_{n1} \\ \hline & \hat{D}_1 L_{21}^2 + \hat{D}_2 & \hat{D}_1 L_{21} L_{31} + \hat{D}_2 L_{32} & . & \hat{D}_1 L_{21} L_{n1} + \hat{D}_2 L_{n2} \\ \hline & & \hat{D}_1 L_{31}^2 + \hat{D}_2 L_{32}^2 + \hat{D}_3 & . & \hat{D}_1 L_{31} L_{n1} + \hat{D}_2 L_{32} L_{n2} + \hat{D}_3 L_{n3} \\ \hline & & & . & . \\ \hline sym & & & & \hat{D}_1 L_{n1}^2 + \hat{D}_2 L_{n2}^2 + \ldots + \hat{D}_n \end{array} \right] \tag{19}$$

In other words, we have the following.

$$\mathbf{L\hat{D}L^T D} = \mathbf{F} \tag{20}$$

Let $\mathbf{LQ}=\mathbf{F}$ (21)

Then $\mathbf{\hat{D}L^T D}=\mathbf{Q}$ (22)

Note that $\mathbf{\hat{D}L^T}$ is of the upper triangular form.

$$\hat{\mathbf{D}}\mathbf{L}^{\mathbf{T}} = \begin{bmatrix} \hat{D}_1 & \hat{D}_1 L_{12} & . & \hat{D}_1 L_{1n} \\ & \hat{D}_2 & . & \hat{D}_2 L_{2n} \\ & & . & . \\ \mathbf{0} & & & \hat{D}_n \end{bmatrix} \tag{23}$$

We can solve Eqns. (21) for $\mathbf{Q}$ through a process known as forward substitution starting with the first equation that has one unknown, Q_1, the second equation that then has one unknown, Q_2 and so on until the last equation that is used to find Q_n. Once $\mathbf{Q}$ has been computed, we can solve Eqns. (22) through a process known as backward substitution by starting with the last equation that has one unknown, D_n, then the second last equation that then has one unknown, D_{n-1}, and so on until the first equation that is used to compute D_1.

By comparing the RHS of Eqn. (19) with the LHS we have the mechanism to compute the $\mathbf{L}$ and $\hat{\mathbf{D}}$ matrices given the $\mathbf{K}$ matrix.

Algorithm

Step 1: Cholesky Factorization. Loop through rows, $i = 1,\ldots,n$.

Step 2: Set $\hat{D}_i = K_{ii} - \sum_{j=1}^{i-1} L_{ij}^2 \hat{D}_j$. If $\hat{D}_i < \varepsilon$, stop. The matrix is not positive definite.

Step 3: For $j = i+1,\ldots,n$, set $L_{ji} = \dfrac{K_{ji} - \sum_{k=1}^{i-1} L_{jk}\hat{D}_k L_{ik}}{\hat{D}_i}$.

Step 4: End loop i. This ends the factorization phase.

Step 5: Forward Substitution. Set $Q_1 = F_1$.

Step 6: For $i = 2,\ldots,n$, set $Q_i = F_i - \sum_{j=1}^{i-1} L_{ij} Q_j$. This ends the Forward Substitution phase.

Step 7: Backward Substitution. Set $D_n = \dfrac{Q_n}{\hat{D}_n}$.

Step 8: For $i = n-1,\ldots,1$, set $D_i = \dfrac{Q_i}{\hat{D}_i} - \sum_{j=i+1}^{n} L_{ij} D_j$. This ends the Backward Substitution phase.

A careful examination of the steps will show that no extra storage is required. The storage locations in $\mathbf{K}$ can be used to store both $\hat{\mathbf{D}}$ and $\mathbf{L}$. Similarly, the storage locations in $\mathbf{D}$ can be used to store the elements of $\mathbf{Q}$.

Example 2

Solve the following set of equations using Cholesky Decomposition method.

$$\begin{bmatrix} 3.5120 & 0.7679 & 0 & 0 & 0 \\ 0.7679 & 3.1520 & 0 & -2 & 0 \\ 0 & 0 & 3.5120 & -0.7679 & 0.7679 \\ 0 & -2 & -0.7679 & 3.1520 & -1.1520 \\ 0 & 0 & 0.7679 & -1.1520 & 3.1520 \end{bmatrix} \begin{Bmatrix} D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 0 \\ -0.04 \\ 0 \end{Bmatrix}$$

Solution

Factorization $(n = 5)$

$i = 1 \Rightarrow \hat{D}_1 = K_{11} = 3.5120$.

$j = 2 \Rightarrow L_{21} = \dfrac{K_{21}}{\hat{D}_1} = \dfrac{0.7679}{3.5120} = 0.21865$. Also, $L_{31} = L_{41} = L_{51} = 0$.

$i = 2 \Rightarrow \hat{D}_2 = K_{22} - L_{21}^2 \hat{D}_1 = 3.1520 - (0.21865)^2 (3.5120) = 2.9841.$

$j = 3 \Rightarrow L_{32} = \dfrac{K_{32} - L_{31}\hat{D}_1 L_{21}}{\hat{D}_2} = 0$.

$j = 4 \Rightarrow L_{42} = \dfrac{K_{42} - L_{41}\hat{D}_1 L_{21}}{\hat{D}_2} = \dfrac{-2-0}{2.9841} = -0.670219$. Also, $L_{52} = 0$.

$i = 3 \Rightarrow \hat{D}_3 = K_{33} - L_{31}^2 \hat{D}_1 - L_{32}^2 \hat{D}_2 = 3.5120 - 0 - 0 = 3.5120$.

$j = 4 \Rightarrow L_{43} = \dfrac{K_{43} - L_{41}\hat{D}_1 L_{31} - L_{42}\hat{D}_2 L_{32}}{\hat{D}_3} = \dfrac{-0.7679 - 0 - 0}{3.5120} = -0.21865.$

$j = 5 \Rightarrow L_{53} = \dfrac{K_{53} - L_{51}\hat{D}_1 L_{31} - L_{52}\hat{D}_2 L_{32}}{\hat{D}_3} = \dfrac{0.7679 - 0 - 0}{3.5120} = 0.21865.$

$$i = 4 \Rightarrow \hat{D}_4 = K_{44} - L_{41}^2\hat{D}_1 - L_{42}^2\hat{D}_2 - L_{43}^2\hat{D}_3$$

$$= 3.1520 - 0 - (-0.670219)^2(2.9841) - (-0.21865)^2(3.5120) = 1.64366$$

$$j = 5 \Rightarrow L_{54} = \frac{K_{54} - L_{51}\hat{D}_1 L_{41} - L_{52}\hat{D}_2 L_{42} - L_{53}\hat{D}_3 L_{43}}{\hat{D}_4}$$

$$= \frac{-1.1520 - 0 - 0 - (0.21865)(3.5120)(-0.21865)}{1.64366} = \text{-}0.598724$$

$$i = 5 \Rightarrow \hat{D}_5 = K_{55} - L_{51}^2\hat{D}_1 - L_{52}^2\hat{D}_2 - L_{53}^2\hat{D}_3 - L_{54}^2\hat{D}_4$$

$$= 3.1520 - 0 - 0 - (0.21865)^2(3.5120) - (-0.598724)^2(1.64366) = 2.3949$$

Forward Substitution $(\mathbf{LQ} = \mathbf{F})$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.21865 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & -0.670219 & -0.21865 & 1 & 0 \\ 0 & 0 & 0.21865 & -0.598724 & 1 \end{bmatrix} \begin{Bmatrix} Q_1 \\ Q_2 \\ Q_3 \\ Q_4 \\ Q_5 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 0 \\ -0.04 \\ 0 \end{Bmatrix}$$

$$i = 1 \Rightarrow Q_1 = F_1 = 0$$

$$i = 2 \Rightarrow Q_2 = F_2 - L_{21}Q_1 = 0$$

$$i = 3 \Rightarrow Q_3 = F_3 - L_{31}Q_1 - L_{32}Q_2 = 0$$

$$i = 4 \Rightarrow Q_4 = F_4 - L_{41}Q_1 - L_{42}Q_2 - L_{43}Q_3 = -0.04$$

$$i = 5 \Rightarrow Q_5 = F_5 - L_{51}Q_1 - L_{52}Q_2 - L_{53}Q_3 - L_{54}Q_4 = 0 - (-0.598724)(-0.04) = -0.023949$$

Backward Substitution $(\hat{\mathbf{D}}\mathbf{L}^{\mathbf{T}}\mathbf{D} = \mathbf{Q})$

$$\begin{bmatrix} 3.5120 & 0.21865 & 0 & 0 & 0 \\ 0 & 2.9841 & 0 & -0.670219 & 0 \\ 0 & 0 & 3.5120 & -0.21865 & 0.21865 \\ 0 & 0 & 0 & 1.64366 & -0.598724 \\ 0 & 0 & 0 & 0 & 2.3949 \end{bmatrix} \begin{Bmatrix} D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 0 \\ -0.04 \\ -0.023949 \end{Bmatrix}$$

$$i = 5 \Rightarrow D_5 = \frac{-0.023949}{2.3949} = -0.01$$

$$i = 4 \Rightarrow D_4 = \frac{Q_4}{\hat{D}_4} - L_{45}D_5 = \frac{-0.04}{1.64366} - (-0.598724)(-0.01) = -0.0303232$$

$$i = 3 \Rightarrow D_3 = \frac{Q_3}{\hat{D}_3} - L_{34}D_4 - L_{35}D_5$$

$$= \frac{0}{3.5120} - (-0.21865)(-0.0303232) - (0.21865)(-0.01) = -0.00444367$$

$$i = 2 \Rightarrow D_2 = \frac{Q_2}{\hat{D}_2} - L_{23}D_3 - L_{24}D_4 - L_{25}D_5$$

$$= \frac{0}{2.9841} - 0 - (-0.670219)(-0.0303232) - 0 = -0.0203232$$

$$i = 1 \Rightarrow D_1 = \frac{Q_1}{\hat{D}_1} - L_{12}D_2 - L_{13}D_3 - L_{14}D_4 - L_{15}D_5$$

$$= \frac{0}{3.5120} - (0.21865)(-0.0203232) - 0 - 0 - 0 = 0.00444367$$

It should also be noted that the Cholesky Decomposition requires no special effort to solve additional RHS vectors since the factorization and the forward/backward substitutions steps are separate. Once the factorization step is completed (**once**), the forward/backward substitutions steps can be repeated as many times as required.

ITERATIVE SOLVERS

Consider a quadratic form $F(\mathbf{d})$ given by

$$F(\mathbf{d}) = \mathbf{d}^T\mathbf{K}\mathbf{d} - 2\mathbf{d}^T\mathbf{f} \tag{24}$$

where $\mathbf{K}$ is a positive definite matrix. Minimizing the quadratic form yields

$$\mathbf{Kd} = \mathbf{f} \tag{25}$$

The idea of minimizing the quadratic form obviously leads to unconstrained optimization techniques. Consider the Steepest Descent Method. Let $\mathbf{s}$ be a vector such that

$$F(\mathbf{d}_k) = F(\mathbf{d}_k + \mathbf{s}) \tag{26}$$

Using Eqn. (24), we have

$$F(\mathbf{d}_k + \mathbf{s}) = (\mathbf{d}_k + \mathbf{s})^T \mathbf{K}\left(\mathbf{d}_k + \mathbf{s}\right) - 2\left(\mathbf{d}_k + \mathbf{s}\right)^T \mathbf{f} \tag{27}$$

If $\mathbf{s}$ is considered to be small, the second order terms can be neglected and the above equation can be simplified as

$$F(\mathbf{d}_k + \mathbf{s}) = F\left(\mathbf{d}_k\right) + \mathbf{s}^T\mathbf{Kd}_k - \mathbf{d}_k^T\mathbf{Ks} - 2\mathbf{s}^T\mathbf{f} \tag{28}$$

However, due to Eqn. (26) we can rewrite the above equation as

$$\mathbf{s}^T\left(\mathbf{Kd}_k - \mathbf{f}\right) + \left(\mathbf{Kd}_k - \mathbf{f}\right)^T \mathbf{s} = 0 \tag{29}$$

The residual vector

$$\mathbf{g}_k = \mathbf{Kd}_k - \mathbf{f} \tag{30}$$

is thus orthogonal to any small $\mathbf{s}$ whose addition does not alter F. The steepest descent direction is the direction of $\mathbf{g}_k$. To minimize F, we can think of moving along the $\mathbf{g}_k$ direction. In other words, we are looking for the point $\mathbf{d}_k + \alpha_k\mathbf{g}_k$ by trying to find the value of α_k that minimizes $F(\mathbf{d}_k + \alpha_k\mathbf{g}_k)$.

$$\begin{aligned} F(\mathbf{d}_{k+1}) &= F(\mathbf{d}_k + \alpha_k\mathbf{g}_k) \\ &= F(\mathbf{d}_k) + \alpha_k\mathbf{g}_k^T\mathbf{Kd}_k + \alpha_k\mathbf{d}_k^T\mathbf{Kg}_k + \alpha_k^2\mathbf{g}_k^T\mathbf{Kg}_k + 2\alpha_k\mathbf{g}_k^T\mathbf{f} \end{aligned} \tag{31}$$

Minimizing

$$\frac{d}{d\alpha_k} F\left(d_k + \alpha_k g_k\right) = 0 \tag{32}$$

yields

$$\mathbf{g}_k^T\mathbf{Kd}_k + \alpha_k\mathbf{g}_k^T\mathbf{Kg}_k - \mathbf{g}_k^T\mathbf{f} = 0 \tag{33}$$

or

$$\alpha_k = \frac{\mathbf{g}_k^T\mathbf{g}_k}{\mathbf{g}_k^T\mathbf{Kg}_k} \tag{34}$$

3.1 Conjugate Gradient Method

The Steepest Descent Method while providing a good start, is not effective. It takes far too many iterations. Error corrected in one iteration is likely to be introduced in the next iteration since successive directions are not mutually orthogonal. The Conjugate Gradient Method avoids this difficulty. Successive solutions are found by

$$\mathbf{d}_{k+1} = \mathbf{d}_k + \alpha_k\mathbf{s}_k \tag{35}$$

where the direction vectors $\mathbf{s}_k$ are computed so that successive residuals are orthogonal to each other as

$$\mathbf{g}_k^T \mathbf{g}_{k+1} = 0 \tag{36}$$

The correction step is computed as follows

$$\begin{aligned} F(\mathbf{d}_{k+1}) &= F(\mathbf{d}_k + \alpha_k \mathbf{s}_k) \\ &= F(\mathbf{d}_k) + \alpha_k \mathbf{s}_k^T \mathbf{K} \mathbf{d}_k + \alpha_k \mathbf{d}_k^T \mathbf{K} \mathbf{s}_k + \alpha_k^2 \mathbf{s}_k^T \mathbf{K} \mathbf{s}_k + 2\alpha_k \mathbf{s}_k^T \mathbf{f} \end{aligned} \tag{37}$$

As before, minimizing F leads to

$$\mathbf{s}_k^T \mathbf{K} \mathbf{d}_k + \alpha_k \mathbf{s}_k^T \mathbf{K} \mathbf{g}_k - \mathbf{s}_k^T \mathbf{f} = 0 \tag{38}$$

or

$$\alpha_k = \frac{\mathbf{s}_k^T \mathbf{s}_k}{\mathbf{s}_k^T \mathbf{K} \mathbf{s}_k} \tag{39}$$

This expression is valid for any direction $\mathbf{s}_k$. Multiplying Eqn. (35) by $\mathbf{K}$ and subtracting $\mathbf{f}$ on both sides yields

$$\mathbf{g}_{k+1} = \mathbf{g}_k + \alpha_k \mathbf{K} \mathbf{s}_k \tag{40}$$

Using the orthogonality relationship in Eqn. (36) and premultiplying both sides of Eqn. (40) by $\mathbf{g}_k$ yields an alternate form of Eqn. (39) (one that ensures that successive directions are orthogonal) as

$$\alpha_k = \frac{\mathbf{g}_k^T \mathbf{g}_k}{\mathbf{g}_k^T \mathbf{K} \mathbf{g}_k} \tag{41}$$

Rewriting Eqn. (38) we have

$$\mathbf{s}_k^T \left(\mathbf{K} \mathbf{d}_k - \mathbf{f} \right) + \alpha_k \mathbf{s}_k^T \mathbf{K} \mathbf{g}_k = 0 \tag{42}$$

Using Eqn. (40) we have

$$\mathbf{s}_k^T \mathbf{g}_{k+1} = 0 \tag{43}$$

To create the search direction $\mathbf{s}_{k+1}$, we start by taking the residual $\mathbf{g}_{k+1}$ as we discussed in the steepest descent method. Adding a component orthogonal to $\mathbf{g}_{k+1}$ just large enough to satisfy Eqn. (35) we have

$$\mathbf{s}_{k+1} = \mathbf{g}_k + \beta_k \mathbf{s}_k \tag{44}$$

where the scaling constant β_k is yet to be determined. Premultiplying both sides by $\mathbf{s}_k^T \mathbf{K}$, Eqn. (35) becomes

$$\mathbf{s}_k^T \mathbf{K} \mathbf{d}_{k+1} = \mathbf{s}_k^T \mathbf{K} \mathbf{g}_{k+1} + \beta_k \mathbf{s}_k^T \mathbf{K} \mathbf{s}_k \tag{45}$$

and the multiplier β_k is given by

$$\beta_k = \frac{\mathbf{s}_k^T \mathbf{K} \mathbf{g}_{k+1} - \mathbf{s}_k^T \mathbf{K} \mathbf{s}_{k+1}}{\mathbf{s}_k^T \mathbf{K} \mathbf{s}_k} \tag{46}$$

Orthogonality of residuals (Eqn. (35)) is the same as requiring Eqns. (39) and (41) to yield the same value for α_k. Premultiplying Eqn. (44) by $\mathbf{g}_{k+1}^T$

$$\mathbf{g}_{k+1}^T \mathbf{s}_{k+1} = \mathbf{g}_{k+1}^T \mathbf{g}_k + \beta_k \mathbf{g}_{k+1}^T \mathbf{s}_k \tag{47}$$

Using Eqn. (43), we have

$$\mathbf{g}_{k+1}^T \mathbf{g}_{k+1} = \mathbf{g}_k^T \mathbf{s}_k \tag{48}$$

or

$$\mathbf{g}_k^T \mathbf{g}_k = \mathbf{g}_k^T \mathbf{s}_k \tag{49}$$

Using this result in Eqn. (41), we have

$$\alpha_k = \frac{\mathbf{s}_k^T \mathbf{g}_k}{\mathbf{g}_k^T \mathbf{K} \mathbf{s}_k} \tag{50}$$

Substitute Eqn. (44) in (55)

$$\alpha_k = \frac{\mathbf{s}_k^T \mathbf{g}_k}{\mathbf{s}_k^T \mathbf{K} \mathbf{s}_k - \beta_{k-1} \mathbf{s}_{k-1}^T \mathbf{K} \mathbf{s}_k} \tag{51}$$

If $\mathbf{s}_k$ and $\mathbf{s}_{k+1}$ are made to be conjugate to each other with respect to $\mathbf{K}$ as in

$$\mathbf{s}_{k+1}^T \mathbf{K} \mathbf{s}_k = 0 \tag{52}$$

then the second term in the denominator of Eqn. (51) vanishes and Eqns. (39) and (41) are identical. This requirement is met by setting

$$\beta_k = -\frac{\mathbf{s}_k^T \mathbf{K} \mathbf{g}_{k+1}}{\mathbf{s}_k^T \mathbf{K} \mathbf{s}_k} \tag{53}$$

Eqn. (52) is the one that gives the method its name.

ALGORITHM (FOR A SINGLE RHS VECTOR)

Start with an initial guess $\mathbf{d}_0$. Set $k = 0$. Set ε as convergence tolerance and $n_{\max}$ as the maximum number of iterations.

1. Construct $\mathbf{g_0} = \mathbf{K}\mathbf{d_0} - \mathbf{f}$ and $\mathbf{s}_0 = -\mathbf{g}_0$.
2. Now form $\alpha_k = \dfrac{\mathbf{g}_k^T \mathbf{g}_k}{\mathbf{s}_k^T \mathbf{K} \mathbf{s}_k}$.
3. Update the solution vector as $\mathbf{d}_{k+1} = \mathbf{d}_k + \alpha_k \mathbf{s}_k$.
4. Now update $\mathbf{g}_{k+1} = \mathbf{g}_k + \alpha_k \mathbf{K} \mathbf{s}_k$.
5. Check if $\mathbf{g}_k^T \mathbf{g}_k \le \varepsilon$ and if so terminate iterations. Otherwise form $\beta_k = \dfrac{\mathbf{g}_{k+1}^T \mathbf{g}_{k+1}}{\mathbf{g}_k^T \mathbf{g}_k}$.

6. Update $\mathbf{s}_{k+1} = -\mathbf{g}_{k+1} + \beta_k \mathbf{s}_k$.
7. Increment k. Check if $k < n_{\max}$. If so, go to Step 2. Otherwise terminate iterations.

3.2 Preconditioned Conjugate Gradient Method

The basic idea in preconditioning is to devise a scheme to accelerate the convergence process. We would like to generate a preconditioning matrix, $\mathbf{B}$ such that

$$\mathbf{B}^{-1}\mathbf{K} \approx \mathbf{I}$$

The area of finding the best pre-conditioner is a very active research area. One of the simplest scheme is called Jacobi preconditioning and is discussed next.

Jacobi Preconditioning: Assume that $\mathbf{B}^{-1} = diag\left(\frac{1}{K_{ii}}\right)$.

Start with an initial guess $\mathbf{d}_0 = 0$. Set $k = 0$. Set ε as convergence tolerance and $n_{\max}$ as the maximum number of iterations.

1. Construct $\mathbf{r_0} = \mathbf{f}$, $\mathbf{z}_0 = \mathbf{B}^{-1}\mathbf{r}_0$ and $\mathbf{s}_0 = \mathbf{z}_0$.
2. Now form $\alpha_k = \dfrac{\mathbf{r}_k^T \mathbf{z}_k}{\mathbf{s}_k^T \mathbf{K} \mathbf{s}_k}$.
3. Update the solution vector as $\mathbf{d}_{k+1} = \mathbf{d}_k + \alpha_k \mathbf{s}_k$.
4. Now update $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k \mathbf{K} \mathbf{s}_k$.
5. Check if $\mathbf{r}_k^T \mathbf{r}_k \le \varepsilon$ and if so terminate iterations. Otherwise form $\beta_k = \dfrac{\mathbf{r}_{k+1}^T \mathbf{z}_{k+1}}{\mathbf{r}_k^T \mathbf{z}_k}$.
6. Form $\mathbf{z}_{i+1} = \mathbf{B}^{-1}\mathbf{r}_{i+1}$
7. Update $\mathbf{s}_{k+1} = \mathbf{z}_{k+1} + \beta_k \mathbf{s}_k$.
8. Increment k. Check if $k \le n_{\max}$. If so, go to Step 2. Otherwise terminate iterations.

Index