An Introduction to Finite Element Analysis

Barna Szabó
Washington University in St. Louis

Ivo Babuška
The University of Texas at Austin

August 6, 2009

"If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is." John von Neumann

Preface

Increasingly, engineering decisions are based on computed information with the expectation that the computed information provides a reliable quantitative estimate of some attributes of a physical system or process. The question of how much reliance on computed information can be justified is being asked with increasing frequency and urgency. Assurance of the reliability of computed information has two key aspects: (a) Selection of a suitable mathematical model and (b) approximation of the solution of the corresponding mathematical problem. The process by which it is ascertained that a mathematical model meets necessary criteria for acceptance (i.e., it is not unsuitable for purposes of analysis) is called *validation*. The process by which it is ascertained that the approximate solution, as well as the data computed from the approximate solution, meet necessary conditions for acceptance, given the goals of computation, is called *verification*. This book addresses the problems of verification and validation.

Obtaining approximate solutions for linear models with guaranteed accuracy was one of the primary objectives of research in the period 1970-1985. An important result was that exponential rates of convergence can be achieved by the *hp*-version of the finite element method for a large and important class of problems that includes problems of elasticity, heat conduction and similar problems. This made it feasible to estimate and control the errors of discretization for many practical problems.

Since the mid-1980s the problems of proper model selection and control of modeling errors came to the forefront of research. The concepts of hierarchic models and modeling strategies have been developed. This subject is now sufficiently mature to make practical applications in some very important areas possible.

The distinguishing feature of this book is that it presents a systematic treatment of verification procedures, illustrated by examples. We believe that users of finite element analysis (FEA) software products must have a basic understanding of how mathematical models are constructed; what are the essential assumptions incorporated in a mathematical model; what is the algorithmic structure of the finite element method; how the discretization parameters affect the accuracy of the finite element solution; how the accuracy of the computed data can be assessed, and how to avoid common pitfalls and mistakes. Our primary objective in assembling the material presented in this book is to provide a basic working knowledge of the finite element method. A commercial

FEA software product called StressCheck[1] is provided with the book to enable readers to perform computational experiments. Another important objective of this book is to prepare readers to follow and understand new developments in the field of FEA through continued self-study.

Engineering students typically take only one course in finite element analysis, consisting of approximately 15 weeks of instruction (45 lecture hours). We organized the material in this book so as to make efficient use of the available time. The book is written in such a way that the prerequisites are minimal. Junior standing in engineering with some background in potential flow and strength of materials are sufficient. For this reason the mathematical content is focused on the introduction of the essential concepts and terminology necessary for understanding applications of FEA in elasticity and heat conduction. Some key theorems are proven in a simple setting.

We would like to thank Dr. Norman F. Knight, Jr. and Dr. Sebastian Nervi for reviewing and commenting on the manuscript.

Barna Szabó
Washington University in St. Louis

Ivo Babuška
The University of Texas at Austin

[1] StressCheck is a trademark of Engineering Software Research and Development, Inc., St. Louis, Missouri, USA.

Contents

Chapter 1

Introduction

Engineering decision-making processes increasingly rely on information computed from approximate solutions of mathematical models. Engineering decisions have legal and ethical implications. The standard applied in legal proceedings in civil cases in the United States is to have opinions, recommendations and decisions be "based upon a reasonable degree of engineering certainty". Codes of ethics of engineering societies impose higher standards. For example the Code of Ethics of the Institute of Electrical and Electronics Engineers (IEEE) requires members "to accept responsibility in making engineering decisions consistent with the safety, health, and welfare of the public, and to disclose promptly factors that might endanger the public or the environment" and "to be honest and realistic in stating claims or estimates based on available data".

An important challenge before the computational engineering community is to establish procedures for creating evidence that will show, with a high degree of certainty, that a mathematical model of some physical reality, formulated for a particular purpose, can in fact represent the physical reality in question with sufficient accuracy to make predictions based on mathematical models useful and justifiable for the purposes of engineering decision-making and the errors in the numerical approximation are sufficiently small. There is a large and rapidly growing body of work on this subject. See, for example, [48], [37]. The formulation and numerical treatment of mathematical models for use in support of engineering decision-making in the field of solid mechanics is addressed in a recently published document issued by the American Society of Mechanical Engineers (ASME) and adopted by the American National Standards Institute (ANSI) [26].

The considerations underlying the selection of mathematical models and methods for estimation and control of modeling errors and the errors of discretization are the two main topics of this book. In this chapter a brief overview is presented and the basic terminology is introduced.

1.1 Numerical simulation

The goal of numerical simulation is to make predictions concerning the response of physical systems to various kinds of excitation and, based on those predictions, make informed decisions. To achieve this goal, mathematical models are defined and the corresponding numerical solutions are computed. Mathematical models should be understood to be idealized representations of reality and should never be confused with the physical reality that they are supposed to represent.

The choice of a mathematical model depends on its intended use: What aspects of physical reality are of interest? What data must be predicted? What accuracy is required? The main elements of numerical simulation and the associated errors are indicated schematically in Fig. 1.1.

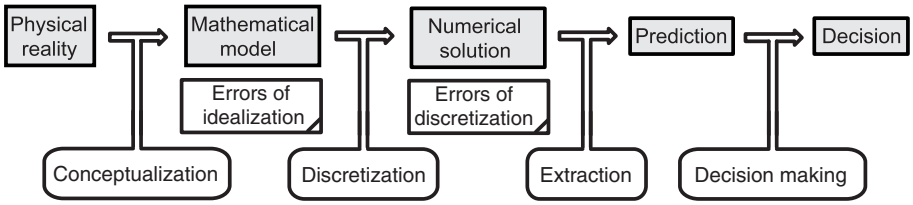

Figure 1.1: The main elements of numerical simulation and the associated errors.

Some errors are associated with the mathematical model and some errors are associated with its numerical solution. These are called errors of idealization and errors of discretization respectively. For predictions to be accurate both kinds of errors have to be controlled. The errors of idealization are also called modeling errors. Conceptualization is a process by which a mathematical model is formulated. Discretization is a process by which the exact solution of the mathematical model is approximated. Extraction is a process by which the data of interest are computed from the approximate solution.

1.1.1 Conceptualization

Mathematical models are operators that transform one set of data, the input, into another set, the output. In solid mechanics, for example, one is typically interested in predicting displacements, strains stresses, stress intensity factors, limit loads, natural frequencies, etc., given a description of the solution domain, constitutive equations and boundary conditions (loading and constraints). Common to all models are the equations that represent the conservation of momentum (in static problems the equations of equilibrium), the strain-displacement relations and constitutive laws.

The end product of conceptualization is a mathematical model. Definition of a mathematical model involves specification of the following:

1. Theoretical formulation. The applicable physical laws, together with certain simplifications, are stated as a mathematical problem in the form

of ordinary or partial differential equations, or extremum principles. For example, the classical differential equation for elastic beams is derived from the assumptions of the theory of elasticity supplemented by the assumption that the transverse variation of the longitudinal components of the displacement vector can be approximated by a linear function without significantly affecting the data of interest, typically the displacements, bending moments, shear forces etc.

2. Specification of the input data. The input data are comprised of the following:

   (a) Data that characterize the solution domain. In engineering practice solution domains are usually constructed by means of computer-aided design (CAD) tools. CAD tools produce idealized representations of real objects. The details of idealization depend on the choice of the CAD tool and the training and preferences of its operator.

   (b) Physical properties (elastic moduli, yield stress, coefficients of thermal expansion, thermal conductivities, etc.)

   (c) Boundary conditions (loads, constraints, prescribed temperatures, etc.)

   (d) Information or assumptions concerning the reference state and the initial conditions

   (e) Uncertainties. When some information needed in the formulation of a mathematical model is unknown then the uncertainty is said to be cognitive (also called epistemic). For example, the magnitude and distribution of residual stresses is usually unknown, some physical properties may be unknown, etc. Statistical uncertainties (also called aleatory uncertainties) are always present: Even when the average values of needed physical properties, loading and other data are known, there are statistical variations, possibly very substantial variations, in these data. Consideration of uncertainties is necessary for proper interpretation of the computed information.

3. Statement of objectives. Definition of the data of interest and the corresponding permissible error tolerances.

Conceptualization involves the application of expert knowledge, virtual experimentation and calibration.

Application of expert knowledge

Depending on the intended use of the model and the required accuracy of prediction, various simplifying assumptions are introduced. For example, the assumptions of the linear theory of elasticity, along with simplifying assumptions concerning the domain and the boundary conditions, are widely used in mechanical and structural engineering applications. In many applications further

simplifications are introduced, resulting in beam, plate, and shell models, planar models, axisymmetric models, each of which impose additional restrictions on what boundary conditions can be specified and what data can be computed from the solution.

In the engineering literature the commonly used simplified models are grouped into separate model classes, called theories. For example, various beam, plate and shell theories have been developed. The formulation of these theories typically involves a statement on the assumed mode of deformation (e.g., plane sections remain plane and normal to the mid-surface of a deformed beam), the relationship between the functions that characterize the deformation and the strain tensor (e.g., the strain is proportional to the curvature and the distance from the neutral axis), application of Hooke's law, and statement of the equations equilibrium.

In undergraduate engineering curricula each model class is presented as a thing in itself and consequently there is a strong predisposition in the engineering community to view each model class as a separate entity. It is much more useful however to view any mathematical model as a special case of a more comprehensive model, rather than a member of a conventionally defined model class. For example, the usual beam, plate and shell models are special cases of a model based on the three-dimensional linear theory of elasticity, which in turn is a special case of a large family of models based on the equations of continuum mechanics that account for a variety of hyperelastic, elasto-plastic and other material laws, large deformation, contact, etc. This is the hierarchic view of mathematical models.

Given the rich variety of choices, model selection for particular applications is a non-trivial problem. The goal of conceptualization is to identify the simplest mathematical model that can provide predictions of the data of interest within a specified range of accuracy.

Conceptualization begins with the formulation of a tentative mathematical model based on expert knowledge. We will call this a working model. The term has the same connotation and meaning as the term working hypothesis. Since subjective judgment is involved, the formulation of the initial working model may differ from expert to expert. Nevertheless, assuming that software tools that allow systematic evaluation of mathematical models with respect to clearly defined objectives are available, it should be possible for experts to arrive at a close agreement on the definition of a mathematical model, given its intended use.

Virtual experimentation

Model selection involves systematic evaluation of the effects of various modeling assumptions on the data of interest and the sensitivity of the data of interest to uncertainties in the input data. This is done through a process called virtual experimentation.

For example, in solid mechanics one usually begins with a working model based on the linear theory of elasticity. The implied assumptions are that the

strain is much smaller than unity, the stress is proportional to the strain, the displacements are so small that equilibrium equations written with respect to the undeformed configuration hold in the deformed configuration also, and the boundary conditions are independent of the displacement function. Once a verified solution is available, it is possible to examine the stress field and determine whether the stress exceeded the proportional limit of the material and whether this affects the data of interest significantly. Similarly, the effects of large deformation on the data of interest can be evaluated. Furthermore, it is possible to test the sensitivity of the data of interest to changes in boundary conditions. Virtual experimentation provides valuable information on the influence of various modeling assumptions on the data of interest.

Calibration

In the process of conceptualization there may be indications that the data of interest are sensitive functions to certain parameters that characterize material behavior or boundary conditions. If those parameters are not available then calibration experiments must be performed for the purpose of determining the needed parameters. In calibration the mathematical model is assumed to be correct and the parameters that characterize the model are selected such that the measured response matches the predicted response.

Example 1.1.1 If the goal of computation is to predict the number of load cycles that cause fatigue failure in a metal part then one or more empirical models must be chosen that require as input stress or strain amplitudes and material parameters. One of the widely used models for the prediction of fatigue life in low cycle fatigue is the general strain-life model:

$$\epsilon_a = \frac{\bar{\sigma}_f}{E}(2N)^b + \bar{\epsilon}_f(2N)^c \tag{1.1}$$

where ϵ_a is the strain amplitude, N is the number of cycles to failure, E is the modulus of elasticity, $\bar{\sigma}_f$ is the fatigue strength coefficient, b is the fatigue strength exponent, $\bar{\epsilon}_f$ is the fatigue ductility coefficient and c is the fatigue ductility exponent. The parameters E, $\bar{\sigma}_f$, b, $\bar{\epsilon}_f$ and c are determined through calibration experiments. See, for example, [54]. Several variants of this model are in use. Standard procedures have been established for calibration experiments for metal fatigue[1].

1.1.2 Validation

Validation is a process by which the predictive capabilities of mathematical models are tested and improved. We will be concerned primarily with problems in solid mechanics for which the predictions can be tested through experiments especially designed for that purpose. This is a very large class of problems that

[1] See, for example, International Organization for Standardization ISO 12106:2003 and ISO 12107:2003.

includes all mathematical models designed for the prediction of the performance of mass-produced items. There are other important problems, such as the effects of earthquakes and other natural disasters, unique design problems, such as dams, siting of nuclear power stations and the like for which the predictions based on mathematical models cannot be tested. In such cases the models are analyzed a posteriori and improved in the light of new information collected following an incident.

Associated with each mathematical model is a modeling error (illustrated schematically in Fig. 1.1). Therefore it is necessary to have a process for testing the predictive capabilities of mathematical models. This process, called validation. is illustrated schematically in Fig. 1.2.

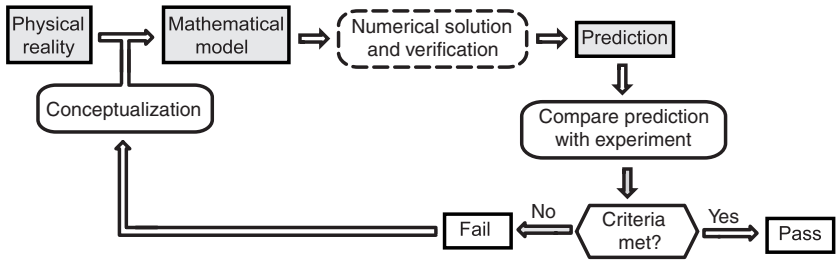

Figure 1.2: Validation.

For a validation experiment one or more metrics and the corresponding criteria are defined. If the predictions meet the criteria then the model is said to have passed the validation test, otherwise the model is rejected.

In large projects, such as the development of an aircraft, a series of validation experiments are performed starting with coupon tests for the determination of physical properties and failure criteria, then progressing to sub-components, components, parts, sub-assemblies and finally the entire assembly. Of course, the cost of experiments increases with complexity and hence the number of experiments decreases with complexity. The goal is to develop sufficiently reliable predictive capabilities such that the outcome of experiments involving sub-assemblies and assemblies will confirm the predictions. Finding problems late in the production cycle is generally very costly.

In evaluating the results of validation experiments it is important to bear in mind the limitations and uncertainties associated with the available information concerning physical systems being modeled:

1. The solution domain is usually assumed to correspond to design specifications ('the blueprint'). In reality, parts, sub-assemblies and assemblies deviate from their specifications and the degree of deviation may not be known, or would be difficult to incorporate into a mathematical model.

2. For many materials the constitutive laws are known imperfectly and only in some average sense and within a narrow range of strain, strain rate, temperature and over a short time interval of loading.

3. The boundary conditions, other than stress-free boundary conditions, are not known with a high degree of precision, even under carefully controlled experimental conditions. The reason for this is that the loading and constraints typically involve mechanical contact which depends on the compliances of the structures that impose the load and constraints (e.g., testing machine, milling machine, assembly rig, etc.) and the physical properties of the contacting surfaces. In other words, the boundary conditions represent the influence of the environment on the mathematical model. The needed information is rarely available. Therefore subjective judgment of the analyst in the formulation of boundary conditions is usually unavoidable.

4. Due to the history of the material prior to manufacturing the parts that will be assembled into a machine or structure, such as casting, quenching, extrusion, rolling, forging, heat treatment, cold forming and machining, initial stresses exist, the magnitude of which can be very substantial. The initial stress state must satisfy the equations of equilibrium and the stress-free boundary conditions but otherwise it is generally unknown.

5. Information concerning the probability distribution of the data that characterize the problem and their covariance functions is rarely available. In general, uncertainties increase with the complexity of models.

Remark 1.1.1 More than one mathematical model may have been proposed with identical objectives and it is possible that more than one mathematical model will meet the validation criteria. In that case the simpler model is preferred.

Remark 1.1.2 Owing to statistical variability in the data and errors in experimental observations comparisons between prediction based on a mathematical model and the outcome of physical experiments must be understood in a statistical sense. The theoretical framework for model selection is based on Bayesian analysis[2]. Specifically, denoting a mathematical model by M, the newly acquired data by D and the background information by I, the probability that the model M is a sufficiently good predictor of the data D, given the background information I, can be written in terms of conditional probabilities:

$$\text{Prob}(M|D,I) \approx \text{Prob}(D|M,I) \times \text{Prob}(M|I). \tag{1.2}$$

In other words, Bayes' theorem relates the probability that a mathematical model is correct, given the measured data D and the background information I, to the probability that the measured data would have been observed if the model were functioning properly. See, for example, [52]. The term $\text{Prob}(M|I)$ is called prior probability. It represents expert opinion about the validity of M prior to coming into possession of some new data D. The term $\text{Prob}(D|M,I)$ is called the likelihood function. In this view competing mathematical models

[2]Thomas Bayes (c. 1702 - 1761).

are assigned probabilities that represent the degree of belief in the reliability of each of the competing models, given the information available prior to acquiring additional information. In the light of the new information, obtained by experiments, the prior probability is updated to obtain the term $\text{Prob}(M|D, I)$, called the posterior probability. An important and highly relevant aspect of Bayes' theorem is that it establishes a framework for improvement of the probability estimate $\text{Prob}(M|D, I)$ based on new data.

1.1.3 Discretization

The finite element method (FEM) is one of the most powerful and widely used numerical methods for finding approximate solutions to mathematical problems formulated so as to simulate the responses of physical systems to various forms of excitation. It is used in various branches of engineering and science, such as elasticity, heat transfer, fluid dynamics, electromagnetism, acoustics, biomechanics, etc.

In the finite element method the solution domain is subdivided into elements of simple geometrical shape, such as triangles, squares, tetrahedra, hexahedra, and a set of basis functions are constructed such that each basis function is non-zero over a small number of elements only. This is called discretization. Details will be given in the following chapters. The set of all functions that can be written as linear combinations of the basis functions is called the finite element space. The accuracy of the data of interest depends on the finite element space and the method used for computing the data from the finite element solution. Associated with the finite element solution are errors of discretization, as indicated in Fig. 1.1.

It is necessary to create finite element spaces such that the data of interest computed from the finite element solution are within acceptable error bounds with respect to their counterparts corresponding to the exact solution of the mathematical model.

The data of interest, such as the maximum displacement, temperature, stress, etc. are computed from the finite element solution u_{FE}. The data of interest will be denoted by $\Phi_i(u_{FE})$, $i = 1, ..., n$ in the following. The objective is to compute $\Phi_i(u_{FE})$ and to ensure that the relative errors are within prescribed tolerances:

$$\frac{|\Phi_i(u_{EX}) - \Phi_i(u_{FE})|}{|\Phi_i(u_{EX})|} \leq \tau_i \tag{1.3}$$

where u_{EX} is the exact solution. Of course u_{EX} is not known in general, however it is known that $\Phi_i(u_{EX})$ is independent of the finite element space. The error in $\Phi_i(u_{FE})$ depends on the finite element space and the method used for computing $\Phi_i(u_{FE})$. The errors of discretization are controlled through suitable enlargement of the finite element spaces, and by various procedures used for computing $\Phi_i(u_{FE})$.

1.1.4 Verification

Verification is concerned with verifying that (a) the input data are correct, (b) the computer code is functioning properly and (c) the errors in the data of interest meet necessary conditions to be within permissible tolerances.

Common errors in input are incorrectly entered data, such as mixed units and errors in data entry. Such errors are easily caught in a careful review of the input data.

The primary responsibility for ensuring that the code is functioning properly rests with the code developers. However, computer codes tend to have programming errors, especially in their less frequently traversed branches, and the user shares in the responsibility of verifying that the code is functioning properly.

In verification accuracy is understood to be with respect to the exact solution of the mathematical model, not with respect to physical reality. The process of verification of the numerical solution is illustrated schematically in Fig. 1.3. The term extraction refers to methods used for computing $\Phi_i(u_{FE})$. Details are presented in the following chapters.

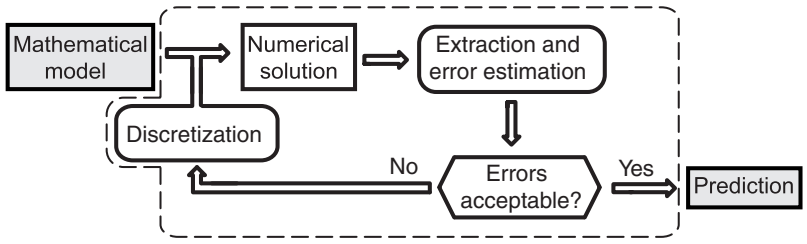

Figure 1.3: Verification of the numerical solution.

Remark 1.1.3 Verification and validation are possible only when the mathematical model is properly formulated with respect to the goals of computation. For example, in linear elasticity the solution domain must not have sharp reentrant corners or edges if the goal of computation is to determine the maximum stress, point constraints and point force can be used only when certain criteria are met, etc. Details are given in the following chapters. Unfortunately, using mathematical models without regard to their limitations is a commonly occurring conceptual error.

Remark 1.1.4 The process illustrated schematically in Fig. 1.1 is often referred to as 'finite element modeling'. This term is unfortunate because it mixes two conceptually different aspects of numerical simulation; the definition of a mathematical model and its numerical solution by the finite element method.

1.1.5 Decision making

As noted at the beginning of Section 1.1, the goal of numerical simulation is to support various engineering decision-making processes. There is an implied expectation of reliability: One could not reasonably base decisions on computed information without believing that the information is sufficiently reliable to support those decisions. Demonstration of the reliability of mathematical models used in support of engineering decision-making is an essential part of any modeling effort. In fact, the role of physical testing is to calibrate and validate mathematical models so that a variety of load cases and design alternatives can be evaluated.

In the following we illustrate the importance of the reliability of numerical simulation processes through brief descriptions of four well-documented examples of the consequences of large errors in prediction either because deficient mathematical models were used or because large errors occurred in the numerical solution. Additional examples can be found in [41], [42]. Undoubtedly, there are many undocumented instances of substantial loss attributable to errors in predictions based on mathematical models.

Example 1.1.2 The Tacoma Narrows Bridge, the first suspension bridge across Puget Sound (Washington State, USA) collapsed on November 7, 1940, four months after its opening. Wind blowing at 68 km/h caused sufficiently large oscillations in the 853 m main span to collapse the span.

Until that time bridges were designed on the basis of equivalent static forces. The possibility that relatively small periodic aerodynamic forces (the effects of Kármán vortices[3]) may become significant was not considered. The Kármán vortices were first analyzed in 1911 and the results were presented in the Göttingen Academy in the same year[4]. The designers were either unaware of those results or did not see their relevance to the Tacoma Narrows Bridge, the failure of which was caused by insufficient torsional stiffness to resist the periodic excitation induced by Kármán vortices.

Example 1.1.3 The roof of the Hartford Civic Center Arena collapsed on January 18, 1978. The roof structure, measuring 91.4 by 109.7 m (300 by 360 ft), was a space frame, an innovative design at that time. It was analyzed using a mathematical model that accounted for linear response only. Furthermore, the connection details were greatly simplified. In linear elastostatic analysis it is assumed that the deformation of a structure is negligibly small and hence it is sufficient to satisfy the equations of equilibrium in the undeformed configuration.

The roof frame was assembled on the ground. Once the roof was lifted into its final position, its deflection was measured to be twice of what was predicted by the mathematical model.

[3] Theodore von Kármán 1881-1963.

[4] Th. von Kármán and L. Edson, *The Wind and Beyond. Theodore von Kármán Pioneer in Aviation and Pathfinder in Space*, Little, Brown & Co., Boston, pp. 211-215, (1967).

> "When notified of this condition, the engineers expressed no concern, explaining that such discrepancies had to be expected in view of the simplifying assumptions of the theoretical calculation"[5].

Subsequent investigation identified that reliance on an oversimplified model that did not represent the connection details properly and failed to account for geometric nonlinearities was the primary cause of failure.

Example 1.1.4 The Vaiont dam, one of the highest dams in the world (262 m) was completed in the Dolomite Region of the Italian Alps, 100 km North of Venice, in 1961. On October 9, 1963, after heavy rains, a massive landslide into the reservoir caused a large wave that over-topped the dam by up to 245 m and swept onto the valley below, resulting in the loss of an estimated 2000 lives[6]. The courts found that, owing to the *predictability* of the landslide, three engineers were criminally responsible for the disaster. The dam withstood the overload caused by the wave. This incident serves as an example of a full scale test of a major structure caused by an unexpected event.

Example 1.1.5 The consequences of large errors of discretization are exemplified by the Sleipner accident. The gravity base structure (GBS) of the Sleipner A offshore platform, made of reinforced concrete, sank during ballast test operations in Gandsfjorden, South of Stavenger, Norway on August 23, 1991. The economic loss was estimated to be 700 million dollars.

The main function of the GBS was to support a platform weighing 56,000 tons. The GBS consisted of 24 caisson cells with a base area of 16,000 m^2. Four cells were elongated to form shafts designed to support the platform. The total concrete volume of the GBS was 75,000 m^3. The accident occurred as the GBS was being lowered to a depth of approximately 99 m. Failure first occurred in two triangular cells, called tri-cells, next to one of the shafts. When the GBS hit the sea bed, seismic events measuring 3 on the Richter scale were recorded in the Stavenger area[7].

There is general agreement among the investigators that the accident was caused by large errors in the finite element analysis, the goal of which was to estimate the requirements for reinforcement of the concrete cells by steel bars:

> "The global finite element analysis gave a 47% underestimation of the shear forces in the tri-cell walls. This error was caused by the use of a coarse finite element mesh with some skewed elements used for analysis of the tri-cell walls"[8].

[5] Levy, M. and Salvadori, M., *Why Buildings Fall Down: How Structures Fail*, W. W. Norton, New York, NY. (2002).

[6] See, for example, Hendron, A. J., and Patten, F. D. The Vaiont Slide. US Corps of Engineers Technical Report GL-85-8 (1985).

[7] Jacobsen, B., "The Loss of the Sleipner A Platform", *Proc. 2nd International Offshore and Polar Engineering Conference*, International Society of Offshore and Polar Engineers, ISBN 1-880653-01-X, Vol. 1, 1992.

[8] Rettedal, W. K., Gudmestad, O. T., and T. Aarum, "Design of Concrete Platforms after Sleipner A-1 Sinking", Proc. 12th International Conference on Offshore Mechanics and Arctic Engineering, Vol. 1, Offshore Technology, pp. 309-310, ASME 1993.

> "A check of the global response analysis revealed serious inaccuracies in the interpretation of results from finite element analyses giving a shear force in a critical section of the cell wall that was less than 60% of the correct value"[9].

1.2 Why is numerical accuracy important?

A number of difficulties associated with accurate representation of a real physical system by mathematical means were noted in Section 1.1.2. Given these difficulties, it may seem reasonable to ask: "If we do not know the input data with sufficient accuracy, then why should we be concerned with the accuracy of the numerical solution?" In answering this question we consider two important areas of application of mathematical models: The application of design rules and the formulation of design rules. It is shown in the following that both the application and the formulation of design rules require estimation and control of the numerical accuracy.

1.2.1 Application of design rules

Design and design certification involve application of existing design rules, established by various codes and conventions. The design rules are typically stated in the form of required minimum factors of safety:

$$FS := \frac{\Phi_{\text{lim}}}{\Phi_{\max}(u_{EX})} \geq (FS)_{\text{design}} \tag{1.4}$$

where FS is the realized factor of safety $\Phi_{\text{lim}} > 0$ is the limiting (not to exceed) value of some entity (such as maximum bending moment, maximum stress, etc.) $\Phi_{\max}(u_{EX}) > 0$ is the exact value of the same entity corresponding to the exact solution of the mathematical model and $(FS)_{\text{design}}$ is the minimum value of the factor of safety specified by the applicable design rules. It is the designer's responsibility to ensure that the applicable design rules are followed.

We will denote by $\Phi_{\max}(u_{FE})$ the value of $\Phi_{\max}$ computed from the finite element solution. Let us suppose that, owing to numerical errors, it is possible to guarantee only that the relative error is not greater than τ:

$$\frac{|\Phi_{\max}(u_{EX}) - \Phi_{\max}(u_{FE})|}{\Phi_{\max}(u_{EX})} \leq \tau \quad 0 \leq \tau < 1$$

in other words, $\Phi_{\max}(u_{FE})$ may underestimate $\Phi_{\max}(u_{EX})$ by 100τ percent. Therefore we have:

$$\Phi_{\max}(u_{EX}) \leq \frac{1}{(1-\tau)}\Phi_{\max}(u_{FE})$$

[9] Holand, I., "The Sleipner Accident" in *From Finite Elements to the Troll Platform - Ivar Holand 70th Anniversary*, K. Bell, editor, ISBN 82-7482-016-9, Department of Structural Engineering, The Norwegian Institute of Technology, Trondheim, Norway, pp. 157-168, 1994.

on substituting this expression into eq. (1.4), we have:

$$\frac{\Phi_{\lim}}{\Phi_{\max}(u_{FE})} \geq \frac{(FS)_{\text{design}}}{1-\tau}. \tag{1.5}$$

On comparing eq. (1.5) with eq. (1.4) it is seen that to compensate for numerical errors in the computation of $\Phi_{\max}(u_{FE})$, it is necessary to increase the required factor of safety to $(FS)_{\text{design}}/(1-\tau)$. For example, if the accuracy of $\Phi_{\max}(u_{FE})$ can be guaranteed to 20% (i.e., $\tau = 0.20$) then $(FS)_{\text{design}}$ must be increased by 25%. Since $(FS)_{\text{design}}$ was chosen conservatively to account for the uncertainties, the economic penalties associated with using an increased factor of safety generally far outweigh the costs associated with guaranteeing the accuracy of the data of interest to within a small relative error (say 5%).

It is necessary to specify the acceptable error tolerance in finite element analysis and to verify that the error is not larger than the specified tolerance.

Remark 1.2.1 In aerospace engineering the design requirements are stated in terms of minimum acceptable margins of safety (MS). By definition: $MS = FS - 1$.

Remark 1.2.2 Economic considerations dictate that the realized factor of safety should not be much larger than $(FS)_{\text{design}}$. This is especially true in aerospace engineering where avoidance of weight penalties dictate upper bounds on the realized factors of safety.

Example 1.2.1 The yield strength in shear of hot rolled 0.2% carbon steel is 165 MPa and the usual factor of safety for static loads is 1.65 (so that the allowable maximum shear stress is 100 MPa)[10]. If the numerical computations could underestimate the maximum shear stress by as much as 20% then the factor of safety would have to be increased to 2.06, that is, the allowable maximum value would be reduced to 80 MPa.

1.2.2 Formulation of design rules

Formulation of design rules involves definition of certain entities Φ_k ($k = 1, 2, \dots$), such as the maximum principal stress, some specific combinations of stress and strain components, etc. that characterize failure and the corresponding limiting values. In the following the subscript k will be dropped and the discussion will be concerned with a generic design rule, that is, the determination of $\Phi_{\lim}$ and evaluation of the associated uncertainties. The factor of safety is determined on the basis of assessment of uncertainties and consideration of the consequences of failure.

Suppose that a hypothesis stating that failure occurs when Φ reaches its critical value $\Phi_{\lim}$ was proposed. First a set of calibration experiments have to be performed with the objective to determine $\Phi_{\lim}$. Second, another set

[10]See, for example, E. P. Popov, Engineering Mechanics of Solids, 2nd. edition, Prentice Hall, Upper Saddle River, NJ., 1998.

of experiments have to be conducted to test whether failure can be predicted on the basis of Φ_{lim}. These are validation experiments. In general Φ cannot be observed directly, therefore it must be inferred from correlations between computed data and experimental observations.

Let Y_{ij} be the ith ideal observation of the jth experiment and let $\phi_i(u_{EX}^{(j)})$ be the corresponding functional[11] computed from the exact solution $u_{EX}^{(j)}$ so that if there were no experimental errors and the mathematical model and the hypothesis were both correct then we would have

$$Y_{ij} - \phi_i(u_{EX}^{(j)}) = 0.$$

Due to experimental errors we actually observe y_{ij} and compare it with $\phi_i(u_{FE}^{(j)})$, the finite element approximation to $\phi_i(u_{EX}^{(j)})$. Let us write

$$Y_{ij} = y_{ij} \pm e_{ij}^{\text{exp}}$$

and

$$\phi_i(u_{EX}^{(j)}) = \phi_i(u_{FE}^{(j)}) \pm e_{ij}^{\text{fea}}$$

where e_{ij}^{exp} (resp. e_{ij}^{fea}) is the experimental (resp. approximation) error. Then:

$$y_{ij} - \phi_i(u_{FE}^{(j)}) = Y_{ij} \mp e_{ij}^{\text{exp}} - \phi_i(u_{EX}^{(j)}) \pm e_{ij}^{\text{fea}}.$$

Using the triangle inequality, we have

$$\underbrace{|y_{ij} - \phi_i(u_{FE}^{(j)})|}_{\text{apparent error}} \leq \underbrace{|Y_{ij} - \phi_i(u_{EX}^{(j)})|}_{\text{true error}} + |e_{ij}^{\text{exp}}| + |e_{ij}^{\text{fea}}|. \tag{1.6}$$

This result shows that in testing a particular hypothesis it is essential to have both the experimental errors and the errors of discretization under control, otherwise it will not be possible to know whether the apparent error is due to an error in the hypothesis, errors in the numerical approximation, or errors in the experiment. Furthermore, means for estimation and control of discretization errors, in terms of the data of interest, must be provided by the computer code.

The aim of experiments needs to include the development of reliable statistical information on the basis of which the factor of safety is established.

1.3 Chapter summary

The principal aim of this book is to present the theoretical and practical considerations relevant to (a) the validation of mathematical models and (b) verification of the data of interest computed from finite element solutions. Some fundamental concepts and basic terminology were introduced:

[11] A functional is a real number defined on a space of functions. In the present context a functional is a real number computed from the exact solution or the finite element solution.

Mathematical model

A mathematical representation of a physical system or process intended for predicting some set of responses is called a mathematical model. Mathematical models transform one set of data, the input, into an other set, the output.

Conceptualization

Conceptualization is a process by which a mathematical model is defined, for a particular application. Conceptualization involves (a) application of expert knowledge, (b) virtual experimentation and (c) calibration.

Discretization

Discretization is a process by which a mathematical problem is formulated that can be solved on digital computers, the solution of which approximates the exact solution of a given mathematical model

Validation

Validation is a process by which the predictive capabilities of mathematical models are tested and improved. Ideally, experiments are performed especially to test whether a mathematical model meets necessary conditions for acceptance from the perspective of its intended use. Validation experiments are evaluated on the basis of one or more metrics and the corresponding criteria. In any important applications of mathematical models the predictions cannot be tested in validation experiments. In such cases the model is analyzed a posteriori in the light of new information collected following an incident.

Verification

Verification is a process by which it is ascertained that the data of interest computed from the approximate solution meet necessary conditions for acceptance. Verification is understood in relation to the exact solution of a mathematical model, not in relation to the physical reality that the mathematical model is supposed to represent.

Errors

Five types of error were discussed: Errors of idealization, errors of discretization, conceptual errors, programming errors and errors in the input data.

Chapter 2

An outline of FEM

In this chapter an outline of the finite element method is presented in one-dimensional setting. It will be generalized to two and three dimensions in subsequent chapters.

Throughout the book the units of physical data will be identified in terms of the standard SI[1] notation. Any consistent set of units may be used, however.

2.1 Mathematical models in one dimension

The formulation of mathematical models will be discussed in Chapter 3. Here a simple mathematical model that will serve as the basis for the discussion of the conceptual and algorithmic aspects of the finite element method is formulated.

2.1.1 The elastic bar

The elastostatic response of an elastic bar to imposed loads is characterized by the axial displacement function $u(x)$. We will assume that the centroidal axis of the bar is coincident with the x-axis. The length of the bar is ℓ. The mathematical model of an elastic bar is based on equations that represent the strain-displacement relationship, the stress-strain relationship and equilibrium:

1. **The strain-displacement relationship.** The total strain is

$$\epsilon_x \equiv \epsilon = \frac{du}{dx} \equiv u'. \tag{2.1}$$

The total strain ϵ is the sum of the mechanical strain ϵ_m and the thermal strain $\epsilon_t = \alpha \mathcal{T}_\Delta$ where $\alpha = \alpha(x) \geq 0$ is the coefficient of thermal expansion ($1/K$ units). Therefore the mechanical strain is:

$$\epsilon_m = \epsilon - \epsilon_t = u' - \alpha \mathcal{T}_\Delta. \tag{2.2}$$

[1] Système International d'Unités (International System of Units).

2. **The stress-strain relationship.** In one dimension Hooke's law states that the stress is proportional to the mechanical strain:

$$\sigma_x \equiv \sigma = E\epsilon_m = E(u' - \alpha\mathcal{T}_\Delta) \tag{2.3}$$

where $E = E(x) > 0$ is the modulus of elasticity (MPa units).

3. **Equilibrium.** It is assumed that the stress is constant over the cross-sectional area. The equilibrium equations are written in terms of the bar force F_b defined by

$$F_b := \int_A \sigma \, dydz = \sigma A = AE(u' - \alpha\mathcal{T}_\Delta) \tag{2.4}$$

where $A = A(x) > 0$ is the area of the cross section. The bar may be subjected to traction forces and/or volume forces T_b in N/m units and tractions exerted by elastic springs:

$$T_s := c(d - u) \tag{2.5}$$

where $d = d(x)$ is displacement imposed on the distributed spring and $c = c(x) \geq 0$ is the spring rate in N/m^2 units. When the force T_b accounts for volume forces then it is understood that the volume forces have been integrated over the cross-sectional area, such that T_b is measured in N/m units.

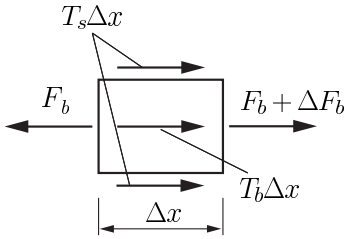

Figure 2.1: Bar element.

Referring to Fig. 2.1, and considering the equilibrium of an isolated part of the bar of length Δx, we write:

$$\Delta F_b + T_b\Delta x + T_s\Delta x = 0$$

Assuming that F_b is a continuous and differentiable function,

$$\Delta F_b = \frac{dF_b}{dx}\Delta x + O(\Delta x).$$

Letting $\Delta x \to 0$, we have the equilibrium equation:

$$\frac{dF_b}{dx} + T_b + T_s = 0. \tag{2.6}$$

On combining equations (2.4), (2.5) and (2.6) the ordinary differential equation that models the mechanical response of elastic bars to applied traction forces is obtained:

$$-\frac{d}{dx}\left(AE\frac{du}{dx}\right) + cu = T_b + cd - \frac{d}{dx}\left(AE\alpha\mathcal{T}_\Delta\right) \quad \text{on } x \in I \tag{2.7}$$

where I represents the set of points x that lie in the interval $0 < x < \ell$. In the following we will write $I = (0, \ell)$.

Boundary conditions

We will be considering linear boundary conditions associated with eq. (2.7). These are shown schematically in Fig. 2.2. A brief description follows.

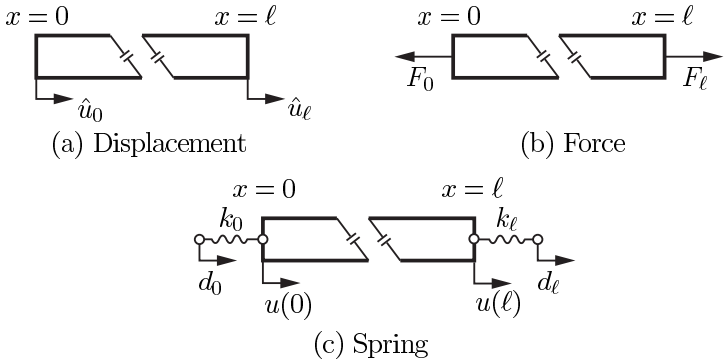

Figure 2.2: Elastic bar. Linear boundary conditions.

1. Displacement boundary conditions, also called kinematic boundary conditions, are shown in Fig. 2.2(a). The given displacement is denoted by $\hat{u}_0$ at $x = 0$ (resp. $\hat{u}_\ell$ at $x = \ell$).

2. Forces, denoted by F_0 and F_ℓ, as indicated in Fig. 2.2(b), may be prescribed at one or both boundary points. The applied forces are positive when tensile:

$$F_0 := [AE(u' - \alpha\mathcal{T}_\Delta)]_{x=0} \qquad F_\ell := [AE(u' - \alpha\mathcal{T}_\Delta)]_{x=\ell}. \tag{2.8}$$

3. Spring boundary conditions are linear relationships between the bar forces F_0 and F_ℓ and the corresponding displacements at the boundary points, as indicated in Fig. 2.2(c):

$$F_0 = [AE(u' - \alpha\mathcal{T}_\Delta)]_{x=0} = k_0(u(0) - d_0) \tag{2.9}$$
$$F_\ell = [AE(u' - \alpha\mathcal{T}_\Delta)]_{x=\ell} = k_\ell(d_\ell - u(\ell)) \tag{2.10}$$

where $k_0 \geq 0$ (resp. $k_\ell \geq 0$) is the spring constant (in N/m units) at $x = 0$ (resp. $x = \ell$) and d_0 (resp. d_ℓ) is a displacement imposed on the spring at $x = 0$ (resp. $x = \ell$).

Of course, the displacement, force and spring boundary conditions may occur in any combination.

Symmetry, antisymmetry and periodicity

The axis of symmetry is a line that passes through mid-point of the interval ℓ and is perpendicular to the x-axis. A scalar function defined on I is said to be symmetric if in symmetrically located points with respect to the axis of symmetry the function has equal values. A scalar function is said to be antisymmetric if in symmetrically located points with respect to the axis of symmetry the function has equal absolute values but opposite sign.

The coefficients $AE(x)$ and $c(x)$ may be symmetric functions with respect to the axis of symmetry. If in such cases $T_b(x)$, $d(x)$, $\alpha T_\Delta(x)$ and the boundary conditions are also symmetric (resp. antisymmetric) then the solution is a symmetric (resp. antisymmetric) function with respect to the axis of symmetry. When the solution is a symmetric or antisymmetric function then the problem can be solved on half of the interval and extended to the entire interval by symmetry or antisymmetry.

We will understand symmetry to mean mirror image symmetry with respect to the axis of symmetry. In the case of the elastic bar the displacement, bar force and traction are vector functions. These vector functions have only one non-zero component which is perpendicular to the axis of symmetry. Examples of symmetric and antisymmetric loading and constraints are shown in Figures 2.3(a) and 2.3(b). In the symmetric case the boundary condition is $u(\ell/2) = 0$. In the antisymmetric case the boundary condition is $F(\ell/2) = 0$.

When $AE(x)$, $c(x)$, $T_b(x)$, $d(x)$ and $\alpha\mathcal{T}_\Delta(x)$ are periodic functions, the length of the period being ℓ, that is, $(AE)_{x=0} = (AE)_{x=\ell}$, $c(0) = c(\ell)$, $T_b(0) = T_b(\ell)$, $d(0) = d(\ell)$, $\alpha\mathcal{T}_\Delta(0) = \alpha\mathcal{T}_\Delta(\ell)$, $u(0) = u(\ell)$ and $F(0) = F(\ell)$ then the solution is a periodic function and the boundary conditions are said to be periodic. The solution obtained for $(0, \ell)$ is extended to $-\infty < x < \infty$. Periodic boundary conditions are illustrated in Fig. 2.3(c).

Remark 2.1.1 The mathematical problem of eq. (2.7), together with specific boundary conditions, is a mathematical model of an elastic bar. This problem is solved with the goal to obtain some desired information, called data of interest, such as the displacement $u(x)$, the axial strain $u'(x)$, or the axial force $AE(u'(x) - \alpha\mathcal{T}_\Delta)$ in all or specific points, or in points where their maxima occurs. Incorporated in the model are the assumptions that $|\epsilon_m| << 1$, $|\epsilon_t| << 1$ and $|\sigma| \leq \sigma_{\rm pl}$ where $\sigma_{\rm pl}$ is the proportional limit of the material.

Example 2.1.1 Consider the problem

$$-(AEu')' + cu = T_b, \qquad u(0) = 0, \quad F_b(\ell) = F_\ell$$

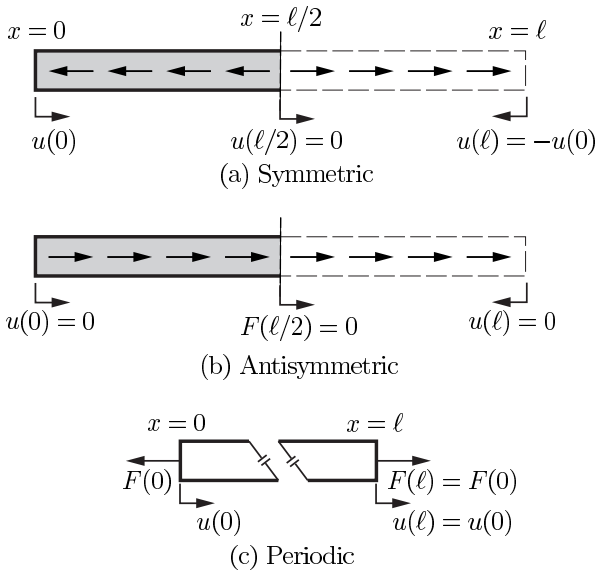

Figure 2.3: Elastic bar. Symmetric, antisymmetric and periodic loading and boundary conditions.

where AE and c are constants and $T_b = b_0 + b_1 x/\ell$, where b_0, b_1 are given constants (in N/mm units). To solve this problem we define $\lambda^2 := c/(AE)$. The general solution can be found in standard texts on ordinary differential equations and engineering mathematics:

$$u = C_1 \cosh \lambda x + C_2 \sinh \lambda x + \frac{b_0}{c} + \frac{b_1}{c}\frac{x}{\ell}. \tag{2.11}$$

From the boundary conditions we find:

$$\begin{aligned} C_1 =& -\frac{b_0}{c} \\ C_2 =& \frac{1}{\lambda \cosh \lambda\ell}\left[\frac{F_\ell}{AE} + \frac{b_0}{c}\lambda \sinh \lambda\ell - \frac{b_1}{c\ell}\right]. \end{aligned}$$

Example 2.1.2 Consider the problem of Example 2.1.1 with periodic boundary conditions:

$$-(AEu')' + cu = T_b, \qquad u(0) = u(\ell), \quad F_b(0) = F_b(\ell)$$

where, as in Example 2.1.1, AE and c are constants and $T_b = b_0 + b_1 x/\ell$, with b_0, b_1 are given constants. The general solution is given by (2.11). On applying the periodic boundary conditions we find:

$$u = \frac{b_1}{2c}\cosh \lambda x + \frac{b_1}{2c}\frac{\sinh \lambda\ell}{1 - \cosh \lambda\ell}\sinh \lambda x + \frac{b_0}{c} + \frac{b_1}{c}\frac{x}{\ell}. \tag{2.12}$$

Exercise 2.1.1 Consider an elastic bar constrained by a distributed spring of stiffness c. Assume that AE, c are constants. The coefficient of thermal expansion is α (constant). The boundary conditions are: $u(0) = 0$, $F_b(\ell) = 0$. The bar is subjected to a temperature change $\mathcal{T}_\Delta(x) = b_0$ (constant). Write down the solution for this problem.

Exercise 2.1.2 Consider an elastic bar constrained by a distributed spring of stiffness c. Assume that AE, c are constants and $u(0) = \hat{u}_0$, $F_b(\ell) = k_\ell(d_\ell - u(\ell))$. Write down the solution for this problem.

2.1.2 Conceptualization

We have formulated mathematical models suitable for predicting static responses of elastic bars. We tacitly assumed that the physical properties and boundary conditions were given. In many practical applications not all of the needed information is available. Therefore it is necessary to perform and interpret calibration experiments. The procedure is illustrated by the following example.

Example 2.1.3 One of the methods used for ensuring that the foundation of a large building is sufficiently stiff to resist the dead and live loads without undergoing excessive settlement is to drive large elastic bars, called piles, into the soil. Suppose that two experts were consulted on the question of how to estimate the stiffness of a pile and both experts agreed that the mathematical model should be based on the following differential equation:

$$-AEu'' + cu = 0, \quad AEu'(0) = F_0,\ u'(\ell) = 0 \tag{2.13}$$

where c represents the action of the soil on the pile. The goal is to predict the displacement u_0 at the top of the pile as a function of the applied axial force. The notation is shown in Fig. 2.4.

Both experts recommended using the nominal value for the modulus of elasticity of steel $E = 200$ GPa however one expert recommended that c should be treated as a constant, the other expert recommended that $c = kx$ where k is a constant should be assumed. In other words, different mathematical models were proposed for the same problem. In the following we refer to these as Model A and Model B, respectively. In order to determine c, an HP305×110 test pile[2] was driven into the soil to the depth of 12.0 m. The cross-sectional area is 1.402×10^{-2} m^2.

A pull test yielded the following results: When the applied force F_0 is 200 kN then the measured upward displacement u_0 is 9.0 mm; at $F_0 = 300$ kN $u_0 = 13.5$ mm; at $F_0 = 400$ kN $u_0 = 18.0$ mm. In other words, the experimental measurements yielded $F_0/u_0 = 22.22$ kN/mm.

[2]This designation indicates that the cross-section is H-shaped, the nominal depth of the cross-section is 305 mm and the mass is approximately 110 kg/m.

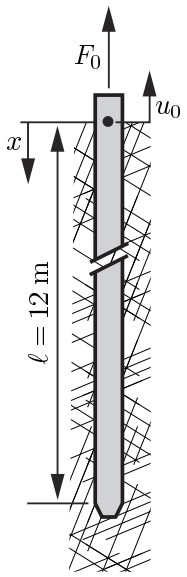

Figure 2.4: Example 2.1.3. Notation.

Calibration of Model A

Equation (2.13) can be re-written as

$$-u'' + \lambda^2 u = 0, \quad u'(0) = F_0/AE, \; u'(\ell) = 0 \quad \text{where } \lambda := \sqrt{\frac{c}{AE}} \tag{2.14}$$

the solution of which is:

$$u(x) = \frac{F_0}{\lambda AE}\left(\sinh \lambda x - \frac{\cosh \lambda \ell}{\sinh \lambda \ell}\cosh \lambda x\right).$$

Note that with the sign convention adopted in Section 3.4.4 and illustrated in Fig. 2.2, the upward displacement is negative, that is $u(0) = -u_0$. Therefore the force-displacement relationship is

$$F_0 = AEu'(0) = AE\frac{\lambda \sinh \lambda \ell}{\cosh \lambda \ell}u_0 \tag{2.15}$$

which can be written as

$$G(\lambda) := \frac{F_0}{u_0} - AE\frac{\lambda \sinh \lambda \ell}{\cosh \lambda \ell} = 0. \tag{2.16}$$

For the three data pairs $F_0/u_0 = 22.22$ kN/mm was measured. We need to find λ such that $G(\lambda) = 0$. Various root finding methods are available. One of the most commonly used methods is the Newton-Raphson method[3]. In this method we

[3]Sir Isaac Newton 1642-1727, Joseph Raphson 1648-1715.

select a trial value for λ, denoted by λ_1, and compute the corresponding $G(\lambda_1)$ and $G'(\lambda_1) := (dG/d\lambda)_{\lambda=\lambda_1}$. The choice of λ_1 must be such that $(dG/d\lambda)_{\lambda=\lambda_1} \neq 0$. We then compute λ_{k+1} from

$$\lambda_{k+1} = \lambda_k - \frac{G(\lambda_k)}{G'(\lambda_k)} \quad \text{for } k = 2, 3, \ldots$$

and continue the process until $\lambda_{k+1} - \lambda_k$ is sufficiently small. By this method we find $\lambda = 2.6112 \times 10^{-2}$ m^{-1} and from the definition of λ given in eq. (2.14) we have $c = 1912$ kN/m^2. This completes the calibration step for Model A.

Calibration of Model B

Calibration of Model B involves solution of the problem

$$-AEu'' + kxu = 0, \quad AEu'(0) = F_0,\ u'(\ell) = 0 \tag{2.17}$$

which will be written as

$$-u'' + \lambda^2 xu = 0, \quad u'(0) = F_0/AE,\ u'(\ell) = 0 \quad \text{where } \lambda := \sqrt{\frac{k}{AE}}. \tag{2.18}$$

This is known as the Airy equation[4], see for example [44]. We will use a Taylor series expansion to find an approximate solution. We denote the nth derivative of u by $D^n u$. The derivatives for $n = 0, 1, \ldots, 7$ are shown in Table 2.1.

Table 2.1: The derivatives of $u(x)$.

n	$D^n u(x)$	$D^n u(0)$
0	u	$-u_0$
1	Du	F_0/AE
2	$\lambda^2 xu$	0
3	$\lambda^2 u + \lambda^2 xDu$	$-\lambda^2 u_0$
4	$2\lambda^2 Du + \lambda^2 xD^2 u$	$2\lambda^2 F_0/AE$
5	$3\lambda^2 D^2 u + \lambda^2 xD^3 u$	0
6	$4\lambda^2 D^3 u + \lambda^2 xD^4 u$	$-4\lambda^4 u_0$
7	$5\lambda^2 D^4 u + \lambda^2 xD^5 u$	$10\lambda^4 F_0/AE$

We see that for $k \geq 3$ we have

$$D^k u = (k-2)\lambda^2 D^{k-3} u + \lambda^2 xD^{k-2} u.$$

The Taylor series expansion of $u(x)$ is:

$$u(x) = -u_0 + \frac{F_0}{AE}x - \frac{\lambda^2}{3!}u_0 x^3 + \frac{2\lambda^2}{4!}\frac{F_0}{AE}x^4 - \frac{4\lambda^4}{6!}u_0 x^6 + \frac{10\lambda^4}{7!}\frac{F_0}{AE}x^7 - \cdots$$

[4]Sir George Biddell Airy 1801-1892.

Letting $u'(\ell) = 0$ we have:

$$0 = \frac{F_0}{AE} - \frac{\lambda^2}{2}u_0\ell^2 + \frac{\lambda^2}{3}\frac{F_0}{AE}\ell^3 - \frac{\lambda^4}{30}u_0\ell^5 + \frac{\lambda^4}{72}\frac{F_0}{AE}\ell^6 - \cdots \tag{2.19}$$

Therefore we need to find λ such that

$$G(\lambda) \approx \frac{F_0}{u_0}\left(\frac{1}{AE} + \frac{\lambda^2\ell^3}{3AE} + \frac{\lambda^4\ell^6}{72AE}\right) - \frac{\lambda^2\ell^2}{2} - \frac{\lambda^4\ell^5}{30} = 0. \tag{2.20}$$

Using the experimental result $F_0/u_0 = 22.22$ kN/mm, we find $\lambda \approx 1.0767 \times 10^{-2}$ m$^{-3/2}$ and hence $k \approx 325.0$ kN/m^3.

In this example the conceptual development of a mathematical model was illustrated in a simple setting. Model A and Model B differ by the definition of the constant c. The characterizing parameters c and k were determined by calibration. Calibration is part of the conceptualization process because definition of the mathematical model depends on information obtained by calibration experiments.

Exercise 2.1.3 Determine whether using 4 significant digits in the estimate $k \approx 325.1$ kN/m^3 in Example 2.1.3 is justified.

2.1.3 Validation

Making a prediction based on a mathematical model concerning the outcome of a physical experiment, then testing to see whether the prediction is correct, is called validation. Validation involves one or more metrics and the corresponding criteria. The metrics and criteria depend on the intended use of the model. For testing the model described in Example 2.1.3 we define the metric to be the ratio F_0/u_0 and the criterion is the corresponding tolerance. Validation is illustrated by the following example.

Example 2.1.4 On examining the pull test data in Example 2.1.3, we see that each 100 kN increment in the applied force resulted in 4.5 mm increment in displacement. Therefore the assumption that the pile is supported by a distributed linear spring is consistent with the available observations. However, it is not possible to determine from these observations how the spring coefficient c varies with x.

Let us assume that a second pile, driven to 8.5 m depth (i.e., $\ell = 8.5$ in Fig. 2.4), is to be tested. Based on Model A and Model B we predict the test results shown in the second and third columns of Table 2.2 and we state our criterion as follows: A model will be rejected if the difference between the predicted and observed values of F_0/u_0 exceeds the tolerance of 5 %.

Let us suppose that we observe the set of displacements shown in the fourth column of Table 2.2. Since the ratio predicted by Model A $(F_0/u_0)^A_{\text{pred}} = 16.0$ kN/mm and the observed ratio $(F_0/u_0)_{\text{obs}} = 11.9$ kN/mm differ by more than 5 percent Model A is rejected. On the other hand, the ratio predicted by Model B $(F_0/u_0)^B_{\text{pred}} = 11.5$ kN/mm and the observed ratio differ by less than 5 percent. Therefore Model B passes the validation test.

Table 2.2: Predicted and observed data.

F_0 Applied kN	u_0 Model A mm	u_0 Model B mm	u_0 Experiment mm
200	12.5	17.4	16.5
300	18.8	26.0	25.1
400	25.0	34.7	33.7

Remark 2.1.2 Example 2.1.4 illustrates some of the difficulties associated with validation of mathematical models. Typically only a very limited number of experimental observations are available. The information being sought, in this case $c(x)$, is not observable directly but must be inferred from some observable information. If force-displacement data were available for one depth only then it would not be possible to decide whether c is constant or not. Based on two pile tests of differing lengths, it was possible to reject the hypothesis that c is a constant and establish that the available information is consistent with linear variation of the form $c = kx$, but it was not possible to establish with certainty that c varies linearly.

The probability that a model adequately represents physical reality increases with the number of successful predictions of the outcomes of independent experiments but the inherent cognitive uncertainty cannot be removed completely by any number of experiments [38]. In fact, it is possible to construct several models that match a given set of observations. In engineering and scientific applications the simplest model is preferred.

Remark 2.1.3 In order to focus on the main points of calibration in Example 2.1.3 and validation Example 2.1.4, the input data and physical observations were treated without consideration of their statistical aspects. Since there are uncertainties in model parameters, comparing predictions with the outcome of experiments should be understood in a statistical sense.

Let us assume that, having considered statistical uncertainties in the input data, we predict a log-normal probability density function (pdf) for the material constant k in Example 2.1.4. Let us assume further that the criterion for rejection was set at the 95 % confidence interval. We make an experimental observation and compute k_{exp}. Let us assume that k_{exp} falls within the 95 % confidence interval. This shows that the outcome of the experiment is consistent with the prediction based on the model at the 95 % confidence level. This should not be interpreted to mean that we are 95 % confident that the model is valid. What this means is that the chance that a valid model would be rejected is 5 %. The chance of rejecting a valid model would be reduced by setting the confidence interval at (say) 99 %, however the chance of not rejecting an invalid model would then be increased.

Exercise 2.1.4 Using the calibration results for Model B in Example 2.1.3, predict the F_0/u_0 ratio for a pile driven to a depth of 17.5 m.

2.1.4 The scalar elliptic boundary value problem in 1D

Equation (2.7) is a second order elliptic ordinary differential equation (ODE). In Chapter 3 it will be shown that the mathematical model of steady state heat conduction in a bar will result in a second order elliptic ODE also. Although the physical meaning of the unknown functions and the coefficients differ, the mathematical problem is essentially the same. For this reason we will focus on the mathematical problem:

$$-\left(\kappa u'\right)' + cu = f(x) \quad \text{on} \quad 0 \le x \le \ell \tag{2.21}$$

where $\kappa(x) \ge \kappa_0 > 0$, $c(x) \ge 0$ and $f(x)$ are bounded functions subject to the restriction that the indicated operations are defined.

The boundary conditions are analogous to those described in Section 2.1.1, however in the mathematical literature they are known by different names. The displacement boundary condition is called essential or Dirichlet boundary condition[5]. The force boundary condition is called Neumann boundary condition[6]. The spring boundary condition is called mixed or Robin boundary condition[7]. The Neumann and Robin boundary conditions are also called natural boundary conditions.

Although eq. (2.21) may be understood to represent an elastic bar, where u is the displacement vector, or heat conduction in a bar, where u is the temperature, a scalar function, symmetry and antisymmetry are treated differently: When u is a scalar function then the symmetry boundary condition is $u'(\ell/2) = 0$ and the antisymmetry condition is $u(\ell/2) = 0$. The symmetric and antisymmetric boundary conditions for the elastic bar are illustrated in Fig. 2.3.

2.2 Approximate solution

A brief introduction to approximation based on minimizing the error of an integral expression is presented in the following.

Consider the problem given by eq. (2.21) with the boundary conditions $u(0) = u(\ell) = 0$ and let us seek to approximate u by u_n, defined as follows:

$$u_n := \sum_{j=1}^{n} a_j \varphi_j(x) \qquad \varphi_j(x) := x^j(\ell - x) \tag{2.22}$$

such that the integral

$$\mathcal{I} := \frac{1}{2}\int_0^\ell \left(\kappa(u' - u_n')^2 + c(u - u_n)^2\right)\, dx \tag{2.23}$$

[5] Johann Peter Gustav Lejeune Dirichlet 1805-1859.
[6] Franz Ernst Neumann 1798-1895.
[7] Victor Gustave Robin 1855-1897.

is minimum. It will be shown in the following that minimization of the error in the sense of this integral will allow us to find an approximation to the exact solution u without knowing u.

The function u_n is called a *trial function*. The functions $\varphi_j(x)$ are called basis functions. Selection of the type and number of basis functions will, of course, influence the error of approximation $u - u_n$. Discussion of this point is postponed in order to keep our focus on the basic algorithmic structure of the method.

Note that $\varphi_j(0) = \varphi_j(\ell) = 0$, hence u_n satisfies the prescribed boundary conditions for any choice of the coefficients a_i. From the minimum condition we have:

$$\frac{\partial \mathcal{I}}{\partial a_i} = 0 : \quad \int_0^\ell (\kappa(u' - u_n')\varphi_i' + c(u - u_n)\varphi_i)\, dx = 0 \qquad i = 1, 2, \ldots, n. \tag{2.24}$$

Recalling the product rule, we write:

$$\kappa u' \varphi_i' = (\kappa u' \varphi_i)' - (\kappa u')' \varphi_i$$

and substitute this expression into eq. (2.24) to obtain:

$$\underbrace{(\kappa u' \varphi_i)_{x=\ell} - (\kappa u' \varphi_i)_{x=0}}_{0} + \int_0^\ell \underbrace{(-(\kappa u')' + cu)}_{f(x)} \varphi_i \, dx - \int_0^\ell (\kappa u_n' \varphi_i' + c u_n \varphi_i)\, dx = 0$$

where the first two terms are zero on account of the boundary conditions. This equation can be written as:

$$\int_0^\ell (\kappa u_n' \varphi_i' + c u_n \varphi_i)\, dx = \int_0^\ell f(x) \varphi_i \, dx \quad i = 1, 2, \ldots, n. \tag{2.25}$$

Observe that eq. (2.25) represents n algebraic equations in the n unknowns a_i. Therefore we are able to compute an approximation to $u(x)$ without knowing $u(x)$, since only the given function $f(x)$ is needed. Specifically, eq. (2.25) is equivalent to

$$[K]\{a\} = \{r\} \tag{2.26}$$

where $\{a\} := \{a_1 \; a_2 \; \ldots \; a_n\}^T$ and the elements of $[K]$ and $\{r\}$ are, respectively;

$$k_{ij} := \int_0^\ell \left(\kappa(x)\varphi_i'(x)\varphi_j'(x) + c(x)\varphi_i(x)\varphi_j(x)\right) dx \tag{2.27}$$

$$r_i := \int_0^\ell f(x)\varphi_i(x)\, dx. \tag{2.28}$$

Example 2.2.1 Consider the problem on $I = (0, \ell)$

$$-u'' + u = x \qquad u(0) = u(\ell) = 0$$

and assume that the goal is to determine $u'(0)$. Let $\ell = 1$. The exact solution of this problem is

$$u = -\frac{2e}{e^2-1}\sinh x + x \quad \text{and therefore} \quad u'(0) = 1 - \frac{2e}{e^2-1} \approx 0.14908$$

where e is the base of the natural logarithm. We will seek to approximate u using the basis functions $\varphi_j(x)$ given in eq. (2.22) with $n = 2$. Therefore

$$k_{11} = \int_0^1 [(\varphi_1')^2 + \varphi_1^2]\,dx = \int_0^1 \left[(1-2x)^2 + x^2(1-x)^2\right]\,dx = \frac{11}{30}$$
$$k_{12} = k_{21} = \int_0^1 [\varphi_1'\varphi_2' + \varphi_1\varphi_2]\,dx = \int_0^1 \left[(1-2x)(2x-3x^2) + x^3(1-x)^2\right]\,dx = \frac{11}{60}$$
$$k_{22} = \int_0^1 [(\varphi_2')^2 + \varphi_2^2]\,dx = \int_0^1 \left[(2x-3x^2)^2 + x^4(1-x)^2\right]\,dx = \frac{1}{7}$$

and

$$r_1 = \int_0^1 x\varphi_1\,dx = \int_0^1 x^2(1-x)\,dx = \frac{1}{12}$$
$$r_2 = \int_0^1 x\varphi_2\,dx = \int_0^1 x^3(1-x)\,dx = \frac{1}{20}.$$

The problem is then to solve the system of linear equations:

$$\begin{bmatrix} 11/30 & 11/60 \\ 11/60 & 1/7 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} 1/12 \\ 1/20 \end{Bmatrix}.$$

On solving we have $a_1 = 0.14588$, and $a_2 = 0.16279$, therefore the approximate solution is:

$$u_n = u_2 = 0.14588x(1-x) + 0.16279x^2(1-x)$$

and hence $u_n'(0) = 0.14588$ and the relative error is:

$$\frac{|u'(0) - u_n'(0)|}{|u'(0)|} = 0.021 \quad (2.1\ \%).$$

In this example the exact solution was known and hence the relative error in the data of interest could be computed. In general the exact solution is not known therefore the relative error in the data of interest has to be estimated. Methods of estimation will be discussed in subsequent chapters.

Exercise 2.2.1 Determine the relative error of $(u_n')_{x=\ell}$ for the problem solved in Example 2.2.1.

Remark 2.2.1 In engineering computations the goal is to determine some data of interest. The data of interest are typically numbers or functions that depend on the solution $u(x)$ and/or its first derivative. For example, if eq. (2.21) is

understood to represent an elastic bar then we may be interested in computing the reaction force at $x = 0$, defined by $F_0 = (\kappa u')_{x=0}$. If, on the other hand, eq. (2.21) is understood to represent heat conduction then we may be interested in the rate of heat flow exiting the bar at $x = 0$ which is defined by $q_0 = -(\kappa u')_{x=0}$.

We will be interested in finding an approximate solution, computing the data of interest from the approximate solution, as illustrated by Example 2.2.1, and estimating the relative error in the data of interest.

Remark 2.2.2 Observe that eq. (2.26) can be obtained by minimizing the quadratic expression

$$\begin{aligned}\pi(u_n) :=& \frac{1}{2}\int_0^\ell \left(\kappa(u_n')^2 + c(u_n)^2\right)\, dx - \int_0^\ell f u_n\, dx \\ =& \frac{1}{2}\{\, a\,\}^T[K]\{\, a\,\} - \{\, a\,\}^T\{\, r\,\} \end{aligned} \tag{2.29}$$

with respect to a_i. Therefore the function that minimizes $\pi(u_n)$ is also closest to the exact solution u in the sense that the error defined by the integral expression of eq. (2.23) is minimized. This method is known as the Rayleigh-Ritz method[8] or simply as the Ritz method. The functional $\pi(u)$ is called potential energy.

Exercise 2.2.2 Compute the coefficients a_1 and a_2 of Example 2.2.1 by minimizing $\pi(u_n)$ with respect to a_1 and a_2.

2.2.1 Basis functions

We defined the polynomial basis functions $\varphi_j(x) := x^j(\ell - x)$, $j = 1, 2, \ldots, n$ in eq. (2.22) and sought to minimize eq. (2.23) with respect to the coefficients a_j of these basis functions. This led to the definition of n algebraic equations in n unknowns, represented by eq. (2.26). The solution of eq. (2.26) is unique, provided that $[K]$ is a non-singular matrix.

In order to ensure that $[K]$ is non-singular, the basis functions must be linearly independent. By definition, a set of functions $\varphi_j(x)$, $(j = 1, 2, \ldots, n)$ are linearly independent if

$$\sum_{j=1}^{n} a_j \varphi_j(x) = 0$$

implies that $a_j = 0$ for $j = 1, 2, \ldots, n$. It is left to the reader to show that $\varphi_j(x)$, $(j = 1, 2, \ldots, n)$ are linearly independent.

Given a set of linearly independent functions $\varphi_j(x)$, $(j = 1, 2, \ldots, n)$, the set of functions S defined by

$$S := \{\, u_n \,|\, u_n = \sum_{j=1}^{n} a_j \varphi_j(x)\}$$

[8] Lord Rayleigh (John William Strutt) 1842-1919, Walter Ritz 1878-1909.

is called the span and $\varphi_j(x)$ are basis functions of S.

We could have defined other polynomial basis functions, for example;

$$u_n := \sum_{i=1}^{n} c_i \psi_i(x) \qquad \psi_i(x) := x(\ell - x)^i. \tag{2.30}$$

When two sets of basis functions $\{\varphi\} := \{\varphi_1\ \varphi_2\ \ldots\ \varphi_n\}^T$ and $\{\psi\} := \{\psi_1\ \psi_2\ \ldots\ \psi_n\}^T$ can be written as

$$\{\psi\} = [B]\{\varphi\} \tag{2.31}$$

where $[B]$ is an invertible matrix of constant coefficients then both sets of basis functions are said to have the same span. The following exercise demonstrates that the approximate solution depends on the span of the basis functions, not on the basis functions.

Exercise 2.2.3 Solve the problem of Example 2.2.1 using the basis functions

$$\psi_1(x) = x(1-x), \qquad \psi_2(x) = x(1-x)(1-2x)$$

and show that the solution $u_2 = b_1\psi_1(x) + b_2\psi_2(x)$ is the same as the solution in Example 2.2.1. In this exercise the span is the set of polynomials of degree 3, subject to the restriction that they vanish in the boundary points.

Exercise 2.2.4 Let $\varphi_i(x) = x^i(\ell - x)$, $\psi_i(x) = x(\ell - x)^i$ and

$$u_n = \sum_{i=1}^{3} a_i \varphi_i(x) = \sum_{i=1}^{3} c_i \psi_i(x).$$

Determine the matrix $[B]$ as defined in eq. (2.31) and, assuming that the values of a_i are given, find an expression for c_i in terms a_i $(i = 1, 2, 3)$ and $[B]$.

2.3 Generalized formulation in one dimension

We have seen in Section 2.2 that it was possible to obtain an approximate solution to a differential equation without knowing the exact solution. This depended on a seemingly fortuitous choice of the integral expression $\mathcal{I}$ and zero boundary conditions, allowing us to replace the unknown exact solution with the known function f following integration by parts. In this section the reasons for the choice of I are explained in a general setting, without restriction on the boundary conditions.

Once again our starting point is eq. (2.21):

$$-\left(\kappa u'\right)' + cu = f(x)$$

subject to boundary conditions to be discussed later. Let us multiply this equation by an arbitrary function $v(x)$ defined on $I = (0, \ell)$ and integrate:

$$\int_0^\ell \Big(-\left(\kappa u'\right)' + cu\Big)\, v\, dx = \int_0^\ell f v\, dx. \tag{2.32}$$

Clearly, if u is the solution of eq. (2.21) then this equation will be satisfied for all v for which the indicated operations are defined. Integrating the first term by parts:

$$-\int_0^\ell (\kappa u')'\, v\, dx = -\int_0^\ell \left[(\kappa u' v)' - \kappa u' v'\right]\, dx =$$
$$-\left[\kappa u' v\right]_{x=\ell} + \left[\kappa u' v\right]_{x=0} + \int_0^\ell \kappa u' v'\, dx$$

we have:

$$\int_0^\ell (\kappa u' v' + cuv)\, dx = \int_0^\ell f v\, dx + \left[\kappa u' v\right]_{x=\ell} - \left[\kappa u' v\right]_{x=0}. \tag{2.33}$$

Note that the integrand $(\kappa u')' v$ became $\kappa u' v'$ plus two boundary terms. This equation will be the starting point for our discussion of the generalized formulation. The specific statement of the generalized formulation for a particular problem depends on the boundary conditions. First, however, some useful definitions and notation are introduced.

2.3.1 Definitions and notation

We denote the set of functions defined on the interval $I = (0, \ell)$ that satisfy the inequality:

$$E(I) := \left\{ u \,\Big|\, \int_0^\ell \left(\kappa (u')^2 + cu^2\right)\, dx \le C < \infty \right\} \tag{2.34}$$

where C is some positive number; $\kappa \ge \kappa_0 > 0$ and $c \ge 0$. $E(I)$ is called the *energy space*. For any $u \in E(I)$ and $v \in E(I)$ the integral expressions in eq. (2.33) are defined. This follows from the Schwarz inequality[9], see Appendix A.

When $u(0) = \hat{u}_0$ and/or $u(\ell) = \hat{u}_\ell$ are specified on the boundaries then the boundary condition is called an *essential* or *Dirichlet* boundary condition. The functions in $E(I)$ that satisfy the essential boundary conditions are called *admissible* functions. The set of all admissible functions is called the *trial space* and is denoted by $\tilde{E}(I)$. This notation should be understood as follows:

(a) If essential boundary conditions are specified at $x = 0$ and $x = \ell$ then

$$\tilde{E}(I) := \{u \,|\, u \in E(I),\ u(0) = \hat{u}_0,\ u(\ell) = \hat{u}_\ell\}. \tag{2.35}$$

Corresponding to $\tilde{E}(I)$ is the *test space* $E^0(I)$ defined as follows:

$$E^0(I) := \{u \,|\, u \in E(I),\ u(0) = 0,\ u(\ell) = 0\}. \tag{2.36}$$

[9] Hermann Amandus Schwarz 1843-1921.

(b) If essential boundary condition is specified only at $x = 0$ then

$$\tilde{E}(I) := \{u \,|\, u \in E(I),\ u(0) = \hat{u}_0\} \tag{2.37}$$

$$E^0(I) := \{u \,|\, u \in E(I),\ u(0) = 0\}. \tag{2.38}$$

(c) If essential boundary condition is specified only at $x = \ell$ then

$$\tilde{E}(I) := \{u \,|\, u \in E(I),\ u(\ell) = \hat{u}_\ell\} \tag{2.39}$$

$$E^0(I) := \{u \,|\, u \in E(I),\ u(\ell) = 0\}. \tag{2.40}$$

(d) If the essential boundary conditions are homogeneous, i.e., $\hat{u}_0 = 0$, $\hat{u}_\ell = 0$ then $\tilde{E}(I) = E^0(I)$.

(e) If essential boundary conditions are not prescribed on either boundary then $\tilde{E}(I) = E^0(I) = E(I)$.

(f) If periodic boundary conditions are prescribed then both the trial and test spaces are:

$$\hat{E}(I) = \{u \,|\, u \in E(I),\ u(0) = u(\ell)\}. \tag{2.41}$$

We are now in a position to discuss generalized formulations for various boundary conditions.

Remark 2.3.1 Note that $\tilde{E}(I)$ is not a linear space. Refer to the Appendix, Section A.2. It is seen that $\tilde{E}(I)$ does not satisfy condition 1 whereas $E^0(I)$ and $\hat{E}(I)$ satisfy all of the conditions of Section A.2.

2.3.2 Essential boundary conditions

Essential boundary conditions are enforced by restriction. This was done in the special case discussed in Section 2.2 where the homogeneous essential boundary conditions $\hat{u}_0 = \hat{u}_\ell = 0$ were used and the basis functions were defined in eq. (2.22) so that the boundary conditions prescribed on u were satisfied for an arbitrary choice of the coefficients a_i.

The known boundary conditions are imposed on the trial functions u and the test function v is set to zero on the boundary points where essential boundary conditions were prescribed. In this way the boundary terms (the terms in the square bracket in eq. (2.33)) vanish and the generalized formulation is stated as follows:
"Find $u \in \tilde{E}(I)$ such that

$$\underbrace{\int_0^\ell (\kappa u'v' + cuv)\, dx}_{B(u,v)} = \underbrace{\int_0^\ell fv\, dx}_{F(v)} \quad \text{for all } v \in E^0(I)". \tag{2.42}$$

We will use the shorthand notation $B(u, v)$ for the left hand side and $F(v)$ for the right hand side, as indicated in eq. (2.42). $B(u, v)$ is a symmetric *bilinear form*,

i.e., it is linear with respect to each of its arguments and $B(u,v) = B(v,u)$ and $F(v)$ is a *linear functional*. The properties of bilinear forms and linear functionals are given in Appendix A.

Alternatively we can select an arbitrary function $u^\star$ from $\tilde{E}(I)$ and write:

$$u = \bar{u} + u^\star \tag{2.43}$$

where $\bar{u} \in E^0(I)$. Clearly, the prescribed boundary conditions are satisfied for any choice $\bar{u} \in E^0(I)$. Substituting eq. (2.43) into eq. (2.33), the generalized formulation can be stated as follows: "Find $\bar{u} \in E^0(I)$ such that

$$\underbrace{\int_0^\ell (\kappa \bar{u}'v' + c\bar{u}v)\, dx}_{B(\bar{u},v)} = \underbrace{\int_0^\ell fv\, dx - \int_0^\ell (\kappa (u^\star)'v' + cu^\star v)\, dx}_{F(v)} \tag{2.44}$$

for all $v \in E^0(I)$".

Example 2.3.1 Let us state the generalized formulation for the following problem:

$$-u'' = (2+x)e^x \qquad u(0) = 1,\ u(2) = -1.$$

In this case $\tilde{E}(I) = \{u \,|\, u \in E(I),\ u(0) = 1,\ u(2) = -1\}$. Let us select $u^\star = 1 - x$ and substitute $u = \bar{u} + u^\star$ into eq. (2.33). The statement of the generalized formulation is now: "Find $\bar{u} \in E^0(I)$ such that $B(\bar{u}, v) = F(v)$ for all $v \in E^0(I)$" where

$$B(\bar{u}, v) := \int_0^2 \bar{u}'v'\, dx, \qquad F(v) := \int_0^2 (2+x)e^x v\, dx - \int_0^2 (-1)v'\, dx.$$

Example 2.3.2 In this example it is shown that eq. (2.33) leads to the same system of equations as obtained in Section 2.2. To obtain an approximation to the solution of eq. (2.21), we substitute u_n from eq. (2.22) for u and similarly substitute v_n for v:

$$v_n := \sum_{i=1}^{n} b_i \varphi_i(x), \quad \varphi_i(x) := x^i(\ell - x)$$

where b_i $i = 1, 2, \ldots, n$ are a set of arbitrary numbers. Since $v_n(0) = v_n(\ell) = 0$, the terms in the square brackets in eq. (2.33) vanish and we have

$$\{b\}^T[K]\{a\} = \{b\}^T\{r\}$$

where $\{b\} := \{b_1\ b_2\ \ldots\ b_n\}^T$ and the definitions of $[K]$ and $\{r\}$ are the same as in eq. (2.27). Equivalently,

$$\{b\}^T\left([K]\{a\} - \{r\}\right) = 0.$$

Since this relationship must hold for any choice of $\{b\}$, we must have $[K]\{a\} = \{r\}$ which is exactly the same as the result obtained in Section 2.2 with k_{ij} (resp. r_i) defined by eq. (2.27) (resp. eq. (2.28)).

Exercise 2.3.1 Show that

(a) $B(u_1 + u_2, v) = B(u_1, v) + B(u_2, v)$

(b) $B(u + v, u + v) = B(u, u) + 2B(u, v) + B(v, v)$.

2.3.3 Neumann boundary conditions

When u' or more commonly $F = \kappa u'$ is prescribed on a boundary then the boundary condition is called a Neumann boundary condition. The treatment of Neumann boundary conditions is straightforward. Let $F_0 = (\kappa u')_{x=0}$ and $F_\ell = (\kappa u')_{x=\ell}$ be given. Substituting F_ℓ and F_0 into eq. (2.33), the generalized formulation is stated as follows: "Find $u \in E(I)$ such that $B(u, v) = F(v)$ for all $v \in E(I)$" where

$$B(u, v) = \int_0^\ell (\kappa u'v' + cuv)\, dx, \quad F(v) = \int_0^\ell fv\, dx + F_\ell v(\ell) - F_0 v(0). \quad (2.45)$$

Note that there are no restrictions on u or v at the endpoints.

Remark 2.3.2 When $c = 0$ and Neumann boundary conditions are prescribed then, since eq. (2.45) must hold for all choices of $v \in E(I)$, it must hold for $v = C$ where C is a constant. Therefore we must have:

$$\int_0^\ell f\, dx + F_\ell - F_0 = 0. \quad (2.46)$$

In other words, f, F_0 and F_ℓ cannot be assigned arbitrarily. The tractions acting on the bar and the bar forces acting on the boundary points must be in equilibrium.

2.3.4 Robin boundary conditions

A linear combination of u' and u is given at the boundary:

$$(\kappa u')_{x=0} = \beta_0(u(0) - U_0)$$
$$(\kappa u')_{x=\ell} = \beta_\ell(U_\ell - u(\ell))$$

where β_0 and β_ℓ are positive numbers. Substituting these expressions into eq. (2.33), the generalized formulation is once again stated as follows: "Find $u \in E(I)$ such that $B(u, v) = F(v)$ for all $v \in E(I)$" where

$$B(u, v) := \int_0^\ell (\kappa u'v' + cuv)\, dx + \beta_0 u(0)v(0) + \beta_\ell u(\ell)v(\ell)$$
$$F(v) := \int_0^\ell fv\, dx + \beta_0 U_0 v(0) + \beta_\ell U_\ell v(\ell).$$

There are no restrictions on u or v at the endpoints. The spring boundary condition described in Section 2.1.1 is a Robin boundary condition.

Example 2.3.3 Any combination of Dirichlet, Neumann and Robin boundary conditions may be prescribed. For example, let us consider the problem

$$-(\kappa u')' + cu = f(x) \qquad u(0) = \hat{u}_0, \quad (\kappa u')_{x=\ell} = \beta_\ell(U_\ell - u(\ell)).$$

In this case $\tilde{E}(I)$ is defined by eq. (2.37), $E^0(I)$ is defined by eq. (2.38) and

$$B(\bar{u}, v) := \int_0^\ell (\kappa \bar{u}' v' + c\bar{u}v)\, dx + \beta_\ell u(\ell) v(\ell)$$
$$F(v) := \int_0^\ell fv\, dx + \beta_\ell U_\ell v(\ell) - \int_0^\ell (\kappa (u^\star)' v' + cu^\star v)\, dx$$

where $u^\star$ is an arbitrary fixed function from $\tilde{E}(I)$. For example, we may select $u^\star = \hat{u}_0(1 - x/\ell)$ or simply $u^\star = \hat{u}_0$.

The generalized formulation of this problem is stated as follows: "Find $\bar{u} \in E^0(I)$ such that $B(\bar{u}, v) = F(v)$ for all $v \in E^0(I)$". The exact solution is then: $u = \bar{u} + u^\star$.

Exercise 2.3.2 State the generalized formulation for the following problem:

$$-(\kappa u')' + cu = f(x) \qquad (\kappa u')_{x=0} = -\hat{q}_0, \quad u(\ell) = 0.$$

Exercise 2.3.3 State the generalized formulation for the following problem:

$$-(\kappa u')' + cu = f(x) \qquad (\kappa u')_{x=0} = \beta_0(u(0) - U_0), \quad u(\ell) = \hat{u}_\ell.$$

Exercise 2.3.4 State the generalized formulation for the following problem:

$$-(\kappa u')' + cu = f(x) \qquad (\kappa u')_{x=0} = -\hat{q}_0, \quad (\kappa u')_{x=\ell} = \beta_\ell(U_\ell - u(\ell)).$$

2.4 Finite element approximations

We have re-cast a differential equation in the form of a generalized formulation which reads: "Find $u \in \tilde{E}(I)$ such that $B(u, v) = F(v)$ for all $v \in E^0(I)$". It may appear that nothing has been gained: This problem is more difficult to solve than the differential equation was, since there are an infinite number of trial functions for u that must be tested against an infinite number of v.

One of the main advantages of the generalized formulation is that it serves as a framework for obtaining approximate solutions. To obtain an approximate solution we construct a finite-dimensional subspace of $E(I)$, as we have done in Section 2.2, where we selected

$$u_n = \sum_{j=1}^{n} a_j \varphi_j(x)$$

with $n = 2$. The family of functions that can be written in this way will be denoted by $S(I)$. The functions $\varphi_j(x)$, called basis functions, will be defined

such that $S(I) \subset E(I)$. The number n is the dimension of $S(I)$. We will use the notation $\tilde{S}(I) := S(I) \cap \tilde{E}(I)$; $S^0(I) = S(I) \cap E^0(I)$. The dimension of space $S^0(I)$, that is, the number of linearly independent functions in $S^0(I)$, is called the number of degrees of freedom and denoted by N.

In the finite element method the space S is constructed by partitioning the solution domain into elements and defining polynomial basis functions on the elements. Approximation spaces so constructed are called *finite element spaces*. A particular partition is called a finite element mesh and will be denoted by Δ. The number of elements will be denoted by $M(\Delta)$. A simple illustration is given in the following example.

Example 2.4.1 Typical finite element basis functions in one dimension are illustrated in Fig. 2.5 where the domain $I = (0, \ell)$ is partitioned into three intervals (i.e., $M(\Delta) = 3$), called finite elements and the polynomial degrees $p_1 = 2$, $p_2 = 1$ and $p_3 = 3$ are assigned. The length of the elements is denoted by $\ell_k,\ k = 1, 2, 3$. There are four node points, labeled $x_i,\ i = 1, 2, 3, 4$. There are seven basis functions, labeled $\varphi_1(x), \ldots, \varphi_7(x)$. The numbering of the basis functions is arbitrary, however it is good practice to number them by polynomial degree: The first four basis functions are piecewise linear. For example,

$$\varphi_2(x) = \begin{cases} \dfrac{x - x_1}{x_2 - x_1} & \text{if } x_1 \le x \le x_2 \\ \dfrac{x_3 - x}{x_3 - x_2} & \text{if } x_2 < x \le x_3 \\ 0 & \text{otherwise.} \end{cases}$$

The basis functions $\varphi_5(x)$, $\varphi_6(x)$ are quadratic functions. For example,

$$\varphi_6(x) = \begin{cases} (x - x_3)(x - x_4) & \text{if } x_3 \le x \le x_4 \\ 0 & \text{otherwise.} \end{cases}$$

The basis function $\varphi_7(x)$ is a cubic polynomial on element 3 and is zero outside of element 3. The finite element space, characterized by the mesh and the polynomial degree of elements as shown in Fig. 2.5, is the set of all functions that can be written in the form:

$$u = \sum_{j=1}^{7} a_j \varphi_j(x). \tag{2.47}$$

Exercise 2.4.1 Refer to Fig. 2.5. Write down the basis function $\varphi_7(x)$.

Exercise 2.4.2 Show that the set of basis functions

$$\begin{aligned} \psi_1(x) := \varphi_1(x) + \varphi_2(x), \quad \psi_2(x) := \varphi_2(x) - \varphi_1(x), \\ \psi_j(x) = \varphi_j(x) \quad j = 3, 4, \ldots, 7 \end{aligned} \tag{2.48}$$

has the same span as the set $\varphi_j(x)$, $j = 1, 2, \ldots, 7$ defined in Example 2.4.1.

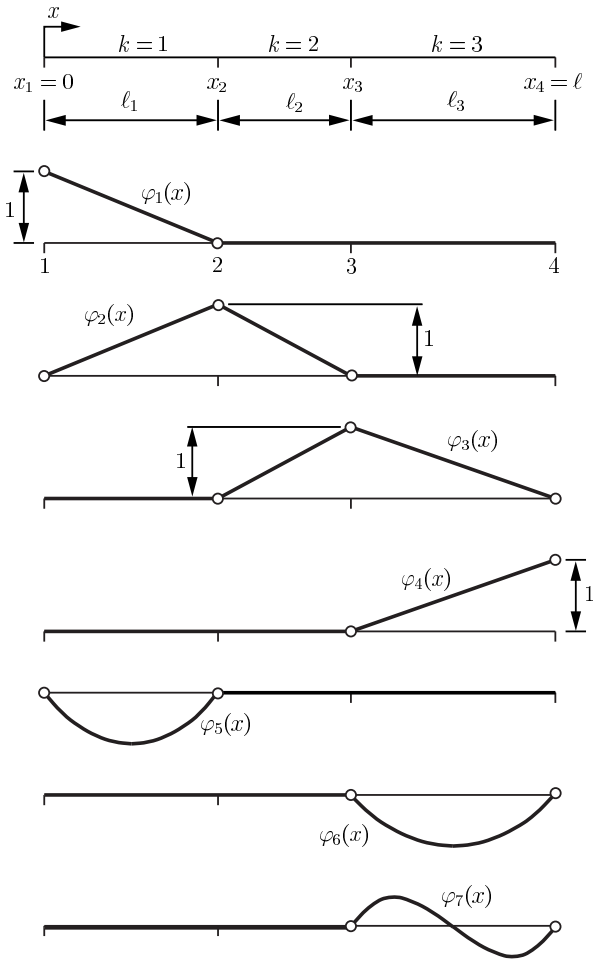

Figure 2.5: Typical finite element basis functions in one dimension.

Remark 2.4.1 The key difference between the original form of the Rayleigh-Ritz method and the finite element method is that in the Rayleigh-Ritz method the basis functions are analytic functions defined on the entire solution domain whereas in the finite element method the basis functions are characterized by piecewise polynomials that are nonzero over a few elements only. In one-dimensional applications, for example, they are nonzero over at most two elements, as seen in Fig. 2.5. This makes it possible to construct algorithms suitable for handling a great variety of problems very efficiently.

The partition of the domain into finite elements makes it possible to construct basis functions analogous to those discussed in Example 2.4.1 on complicated domains such as shown in Fig. 2.6.

In the finite element method the approximating functions are piecewise poly-

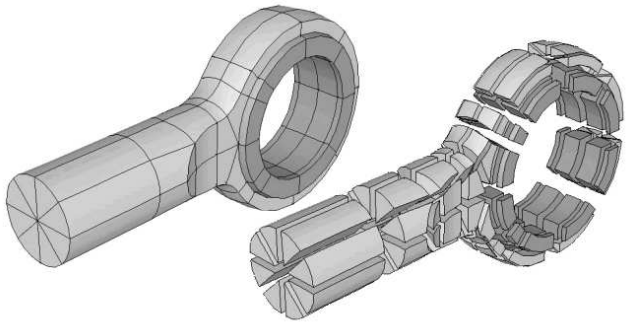

Figure 2.6: Example of a 3-dimensional finite element mesh.

nomials. There are two reasons for this: Piecewise polynomials are advantageous from the point of view of implementation and they have favorable approximation properties.

2.4.1 Error measures and norms

Since we will be solving various problems approximately, we need to ask; what is the error of approximation? Various quantitative measures of error are used in connection with finite element analyses. The most useful measure is the relative error in terms of the data of interest which we computed in Example 2.2.1

There are other measures, called norms, useful for measuring the quality of the approximate solution. Three norms are defined for functions of a single variable in the following. Their generalization to two and three dimensions is straightforward. Norms are analogous to the length of a vector in Euclidean space[10] . The definitive properties of norms are listed in the Appendix, Section A.1.

The energy norm

It is natural to use the energy norm in connection with the formulation discussed in Section 2.2 (page 27) because the coefficients a_j in eq. (2.22) are determined in such a way that the error measured in the energy norm is minimal. This is Theorem 2.4.2 (see page 41). By definition:

$$\|u\|_E := \sqrt{\frac{1}{2}B(u,u)}. \tag{2.49}$$

In Example 2.2.1

$$B(u,u) = \int_0^\ell \left(\kappa(u')^2 + cu^2\right)\, dx$$

[10] Euclid of Alexandria about 325-265 BC.

and $\|u - u_n\|_E = 1.01606 \times 10^{-3}$ and $\|u\|_E = 0.10074$. Therefore the percent relative error measured in energy norm is

$$(e_r)_E := \frac{\|u - u_n\|_E}{\|u\|_E} = 1.01\ \%.$$

The maximum norm

The maximum norm of a continuous function $u(x)$ defined on the interval $\bar{I}$ is:

$$\|u\|_{\max} := \max_{x \in \bar{I}} |u(x)| \tag{2.50}$$

and the relative error in maximum norm is defined by:

$$(e_r)_{\max} := \frac{\|u - u_n\|_{\max}}{\|u\|_{\max}}. \tag{2.51}$$

Often the percent relative error is given. The maximum norm is usually approximated by computing the maximum on some fine grid. The abscissa at which $\|u - u_n\|_{\max}$ is computed may be different from the abscissa at which the reference value $\|u\|_{\max}$ was computed.

In Example 2.2.1, using 100 equally spaced grid points, $\max|u - u_n| = 2.3 \times 10^{-4}$ at $x = 0.51$; $\max|u| = 5.83 \times 10^{-2}$ at $x = 0.60$. Therefore the relative error is $(e_r)_{\max} = 0.39\ \%$.

The L_2 norm

The L_2 norm of a function u defined on the interval $I = (0, \ell)$ is:

$$\|u\|_{L_2} := \sqrt{\int_0^\ell u^2\, dx}. \tag{2.52}$$

The definition of relative error in L_2 norm is analogous to eq. (2.51). In Example 2.2.1 $\|u - u_n\|_{L_2} = 1.487 \times 10^{-4}$ and $\|u\|_{L_2} = 4.183 \times 10^{-2}$. Therefore the percent relative error measured in L_2 norm is

$$(e_r)_{L_2} := \frac{\|u - u_n\|_{L_2}}{\|u\|_{L_2}} = 0.36\ \%.$$

Remark 2.4.2 The error depends on the norm in which it is measured. The choice of the norm depends on the purpose of computation.

Exercise 2.4.3 Obtain an approximate solution for the problem

$$-u'' + u = 1, \qquad u(0) = u(1) = 0$$

by minimizing π given by eq. (2.29). Use $n = 2$ and the same basis functions as in Example 2.2.1. Determine the exact solution and compute the relative error in maximum norm and in energy norm.

2.4.2 The error of approximation in energy norm

In Section 2.2 we minimized the integral expression (2.23) to obtain an approximate solution to the problem (2.21). We are now in a position to generalize this to any combination of the three kinds of boundary conditions discussed in Section 2.3.

In the following we will denote the exact solution by u_{EX} and the finite element solution by u_{FE}. The approximation problem is stated as follows: "Find $u_{FE} \in \tilde{S}(I)$ such that $B(u_{FE}, v) = F(v)$ for all $v \in S^0(I)$".

Theorem 2.4.1 The error of approximation $e := u_{EX} - u_{FE}$ is orthogonal to all test functions in $S^0(I)$ in the following sense:

$$B(e, v) = 0 \quad \text{for all } v \in S^0(I). \tag{2.53}$$

This is a basic property of the error of approximation, known as the Galerkin orthogonality[11]. Proof: Since $S^0(I) \subset E^0(I)$,

$$\begin{aligned} B(u_{EX}, v) &= F(v) \quad \text{for all } v \in S^0(I) \\ B(u_{FE}, v) &= F(v) \quad \text{for all } v \in S^0(I). \end{aligned}$$

On subtracting the second equation from the first eq. (2.53) is obtained. □

An important theorem is proven in the following. This theorem establishes that the finite element method will select the coefficients of the basis functions in such a way that the energy norm of the error $\|e\|_E$ will be minimum.

Theorem 2.4.2 The finite element solution minimizes the error in energy norm on the space $\tilde{S}(I)$:

$$\|u_{EX} - u_{FE}\|_E = \min_{u \in \tilde{S}(I)} \|u_{EX} - u\|_E. \tag{2.54}$$

We have seen a direct application of this theorem in Section 2.2. Once again we write $e := u_{EX} - u_{FE}$. For an arbitrary $v \in S^0(I)$, $\|v\|_E \neq 0$, we have

$$\|e + v\|_E^2 = \frac{1}{2} B(e + v, e + v) = \frac{1}{2} B(e, e) + \underbrace{B(e, v)}_{0} + \underbrace{\frac{1}{2} B(v, v)}_{\|v\|_E^2 > 0}.$$

By eq. (2.53) $B(e, v) = 0$ therefore for any $\|v\|_E \neq 0$ we have $\|e + v\|_E^2 > \|e\|_E^2$ which was to be proven. □

This theorem shows that the selection $S(I)$ is of crucial importance, since the error of approximation is determined by $S(I)$. This theorem also shows that if u_{EX} happens to lie in $S(I)$ then $u_{FE} = u_{EX}$. Furthermore, this theorem shows that if we construct a sequence of finite element spaces $S_1 \subset S_2 \subset \cdots \subset S_m$ and compute the corresponding finite element solutions $u_{FE}^{(1)}$, $u_{FE}^{(2)}, \ldots, u_{FE}^{(m)}$, then the error measured in energy norm will decrease monotonically with respect to increasing m.

[11] Boris Grigorievich Galerkin, 1871-1945.

2.5 FEM in one dimension

In this section the key algorithmic procedures common to all finite element computer programs are outlined in the simplest setting. Although the discussion covers the one-dimensional case only, analogous procedures apply to two- and three-dimensional problems.

2.5.1 The standard element

In order make the computation of the coefficient matrices and load vectors suitable for implementation in a computer program, the computations are performed element by element. In one dimension the kth element is characterized by the node points x_k and x_{k+1} of the mesh Δ. The mesh is the set of elements $I_k := \{x \,|\, x_k < x < x_{k+1}\}$, $k = 1, 2, \ldots, M(\Delta)$ and $\ell_k := x_{k+1} - x_k$ is the size of element k.

In order to standardize the element-level computations, a standard element $I_{\rm st}$ is defined:

$$I_{\rm st} := \{\xi \,|\, -1 < \xi < +1\}. \tag{2.55}$$

The standard element is mapped into the kth element by the mapping function:

$$x = Q_k(\xi) := \frac{1-\xi}{2} x_k + \frac{1+\xi}{2} x_{k+1} \qquad \xi \in I_{\rm st}. \tag{2.56}$$

The inverse map is:

$$\xi = Q_k^{-1}(x) := \frac{2x - x_k - x_{k+1}}{x_{k+1} - x_k} \qquad x \in I_k. \tag{2.57}$$

Remark 2.5.1 The mapping of the standard element $I_{\rm st}$ onto the "real" element I_k is not unique. For example, the mapping

$$x = \frac{1}{2}\xi(\xi - 1)x_k + \frac{1}{2}\xi(1+\xi)x_{k+1} \qquad \xi \in I_{\rm st}$$

would serve the same purpose as eq. (2.56). It will be shown, however, that in general the simplest mapping functions are preferable. In special cases there are some exceptions, however.

2.5.2 The standard polynomial space

The polynomials of degree p defined on the standard element $I_{\rm st}$ will be denoted by $\mathcal{S}^p(I_{\rm st})$. The basis functions that span $\mathcal{S}^p(I_{\rm st})$ are usually called *shape functions*. We will discuss two types of shape functions used in FEA programs: Lagrange[12] and *hierarchic* shape functions that are scaled integrals of Legendre[13] polynomials. We will use the notation $N_i(\xi)$ $(i = 1, 2, \ldots, p+1)$ for both

[12] Joseph-Louis Lagrange 1736-1813.
[13] Adrien-Marie Legendre 1752-1833.

types of shape functions. For $\mathcal{S}^1(I_{\rm st})$ the shape functions for both sets are:

$$N_1(\xi) := \frac{1-\xi}{2} \tag{2.58}$$

$$N_2(\xi) := \frac{1+\xi}{2}. \tag{2.59}$$

Lagrange shape functions

The standard domain is partitioned into p sub-intervals. The lengths of the sub-intervals may vary. The node points are: $\xi_1 = -1$, $\xi_2 = 1$ and $-1 < \xi_3 < \xi_4 < \cdots < \xi_{p+1} < 1$. The shape functions for $\mathcal{S}^p(I_{\rm st})$ are:

$$N_i(\xi) := \prod_{\substack{k=1 \\ k\neq i}}^{p+1} \frac{\xi - \xi_k}{\xi_i - \xi_k} \qquad i = 1, 2, \ldots, p+1 \qquad \xi \in I_{\rm st}. \tag{2.60}$$

Note that

$$N_i(\xi_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \qquad \text{and} \qquad \sum_{i=1}^{p+1} N_i(\xi) = 1. \tag{2.61}$$

For example, for $p = 2$ the equally spaced node points are $\xi_1 = -1$, $\xi_2 = 1$, $\xi_3 = 0$ and the three Lagrange shape functions as illustrated in Fig. 2.7.

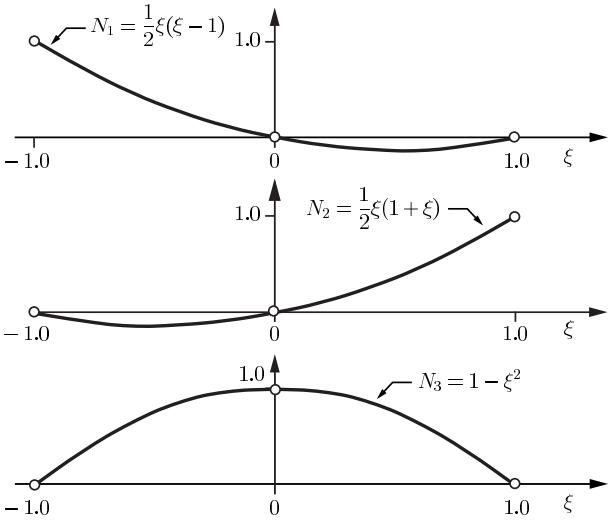

Figure 2.7: Lagrange shape functions in one dimension, $p = 2$.

Exercise 2.5.1 Write down the Lagrange shape functions for $\mathcal{S}^3(I_{\rm st})$ using equally spaced node points.

Hierarchic shape functions based on Legendre polynomials

It is advantageous to retain the definitions (2.58) and (2.59) for $p = 1$ and then for $p \geq 2$ define the shape functions as follows:

$$N_i(\xi) = \sqrt{\frac{2i-3}{2}} \int_{-1}^{\xi} P_{i-2}(t)\, dt \quad i = 3, 4, \ldots, p+1 \tag{2.62}$$

where $P_i(t)$ are the Legendre polynomials. These shape functions have the following important properties:

(a) Orthogonality. For $i, j \geq 3$:

$$\int_{-1}^{+1} \frac{dN_i}{d\xi} \frac{dN_j}{d\xi}\, d\xi = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases} \tag{2.63}$$

This property follows directly from the orthogonality of Legendre polynomials, see eq. (A.13) in the Appendix.

(b) The shape functions of $\mathcal{S}^{p-1}(I_{\text{st}})$ are a subset of the shape functions of $\mathcal{S}^{p}(I_{\text{st}})$. Shape functions that have this property are called hierarchic shape functions.

(c) These shape functions vanish at the endpoints of I_{st}: $N_i(-1) = N_i(+1) = 0$ for $i \geq 3$.

The first five hierarchic shape functions are shown in Fig. 2.8. Observe that all roots lie in I_{st}.

Exercise 2.5.2 Show that for the hierarchic shape functions, defined by eq. (2.62), $N_i(-1) = N_i(+1) = 0$ for $i \geq 3$.

Exercise 2.5.3 Show that the hierarchic shape functions defined by eq. (2.62) can be written in the form:

$$N_i(\xi) = \frac{1}{\sqrt{2(2i-3)}} \left(P_{i-1}(\xi) - P_{i-3}(\xi)\right) \qquad i = 3, 4, \ldots \tag{2.64}$$

Hint: Use eq. (A.12) in Appendix A.

2.5.3 Finite element spaces

We are now in a position to provide a precise definition of finite element spaces in one-dimension: A finite element space S is a set of continuous functions characterized by the finite element mesh Δ, the assigned polynomial degrees p_k and the mapping functions Q_k, $k = 1, 2, \ldots, M(\Delta)$. Specifically,

$$S = S(\Delta, \mathbf{p}, \mathbf{Q}) = \{u \,|\, u \in E(I),\ u(Q_k(\xi)) \in \mathcal{S}^{p_k}(I_{\text{st}}),\ k = 1, 2, \ldots, M(\Delta)\} \tag{2.65}$$

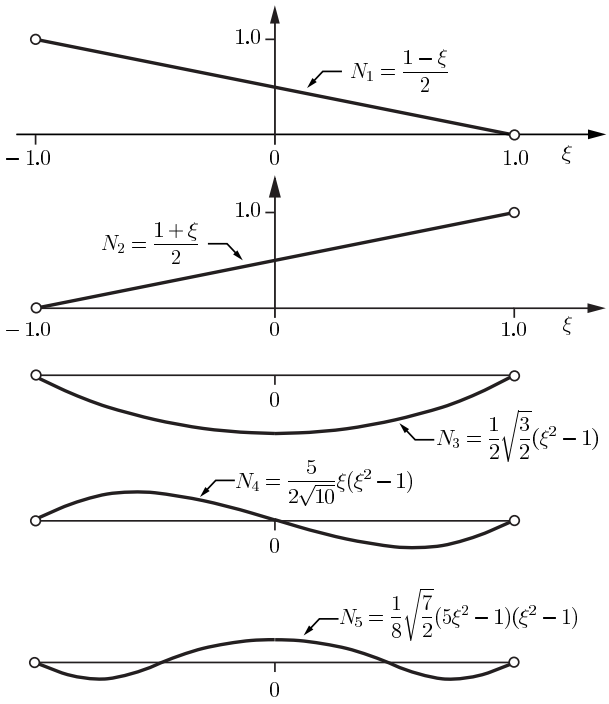

Figure 2.8: The first five hierarchic shape functions in one dimension.

where $\mathbf{p}$ and $\mathbf{Q}$ represent, respectively, the arrays of the assigned polynomial degrees and the mapping functions. The expression $u(Q_k(\xi)) \in \mathcal{S}^{p_k}(I_{\text{st}})$ indicates that on element I_k $u(x)$ is mapped from the standard polynomial space $\mathcal{S}^{p_k}(I_{\text{st}})$. If linear mapping is used then the finite element space is comprised of piecewise polynomial functions.

It is shown in Section 2.6.4 that if $u \in E(I)$ then u must be a continuous function. Of course, it would be possible to enforce levels of continuity higher than the minimum required. For example, we could require the first derivatives of the basis functions to be continuous also. However enforcement of higher levels of continuity than the minimum required is a restriction on the space of admissible functions which, in view of Theorem 2.4.2, is detrimental to the generality and good overall approximation properties of the method. For this reason we will be concerned with finite element spaces that are exactly and minimally continuous. In other words, not more than the minimal continuity required for satisfying the condition that $u \in E(I)$ will be enforced.

We will be concerned with approximations based sequences of finite element spaces created by systematic mesh refinement or increase of the polynomial degree of elements, or a combination of both. Systematic enlargement of finite element spaces by mesh refinement, increase of the polynomial degree(s) of elements or a combination of both is called, respectively, h-extension, p-extension

and *hp*-extension. When a sequence of finite element spaces has the property that $S_1 \subset S_2 \subset S_3 \subset \cdots S_n$ then the sequence is called hierarchic sequence.

A family of meshes Δ_k is said to be quasiuniform if there exist positive constants C_1, C_2, independent of k, such that

$$C_1 \leq \frac{h_{\max}^{(k)}}{h_{\min}^{(k)}} \leq C_2, \quad k = 1, 2, \ldots \tag{2.66}$$

where $h_{\max}^{(k)}$ (resp. $h_{\min}^{(k)}$) is the diameter of the largest (resp. smallest) element in mesh Δ_k.

A mesh is geometrically graded toward the point $x = 0$ on the interval $0 < x < \ell$ if the node points are located as follows:

$$x_k = \begin{cases} 0 & \text{for } k = 1 \\ q^{M(\Delta)+1-k}\ell & \text{for } k = 2, 3, \ldots, M(\Delta)+1 \end{cases}$$

where $0 < q < 1$ is called grading factor or common factor. Such meshes are called geometric meshes.

A mesh is said to be a radical mesh if on the interval $0 < x < \ell$ the node points are located by

$$x_k = \left(\frac{k-1}{M(\Delta)}\right)^{\theta} \ell \quad \theta > 1, \quad k = 1, 2, \ldots, M(\Delta)+1.$$

It will be shown in Chapter 6 that for a large and important class of problems the ideal meshes are geometric meshes when the mesh is fixed and the polynomial degree of elements is increased and the ideal meshes are radical meshes when the polynomial degree is fixed and the number of elements is increased.

Exercise 2.5.4 Consider the family of mapping functions

$$x = Q_k(\alpha, \xi) := \frac{1}{2}\xi(\xi-1)x_k + (1-\xi^2)(\alpha x_k + (1-\alpha)x_{k+1}) + \frac{1}{2}\xi(1+\xi)x_{k+1}$$

where $1/4 < \alpha < 3/4$. Show that letting $\alpha = 1/2$ the shape function N_2 is mapped into $(x - x_k)/\ell_k$ and all mapped shape functions are polynomials. Show also that letting $\alpha = 3/4$ the shape function N_2 is mapped into $\sqrt{(x-x_k)/\ell_k}$. In one dimension this mapping is not admissible because the mapped shape function does not lie in the energy space. Similar mappings are admissible however in two and three dimensions where elements mapped by analogous functions are called quarter-point elements.

This exercise illustrates that when the mapping is linear ($\alpha = 1/2$) then the mapped shape functions are polynomials.

2.5.4 Computation of the coefficient matrices

In the finite element method the coefficient matrices are computed element by element. These computations produce element level matrices that are "assembled" in a separate step. The procedure is outlined in the following.

Computation of the stiffness matrix

The first term of the bilinear form $B(u_n, v_n)$ is computed as a sum of integrals over the elements

$$\int_0^\ell \kappa(x) u_n' v_n' \, dx = \sum_{k=1}^{M(\Delta)} \int_{x_k}^{x_{k+1}} \kappa(x) u_n' v_n' \, dx.$$

We will be concerned with the evaluation of the integral

$$\int_{x_k}^{x_{k+1}} \kappa(x) u_n' v_n' \, dx = \int_{x_k}^{x_{k+1}} \kappa(x) \left(\sum_{j=1}^{p_k+1} a_j \frac{dN_j}{dx} \right) \left(\sum_{i=1}^{p_k+1} b_i \frac{dN_i}{dx} \right) dx.$$

The shape functions N_i are defined on the standard domain $I_{\rm st}$. Therefore the indicated differential operations cannot be performed directly. Using

$$\frac{d}{dx} = \frac{d}{d\xi}\frac{d\xi}{dx} = \frac{2}{x_{k+1} - x_k}\frac{d}{d\xi} \equiv \frac{2}{\ell_k}\frac{d}{d\xi} \quad \text{and} \quad dx = \frac{x_{k+1} - x_k}{2} d\xi \equiv \frac{\ell_k}{2} \, d\xi$$

where $\ell_k := x_{k+1} - x_k$ is the length of the kth element, we have:

$$\int_{x_k}^{x_{k+1}} \kappa(x) u_n' v_n' \, dx = \frac{2}{\ell_k} \int_{-1}^{+1} \kappa(Q_k(\xi)) \left(\sum_{j=1}^{p_k+1} a_j \frac{dN_j}{d\xi} \right) \left(\sum_{i=1}^{p_k+1} b_i \frac{dN_i}{d\xi} \right) d\xi.$$

Defining:

$$k_{ij}^{(k)} := \frac{2}{\ell_k} \int_{-1}^{+1} \kappa(Q_k(\xi)) \frac{dN_i}{d\xi} \frac{dN_j}{d\xi} \, d\xi \tag{2.67}$$

the following expression is obtained:

$$\int_{x_k}^{x_{k+1}} \kappa(x) u_n' v_n' \, dx = \sum_{i=1}^{p_k+1} \sum_{j=1}^{p_k+1} k_{ij}^{(k)} a_j b_i \equiv \{b\}^T [K^{(k)}] \{a\}. \tag{2.68}$$

The terms of the coefficient matrix $k_{ij}^{(k)}$ are computable from the mapping, the definition of the shape functions and the function $\kappa(x)$. The matrix $[K^{(k)}]$ is called the element stiffness matrix. Observe that $k_{ij}^{(k)} = k_{ji}^{(k)}$, hence $[K^{(k)}]$ is symmetric. This follows directly from the symmetry of $B(u, v)$ and the fact that $u_n \in S$, $v_n \in S$ and the same basis functions are used for u_n and v_n.

In the finite element method the integrals are evaluated by numerical integration. Numerical integration is discussed in Appendix B. In the important special case where $\kappa(x) = \kappa_k$ is constant on I_k it is possible to compute $[K^{(k)}]$ once and for all. This is illustrated by the following example.

Example 2.5.1 When $\kappa(x) = \kappa_k$ is constant on I_k and the hierarchic shape functions defined in Section 2.5.2 are used then, except for the first two rows

and columns, the elemental stiffness matrix is perfectly diagonal:

$$[K^{(k)}] = \frac{2\kappa_k}{\ell_k} \begin{bmatrix} 1/2 & -1/2 & 0 & 0 & \cdots & 0 \\ & 1/2 & 0 & 0 & & 0 \\ & & 1 & 0 & & 0 \\ & & & 1 & & 0 \\ & (\text{sym.}) & & & \ddots & \vdots \\ & & & & & 1 \end{bmatrix}. \tag{2.69}$$

Exercise 2.5.5 Assume that $\kappa(x) = \kappa_k$ is constant on I_k. Using the Lagrange polynomials defined in Section 2.5.2 for $p = 2$, compute $k_{11}^{(k)}$ and $k_{13}^{(k)}$ in terms of κ_k and ℓ_k.

Computation of the Gram matrix

The second term of the bilinear form is also computed as a sum of integrals over the elements:

$$\int_0^\ell c(x)u_n v_n \, dx = \sum_{k=1}^{M(\Delta)} \int_{x_k}^{x_{k+1}} c(x)u_n v_n \, dx.$$

We will be concerned with evaluation of the integral

$$\begin{aligned} \int_{x_k}^{x_{k+1}} c(x)u_n v_n \, dx &= \int_{x_k}^{x_{k+1}} c(x) \left(\sum_{j=1}^{p_k+1} a_j N_j \right) \left(\sum_{i=1}^{p_k+1} b_i N_i \right) dx \\ &= \frac{\ell_k}{2} \int_{-1}^{+1} c(Q_k(\xi)) \left(\sum_{j=1}^{p_k+1} a_j N_j \right) \left(\sum_{i=1}^{p_k+1} b_i N_i \right) d\xi. \end{aligned}$$

Defining:

$$m_{ij}^{(k)} := \frac{\ell_k}{2} \int_{-1}^{+1} c(Q_k(\xi)) N_i N_j \, d\xi \tag{2.70}$$

the following expression is obtained:

$$\int_{x_k}^{x_{k+1}} c(x)u_n v_n \, dx = \sum_{i=1}^{p_k+1} \sum_{j=1}^{p_k+1} m_{ij}^{(k)} a_j b_i = \{b\}^T [M^{(k)}] \{a\} \tag{2.71}$$

where $\{a\} := \{a_1 \ a_2 \ \ldots \ a_{p_k+1}\}^T$, $\{b\}^T := \{b_1 \ b_2 \ \ldots \ b_{p_k+1}\}$ and

$$[M^{(k)}] := \begin{bmatrix} m_{11}^{(k)} & m_{12}^{(k)} & \cdots & m_{1,p_k+1} \\ m_{21}^{(k)} & m_{22}^{(k)} & \cdots & m_{2,p_k+1} \\ \vdots & & \ddots & \vdots \\ m_{p_k+1,1}^{(k)} & m_{p_k+1,2}^{(k)} & \cdots & m_{p_k+1,p_k+1} \end{bmatrix}.$$

The terms of the coefficient matrix $m_{ij}^{(k)}$ are computable from the mapping, the definition of the shape functions and the function $c(x)$. The matrix $[M^{(k)}]$ is called the elemental Gram matrix[14] or the elemental mass matrix. Observe that $[M^{(k)}]$ is symmetric. In the important special case where $c(x) = c_k$ is constant on I_k it is possible to compute $[M^{(k)}]$ once and for all. This is illustrated by the following example.

Example 2.5.2 When $c(x) = c_k$ is constant on I_k and the hierarchic shape functions defined in Section 2.5.2 are used then the elemental Gram matrix is strongly diagonal. For example, for $p_k = 5$ the elemental Gram matrix is:

$$\left[M^{(k)}\right] = \frac{c_k \ell_k}{2} \begin{bmatrix} 2/3 & 1/3 & -1/\sqrt{6} & 1/3\sqrt{10} & 0 & 0 \\ & 2/3 & -1/\sqrt{6} & -1/3\sqrt{10} & 0 & 0 \\ & & 2/5 & 0 & -1/5\sqrt{21} & 0 \\ & & & 2/21 & 0 & -1/7\sqrt{45} \\ & & \text{(sym.)} & & 2/45 & 0 \\ & & & & & 2/77 \end{bmatrix} \tag{2.72}$$

Remark 2.5.2 For $p_k \geq 2$ a simple closed form expression can be obtained for the diagonal terms and the off-diagonal terms. Using eq. (2.64) it can be shown that:

$$\begin{aligned} m_{ii}^{(k)} =& \frac{c_k \ell_k}{2} \frac{1}{2(2i-3)} \int_{-1}^{+1} (P_{i-1}(\xi) - P_{i-3}(\xi))^2 \, d\xi \\ =& \frac{c_k \ell_k}{2} \frac{2}{(2i-1)(2i-5)}, \qquad i \geq 3 \end{aligned} \tag{2.73}$$

and all off-diagonal terms are zero for $i \geq 3$, with the exceptions:

$$m_{i,i+2}^{(k)} = m_{i+2,i}^{(k)} = -\frac{c_k \ell_k}{2} \frac{1}{(2i-1)\sqrt{(2i-3)(2i+1)}}, \quad i \geq 3. \tag{2.74}$$

Exercise 2.5.6 Assume that $c(x) = c_k$ is constant on I_k. Using the Lagrange shape functions defined in Section 2.5.2 for $p = 2$, compute $m_{11}^{(k)}$ and $m_{13}^{(k)}$ in terms of c_k and ℓ_k.

2.5.5 Computation of the right hand side vector

Computation of the right hand side vector involves numerical evaluation of the functional $F(v)$, given that $v \in S^0$. In particular, we write:

$$F(v_n) = \int_0^\ell f(x) v_n \, dx = \sum_{k=1}^{M(\Delta)} \int_{x_k}^{x_{k+1}} f(x) v_n \, dx.$$

[14] Jörgen Pedersen Gram 1850-1916.

The element-level integral is computed from the definition of v_n on I_k:

$$\int_{x_k}^{x_{k+1}} f(x)v_n\,dx = \frac{\ell_k}{2}\int_{-1}^{+1} f(Q_k(\xi))\left(\sum_{i=1}^{p_{k+1}} b_i^{(k)}N_i\right)d\xi = \sum_{i=1}^{p_{k+1}} b_i^{(k)}r_i^{(k)} \quad (2.75)$$

where

$$r_i^{(k)} := \frac{\ell_k}{2}\int_{-1}^{+1} f(Q_k(\xi))N_i(\xi)\,d\xi \quad (2.76)$$

which can be computed from the mapping, the given function $f(x)$ and the definition of the shape functions.

Example 2.5.3 Let us assume that $f(x)$ is a linear function on I_k. In this case $f(x)$ can be written as

$$f(x) = \frac{1-\xi}{2}f(x_k) + \frac{1+\xi}{2}f(x_{k+1}) = f(x_k)N_1(\xi) + f(x_{k+1})N_2(\xi)$$

and, assuming that the hierarchic shape functions defined in Section 2.5.2 are used,

$$\begin{aligned}
r_1^{(k)} =& f(x_k)\frac{\ell_k}{2}\int_{-1}^{+1} N_1^2\,d\xi + f(x_{k+1})\frac{\ell_k}{2}\int_{-1}^{+1} N_1N_2\,d\xi = \frac{\ell_k}{6}\left(2f(x_k) + f(x_{k+1})\right)\\
r_2^{(k)} =& f(x_k)\frac{\ell_k}{2}\int_{-1}^{+1} N_1N_2\,d\xi + f(x_{k+1})\frac{\ell_k}{2}\int_{-1}^{+1} N_2^2\,d\xi = \frac{\ell_k}{6}\left(f(x_k) + 2f(x_{k+1})\right)\\
r_3^{(k)} =& f(x_k)\frac{\ell_k}{2}\int_{-1}^{+1} N_1N_3\,d\xi + f(x_{k+1})\frac{\ell_k}{2}\int_{-1}^{+1} N_2N_3\,d\xi\\
=& -\frac{\ell_k}{6}\sqrt{\frac{3}{2}}\left(f(x_k) + f(x_{k+1})\right).
\end{aligned}$$

Exercise 2.5.7 Assume that $f(x)$ is a linear function on I_k and the hierarchic shape functions defined in Section 2.5.2 are used. Compute $r_4^{(k)}$ and show that $r_i^{(k)} = 0$ for $i > 4$. Hint: Make use of eq. (2.64).

Exercise 2.5.8 Let

$$f(x) = f_k \sin\frac{x - x_k}{\ell_k}\pi \quad x \in I_k$$

where f_k is a constant. Compute $r_5^{(k)}$ numerically in terms of f_k and ℓ_k using 3, 4 and 5 Gauss points. See Appendix B. Use the hierarchic basis functions defined in Section 2.5.2.

Exercise 2.5.9 Assume that $f(x)$ is a linear function on I and the Lagrange shape functions defined in Section 2.5.2 for $p = 2$ are used, compute $r_1^{(k)}$.

Loading by a concentrated force

A concentrated force F_0 acting on an elastic bar at $x = x_0$ is understood as a surface loading $T(x)$ defined by

$$T(x) = \begin{cases} 0 & \text{if } 0 < x < x_0 - \Delta x/2 \\ F_0/\Delta x & \text{if } x_0 - \Delta x/2 \le x \le x_0 + \Delta x/2 \\ 0 & \text{if } x_0 + \Delta x/2 < x < \ell \end{cases}$$

where $\Delta x \to 0$. In the following we make use of the result, obtained in Section 2.6.4, that in one dimension if $v \in E(I)$ then v is a continuous function. Writing

$$\int_0^\ell T(x)v\,dx = \int_{x_0-\Delta x/2}^{x_0+\Delta x/2} \frac{F_0}{\Delta x} v\,dx$$

and since $v \in E(I)$ is continuous (see proof in Section 2.6.4), we have:

$$\lim_{\Delta x\to 0} \int_{x_0-\Delta x/2}^{x_0+\Delta x/2} \frac{F_0}{\Delta x} v\,dx = F_0 v(x_0). \tag{2.77}$$

The computation of the right hand side terms corresponding to a concentrated force F_0 involves identification of the element I_k in which x_0 lies. It is then necessary find $\xi_0 = Q_k^{-1}(x_0)$ and compute

$$r_i^{(k)} = F_0 N_i(\xi_0) \qquad i = 1, 2, \ldots, p_k + 1. \tag{2.78}$$

If x_0 is a node point then either element sharing that node point may be chosen. The reason for this is discussed in Section 2.5.6.

Thermal loading

When an elastic bar is subjected to a temperature change $\mathcal{T}_\Delta := \mathcal{T} - \mathcal{T}_0$ then $F(v)$ includes the term

$$\int_0^\ell AE\alpha\mathcal{T}_\Delta \frac{dv}{dx}\,dx = \sum_{k=1}^{M(\Delta)} \int_{x_k}^{x_{k+1}} AE\alpha\mathcal{T}_\Delta \frac{dv}{dx}\,dx.$$

On the kth element:

$$\int_{x_k}^{x_{k+1}} AE\alpha\mathcal{T}_\Delta \frac{dv}{dx}\,dx = \int_{-1}^{+1} AE\alpha\mathcal{T}_\Delta \frac{dv}{d\xi}\,d\xi = \sum_{i=1}^{p_k+1} b_i \bar{r}_i^{(k)}$$

where

$$\bar{r}_i^{(k)} := \int_{-1}^{+1} \underbrace{AE\alpha\mathcal{T}_\Delta}_{x\to Q_k(\xi)} \frac{dN_i}{d\xi}\,d\xi, \qquad i = 1, 2, \ldots, p_k + 1$$

is the element-level right hand side vector corresponding to the temperature change.

Remark 2.5.3 When $AE\alpha\mathcal{T}_\Delta = \tau_k$ is constant on I_k and the hierarchic basis functions defined in Section 2.5.2 are used then $\bar{r}_1^{(k)} = -\tau_k$, $\bar{r}_2^{(k)} = \tau_k$ and $\bar{r}_i^{(k)} = 0$ for $i \geq 3$.

Example 2.5.4 Assuming that $AE\alpha\mathcal{T}_\Delta = \tau$ is a linear function on I_k and the hierarchic basis functions defined in Section 2.5.2 are used then

$$\bar{r}_1^{(k)} = \int_{-1}^{+1} \left(\frac{1-\xi}{2}\tau(x_k) + \frac{1+\xi}{2}\tau(x_{k+1}) \right) \left(-\frac{1}{2} \right) d\xi = -\frac{1}{2}(\tau(x_k) + \tau(x_{k+1})).$$

Exercise 2.5.10 Assuming that $AE\alpha\mathcal{T}_\Delta = \tau$ is a linear function on I_k and the hierarchic basis functions defined in Section 2.5.2 are used, determine $\bar{r}_3^{(k)}$.

2.5.6 Assembly

The bilinear form was computed element by element, resulting in the expressions shown in eq. (2.68), (2.71). These element-level expressions must be summed to obtain the coefficient matrix for the entire problem:

$$B(u_n, v_n) = \sum_{k=1}^{M(\Delta)} \{b^k\}^T \left([K^{(k)}] + [M^{(k)}] \right) \{a^k\}.$$

Similarly, the linear functional was computed element by element, resulting in the expressions shown in eq. (2.75), which may be augmented by the terms corresponding to concentrated forces applied at nodes, thermal loads, and essential boundary conditions, the treatment of which is discussed in Section 2.5.7.

$$F(v_n) = \sum_{k=1}^{M(\Delta)} \{b^k\}^T \{r^{(k)}\}.$$

The indicated summations are performed in the assembly process.

Prior to the summation a unique identifying number must be assigned to each basis function and the corresponding coefficients. At the element level the shape functions and their coefficients are numbered from 1 to p_{k+1}. At the domain level the basis functions and their coefficients are numbered from 1 to n, where n is the dimension of the finite element space $S(I, \Delta, \mathbf{p}, \mathbf{Q})$. The numbering of the basis functions at the domain level is arbitrary, however the numbering influences the structure of the assembled coefficient matrix.

Since the basis functions must be continuous on the domain I, the renumbering must be consistent with the requirement of continuity. Typically the basis functions associated with nodes (i.e., the basis functions which are non-zero at the nodes) are numbered first. The node number is assigned to the basis function. For example, as shown in Fig. 2.9, node k is shared by element I_{k-1} and element I_k. The variable $\varphi_k(x)$ is comprised of two linear segments. One of the segments is mapped from the shape function N_2, the other from the

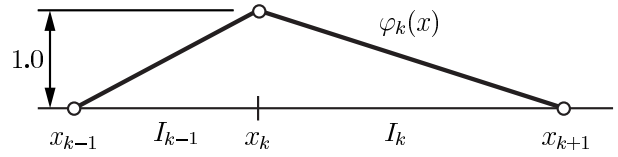

Figure 2.9: Typical nodal basis function.

shape function N_1:

$$\varphi_k(x) = \begin{cases} N_2(Q_{k-1}^{-1}(x)) & \text{for } x \in I_{k-1} \\ N_1(Q_k^{-1}(x)) & \text{for } x \in I_k. \end{cases}$$

Therefore the local (i.e., element-level) variables a_2, b_2 associated with element I_{k-1} and the local variables a_1, b_1 associated with element I_k must be assigned the same 'global' number k. The basis functions mapped from the internal shape functions are zero in the node points, hence automatically satisfy the continuity condition on I.

The numbering of the basis functions is arbitrary, however, the numbering scheme influences the structure of the coefficient matrix and hence, depending on the choice of the solver, may affect the solution time. This point is discussed further in Section 2.5.8.

Example 2.5.5 Consider the 3-element mesh shown in Fig. 2.5. The p-distribution is $\{2,\ 1,\ 3\}$. The basis functions, mapped from the hierarchic shape functions, are illustrated in Fig. 2.5. Note that the basis functions are numbered first by nodes then by p-level. Assigning the global numbering to the coefficients a_i, b_i, $i = 1, 2, \ldots, 7$, but retaining the local numbering for the stiffness coefficients, with the element numbers indicated by superscripts, we have:

$$\begin{aligned} B(u_n, v_n) =& \{b_1\ b_2\ b_5\} \begin{bmatrix} c_{11}^{(1)} & c_{12}^{(1)} & c_{13}^{(1)} \\ c_{21}^{(1)} & c_{22}^{(1)} & c_{23}^{(1)} \\ c_{31}^{(1)} & c_{32}^{(1)} & c_{33}^{(1)} \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \\ a_5 \end{Bmatrix} + \\ & \{b_2\ b_3\} \begin{bmatrix} c_{11}^{(2)} & c_{12}^{(2)} \\ c_{21}^{(2)} & c_{22}^{(2)} \end{bmatrix} \begin{Bmatrix} a_2 \\ a_3 \end{Bmatrix} + \\ & \{b_3\ b_4\ b_6\ b_7\} \begin{bmatrix} c_{11}^{(3)} & c_{12}^{(3)} & c_{13}^{(3)} & c_{14}^{(3)} \\ c_{21}^{(3)} & c_{22}^{(3)} & c_{23}^{(3)} & c_{24}^{(3)} \\ c_{31}^{(3)} & c_{32}^{(3)} & c_{33}^{(3)} & c_{34}^{(3)} \\ c_{41}^{(3)} & c_{42}^{(3)} & c_{43}^{(3)} & c_{44}^{(3)} \end{bmatrix} \begin{Bmatrix} a_3 \\ a_4 \\ a_6 \\ a_7 \end{Bmatrix} \end{aligned}$$

where $c_{ij}^{(k)} := k_{ij}^{(k)} + m_{ij}^{(k)}$. Of course, the element-level matrices are symmetric. The full matrices are displayed for purposes of clarity only. The 7×7 coefficient

matrix is obtained by performing the summation:

$$B(u_n, v_n) = \sum_{j=1}^{7}\sum_{i=1}^{7} c_{ij}a_j b_i = \{b_1\ b_2\ \cdots\ b_7\}\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{17} \\ c_{21} & c_{22} & & c_{27} \\ \vdots & & & \vdots \\ c_{71} & c_{72} & \cdots & c_{77} \end{bmatrix}\begin{Bmatrix} a_1 \\ a_2 \\ \vdots \\ a_7 \end{Bmatrix}$$

$$\equiv \{b\}^T[C]\{a\}$$

where (for example) $c_{11} = c_{11}^{(1)}$, $c_{22} = c_{22}^{(1)} + c_{11}^{(2)}$, $c_{36} = c_{13}^{(3)}$, etc. The matrix $[C]$ is called the unconstrained coefficient matrix or unconstrained global stiffness matrix. This matrix needs to be modified in order to account for the restrictions associated with the essential boundary conditions. This point will be discussed in Section 2.5.7.

The assembly of the right-hand side vector from the element-level right-hand side vectors is analogous to the procedure just described:

$$F(v_n) = \{b_1\ b_2\ b_5\}\begin{Bmatrix} r_1^{(1)} \\ r_2^{(1)} \\ r_3^{(1)} \end{Bmatrix} + \{b_2\ b_3\}\begin{Bmatrix} r_1^{(2)} \\ r_2^{(2)} \end{Bmatrix} + \{b_3\ b_4\ b_6\ b_7\}\begin{Bmatrix} r_1^{(3)} \\ r_2^{(3)} \\ r_3^{(3)} \\ r_4^{(3)} \end{Bmatrix}$$

$$= \{b_1\ b_2\ \cdots\ b_7\}\begin{Bmatrix} r_1 \\ r_2 \\ \vdots \\ r_7 \end{Bmatrix}$$

where (for example) $r_2 = r_2^{(1)} + r_1^{(2)}$.

Exercise 2.5.11 Assume that $\kappa(x) = \kappa_k$ is constant on each of the three elements shown in Fig. 2.5 and $c(x) = 0$ and hierarchic shape functions are used. Write down the assembled coefficient matrix in terms of κ_k, ℓ_k. Let $\mathbf{p} = \{2\ 1\ 3\}$.

Exercise 2.5.12 Let

$$f(x) = \begin{cases} C_1 + C_2 x & \text{for } x \in I_1 \\ C_3 & \text{for } x \in I_2 \\ 0 & \text{for } x \in I_3 \end{cases}$$

where C_1, C_2, C_3 are constants. Using the basis functions defined in Example 2.5.5, determine elements r_2 and r_3 of the assembled right hand side vector.

Exercise 2.5.13 If a concentrated force is acting on a node point then either element sharing that node point may be chosen for computing the element-level right hand side vector (load vector). Explain why.

Exercise 2.5.14 The elastic bar shown in Fig. 2.10 consists of a tapered and a prismatic section. The cross-sectional area of the tapered section varies linearly from A_a to A_b. The modulus of elasticity E and the spring coefficient c are constants. The boundary conditions are: $F_a = F_c = 0$. Let $T_b(x) = T_b$ constant. Using two elements and $p = 1$ on both elements, write down each term of the

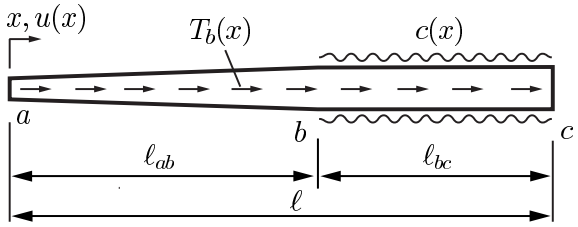

Figure 2.10: Elastic bar. Notation.

assembled coefficient matrix and load vector for this problem.

2.5.7 Treatment of the essential boundary conditions

When essential boundary conditions are specified then it is necessary to modify the assembled coefficient matrix $[C]$ and the right hand side vector $\{r\}$ using the procedure described in Section 2.3.2. For example, when using the hierarchic shape functions, we select $u^\star$ as follows:

$$u^\star = \begin{cases} \hat{u}_0 N_1(Q_1^{-1}(x)) & \text{for } x \in I_1 \\ \hat{u}_\ell N_2(Q_M^{-1}(x)) & \text{for } x \in I_M \\ 0 & \text{elsewhere} \end{cases}$$

where $M = M(\Delta)$ is the number of elements. Such a choice is illustrated in Fig. 2.11.

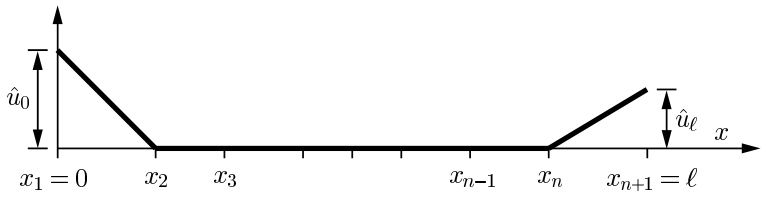

Figure 2.11: Typical choice of the function $u^\star$ in one dimension.

It is necessary to compute the integral

$$\int_0^\ell (\kappa(x)(u^\star)'v' + c(x)u^\star v)\ dx =$$
$$\hat{u}_0 \int_{x_1}^{x_2} \left(\kappa(x)N_1'(Q_1^{-1}(x))v' + c(x)N_1(Q_1^{-1}(x))v\right)\ dx +$$
$$\hat{u}_\ell \int_{x_n}^{x_{n+1}} \left(\kappa(x)N_2'(Q_M^{-1}(x))v' + c(x)N_2(Q_M^{-1}(x))v\right)\ dx.$$

Noting that

$$\int_{x_1}^{x_2} \kappa(x)(u^\star)'v'\,dx = \hat{u}_0 \frac{2}{\ell_1} \int_{-1}^{+1} \kappa(Q_1(\xi)) \frac{dN_1}{d\xi} \sum_{i=1}^{p_1+1} b_i^{(1)} \frac{dN_i}{d\xi}\,d\xi$$
$$= \hat{u}_0 \sum_{i=1}^{p_1+1} k_{i1}^{(1)} b_i^{(1)}$$

and, upon treating the other terms similarly, we have:

$$\int_0^\ell (\kappa(x)(u^\star)'v' + c(x)u^\star v)\ dx = \hat{u}_0 \sum_{i=1}^{p_1+1} \left(k_{i1}^{(1)} + m_{i1}^{(1)}\right) b_i^{(1)} +$$
$$\hat{u}_\ell \sum_{i=1}^{p_M+1} \left(k_{i2}^{(M)} + m_{i2}^{(M)}\right) b_i^{(M)}.$$

These expressions represent the first columns of the stiffness and Gram matrices of element 1 multiplied by $\hat{u}_0$ and the second columns of the stiffness and Gram matrices of element M multiplied by $\hat{u}_\ell$. Therefore enforcement of non-zero essential boundary conditions involves multiplication of the first columns of the stiffness and Gram matrices of element 1 (resp. element $M(\Delta)$) by $\hat{u}_0$ (resp. $\hat{u}_\ell$) and subtracting these columns from the right hand side vector. Furthermore, since $b_1^{(1)} = b_2^{(M)} = 0$, the corresponding rows in the assembled coefficient matrix and right hand side vector are deleted. Therefore the number of equations is reduced by the number of essential boundary conditions specified. The number of equations following enforcement of the essential boundary conditions is the number of degrees of freedom N.

Example 2.5.6 Let us consider the following problem on $I = (0, 2)$:

$$-u'' + 12u = 0 \qquad u(0) = 1.25,\ u(2) = 5.50 \tag{2.79}$$

and let us use a uniform mesh with 4 elements, $\mathbf{p} = \{1\ 1\ 1\ 1\}$. Using eq. (2.69) and (2.72) the unconstrained coefficient matrix $[C]$ is readily assembled:

$$[C] = \begin{bmatrix} 4 & -1 & 0 & 0 & 0 \\ -1 & 8 & -1 & 0 & 0 \\ 0 & -1 & 8 & -1 & 0 \\ 0 & 0 & -1 & 8 & -1 \\ 0 & 0 & 0 & -1 & 4 \end{bmatrix}.$$

Multiplying the first column by $a_1 = 1.25$, the fifth column by $a_5 = 5.50$ and subtracting these columns from the right hand side vector (which is the zero vector in this case), we have:

$$\begin{bmatrix} 8 & -1 & 0 \\ -1 & 8 & -1 \\ 0 & -1 & 8 \end{bmatrix} \begin{Bmatrix} a_2 \\ a_3 \\ a_4 \end{Bmatrix} = \begin{Bmatrix} 1.25 \\ 0 \\ 5.50 \end{Bmatrix}.$$

The number of degrees of freedom is 3. On solving we find $a_2 = 0.169859$, $a_3 = 0.108871$, $a_4 = 0.701109$.

Exercise 2.5.15 Using one element, $p = 4$ and hierarchic basis functions, write down the system of equations following enforcement of the essential boundary conditions for the problem in Example 2.5.6.

Example 2.5.7 Let us consider the problem on $(0, \ell)$:

$$-\kappa_0 u'' = f(x), \qquad \kappa_0 u'(0) = F_0,\ \kappa_0 u'(\ell) = F_\ell$$

where κ_0 is a constant. Using one element of degree p and the linear mapping given by eq. (2.56), the system of equations is

$$\frac{\kappa_0}{\ell} \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ -1 & 1 & 0 & \dots & 0 \\ 0 & 0 & 2 & \dots & 0 \\ \vdots & & & \ddots & \\ 0 & \dots & & & 2 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_{p+1} \end{Bmatrix} = \begin{Bmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_{p+1} \end{Bmatrix}$$

where the coefficient matrix is from eq. (2.69) and the right-hand side vector is from equations(2.76) and (2.78):

$$\begin{aligned} r_1 =& \frac{\ell}{2} \int_{-1}^{+1} f(Q(\xi)) N_1(\xi)\, d\xi - F_0 \\ r_2 =& \frac{\ell}{2} \int_{-1}^{+1} f(Q(\xi)) N_2(\xi)\, d\xi + F_\ell \\ r_i =& \frac{\ell}{2} \int_{-1}^{+1} f(Q(\xi)) N_i(\xi)\, d\xi, \qquad i = 3, 4, \dots, p+1. \end{aligned}$$

Observe that the second row is -1 times the first row, hence the coefficient matrix is singular. Therefore this problem does not have a unique solution and solutions exist only if the equations are consistent. On adding the first and second rows we have:

$$r_1 + r_2 = \frac{\ell}{2} \int_{-1}^{+1} f(Q(\xi)) \underbrace{(N_1(\xi) + N_2(\xi))}_{=1}\, d\xi - F_0 + F_\ell = 0$$

which is equivalent to

$$\int_0^\ell f(x)\,dx - F_0 + F_\ell = 0. \tag{2.80}$$

This is a re-statement of eq. (2.46). In solving such problems we assign an arbitrary value to a_1 or a_2. The treatment of the problem is then similar to that discussed in Example 2.5.6.

2.5.8 Solution

Following assembly of the coefficient matrix and enforcement of the essential boundary conditions (when applicable) the system of simultaneous equations is solved by one of several methods designed to exploit the symmetry and sparsity of the coefficient matrix. The solvers are classified into two main categories; direct and iterative solvers. The choice of solver in a particular application is influenced by the size of the problem and the available computational resources. Generally speaking, for small problems (a few thousand degrees of freedom) either direct or iterative solvers may be used. In most finite element analysis problems the memory requirements for direct solvers rapidly increase with the size of the problem therefore for large problems (more than 100,000 degrees of freedom) iterative solvers are used. For problems of intermediate size the choice of the solver depends on the rate of convergence of the iteration process and the number of right-hand side vectors. Discussion of the various solvers used in FEA is beyond the scope of this introductory exposition. We refer to standard texts, such as [19] and [24].

When a direct solver is used then the numbering of the basis functions should be optimized so that the solver can exploit the sparsity of the coefficient matrix. Commercial finite element codes offer a variety of solvers and optimization schemes for numbering the basis functions or the elements, depending on the choice of the solver.

The solver produces the N unknown coefficients of the basis functions. The finite element solution is then

$$u_{FE} = \bar{u} + u^\star = \sum_{i=1}^{N+N^\star} a_i \varphi_i(x)$$

where $N^\star$ is the number of coefficients determined by the essential boundary conditions. If no essential boundary conditions were specified then $N^\star = 0$. For instance, in Example 2.5.6 $N = 3$, $N^\star = 2$. The exact and finite element solutions are shown in Fig. 2.12.

In the assembly process the element-level numbering scheme, ranging from 1 to $p_k + 1$, was replaced by the global numbering scheme. This is now reversed and the element-level coefficients $a_i^{(k)}$ are stored with the topological data and other information, such as material properties, for each element. At the end of

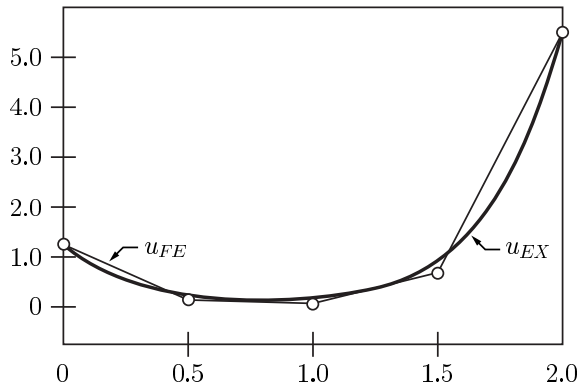

Figure 2.12: The exact and finite element solutions for Example 2.5.6.

the solution process the finite element solution is available in the form

$$u_{FE} = \sum_{i=1}^{p_k+1} a_i^{(k)} N_i(\xi) \qquad k = 1, 2, \ldots, M(\Delta). \tag{2.81}$$

The array of coefficients in eq. (2.81), the corresponding shape functions and the mapping functions contain all information generated by a finite element solution.

2.5.9 Post-solution operations

The data of interest are typically functionals that depend on u and u'. These are computed from the finite element solution in the post-solution operations.

Computation of $u_{FE}(x_0)$

The computation of u_{FE} at some arbitrary point x_0 involves a search to identify the element I_k in which point x_0 lies and, using the inverse map given by eq. (2.57), identification of the point $\xi_0 \in I_{\text{st}}$ corresponding to x_0. Then $u_{FE}(x_0)$ is computed from

$$u_{FE}(x_0) = \sum_{i=1}^{p_k+1} a_i^{(k)} N_i(\xi_0).$$

Direct computation of $u'_{FE}(x_0)$

In many applications the first derivative of the solution is of interest. The first derivatives are computed from:

$$\left(\frac{du_{FE}}{dx}\right)_{x=x_0} = \frac{2}{\ell_k} \sum_{i=1}^{p_k+1} a_i^{(k)} \left(\frac{dN_i}{d\xi}\right)_{\xi=\xi_0}. \tag{2.82}$$

Computation of the higher derivatives is analogous.

Indirect computation of $u'_{FE}(x_0)$ in node points

The first derivative in node points is usually determined indirectly from the generalized formulation given by eq. (2.33). For example, to compute the first derivative at node x_k from the finite element solution computed for element k we select $v = N_1(\xi)$ and use eq. (2.33):

$$\int_{x_k}^{x_{k+1}} (\kappa u'_{FE} v' + c u_{FE} v)\, dx = \int_{x_k}^{x_{k+1}} f v\, dx + [\kappa u'_{FE} v]_{x=x_{k+1}} - [\kappa u'_{FE} v]_{x=x_k}\,. \tag{2.83}$$

Test functions used in post-solution operations for the computation of a particular functional are called extraction functions. Here $v = N_1(Q_k^{-1}(x))$ is an extraction function for the functional $-[\kappa u'_{FE}]_{x=x_k}$. This is because $v(x_k) = 1$ and $v(x_{k+1}) = 0$ and hence

$$\begin{aligned} -[\kappa u'_{FE}]_{x=x_k} &= \int_{x_k}^{x_{k+1}} (\kappa u'_{FE} v' + c u_{FE} v)\, dx - \int_{x_k}^{x_{k+1}} f v\, dx \\ &= \sum_{j=1}^{p_k+1} c_{1j}^{(k)} a_j^{(k)} - r_1^{(k)}. \end{aligned} \tag{2.84}$$

This result shows that the first derivative at node x_k can be computed from the element-level coefficient matrix and the right-hand side vector.

Example 2.5.8 For the problem solved in Example 2.5.6 let us compute (a) the exact value of $u'(0)$; (b) $u'_{FE}(0)$ from eq. (2.82), and (c) $u'_{FE}(0)$ from eq. (2.84) using $v = N_1(\xi)$.
The exact solution of eq. (2.79) is:

$$u_{EX} = 1.25 \cosh \sqrt{12}\, x + \frac{5.50 - 1.25 \cosh \sqrt{48}}{\sinh \sqrt{48}} \sinh \sqrt{12}\, x. \tag{2.85}$$

Therefore:

$$u'_{EX}(0) = \sqrt{12}\, \frac{5.50 - 1.25 \cosh \sqrt{48}}{\sinh \sqrt{48}} = -4.292801.$$

Using the direct method of eq. (2.82) and the results from Example 2.5.6,

$$u'_{FE}(0) = 2.00(-a_1^{(1)} + a_2^{(1)}) = 2.00(-1.25 + 0.169859) = -2.160282.$$

The relative error is 49.7 %. Using the indirect method of eq. (2.84) and the results from Example 2.5.6,

$$[-u'_{FE}]_{x=0} = 4a_1 - a_2 = 4 \times 1.25 - 0.169859 = 4.830141$$

we have $u'_{FE}(0) = -4.830141$ (12.5 % error).

This example illustrates that it is much more efficient to compute the derivatives by the indirect method than by the direct method. For further details refer to Section 2.8.

Exercise 2.5.16 Compute $u'_{FE}(2)$ for Example 2.5.6 using the direct method given by eq. (2.82) and the indirect method. Compute the relative errors.

Exercise 2.5.17 Two elastic bars of length ℓ are constrained at the ends A and B and supported by a distributed spring characterized by the spring coefficient c (constant). Assume that the cross sectional area A, the modulus of elasticity E and the coefficient of thermal expansion α are constants. The bars are separated by a gap Δ at a reference temperature $\mathcal{T}_0$ (constant). The temperature of both bars is uniformly increased. The goal of computation is to determine the reaction forces developed at the supports A and B as a function of the temperature $\mathcal{T} > \mathcal{T}_0$.

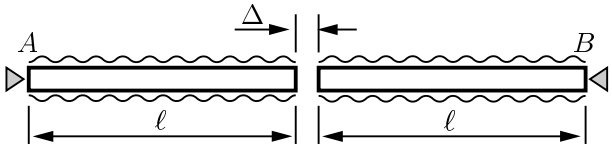

Figure 2.13: Exercise 2.5.17. Notation.

(a) Explain how you would determine the temperature $\mathcal{T}_c$ at which the gap closes. (b) Assuming that $\mathcal{T} \leq \mathcal{T}_c$, explain how you would compute the reaction force at A by the direct and indirect methods. (c) Assuming that $\mathcal{T} > \mathcal{T}_c$, explain how you would compute the reaction force at A by the direct and indirect methods.

Exercise 2.5.18 Write an ad hoc computer program for solving the problem in Example 2.5.6 for an arbitrary number of elements. Plot the relative errors for $u'_{FE}(2)$ computed by (a) the direct method and (b) the indirect method versus the number of elements. Compare the number of elements needed for reducing the relative error to under 1 %.

Nodal forces

In the finite element method the term $-[\kappa u'_{FE}]_{x=x_k}$ (resp. $[\kappa u'_{FE}]_{x=x_{k+1}}$) is interpreted as a 'nodal force' $F_1^{(k)}$ acting on node 1 (resp. $F_2^{(k)}$ acting on node 2) of element k as shown in Fig. 2.14.

Note that the convention for nodal forces is different from the convention for the bar force: Whereas the bar force is positive when tensile, a nodal force is positive when acting in the direction of the positive x axis.

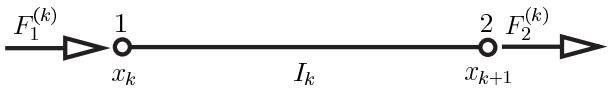

Figure 2.14: Convention for nodal forces.

Exercise 2.5.19 Show that when $c(x) = 0$ on I_k then $F_1^{(k)} + F_2^{(k)} = 0$.

Exercise 2.5.20 Show that the sum of nodal forces computed for node k from element $k-1$ and from element k is zero unless a concentrated force F_k is acting on node k, in which case the sum is equal to F_k. Illustrate this by a simple example and demonstrate that the sum of forces acting on a node point is zero. Hint: Let $v = \varphi_k(x)$ shown in Fig. 2.9.

Remark 2.5.4 We have seen in Theorem 2.4.2 that the error in energy norm depends on the choice of the finite element space S, which depends on the choice of discretization characterized by the mesh Δ, the polynomial degrees assigned to the elements $\mathbf{p}$ and the mapping functions $\mathbf{Q}$. Of course, this is true for all data computed from the finite element solution. It was noted in Section 1.1.4 that it is necessary to verify that the data of interest are substantially independent of the discretization. It is of great practical importance to ensure that the relative errors in the data of interest are within acceptable bounds. Procedures for error estimation are discussed in Section 2.7.

Remark 2.5.5 Let us consider the problem of eq. (2.21) with and $c = 0$ and $\kappa(x) = \kappa_k$ constant on I_k. Let us assume that the mapping is linear and $u(0) = \hat{u}_0$ is one of the boundary conditions. In this case the finite element solution in the node points is the exact solution. To show this we define $\varphi_k(x) \in S^0$ as follows:

$$\begin{aligned}\varphi_1(x) =& 1 - \frac{x}{\ell_1} \quad \text{for} \quad 0 \le x < x_2 = \ell_1\\ \varphi_{M(\Delta)+1}(x) =& 1 - \frac{x - x_{M(\Delta)}}{\ell_{M(\Delta)}} \quad \text{for} \quad x_{M(\Delta)} \le x \le x_{M(\Delta)+1} = \ell\end{aligned}$$

and for $k = 2, 3, \ldots, M(\Delta)$

$$\varphi_k(x) = \begin{cases}(x - x_{k-1})/\ell_{k-1} & \text{for } x_{k-1} \le x \le x_k\\ 1 - (x - x_k)/\ell_k & \text{for } x_k < x \le x_{k+1}.\end{cases}$$

Using the Galerkin orthogonality we have:

$$\begin{aligned}B(e, \varphi_1) =& -\frac{\kappa_1}{\ell_1}\int_{x_1}^{x_2} e'\,dx = 0\\ B(e, \varphi_k) =& \frac{\kappa_{k-1}}{\ell_{k-1}}\int_{x_{k-1}}^{x_k} e'\,dx - \frac{\kappa_k}{\ell_k}\int_{x_k}^{x_{k+1}} e'\,dx = 0, \quad k = 2, \ldots, M(\Delta)\\ B(e, \varphi_{M(\Delta)+1}) =& \frac{\kappa_{M(\Delta)}}{\ell_{M(\Delta)}}\int_{x_{M(\Delta)}}^{x_{M(\Delta)+1}} e'\,dx = 0\end{aligned}$$

where $e := u_{EX} - u_{FE}$. Therefore we have the system of $M(\Delta) + 1$ equations,

the coefficient matrix of which is tridiagonal:

$$\begin{aligned}
\frac{\kappa_1}{\ell_1}e(x_1) - \frac{\kappa_1}{\ell_1}e(x_2) &= 0 \\
&\vdots \\
-\frac{\kappa_{k-1}}{\ell_{k-1}}e(x_{k-1}) + \left(\frac{\kappa_{k-1}}{\ell_{k-1}} + \frac{\kappa_k}{\ell_k}\right)e(x_k) - \frac{\kappa_k}{\ell_k}e(x_{k+1}) &= 0 \\
&\vdots \\
-\frac{\kappa_{M(\Delta)}}{\ell_{M(\Delta)}}e(x_{M(\Delta)}) + \frac{\kappa_{M(\Delta)}}{\ell_{M(\Delta)}}e(x_{M(\Delta)+1}) &= 0.
\end{aligned}$$

We have assumed that an essential boundary condition is prescribed at $x = 0$ (node 1). Then $e(x_1) = 0$ and hence $e(x_k) = 0$ for $k = 1, 2, \ldots, M(\Delta) + 1$. Alternatively when an essential boundary condition is prescribed at $x = \ell$ (node $M(\Delta)+1$) then $e(x_{M(\Delta)+1}) = 0$ and hence $e(x_k) = 0$ for $k = 1, 2, \ldots, M(\Delta)+1$.

Exercise 2.5.21 Show that under the assumptions of Remark 2.5.5 $u_{FE}(x_k) = u_{EX}(x_k)$ when natural boundary conditions are prescribed.

Exercise 2.5.22 Consider the problem $-u'' = \alpha(\alpha - 1)x^{\alpha-2}$, $\alpha > 1$, on $I = (0, 1)$ with the boundary conditions $u(0) = u(1) = 0$. The goal is to estimate $u'(0)$ using the finite element method within an error tolerance of 1%. Let $\alpha = 1.05$. Using a uniform finite element mesh and $p = 1$, what is the number of elements $M(\Delta)$ needed when (a) $u'(0)$ is computed by the direct method and (b) $u'(0)$ is computed by the nodal force method? Hint: Make use of the fact that $u_{FE}(x_k) = u_{EX}(x_k)$ for $k = 1, 2, \ldots, M(\Delta) + 1$, see Remark 2.5.5.

2.6 Properties of the generalized formulation

Some of the key properties of the generalized formulation and the finite element solution are presented in the following. Although these properties are presented in the one-dimensional setting only, they are applicable to two and three dimensions as well, unless noted otherwise.

2.6.1 Uniqueness

The model problem given by eq. (2.21) has been replaced by the corresponding generalized formulation. The following theorem establishes that the solution of the generalized formulation is unique.

Theorem 2.6.1 The function $u \in \tilde{E}(I)$ that satisfies $B(u, v) = F(v)$ for all $v \in E^0(I)$ is unique in the space $E(I)$.

The theorem is proven by contradiction: Assume that there are two functions u_1, u_2 in $\tilde{E}(I)$, $u_1 \neq u_2$ that satisfy

$$B(u_1, v) = F(v)$$
$$B(u_2, v) = F(v)$$

for all $v \in E^0(I)$. Subtracting the second equation from the first we have

$$B(u_1 - u_2, v) = 0 \quad \text{for all } v \in E^0(I).$$

Since $(u_1 - u_2) \in E^0(I)$, we may select $v = u_1 - u_2$ in which case $B(u_1 - u_2, u_1 - u_2) = 0$. In view of eq. (2.49) this is equivalent to $\|u_1 - u_2\|_E = 0$ which contradicts the assumption that $u_1 \neq u_2$ in $\tilde{E}(I)$. □

Remark 2.6.1 Note that uniqueness is understood in the space $E(I)$. Suppose that $c = 0$ in eq. (2.21) and Neumann boundary conditions are specified, subject to eq. (2.46). Then if u_1 is a solution then $u_2 := u_1 + C$ is also a solution, where C is an arbitrary constant. In this case the energy norm cannot distinguish between two functions that differ by an arbitrary constant. This is seen directly from the definition of the energy norm, given by eq. (2.49):

$$\|u_1 - u_2\|_E^2 = \frac{1}{2}\int_0^\ell (u_1' - u_2')^2 \, dx = 0.$$

Therefore the solution can be determined only up to an arbitrary constant[15]. In mechanics this has a simple physical interpretation: The constant C represents rigid body displacement.

2.6.2 Potential energy

An important property of the generalized formulation is that the solution minimizes a quadratic functional, called the potential energy. This is proven by the following theorem.

Theorem 2.6.2 The function $u \in \tilde{E}(I)$ that satisfies $B(u, v) = F(v)$ for all $v \in E^0(I)$ minimizes the quadratic functional $\pi(u)$, called the potential energy;

$$\pi(u) := \frac{1}{2}B(u, u) - F(u) \tag{2.86}$$

on the space $\tilde{E}(I)$.
Proof: For any $v \in E^0(I)$, $\|v\|_E \neq 0$ we have:

$$\begin{aligned}
\pi(u + v) =& \frac{1}{2}B(u + v, u + v) - F(u + v) \\
=& \frac{1}{2}B(u, u) + B(u, v) + \frac{1}{2}B(v, v) - F(u) - F(v) \\
=& \pi(u) + \underbrace{B(u, v) - F(v)}_{0} + \frac{1}{2}\underbrace{B(v, v)}_{>0}.
\end{aligned}$$

[15] Such a norm is called a "seminorm".

Therefore any admissible nonzero perturbation of u will increase $\pi(u)$. $\square$

This theorem is known as the principle of minimum potential energy.

Remark 2.6.2 Whereas the strain energy is always positive, the potential energy may be positive, negative or zero.

2.6.3 Error in energy norm

The relationship between the error in energy norm and the error in potential energy is established by the following theorem. One of the methods used for estimating the error $e = u_{EX} - u_{FE}$ in energy norm is based on this theorem.

Theorem 2.6.3

$$\|u_{EX} - u_{FE}\|_E^2 = \pi(u_{FE}) - \pi(u_{EX}). \tag{2.87}$$

Proof: Writing $e = u_{EX} - u_{FE}$ and noting that $e \in E^0(I)$, we have:

$$\begin{aligned}
\pi(u_{FE}) =& \pi(u_{EX} - e) = \frac{1}{2}B(u_{EX} - e, u_{EX} - e) - F(u_{EX} - e) \\
=& \frac{1}{2}B(u_{EX}, u_{EX}) - F(u_{EX}) \underbrace{-B(u_{EX}, e) + F(e)}_{0} + \frac{1}{2}B(e, e) \\
=& \pi(u_{EX}) + \|e\|_E^2
\end{aligned}$$

which is the same as eq. (2.87). $\square$

2.6.4 Continuity

By definition, $u(x)$ is continuous on $\bar{I} := \{0 \le x \le \ell\}$ if for any $\epsilon > 0$ we can find a $\delta(\epsilon)$ such that

$$|u(x_2) - u(x_1)| \le \epsilon \quad \text{if} \quad |x_2 - x_1| < \delta(\epsilon), \quad x_1, x_2 \in \bar{I}. \tag{2.88}$$

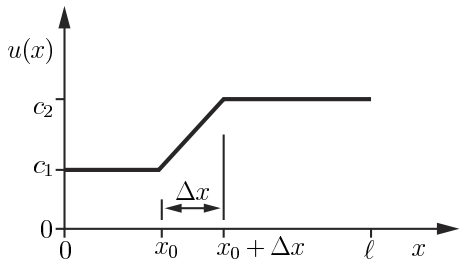

Figure 2.15: Notation.

First we show by an example that a discontinuous function cannot lie in the energy space. Specifically, let us consider the continuous function $u(x)$ shown in Fig. 2.15 and compute the integral

$$\int_0^\ell (u')^2\,dx = \int_{x_0}^{x_0+\Delta x} \left(\frac{c_2-c_1}{\Delta x}\right)^2 dx = \frac{(c_2-c_1)^2}{\Delta x}.$$

At $\Delta x = 0$ the function becomes discontinuous and the value of the integral is infinity.

With one qualification, any function $u(x) \in E(I)$ is continuous and bounded on $\bar{I}$ by the energy norm $\|u\|_E$. The qualification is that when $c(x) = 0$ and Neumann boundary conditions are specified, subject to eq. (2.46), then $u = C$ is a solution (where C is an arbitrary constant) hence $u(x)$ is not bounded by the energy norm. See Remark 2.6.1. The statement that $u(x) \in E(I)$ is bounded on $\bar{I}$ by $\|u\|_E$ is understood to mean that there exists a constant C, independent of $u(x)$, such that for any $x \in \bar{I}$ the following inequality holds: $|u| \le C\|u\|_E$.

We now prove continuity for the case $\kappa = 1$, $c = 0$. Let us assume that $x_2 \ge x_1$. Then:

$$u(x_2) - u(x_1) = \int_{x_1}^{x_2} u'(x)dx \tag{2.89}$$

where the prime ($'$) represents differentiation with respect to x. Applying the Schwarz inequality (see Appendix A, Section A.8) we have:

$$|u(x_2)-u(x_1)| \le \left(\int_{x_1}^{x_2} dx\right)^{1/2} \left(\int_{x_1}^{x_2} (u'(x))^2 dx\right)^{1/2} = |x_2-x_1|^{1/2}\,\sqrt{2}\,\|u\|_E. \tag{2.90}$$

Therefore if we select $\delta(\epsilon) < \epsilon^2/2\,\|u\|_E^2$ then the condition of continuity is satisfied. That $u(x)$ is bounded on $\bar{I}$ by $\|u\|_E$ follows directly from (2.90) for the case $u(x_1) = 0$. This theorem holds in one dimension only.

Remark 2.6.3 Whereas all functions $u \in E(I)$ are continuous and bounded, du/dx has to be neither continuous nor bounded. For this reason Neumann boundary conditions cannot be enforced by restriction.

Exercise 2.6.1 Consider functions of the form $u(x) = x^\alpha$ on the interval $0 < x < \ell$. Show that u is in the energy space only if $\alpha > 1/2$.

2.6.5 Convergence in energy norm

In the example discussed in Section 2.2 the basis functions were selected to be polynomials that satisfied the homogeneous essential boundary conditions. The tacit assumption was that in some sense $u_n \to u$ as $n \to \infty$. Convergence in a normed linear space X is understood to mean that for any $\epsilon > 0$ there is an n, dependent on ϵ, such that $\|u - u_n\|_X < \epsilon$.

In the following we assume that $u''(x)$ is continuous and $\|u''(x)\|_{\max} \le C < \infty$ on $\bar{I} := [0, \ell]$. Let us partition $\bar{I}$ into n segments and denote the kth node

point by x_k. Let us interpolate $u(x)$ by a piecewise linear function $\bar{u}_n(x)$ such that $\bar{u}_n(x_k) = u(x_k)$. An example where $n = 4$ is shown in Fig. 2.16a. On the kth sub-interval $I_k := (x_k, x_{k+1})$ we have:

$$\bar{u}_n = \frac{x - x_k}{h_k} u(x_{k+1}) - \frac{x - x_{k+1}}{h_k} u(x_k) \qquad x \in I_k \qquad k = 1, 2, \ldots, n$$

where $h_k := x_{k+1} - x_k$ is the length of the kth element. The maximum value of h_k will be denoted by h.

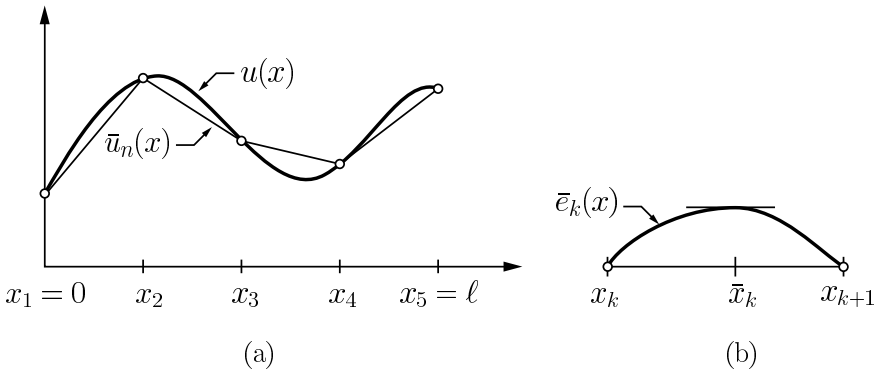

Figure 2.16: Linear interpolation. Notation.

We first show that

$$\|u - \bar{u}_n\|_{\max} \leq \frac{h^2}{8} \|u''\|_{\max}. \tag{2.91}$$

To prove this inequality, we consider the error on the kth sub-interval, shown in Fig. 2.16b. Since $\bar{e}_k(x) := u(x) - \bar{u}_n(x)$ for $x \in I_k$, $\bar{e}_k(x)$ vanishes at the endpoints of the sub-interval I_k. Furthermore, by assumption u'' is continuous on I_k, therefore there is a point $\bar{x}_k$ where $|\bar{e}_k|$ is maximum and $\bar{e}'_k(x_k) = 0$. We expand $\bar{e}_k$ about this point into a Taylor series. Let us assume that $x_{k+1} - \bar{x}_k \leq h_k/2$ and write

$$\bar{e}_k(x_{k+1}) = 0 = \bar{e}_k(\bar{x}_k) + \underbrace{\bar{e}'_k(\bar{x}_k)}_{0}(x_{k+1} - \bar{x}_k) + \frac{1}{2}\bar{e}''_k(\xi_k)(x_{k+1} - \bar{x}_k)^2 \qquad \xi \in I_k$$

where the last term is the error term of the Taylor expansion. From this relationship we have:

$$\max_{x \in I_k} |\bar{e}_k(x)| = |\bar{e}_k(\bar{x}_k)| = \frac{1}{2}|\bar{e}''_k(\xi_k)|(x_{k+1} - \bar{x}_k)^2 \qquad \xi_k \in I_k.$$

Since $\bar{u}''_n = 0$ we have $\bar{e}''_k = u''$. Also, since $x_{k+1} - \bar{x}_k \leq h_k/2$ we have:

$$\max_{x \in I_k} |\bar{e}_k| \leq \frac{h_k^2}{8} \|u''\|_{\max} \qquad k = 1, 2, \ldots, n.$$

If $x_{k+1} - \bar{x}_k > h_k/2$ then we express $\bar{e}_k(x_k)$ and obtain the same result. On replacing h_k with h, eq. (2.91) follows directly from this result.

The estimate for $\|\bar{e}'_k\|_{L_2(I_k)}$ is obtained from the Taylor expansion of $\bar{e}'_k(x)$ at $x = \bar{x}_k$:

$$\bar{e}'_k(x) = \underbrace{\bar{e}'_k(\bar{x}_k)}_{0} + \bar{e}''_k(\xi_k)(x - \bar{x}_k) = u''(\xi_k)(x - \bar{x}_k).$$

Therefore:

$$|\bar{e}'_k| \leq \|u''\|_{\max}|x - x_k|$$

and

$$\int_{x_k}^{x_{k+1}} (\bar{e}'_k)^2\, dx \leq \|u''\|^2_{\max} \int_{x_k}^{x_{k+1}} (x - x_k)^2\, dx = \ \|u''\|^2_{\max} \frac{h_k^3}{3}.$$

On the entire domain:

$$\int_0^{\ell} (\bar{e}'_k)^2\, dx \leq \|u''\|^2_{\max} \sum_{k=1}^{M(\Delta)} \frac{h_k^3}{3} \leq \frac{1}{3} h^2 \|u''\|^2_{\max} \sum_{k=1}^{M(\Delta)} h_k.$$

Since $\sum_{k=1}^{M(\Delta)} h_k = \ell$, we have

$$\int_0^{\ell} (\bar{e}'_k)^2\, dx \equiv \|u' - \bar{u}'_n\|^2_{L_2(I)} \leq h^2 \frac{\ell}{3} \|u''\|^2_{\max}. \tag{2.92}$$

We are now in a position to obtain an estimate for the error of the interpolant in the energy norm. By definition,

$$\|u - \bar{u}_n\|^2_{E(I)} = \frac{1}{2} \int_0^{\ell} \left(\kappa (u' - \bar{u}'_n)^2 + c(u - \bar{u}_n)^2 \right)\, dx.$$

Using eq. (2.91) and eq. (2.92), for sufficiently small h we have:

$$\|u - \bar{u}_n\|_{E(I)} \leq Ch \|u''\|_{\max}$$

where C is a constant that depends on κ, c, ℓ but is independent of h and u.

Theorem 2.4.2 (see p. 41) states that the finite element solution minimizes the error in energy norm on the space $\tilde{S}(I)$. Therefore $\|u - u_n\|_{E(I)} \leq \|u - \bar{u}_n\|_{E(I)}$ and

$$\|u - u_n\|_{E(I)} \leq Ch \|u''\|_{\max}. \tag{2.93}$$

Error estimates of this type are called a priori estimates[16]. This estimate shows that the finite element solution converges to the exact solution in energy norm given the assumption that u'' is continuous and bounded on $\bar{I}$. It also shows how fast the error in energy norm decreases as the mesh is refined so that the size of the largest element h approaches zero. This estimate holds for all

[16] A priori estimates are obtained through deductive reasoning, based on certain characteristics of a problem class. In this instance the problem class is characterized by $u''(x)$ being continuous and bounded on $\bar{I}$.

h, and for sufficiently small h the inequality (2.93) becomes an approximate equality. Convergence can be proven for any $u \in E(I)$ see, for example, [49]. A brief overview of priori estimates for two and three dimensional problems is presented in Section 6.4.

Remark 2.6.4 It is possible to prove convergence of the finite element solution in other norms, such as $\|u' - u'_n\|_{\max}$ and $\|u' - u'_n\|_{L_2(I)}$ but this is beyond the scope of this book.

Exercise 2.6.2 The estimate (2.93) was derived for linear shape functions. Obtain an analogous estimate for quadratic shape functions under the assumption that $|u'''(x)| \leq C < \infty$.

Exercise 2.6.3 Repeat Exercise 2.6.2 assuming that $|u''(x)| \leq C < \infty$ but $|u'''(x)|$ can be arbitrarily large.

2.7 Error estimation based on extrapolation

Computed values of the potential energy corresponding to a hierarchic sequence of finite element spaces can be used for estimating the error in energy norm by extrapolation. By Theorem 2.6.3 we have;

$$\|u_{EX} - u_{FE}\|_E^2 = \pi(u_{FE}) - \pi(u_{EX}). \tag{2.94}$$

In Section 6.4 it will be shown that for a large and important class of problems the error in energy norm is proportional to $N^{-\beta}$ when N is sufficiently large:

$$\|u_{EX} - u_{FE}\|_E \approx \frac{k}{N^\beta} \tag{2.95}$$

where k is some positive constant. An error estimate can be based on this relationship. Using eq. (2.94) we have:

$$\pi(u_{FE}) - \pi(u_{EX}) \approx \frac{k^2}{N^{2\beta}}. \tag{2.96}$$

There are three unknowns: $\pi(u_{EX})$, k and β. Assume that we have a sequence of solutions corresponding to $S_{p-2} \subset S_{p-1} \subset S_p$. Let us denote the corresponding potential energy values by π_{p-2}, π_{p-1}, π_p and the degrees of freedom by N_{p-2}, N_{p-1}, N_p. We will denote $\pi := \pi(u_{EX})$. With this notation we have:

$$\pi_p - \pi \approx \frac{k^2}{N_p^{2\beta}} \tag{2.97}$$

$$\pi_{p-1} - \pi \approx \frac{k^2}{N_{p-1}^{2\beta}}. \tag{2.98}$$

On dividing eq. (2.97) with eq. (2.98), we have

$$\frac{\pi_p - \pi}{\pi_{p-1} - \pi} \approx \left(\frac{N_{p-1}}{N_p}\right)^{2\beta} \quad \text{or} \quad \log \frac{\pi_p - \pi}{\pi_{p-1} - \pi} \approx 2\beta \log \frac{N_{p-1}}{N_p}.$$

Repeating for $p-1$ and $p-2$, it is possible to eliminate 2β to obtain:

$$\frac{\pi_p - \pi}{\pi_{p-1} - \pi} \approx \left(\frac{\pi_{p-1} - \pi}{\pi_{p-2} - \pi} \right)^Q \quad \text{where} \quad Q := \log \frac{N_{p-1}}{N_p} \left(\log \frac{N_{p-2}}{N_{p-1}} \right)^{-1}. \tag{2.99}$$

This equation can be solved for π to obtain an estimate of the exact value of the potential energy. With the estimated value of π it is possible to estimate the relative error in energy norm using eq. (2.94). By definition, the relative error in energy norm is:

$$(e_r)_E = \frac{\|u_{EX} - u_{FE}\|_E}{\|u_{EX}\|_E} \tag{2.100}$$

where $\|u_{EX}\|_E$ is estimated from u_{FE} corresponding to the finite element space of the largest number of degrees of freedom:

$$\|u_{EX}\|_E \approx \sqrt{\frac{1}{2} B(u_{FE}, u_{FE})}.$$

The relative error is usually reported as percent error. This estimator has been tested against a number of problems for which the exact solutions are known. It was found that the estimator works well for a wide range of problems, including most problems of practical interest, however it cannot be guaranteed to work for all problems.

The quality of an estimator is measured by the effectivity index θ, defined as the estimated error divided by the true error:

$$\theta := \frac{\left(\|u_{EX} - u_{FE}\|_E \right)_{\text{est.'d}}}{\left(\|u_{EX} - u_{EX}\|_E \right)_{\text{true}}}. \tag{2.101}$$

Of course, the effectivity index can be computed only for those problems for which the exact solution is known. Evaluation of an estimator involves the solution of a variety of such problems. An estimator is generally considered to be reliable if $0.8 < \theta < 1.2$ for most problems.

Remark 2.7.1 Let us divide eq. (2.95) by $\|\mathbf{u}_{EX}\|_E$ to obtain

$$\frac{\|u_{EX} - u_{FE}\|_E}{\|\mathbf{u}_{EX}\|_E} = (e_r)_E \approx \frac{\bar{k}}{N^\beta}$$

where $\bar{k} := k/\|\mathbf{u}_{EX}\|_E$. Therefore:

$$\log (e_r)_E \approx \log \bar{k} - \beta \log N.$$

Therefore if we plot $(e_r)_E$ vs. N on log-log scale then, if the assumption (2.95) is correct, then we will see a straight line with the slope $-\beta$. The convergence (2.95) is called algebraic convergence and β is called the rate of convergence.

2.8 Extraction methods

We have seen in Example 2.5.8 that the indirect method used for computing $u'_{FE}(0)$ yielded a much smaller error than the direct method. The reasons for this are discussed in the following.

Let us be interested in computing some functional Q and let us assume that we can find a function w such that the exact value of Q, denoted by Q_{EX}, is:

$$Q_{EX} = B(u_{EX}, w) - F(w). \tag{2.102}$$

The function w is called extraction function. It does not have to lie in the energy space, however the indicated operations must be defined. The finite element approximation of Q_{EX} is denoted by Q_{FE}:

$$Q_{FE} = B(u_{FE}, w) - F(w). \tag{2.103}$$

Therefore the error in Q is:

$$Q_{EX} - Q_{FE} = B(e_u, w) \tag{2.104}$$

where $e_u := u_{EX} - u_{FE}$. Let g_{EX} be the projection of w onto the test space $E^0(I)$, that is,

$$B(g_{EX}, v) = B(w, v) \qquad \text{for all } v \in E^0(I) \tag{2.105}$$

and let g_{FE} be the projection of g_{EX} onto the test space $S^0(I)$:

$$B(g_{FE}, v) = B(g_{EX}, v) \qquad \text{for all } v \in S^0(I). \tag{2.106}$$

In view of eq. (2.105) and eq. (2.53) we can write eq. (2.104) as follows:

$$Q_{EX} - Q_{FE} = B(e_u, g_{EX}) - \underbrace{B(e_u, g_{FE})}_{0} \tag{2.107}$$

where the second term on the right hand side is zero on account of the Galerkin orthogonality (2.53). Therefore denoting $e_g := g_{EX} - g_{FE}$ we have:

$$Q_{EX} - Q_{FE} = B(e_u, e_g). \tag{2.108}$$

Using the Schwarz inequality (see Section A.8) we have:

$$|Q_{EX} - Q_{FE}| = |B(e_u, e_g)| \leq 2\|e_u\|_{E(I)}\|e_g\|_{E(I)}. \tag{2.109}$$

In other words, the error in Q depends on the error of the solution e_u and the error e_g. Therefore the finite element space has to be designed such that both errors are small. Note that g_{EX} does not have to be known, it is of theoretical importance only.

If $\|e_g\|_{E(I)}$ converges to zero at approximately the same rate as $\|e_u\|_{E(I)}$ then $|Q_{EX} - Q_{FE}|$ converges to zero at about the same rate as the error in strain energy, which is twice the rate of convergence of error in energy norm. A method of computation for some functional is said to be superconvergent when the data of interest converge to their limit value at approximately the same rate as the strain energy.

Remark 2.8.1 Superconvergent methods of utilize extraction functions functions constructed so as to approximate the appropriate Green's function in the neighborhood of the point of extraction. For details we refer to [4].

2.9 Laboratory exercises

The following chapters deal with the finite element method in two and three dimensions. To perform the exercises, the reader will need to use the finite element analysis software product StressCheck[17], which is provided with this book, or some other FEA software product that has h- and p-extension capabilities.

At this point the reader should become acquainted with the key features of StressCheck. The best way to start is to read the Getting Started Guide, which can be found under the main menu heading 'Help'. Chapter 2 of this guide provides information about the most important features of the user interface. Chapter 4 is a tutorial that provides information about the preparation of input data for problems in two- and three-dimensional elasticity, execution of the solution, and post-processing procedures.

Chapter 3 of the Getting Started Guide provides an overview of StressCheck's Handbook Library. The Handbook Library contains a number of problems defined in terms of parameters in much the same way as in conventional engineering handbooks. The principal difference is that here the solutions are obtained by the finite element method. This allows problems of a far greater complexity and variety to be formulated in terms of parameters. The finite element meshes change automatically with the geometric parameters, therefore handbook users do not need to be concerned with mesh generation. The reader is encouraged to explore the handbook library, and use it for guidance when formulating and solving exercise problems in the following chapters.

Having gained some familiarity with StressCheck, the reader will find detailed information in the Master Guide[18]. The Master Guide is comprised of four parts: Part 1, the Users' Guide, provides detailed information about the user interface, post-processing and the handbook framework. Part 2, the Modeling Guide, explains procedures for the creation of geometric entities in two and three dimensions and for automatic generation of finite element meshes. Part 3, the Analysis Guide, provides instructions on the preparation of data for the various types of analyses supported by StressCheck. Part 4, the Advanced Guide, provides information about fracture mechanics applications, nonlinear analysis procedures, the solvers, and other topics that are of interest to advanced users. Specific topics can be located by means of the Index which can be accessed through the Bookmarks section of the Master Guide.

[17] StressCheck is a trademark of Engineering Software Research and Development, Inc., St. Louis, Missouri.

[18] The Master Guide can be found under the main menu heading 'Help'.

2.10 Chapter summary

Some fundamental concepts, procedures and definitions, essential for understanding the finite element method, were presented in a simple setting:

1. The generalized formulation, its dependence on the boundary conditions, treatment of natural and essential conditions, definitions of the energy space, various norms and the potential energy are fundamental to the finite element method.

2. The approximate solution and hence the error of approximation is determined by the finite element space characterized by the finite element mesh, the polynomial degrees of the elements and the mapping functions.

3. The finite element solution is unique and minimizes the error in energy norm, see Theorem 2.4.2 on page 41.

4. All information generated by the finite element method resides in the standard basis functions, called shape functions, their coefficients and the mapping functions.

5. The errors in the data of interest depend on how the data are computed from the finite element solution. In computing the first derivative the indirect method was substantially more accurate than the direct method.