Introduction into Finite Elements

Fred Vermolen and Domenico Lahaye
Delft University of Technology
Delft Institute for Applied Mathematics
Section of Numerical Analysis
Version 2.1
February 26, 2008
Preface

These lecture notes are used for the course Finite Elements (wi3098LR) for BSc-students of the department of aerospace engineering with a minor in applied mathematics, and for students visiting the Delft University of Technology in the framework of ATHENS. For the interested student some references are given for further reading, but note that the literature survey is far from complete.

The notes provide an introduction into the Finite Element method applied to Partial Differential Equations. The treated problems are classical. The treatment of the Finite Element Method is mathematical, but without the concept of Hilbert and Sobolev spaces, which are fundamental in a rigorous treatment of the Finite Element Method. Recently, a more detailed and comprehensive treatment appeared by van Kan et al. [2006], which is available from the Delft University Press.

Further, we recommend all students of this course to attend the lectures since the lecture notes do not aim at being complete.

We wish everybody good luck with the material and the course! Further, we would like to thank Caroline van der Lee for the beautiful typesetting of this document. Finally, we thank Fons Daalderop and Martin van Gijzen for their critical reading of these lecture notes and valuable suggestions to improve it.

- Alfonzo e Martino, grazie!

Fred Vermolen and Domenico Lahaye
Delft Institute of Applied Mathematics
Delft University of Technology
F.J.Vermolen@tudelft.nl

©Copyright Fred J. Vermolen, 2008
## Contents

**Preface**

1 Model Equations
   1.1 Introduction .................................................. 1
   1.2 Mathematical Preliminaries ................................... 5

2 Minimization Problems ............................................. 9
   2.1 Example .......................................................... 9
   2.2 A general 1-dimensional minimization problem .............. 13
   2.3 A 2-dimensional minimization problem ....................... 15
   2.4 Some formalization ............................................ 17

3 Variational Formulation and Differential equations ............ 20
   3.1 Weak forms ..................................................... 20
   3.2 Which weak formulation? ....................................... 24
   3.3 Mathematical considerations: existence and uniqueness ..... 25

4 Galerkin’s Finite Element method ................................ 28
   4.1 The principle of Galerkin’s method .......................... 28
   4.2 Motivation of piecewise linear basis-functions ............. 31
   4.3 Evaluation of a one-dimensional example .................... 32
   4.4 Ritz’ method of Finite Elements by a simple example ...... 34
   4.5 The treatment of a non-homogeneous Dirichlet BC .......... 35
   4.6 A time-dependent example ..................................... 37
   4.7 The principle of element matrices and vectors ............. 38
   4.8 Numerical integration ......................................... 42
   4.9 Error considerations .......................................... 43

5 Time dependent problems: numerical methods .................... 46
   5.1 Time-integration methods ..................................... 46
   5.2 Accuracy of Time-integration methods ....................... 47
   5.3 Time-integration of PDE’s ................................... 49
5.3.1 The Heat equation ................................................. 49
5.3.2 The wave equation ............................................. 51
5.4 Stability analysis ................................................. 53

6 Numerical linear algebra ............................................. 55
  6.1 Basics ................................................................. 55
  6.1.1 Eigenvalues and eigenvectors of $A$ ......................... 56
  6.1.2 Diagonalization of $A$ ........................................... 57
  6.1.3 Eigenvalues of a symmetric matrix ......................... 57
  6.2 Solution of systems of linear equations ....................... 58
  6.2.1 The $LU$-decomposition ...................................... 58
  6.2.2 Basic, classical iterative solvers ......................... 62
  6.3 Convergence properties of iterative methods ................. 63
  6.4 Gradient methods for $Ax = b$ ................................. 68

Bibliography ............................................................. 71
Model Equations

1.1 Introduction

In technology and engineering there is much to model and predict. The predictions are often based on mathematical models. The models include both physical/chemical assumptions and mathematical computation techniques. The answers from the models should be reliable, hence accuracy, consistency and stability are important issues in the analysis of the models.

Since mathematical models can become complicated, numerical (approximation) techniques are often necessary to find a prediction. The numerical techniques give unrealistic predictions when they are not used properly.

This course aims at:

- Interpretation of numerical results;
- Analysis of numerical methods;
- Analysis of mathematical models to get a qualitative view of the prediction without actually computing it.

This course will include items of:

- Linear Algebra (matrix computation),
- Numerical analysis,
- Calculus,
- Differential equations.

Before we sum the partial differential equations, which will be the basis of the course, we introduce some notation:
[1] A position vector in a two-dimensional plane and three-dimensional space is denoted as a row of respectively two and three numbers $a_1$, $a_2$ and $a_3$ separated by a comma, \(i.e.\)

\[
\bar{a} = <a_1, a_2> \quad \text{in 2-D, or} \quad \bar{a} = <a_1, a_2, a_3> \quad \text{in 3D}.
\]

The overbar is used to indicate that we mean a position vector. The coefficients $a_1$, $a_2$ and $a_3$ represent the $x$, $y$ and $z$ components of the vector.

[2] A vector in the linear algebra sense can have a length $n$, where $n$ is a positive integer, is denoted by a column, \(i.e.\)

\[
\bar{b} = \begin{pmatrix}
    b_1 \\
    b_2 \\
    b_3 \\
    \vdots \\
    b_n
\end{pmatrix}.
\]

The underbar is used to indicate that we mean a vector in the linear algebra sense. The vector may represent a result as the solution of a linear system of equations. This will be used in the later chapters of the notes.

[3] We introduce the gradient of a scalar function as the vector containing the directional derivatives with respect to the co-ordinate axes, \(i.e.\) for a function $f = f(x, y)$ and a function $g = g(x, y, z)$:

\[
\nabla f = <\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}> \quad \text{and} \quad \nabla g = <\frac{\partial g}{\partial x}, \frac{\partial g}{\partial y}, \frac{\partial g}{\partial z}>.
\]

[4] The laplacian is defined as follows:

\[
\Delta(\cdot) := \nabla \cdot \nabla(\cdot)
\]

\[
= \left\langle \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right\rangle \cdot \left\langle \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right\rangle (\cdot)
\]

\[
= \frac{\partial^2}{\partial x^2} (\cdot) + \frac{\partial^2}{\partial y^2} (\cdot) + \frac{\partial^2}{\partial z^2} (\cdot). \quad (1.1)
\]

This is the sum of the second derivatives of a function with respect to all the co-ordinate axes.

We will consider the following PDE’s (Partial Differential Equations) as a basis for many mathematical models:
− ∆u = f \quad \text{Poisson, Laplace (f = 0)} \quad (1.2) \\
− \nabla \cdot (D(u)\nabla u) = f \quad \text{non-linear Poisson} \quad (1.3) \\
\frac{\partial u}{\partial t} = \nabla \cdot (D(u)\nabla u) \quad \text{diffusion} \quad (1.4) \\
\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0 \quad \text{Buckley-Leveret / two-phase flow} \quad (1.5) \\
\frac{\partial u}{\partial t} + \nabla \cdot (\bar{q}u) = \Delta u \quad \text{convection-diffusion} \quad (1.6) \\
\frac{\partial^2 u}{\partial t^2} = c^2 \Delta u \quad \text{wave propagation} \quad (1.7)

The above PDE’s originate from various physical or biological models. In order to have a unique (well defined) solution to the PDE, we need boundary and initial conditions. We will treat the most common boundary conditions briefly. Let us assume that \( u \) satisfies a PDE within the domain of computation \( \Omega \).

[1] The first boundary condition concerns the situation in which the solution is given at the boundary, that is
\[
\quad u = g_1(x), \quad \text{for } x \in \partial \Omega_D. \quad (1.8)
\]
The solution is well defined at \( \Omega_D \) which represents (part of) the boundary of the domain \( \Omega \). This boundary condition is referred to as a \textit{Dirichlet boundary condition}. Physically, this could correspond to a prescription of the temperature at that (part of the) boundary.

[2] The second boundary condition concerns the case in which the normal derivative is prescribed, that is
\[
\quad \frac{\partial u}{\partial n} = g_2(x), \quad \text{for } x \in \partial \Omega_N. \quad (1.9)
\]
This boundary condition is referred to as a \textit{Neumann boundary condition}. Physically, this could correspond to a prescription of the flux of matter or heat on this (part of the) boundary. This condition is also referred to as a flux boundary condition.

[3] The third boundary condition concerns the case in which a linear combination of the solution and its normal derivative is specified, that is
\[
\quad \sigma u + \frac{\partial u}{\partial n} = g_3(x), \quad x \in \Omega_R. \quad (1.10)
\]
Here \( \sigma \) is a nonnegative real number. This boundary condition is referred to as a \textit{Robin boundary condition}. People also call it a mixed boundary condition. Physically, it may correspond to radiation.
The time dependent problems also need one *initial condition* if the maximum order of the time derivative is one and more initial initial conditions of the maximum order of the time derivative is larger than one. We will discuss the boundary and initial conditions for three standard PDE’s.

**Poisson’s equation**

Consider Poisson’s equation

$$-\Delta u = f(x), \quad (1.11)$$

where \( f(x) \) is a given function. For this PDE, we have *precisely* one boundary condition on each point on the boundary. In the mathematical literature, this equation is referred to as elliptic.

**The heat equation**

Consider the heat equation

$$\frac{\partial u}{\partial t} - \Delta u = f(x), \quad (1.12)$$

then we need one boundary condition on each point on the boundary like in the previous example. Furthermore, due to the time derivative, we need an initial condition to hold at each point of the domain of computation \( \Omega \):

$$u(x, t) = u_0(x), \quad \text{for } x \in \Omega. \quad (1.13)$$

Here \( u_0(x) \) is a known function. In the mathematical literature, the above equation is referred to as parabolic.

**The wave equation**

Consider the wave equation

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = f(x), \quad (1.14)$$

then we need one boundary condition on each point on the boundary like in the previous example. Furthermore, due to the second order time derivative, we need two initial conditions to hold at each point of the domain of computation \( \Omega \):

$$u(x, t) = u_0(x), \quad \text{for } x \in \Omega,$$

$$\frac{\partial u}{\partial t}(x, t) = v_0(x), \quad \text{for } x \in \Omega. \quad (1.15)$$

Here \( u_0(x) \) and \( v_0(x) \) are known functions. In the mathematical literature, the above equation is referred to as hyperbolic.

The classification of the PDE’s into hyperbolic, parabolic and elliptic only applies for second order PDE’s and has been inspired by quadratic curves. The classification is not treated in more detail.
Let $\Omega \subset \mathbb{R}^3$, $\partial \Omega = \hat{\Omega} \setminus \Omega$ (the boundary of $\Omega$), and $\vec{n}$ is the outward unit normal on $\partial \Omega$ with $\|\vec{n}\| = 1$ (Figure 1.1). Further $\hat{\Omega}$ is regular and closed. Then:

1. Divergence theorem of Gauss.

Let $\hat{\Omega} \subset \mathbb{R}^3$ be a regular closed body with outward unit normal vector $\vec{n}$ ($\|\vec{n}\| = 1$) on boundary $\partial \Omega$, and let $\vec{F}(x, y, z)$ be a differentiable vector-function on $\hat{\Omega}$, then

$$\int_{\hat{\Omega}} \nabla \cdot \vec{F}(x, y, z)dV = \int_{\partial \Omega} \vec{F}(x, y, z) \cdot \vec{n}dS, \quad (1.16)$$

or shortly

$$\int_{\hat{\Omega}} \nabla \cdot \vec{F}dV = \int_{\partial \Omega} \vec{F} \cdot \vec{n}dS. \quad (1.17)$$

2. Circulation theorem of Stokes.

Let $S \subset \mathbb{R}^3$ be a closed regular oriented surface bounded by curve $\Gamma$ with
1.2 Mathematical Preliminaries

Figure 1.3: A domain $\Omega$ in $\mathbb{R}^2$ with boundary $\Gamma$

positive orientation (figure 1.2), then

$$\iint_S (\nabla \times \bar{F}(x, y, z)) \cdot \bar{n} dS = \int_\Gamma \bar{F}(x, y, z) \cdot d\bar{r}, \quad (1.18)$$

or shortly

$$\iint_S \text{curl} \bar{F} \cdot \bar{n} dS = \int_\Gamma \bar{F} \cdot d\bar{r}. \quad (1.19)$$

here $\bar{n}$ is the unit normal and $\bar{F}$ is assumed to be differentiable over $S$.

[3] Divergence theorem for $\mathbb{R}^2$

Let $\Omega \subset \mathbb{R}^2$ be a regular closed area in $\mathbb{R}^2$ with outward unit normal vector $\bar{n}$ ($||\bar{n}|| = 1$) on boundary $\Gamma$ (figure 1.3), and let $\bar{F}(x, y)$ be a differentiable vector-function on $\Omega$, then

$$\iint_\Omega \nabla \cdot \bar{F}(x, y) dA = \oint_\Gamma \bar{F}(x, y) \cdot \bar{n} ds. \quad (1.20)$$

Proof of the above theorems can be found in: Adams [1999], Steward [2003], Almering e.a. [1987], ...

Examples:

[1]

$$\bar{F}(x, y, z) = < x, y, z >, \quad (1.21)$$

$$\nabla \cdot \bar{F} = 3, \quad (1.22)$$

$$\Omega := \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 \leq 1\}, \quad (1.23)$$

$$\partial \Omega := \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}, \quad (1.24)$$

$$\Rightarrow \iiint_\Omega \nabla \cdot \bar{F} dV = 3 \iiint_\Omega dV = 3V(\Omega) = 3 \frac{4}{3} \pi 1^3 = 4\pi. \quad (1.25)$$

Furthermore on $\partial \Omega$

$$\bar{n} = \frac{< x, y, z >}{\sqrt{x^2 + y^2 + z^2}} \quad \partial \Omega := \{(x, y, z) \in \mathbb{R} : x^2 + y^2 + z^2 = 1\}, \quad (1.26)$$
\[
\iint_{\partial \Omega} \bar{F} \cdot \bar{n} dS = \iint_{\partial \Omega} < x, y, z > \cdot < x, y, z > \frac{1}{\sqrt{x^2 + y^2 + z^2}} dS,
\]
\[
= \iint_{\partial \Omega} \frac{x^2 + y^2 + z^2}{\sqrt{x^2 + y^2 + z^2}} dS,
\]
\[
= \iint_{\partial \Omega} \sqrt{x^2 + y^2 + z^2} dS = \iint_{\partial \Omega} dS = A(\partial \Omega) = 4\pi. \quad (1.27)
\]

[2] Let:
\[
\bar{F}(x, y, z) = < x, y, z >, \quad (1.28)
\]
\[
\bar{\nabla} \times \bar{F} = \begin{vmatrix}
\hat{i} & \hat{j} & \hat{k} \\
\partial_x & \partial_y & \partial_z \\
x & y & z
\end{vmatrix} = 0, \quad (1.29)
\]
\[
\Omega := \{ (x, y, z) \in \mathbb{R}^3 : x^2 + y^2 \leq 1, z = 0 \},
\]
\[
\Gamma := \{ (x, y, z) \in \mathbb{R}^3 : x^2 + y^2 = 1, z = 0 \},
\]
\[
\bar{n} = (0, 0, 1), \quad (1.32)
\]
\[
\Rightarrow \iint_{\Omega} \nabla \times \bar{F} \cdot \bar{n} dS = 0. \quad (1.33)
\]

Further on \( \Gamma := \{ <x, y, z> \in \mathbb{R}^3 : x^2 + y^2 = 1, z = 0 \} \),
\[
\int_{\Gamma} \bar{F} \cdot d\bar{r} = \int_{\Gamma} \bar{F}(x(t), y(t), z(t)) \cdot d\bar{r},
\]
\[
x(t) = \cos(t), y(t) = \sin(t), z(t) = 0,
\]
\[
\Rightarrow d\bar{r} = \bar{r}'(t) dt = < -\sin(t), \cos(t), 0 > dt,
\]
\[
\Rightarrow \int_{\Gamma} \bar{F} \cdot d\bar{r} = \int_{0}^{2\pi} < \cos(t), \sin(t), 0 > \cdot < -\sin(t), \cos(t), 0 > dt = 0.
\]

[3] Consider:
\[
\bar{F}(x, y) = < x, y >, \quad (1.38)
\]
\[
\nabla \cdot \bar{F} = 2, \quad (1.39)
\]
\[
\Omega := \{ (x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1 \},
\]
\[
\Gamma := \{ (x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1 \},
\]
\[
\Rightarrow \iint_{\Omega} \nabla \cdot \bar{F} dA = 2 \int_{\Omega} dA = 2A(\Omega) = 2\pi. \quad (1.42)
\]
Furthermore on $\Gamma$

$$\vec{n} = < x, y > \frac{1}{\sqrt{x^2 + y^2}},$$

(1.43)

$$\int_{\Gamma} F \cdot \vec{n} dS = \int_{\partial \Omega} < x, y > \cdot < x, y > \frac{1}{\sqrt{x^2 + y^2}} dS,$$

(1.44)

$$= \int_{\Gamma} \frac{x^2 + y^2}{\sqrt{x^2 + y^2}} dS \int_{\Gamma} \frac{x^2 + y^2}{\sqrt{x^2 + y^2}} dS = \int_{\Gamma} dS = 2\pi.$$  

(1.45)


$$\begin{cases} 
-\Delta u = f & \text{on } \Omega \subset \mathbb{R}^3, \\
\frac{\partial u}{\partial n} = g_1 & \text{on } \partial \Omega, \partial \Omega = \bar{\Omega} \setminus \Omega.
\end{cases}$$

(1.46)

$$-\int_{\Omega} \Delta u dV = \int_{\Omega} f dV,$$

(1.47)

$$-\int_{\Omega} \nabla \cdot \nabla u dV = -\int_{\partial \Omega} \vec{n} \cdot \nabla u dS = -\int_{\partial \Omega} \frac{\partial u}{\partial n} dS = \int_{\Omega} f dV,$$

(1.48)

$$-\int_{\partial \Omega} g_1 dS = \int_{\Omega} f dV.$$  

(1.49)

The above condition is used as a check whether problem (1.46) has a solution at all. If the above relation (1.49) does not hold, then (1.46) is ill-posed (a mathematical word for nonsense). Relation (1.49) can be used for a check between $f$ and $g$.

**Exercise 1.2.1.** Does (1.50), specified below, have a solution?

$$\begin{cases} 
-\Delta u = 1 & \text{on } \Omega := \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 < 1\}, \\
\frac{\partial u}{\partial n} = 0 & \text{on } \partial \Omega := \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}.
\end{cases}$$

(1.50)

**Remark.** An important question that concerns a mathematical problem is whether a solution for the problem exists (existence), and if so, is the solution then unique (uniqueness)?
Minimization Problems

The application of the Finite Element Method (FEM) is very general. Further, from a mathematical point of view this method is elegant, however also more complicated than the method of Finite Differences or Finite Volumes. The most important reasons for the application of finite element methods is the ease of the use of odd-shaped domains, local mesh refinement, unstructured grids, jumps in coefficients and use of parallel computing.

The principle of the FEM is explained in two steps

[1] minimization problem,


Ritz’s FEM method is based on the solution of a discrete minimization problem. The method cannot be applied to all PDE’s. The method of Galerkin, based on a variational principle, can always be applied and, hence, is more general. The method of Galerkin will be treated in more detail in this course in Chapter 4. We will briefly treat some principles of minimization to motivate the variational principle in the Galerkin’s FEM. We will first see that for some cases minimization problems correspond to PDE’s, i.e. the solution of a minimization problem can also be a solution of a PDE with appropriate Boundary Conditions (BC). Later in Chapter 3, we will introduce “weak forms” of PDE’s. After this, we have sufficient background to actually concentrate on the Finite Element Method to be treated in Chapter 4.

We will see that a minimization problem and a (partial) differential equation sometimes have the same solution under strict smoothness requirements.

2.1 Example

A minimization problem can be (figure 2.1):
Find a smooth curve such that its length between points \((a, y_a)\) and \((b, y_b)\) is minimal.

In this course a function is smooth if all (partial) derivatives that we need exist. In more advanced courses Sobolev spaces are used to specify the necessary smoothness conditions. Let us not be bothered by this in this course. We will formulate this problem in a more mathematical way: the length of the curve is given by

\[
    l(y) = \int_{a}^{b} \sqrt{1 + f'(x)^2} \, dx, \quad \text{for any } y = f(x).
\]

Mathematically

**Minimize the integral** \(l(y)\) **over the class of smooth functions which satisfy**
\(y(a) = y_a\) and \(y(b) = y_b\).

or even more mathematically

Find \(\tilde{y}\), which is continuous and differentiable and satisfies \(\tilde{y}(a) = y_a\), \(\tilde{y}(b) = y_b\) such that

\[
    l(\tilde{y}) \leq l(y),
\]

for all \(y\) smooth and \(y(a) = y_a\), \(y(b) = y_b\).

Many physical laws are based on minimization problems:

- minimum potential energy,
- Hamilton’s principle in quantum mechanics,
- optical length,
- . . .

We go back to example 1:
Find $\tilde{y} \in Y := \{ y \text{ continuous, differentiable: } y(a) = y_a, y(b) = y_b \}$ such that $l(\tilde{y}) \leq l(y) \quad \forall y \in Y$.

The solution $\tilde{y}(x)$ is perturbed with a function $v(x)$, with $v(a) = 0$ and $v(b) = 0$, where $v(x)$ is continuous and differentiable. Given any $\varepsilon \in \mathbb{R}$, with $\varepsilon$ a real constant, then we put $y(x) = \tilde{y}(x) + \varepsilon v(x)$. Further

- $y(a) = \tilde{y}(a) + \varepsilon v(a) = y(a) + 0 = y_a$.
- $y(b) = \tilde{y}(b) + \varepsilon v(b) = y(b) + 0 = y_b$.
- since $\tilde{y}(x)$ and $v(x)$ are smooth and $\varepsilon$ is a constant, it follows that $y(x) = \tilde{y}(x) + \varepsilon v(x)$ is also smooth.

This implies that $\tilde{y}(x) \in Y$ and $y(x) \in Y$. Further we note that we search $\tilde{y}(x)$ such that

$$l(\tilde{y}(x)) \leq l(\tilde{y}(x) + \varepsilon v(x)), \quad \forall v(x) \text{ smooth and } v(a) = 0 = v(b).$$

Further $l(\tilde{y}(x) + \varepsilon v(x))$ is minimal for $\varepsilon = 0$. Hence $\frac{d}{d\varepsilon} l(\tilde{y}(x) + \varepsilon v(x))|_{\varepsilon=0} = 0 \quad \forall v \in V$. Where $V := \{v \text{ smooth: } v(a) = 0 = v(b)\}$. We will proceed with this:

$$l(\tilde{y} + \varepsilon v) = \int_a^b \sqrt{1 + (\tilde{y}'(x) + \varepsilon v'(x))^2} dx, \quad (2.3)$$

$$\frac{d}{d\varepsilon} l(\tilde{y} + \varepsilon v) = \frac{d}{d\varepsilon} \int_a^b \sqrt{1 + (\tilde{y}'(x) + \varepsilon v'(x))^2} dx, \quad (2.4)$$

$$\left. \frac{d}{d\varepsilon} l(\tilde{y} + \varepsilon v) \right|_{\varepsilon=0} = \int_a^b \frac{(\tilde{y}' + \varepsilon v') v'}{\sqrt{1 + (\tilde{y}' + \varepsilon v')^2}} dx = 0 \quad \forall v \in V.$$ (2.5)

Using partial integration, the above expression is written as:

$$\left[ \frac{\tilde{y}'v}{\sqrt{1 + (\tilde{y}')^2}} \right]_a^b - \int_a^b \frac{\tilde{y}'}{\sqrt{1 + (\tilde{y}')^2}} v dx = 0. \quad (2.8)$$

Since $v(a) = 0 = v(b)$, the first term in the left-hand side vanishes. The differentiation in the 2nd term of the left-hand side is done by the use of the chain rule:
\[
\frac{d}{dx} \left[ \frac{\tilde{y}'}{\sqrt{1 + (\tilde{y}')^2}} \right] = \frac{d}{d\tilde{y}'} \left[ \frac{\tilde{y}'}{\sqrt{1 + (\tilde{y}')^2}} \right] \cdot \tilde{y}'' = \frac{\tilde{y}''}{(1 + (\tilde{y}')^2)^{3/2}}. \tag{2.9}
\]

Hence one obtains:
\[
\int_a^b \frac{\tilde{y}''}{(1 + (\tilde{y}')^2)^{3/2}} \cdot vdx = 0 \quad \forall v \in V. \tag{2.10}
\]

One easily sees that \(\tilde{y}'' = 0\) on \(x \in (a, b)\) satisfies the above requirement. This implies that
\[
\begin{aligned}
\tilde{y}'' &= 0 \quad \text{for} \quad x \in (a, b), \\
\tilde{y}(a) &= y_a, \tilde{y}(b) = y_b.
\end{aligned} \tag{2.11}
\]

The minimization is now replaced by a differential equation. Note that we require \(\tilde{y}'' = 0\) and that for the differential equation to hold the second derivative must be continuous (and of course exist). Where, for the minimization problem we only need \(y'\) to exist. The differential equation puts an additional requirement on \(\tilde{y}\). Hence the conditions for the solution of the minimization problem are weaker than the conditions for the differential equation. Further, we see that \(\tilde{y}'' = 0\) satisfies (2.10), we did not prove yet that \(\tilde{y}'' = 0\) is the only possibility!

For now we assume that \(\tilde{y}'' = 0\) is the only possibility, later we will prove this fact in a lemma, which was proved for the first time by DuBois and Reymond. This can also be found in the books of van Kan et al. [2006] and Strang and Fix [1973] for instance.

\[
\tilde{y}'' = 0 \Rightarrow \tilde{y}(x) = Ax + B \quad (\text{linear}). \tag{2.12}
\]

Combination with the boundary conditions gives
\[
\tilde{y}(x) = y_a + \frac{y_b - y_a}{b - a}(x - a). \tag{2.13}
\]

This is the straight line that minimizes the length of the smooth curve that connects \((a, y_a)\) and \((b, y_b)\). Of course, the result is trivial but the relation between a minimization problem and a differential equation is clearly visible.

Now we will establish the Raymond/DuBois lemma:

**Lemma 2.1.1.** Let \(F(x)\) be continuous over \([a, b]\), i.e. \(F \in C^0[a, b]\) and suppose that
\[
\int_a^b F(x)v(x)dx = 0, \quad \forall v \in C^0[a, b] \quad \text{and} \quad v(a) = 0 = v(b)
\]
Then,
\[
F(x) = 0 \quad \text{for} \quad x \in [a, b].
\]
2.2 A general 1-dimensional minimization problem

Proof 2.1.1. We use contradiction as an argument. Suppose that $F(x_0) > 0$ for any $x_0 \in (a, b)$, note that the case $F(x_0) < 0$ is similar, then since $F(x)$ is continuous, there exists a $\delta > 0$ such that

$$F(x) > 0 \text{ whenever } |x - x_0| < \delta.$$ 

Now we choose (this is a trick)

$$v(x) = \begin{cases} 
(x - x_0 - \delta)(x - x_0 + \delta), & |x - x_0| < \delta, \\
0, & |x - x_0| > 0.
\end{cases}$$

Note that $v(a) = 0 = v(b)$ and that $v$ is differentiable and continuous over $[a, b]$. This implies that

$$\int_a^b F(x)v(x)dx = \int_{x_0-\delta}^{x_0+\delta} F(x)(x - x_0 - \delta)(x - x_0 + \delta)dx > 0.$$ 

This violates the hypothesis

$$\int_a^b F(x)v(x)dx = 0, \quad \forall v.$$ 

Hence $F(x) > 0$ is impossible at any point within $[a, b]$. Further the case $F(x) < 0$ is dealt with similarly. We conclude that $F(x) = 0$ on $[a, b]$. \qed

By the use of Lemma 2.1.1 is follows that $\tilde{y}'' = 0$ is the only possibility for condition (relation) (2.10). We will generalize this a little.

2.2 A general 1-dimensional minimization problem

Given the following minimization problem; given $f$ continuous with continuous partial derivatives

Minimize:

$$I(u) := \int_{x_0}^{x_1} f(x, u, u')dx,$$

over all $u \in C^1[x_0, x_1]$ with $u(x_0) = u_0$.

We will find a differential equation for this minimum $\hat{u}$, for which $\hat{u}(x_0) = u_0$.

$$I(\hat{u}) \leq I(u) \text{ for all } u \text{ continuous, differentiable and } u(x_0) = u_0.$$ 

Hence $\frac{d}{d\varepsilon} I(\hat{u} + \varepsilon v)|_{\varepsilon} = 0$, with $v$ differentiable and $\hat{u}(x_0) + \varepsilon v(x_0) = u_0$, so $v(x_0) = 0$. 

2.2 A general 1-dimensional minimization problem

\[ I(\hat{u} + \varepsilon v) = \int_{x_0}^{x_1} f(x, \hat{u} + \varepsilon v, \hat{u}' + \varepsilon v') dx, \]  
(2.15)

\[ \frac{d}{d\varepsilon} I(\hat{u} + \varepsilon v) = \int_{x_0}^{x_1} \left\{ \frac{\partial f}{\partial \hat{u}}(x, \hat{u} + \varepsilon v, \hat{u}' + \varepsilon v') v + \frac{\partial f}{\partial \hat{u}'}(x, \hat{u} + \varepsilon v, \hat{u}' + \varepsilon v') v' \right\} dx. \]  
(2.16)

Furthermore, since \( \frac{d}{d\varepsilon} I(\hat{u} + \varepsilon v)|_{\varepsilon=0} = 0 \), we obtain

\[ \int_{x_0}^{x_1} \left\{ \frac{\partial f}{\partial \hat{u}}(x, \hat{u}, \hat{u}') v + \frac{\partial f}{\partial \hat{u}'}(x, \hat{u}, \hat{u}') v' \right\} dx = 0, \]  
(2.17)

for all \( v \) smooth and \( v(x_0) = 0 \). Partial integration of the second term gives:

\[ \int_{x_0}^{x_1} \frac{\partial f}{\partial \hat{u}} v dx + \left[ \frac{\partial f}{\partial \hat{u}'} v \right]_{x_0}^{x_1} - \int_{x_0}^{x_1} \frac{\partial}{\partial x} \left( \frac{\partial f}{\partial \hat{u}'} \right) v dx = 0. \]  
(2.18)

In other words we have for all smooth functions \( v(x) \) smooth and \( v(x_0) = 0 \):

\[ \int_{x_0}^{x_1} \left\{ \frac{\partial f}{\partial \hat{u}} - \frac{\partial}{\partial x} \left( \frac{\partial f}{\partial \hat{u}'} \right) \right\} v dx = -\frac{\partial f}{\partial \hat{u}'} v |_{x_1}. \]  
(2.19)

Note that the above relation holds for every \( v \) with \( v(x_0) = 0 \). If we refine this class to both \( v(x_0) = 0 \) and \( v(x_1) = 0 \), we see that

\[ \int_{x_0}^{x_1} \left\{ \frac{\partial f}{\partial \hat{u}} - \frac{\partial}{\partial x} \left( \frac{\partial f}{\partial \hat{u}'} \right) \right\} v dx = 0, \]  
(2.20)

for all smooth \( v(x) \), with \( v(x_0) = 0 = v(x_1) \). From Lemma 2.1.1 then follows

\[ \int_{x_0}^{x_1} \left\{ \frac{\partial f}{\partial \hat{u}} - \frac{\partial}{\partial x} \left( \frac{\partial f}{\partial \hat{u}'} \right) \right\} v dx = 0, \quad \forall v : v(x_0) = 0 = v(x_1), \]  
(2.21)

\[ \Rightarrow \frac{\partial f}{\partial \hat{u}} - \frac{\partial}{\partial x} \left( \frac{\partial f}{\partial \hat{u}'} \right) = 0 \quad \text{for} \quad x \in (x_0, x_1). \]  
(2.22)

Since relation (2.22) holds it follows that equation (2.19) implies for every \( v(x_0) = 0 \) (now we drop the condition \( v(x_1) = 0 \)) that

\[ \frac{\partial f}{\partial \hat{u}'}(x_1, \hat{u}(x_1), \hat{u}'(x_1)) = 0. \]  
(2.23)

This is an additional boundary condition for the differential equation. It was not stated in the minimization problem. Since it is only required in the
differential equation, the condition is called a natural boundary condition. Hence the minimization problem gives

\[
\begin{cases}
  u(x_0) = u_0, \\
  \frac{\partial f}{\partial u} - \frac{\partial}{\partial x} \left( \frac{\partial f}{\partial u'} \right) = 0 \text{ on } (x_0, x_1), \\
  \frac{\partial f}{\partial u'}(x_1, u(x_1), u'(x_1)) = 0.
\end{cases}
\] (2.24)

The boundary condition \(u(x_0) = u_0\) is called an essential boundary condition: it concerns both the minimization problem and differential equation. The differential equation (2.22) is also referred to as the Euler-Lagrange equation.

### 2.3 A 2-dimensional minimization problem

First, we note that Dubois’s Lemma 2.1.1 can also be proved for 2 dimensional and 3 dimensional situations.

**Lemma 2.3.1.** Let \(F(x, y)\) be continuous over \(\Omega \subset \mathbb{R}^2\) with \(\Gamma\) as its boundary, i.e. \(F \in C^0(\Omega)\) and suppose that

\[
\int_{\Omega} F(x, y)v(x, y)dA = 0, \quad \forall v(x, y) \in C^0(\Omega) \quad \text{with} \quad v(x, y)|_{\Gamma} = 0.
\]

Then

\[
F(x, y) = 0 \quad \text{on} \quad (x, y) \in \Omega \cup \Gamma.
\]

To treat two dimensional minimization problems, we use the following example:

\[
\min_{u \in U} \int_{\Omega} \left\{ \frac{1}{2} \| \bar{\nabla} u \|^2 - f u \right\} dA + \int_{\Gamma_2} \alpha ud\mathbf{s} =: \min_{u \in U} J(u),
\]

(2.25)

where \(U\) is the set of functions with a continuous derivative on \(\Omega\) and where \(u|_{\Gamma_1} = u_1\):

\[
U := \{u \text{ smooth : } u|_{\Gamma_1} = u_1\}. \quad (2.26)
\]

Further, we assume that the domain of computation \(\Omega\) is bounded by \(\Gamma_1 \cup \Gamma_2\), where \(\Gamma_1\) and \(\Gamma_2\) do not overlap. Again we search \(\hat{u} \in U\) such that \(J(\hat{u}) \leq J(u)\), \(\forall u \in U\). We put \(J(\hat{u} + \varepsilon v) \geq J(\hat{u})\) \(\forall v \in V := \{v \text{ smooth : } v|_{\Gamma_1} = 0\}\) (why is \(v = 0\) on \(\Gamma_1\)?)

\[
\frac{d}{d\varepsilon} J(\hat{u} + \varepsilon v)|_{\varepsilon=0} = 0, \quad (2.27)
\]

hence \(J(\hat{u} + \varepsilon v)\) should be minimal for \(\varepsilon = 0\).
\[ J(\hat{u} + \varepsilon v) = \int_{\Omega} \left\{ \frac{1}{2} \nabla (\hat{u} + \varepsilon v) \cdot \nabla (\hat{u} + \varepsilon v) - (\hat{u} + \varepsilon v)f \right\} dA + \int_{\Gamma_2} (\hat{u} + \varepsilon v) \alpha ds, \]

\[ = \int_{\Omega} \left\{ \frac{1}{2} \nabla \hat{u} \cdot \nabla \hat{u} + \varepsilon \nabla v \cdot \nabla \hat{u} + \frac{1}{2} \varepsilon^2 \nabla v \cdot \nabla v - f \hat{u} - \varepsilon f v \right\} dA \]

\[ + \int_{\Gamma_2} (\alpha \hat{u} + \alpha \varepsilon v) ds, \quad (2.28) \]

\[ \frac{d}{d\varepsilon} J(\hat{u} + \varepsilon v) = \int_{\Omega} \left\{ \nabla v \cdot \nabla \hat{u} + \varepsilon \| \nabla v \|^2 - f \right\} dA + \int_{\Gamma_2} \alpha v ds, \quad (2.29) \]

\[ \frac{d}{d\varepsilon} J(\hat{u} + \varepsilon v)|_{\varepsilon = 0} = \int_{\Omega} \nabla v \cdot \nabla \hat{u} dA - \int_{\Omega} f v dA + \int_{\Gamma_2} \alpha v ds = 0 \quad \text{for all } v|_{\Gamma_1} = 0. \quad (2.30) \]

Later we will call equation (2.30) a weak formulation. The Product Rule of differentiation, using

\[ \int_{\Omega} \nabla v \cdot \nabla \hat{u} dA = - \int_{\Omega} v \Delta \hat{u} dA + \int_{\Omega} \nabla \cdot (\nabla \hat{u} v) dA, \quad (2.31) \]

(from \( \nabla \cdot (v \nabla \hat{u}) = v \Delta \hat{u} + \nabla v \cdot \nabla \hat{u} \)), it follows for equation (2.30):

\[ \int_{\Omega} \nabla \cdot [v \nabla \hat{u}] dA - \int_{\Omega} v \Delta \hat{u} dA - \int_{\Omega} f v dA + \int_{\Gamma_2} \alpha v ds = 0. \quad (2.32) \]

Making use of the two-dimensional Divergence theorem in the first term of the above equation, gives

\[ \int_{\Gamma_1 \cup \Gamma_2} \bar{n} \cdot [v \nabla \hat{u}] ds + \int_{\Gamma_2} \alpha v ds = \int_{\Omega} [\Delta \hat{u} + f] v dA \quad \forall v \in V. \quad (2.33) \]

Since \( \hat{u} = u_1 \) on \( \Gamma_1 \), we have \( v = 0 \) on \( \Gamma_1 \) and therefore

\[ \int_{\Gamma_1} v \frac{\partial \hat{u}}{\partial n} ds = 0, \quad (2.34) \]

hence

\[ \int_{\Gamma_2} [\frac{\partial \hat{u}}{\partial n} + \alpha] v ds = \int_{\Omega} [\Delta \hat{u} + f] v dA \quad \forall v \in V. \quad (2.35) \]

In other words we obtain from the minimization problem:

Find \( \hat{u} \in U \) such that

\[ \int_{\Omega} [\Delta \hat{u} + f] v dA = \int_{\Gamma_2} [\frac{\partial \hat{u}}{\partial n} + \alpha] v ds \quad \forall v \in V. \quad (2.36) \]

Hence using the DuBois twice:
2.4 Some formalization

[1] Take \( v|_{\Gamma_2} = 0 \), then from DuBois’ Lemma it follows
\[
\Delta \hat{u} + f = 0;
\]

[2] The previous item implies
\[
\int_{\Gamma_2} \left[ \frac{\partial \hat{u}}{\partial n} + \alpha \right] ds = 0, \forall v \in V,
\]
then, once again DuBois gives the answer
\[
\frac{\partial \hat{u}}{\partial n} + \alpha = 0.
\]

Hence, one obtains
\[
\begin{cases}
-\Delta u = f, \\
u|_{\Gamma_1} = u_1 \quad \text{on} \quad \Gamma_1 \quad \text{(Dirichlet)}, \\
\frac{\partial \hat{u}}{\partial n}|_{\Gamma_2} = -\alpha \quad \text{on} \quad \Gamma_2 \quad \text{(Neumann)}.
\end{cases}
\tag{2.37}
\]

The solution of this problem coincides with the minimization problem. The Dirichlet BC is an “essential” BC, the Neumann BC is a new condition ‘introduced’ by use of DuBois: “Natural BC”. The natural BC did not appear in the minimization problem.

**Exercise 2.3.1.** Given
\[
J(u) = \frac{1}{2} \int_{\Omega} \| \nabla u \|^2 dA + \int_{\Gamma_2} \frac{1}{2} gu^2 - hu \} ds,
\tag{2.38}
\]
with \( u|_{\Gamma_1} = 0 \) show that \( J(u) \) is minimized by the solution of the PDE
\[
\begin{cases}
\Delta u = 0 \quad \text{on} \quad \Omega, \\
\frac{\partial u}{\partial n} + gu = h \quad \text{on} \quad \Gamma_2, \\
u|_{\Gamma_1} = 0.
\end{cases}
\tag{2.39}
\]

*Hint:* use \( J(u + \varepsilon v) \geq J(u) \) when \( u \) is the solution. Set \( \frac{d}{d\varepsilon} J(u + \varepsilon v)|_{\varepsilon=0} = 0 \).

2.4 Some formalization

**Definition 2.4.1.** A functional is a map from a real vector-space (of functions) onto \( \mathbb{R} \).

**Example.** Let \( f(x) \) be a given function, then
\[
J(u) = \int_{x_0}^{x_1} \left\{ \left( \frac{du}{dx} \right)^2 + fu \right\} dx,
\tag{2.40}
\]
is a functional:

- Input is $u = u(x)$: a whole curve.
- Output $\in \mathbb{R}$.

Further $f(x)$ is a known, fixed function of $x$.

The following theorem answers the question whether the solution of a linear (partial) differential problem is also the solution of a minimization problem.

**Theorem 2.4.2.** Let $\mathcal{L}(\cdot)$ be a linear operator, for instance a minus Laplace operator ($\mathcal{L}u := -\Delta u$) with homogeneous boundary conditions, then the function, $u_0$, which is the solution of $\mathcal{L}u_0 = f$, minimizes

$$J(u) = \int_{\Omega} \left( \frac{1}{2} u \mathcal{L}u - uf \right) dA,$$

(2.41)
on $U$, if the following conditions are satisfied

- $\mathcal{L}$ is positive, that is, $\int_{\Omega} u \mathcal{L}udA > 0 \ \forall u \in U$, for which $u \neq 0$ for at least one element of $\Omega$.
- $\mathcal{L}$ is self-adjoint (symmetric), that is, $\int_{\Omega} v \mathcal{L}udA = \int_{\Omega} u \mathcal{L}vdA \ \forall u, v \in U$.

The above remark is proved. We refer to Kreyszig [1989] for a proof for more general cases concerning non-homogeneous boundary conditions.

**Proof:** We note that

$$J(u) = \int_{\Omega} \left( \frac{1}{2} u \mathcal{L}u - uf \right) dA =$$

$$\frac{1}{2} \int_{\Omega} (u - u_0) \mathcal{L}(u - u_0) dA - \int_{\Omega} uf dA + \frac{1}{2} \int_{\Omega} (u_0 \mathcal{L}u + u \mathcal{L}u_0 - u_0 \mathcal{L}u_0) dA.$$  

(2.42)

Since $\mathcal{L}$ is self-adjoint, and $\mathcal{L}u_0 = f$, the above relation can be written as

$$J(u) = \frac{1}{2} \int_{\Omega} (u - u_0) \mathcal{L}(u - u_0) dA - \frac{1}{2} \int_{\Omega} u_0 \mathcal{L}u_0 dA \geq -\frac{1}{2} \int_{\Omega} u_0 \mathcal{L}u_0 dA. \quad (2.43)$$

The last inequality follows since the operator $\mathcal{L}$ is positive and equality holds if $u = u_0$. Hence, $u_0$ minimizes $J(u)$. $\square$

In the above theorem, we showed that positivity and self-adjointness are sufficient conditions for the existence of a minimization problem. Strictly speaking, if self-adjointness (symmetry) is violated, we did not demonstrate that no minimization problem exists. However, in practice, positivity and symmetry are also necessary conditions for the existence of a minimization problem.

**Examples:**
2.4 Some formalization

[1] For \( -\Delta u = f \) a minimization problem can be formulated when \( u|_{\partial\Omega} = 0 = v|_{\partial\Omega} \).

[2] For \( \Delta u - \bar{q} \cdot \nabla u = f \) no minimization problem can be formulated for \( u|_{\partial\Omega} = 0 \).

Reasons:

[1]

\[
-\int_{\Omega} v\Delta u dA = -\int_{\Omega} \nabla \cdot [v \nabla u] dA + \int_{\Omega} \nabla v \cdot \nabla u dA,
\]

\[
= -\int_{\partial\Omega} \bar{n} \cdot v \nabla u dS + \int_{\Omega} \nabla v \cdot \nabla u dA,
\]

\[
= + \int_{\Omega} \nabla v \cdot \nabla u dA \quad (\text{since} \quad u|_{\partial\Omega} = 0 = v|_{\partial\Omega}),
\]

\[
-\int_{\Omega} u\Delta v dA = -\int_{\Omega} \nabla \cdot [u \nabla v] dA + \int_{\Omega} \nabla u \cdot \nabla v dA,
\]

\[
= + \int_{\Omega} \nabla u \nabla v dA \quad (\text{since} \quad u|_{\partial\Omega} = 0 = v|_{\partial\Omega}).
\]

Hence \(-\Delta(\cdot)\) is self-adjoint, i.e.

\[
\int_{\Omega} v\Delta u dA = \int_{\Omega} u\Delta v dA.
\]

And

\[
-\int_{\Omega} u\Delta u dA = \int_{\Omega} \| \nabla u \|^2 dA \geq 0,
\]

hence \(-\Delta(\cdot)\) is positive.

[2] Let \( \bar{q} \) be a given constant vector and \( u|_{\partial\Omega} = 0 = v|_{\partial\Omega} \), then

\[
\int_{\Omega} v\bar{q} \cdot \nabla u dA = \int_{\partial\Omega} u\bar{q} \cdot \bar{n} dS - \int_{\Omega} u\bar{q} \cdot \nabla v dA = -\int_{\Omega} u\bar{q} \cdot \nabla v dA.
\]

Hence,

\[
\int_{\Omega} v\bar{q} \cdot \nabla u dA \neq \int_{\Omega} u\bar{q} \cdot \nabla v dA,
\]

(not self-adjoint).
Variational Formulation and Differential equations

In the previous chapter we saw the relation between minimization problems and (partial) differential equations. It was demonstrated that if the differential operator is positive and self-adjoint, then, such a minimization problem exist. Ritz’ finite element method is based on the numerical solution of a minimization problem. To solve problems with differential operators that do not satisfy these requirements, the so-called weak form is introduced. The differential equation is written as a weak form and then a numerical solution to this weak form is determined. This method is more generally applicable and it is the backbone of Galerkin’s finite element method.

3.1 Weak forms

Consider the following minimization problem on domain $\Omega$ with boundaries $\partial \Omega = \Gamma_1 \cup \Gamma_2$:

$$\begin{align*}
\text{Find } \hat{u} \text{ smooth, such that } \hat{u}|_{\Gamma_1} &= g, \\
J(\hat{u}) &\leq J(u) \text{ for all smooth } u \text{ with } u|_{\Gamma_1} = g, \\
\text{where } J(u) &:= \frac{1}{2} \int_{\Omega} \|\nabla u\|^2 dA.
\end{align*}$$

(3.1)

Using $u = \hat{u} + \varepsilon v$ for all smooth $v$ with $v|_{\Gamma_1} = 0$, it follows that the solution of equation (3.1) coincides with:

$$\begin{align*}
\text{Find } \hat{u} \text{ smooth, such that } \hat{u}|_{\Gamma_1} &= g, \\
\int_{\Omega} \nabla \hat{u} \cdot \nabla v dA &= 0 \text{ for all smooth } v \text{ with } v|_{\Gamma_1} = 0.
\end{align*}$$

(3.2)

It can be demonstrated that the solution of the above exists and that it is unique. We suppose that the boundary of $\Omega$ is given by $\Gamma_1 \cup \Gamma_2$. The problems
(3.1) and (3.2) have the same solution. The Product Rule for differentiation applied to $\int_\Omega \vec{\nabla} u \cdot \vec{\nabla} v dA$ gives:

$$\int_\Omega \vec{\nabla} \cdot [v \vec{\nabla} u] dA - \int_\Omega v \Delta u dA = 0 \quad \forall v \quad \text{with} \quad v|_{\Gamma_1} = 0,$$

or

$$\int_{\Gamma_1} v \frac{\partial u}{\partial n} ds + \int_{\Gamma_2} v \frac{\partial u}{\partial n} ds = \int_\Omega v \Delta u dA \quad \forall v|_{\Gamma_1} = 0 \quad \text{(with} \ v \text{ smooth)},$$

with $v|_{\Gamma_1} = 0$, follows

$$\int_{\Gamma_2} v \frac{\partial u}{\partial n} dS = \int_\Omega v \Delta u dA \quad \forall v|_{\Gamma_1} = 0 \quad \text{with} \ v \text{ smooth}. \quad (3.5)$$

Suppose that $v|_{\Gamma_2} = 0$ besides $v|_{\Gamma_1} = 0$, then Du Bois’s Lemma gives $\Delta u = 0$ on $\Omega$. When we release $v|_{\Gamma_2} = 0$ and use $\Delta u = 0$ on $\Omega$, we obtain the following natural boundary condition $\frac{\partial u}{\partial n}|_{\Gamma_2} = 0$. Hence smooth solutions of (3.1) and (3.2) satisfy:

$$\begin{cases}
-\Delta u = 0, \\
u|_{\Gamma_1} = g, \\
\frac{\partial u}{\partial n}|_{\Gamma_2} = 0.
\end{cases} \quad (3.6)$$

When $\Delta u$ exists within $\Omega$, then the solutions of equations (3.1), (3.2) and (3.6) are the equal. (3.6) contains a PDE, (3.1) is its corresponding minimization problem and (3.2) is called a ‘variational formulation’ or a ‘weak form' of PDE (3.6).

So far we went from a weak form to a problem with a PDE. In practice, one often goes the other way around. Since Finite Element Methods are based on either the solution of a minimization problem (such as (3.1)) or a weak form (as in (3.2)), we would like to go from a PDE to a weak from. Further, the condition $v|_{\Gamma_1} = 0$ because of $u|_{\Gamma_1} = g$ and hence prescribed, originates from the use of a minimization problem.

Solving of (3.6) by a numerical solution of the representation of (3.2) is referred to as Galerkin’s method. Whereas, acquiring the numerical solution of a representation of (3.1) is called Ritz’ method. Galerkin’s method is most general: it can always be applied. It doesn’t matter whether differential operators are self-adjoint or positive. Therefore, this method will be treated in more detail. The study of minimization problems was needed to motivate the condition $v = 0$ on locations where $u$ is prescribed (by an essential condition).

A mayor advantage of the weak form (3.2) is the fact it is easier to prove existence and uniqueness of (3.2) than for one satisfying (3.6) with boundary conditions. It is clear that a solution of (3.6) always is always a solution of (3.2). A solution of the PDE (3.6) always needs the second order derivatives to
exist, whereas in the solution of (3.2) only the integrals have to exist. For the solutions of the weak form, it may be possible that the second order derivatives do not exist at all. For that reason, the term weak form or weak solution is used for the problem and its solution respectively.

The function \( v \) is commonly referred to as a 'test function'. Let’s go from (3.6) to (3.2). Given \( \Delta u = 0 \iff v \Delta u = 0 \) for all \( v|_{\Gamma_1} = 0 \) (reason is that \( u|_{\Gamma_1} = g \) is prescribed!) then

\[
\int_{\Omega} v \Delta u \, dA = 0, \tag{3.7}
\]

\[
\int_{\Omega} \nabla \cdot [v \nabla u] \, dA - \int_{\Omega} \nabla u \cdot \nabla vdA = 0 \quad \forall v|_{\Gamma_1} = 0. \tag{3.8}
\]

The Product Rule for differentiation was used here. Using the Divergence Theorem, this gives (since \( v|_{\Gamma_1} = 0 \))

\[
\int_{\Gamma_2} v \frac{\partial u}{\partial n} \, ds - \int_{\Omega} \nabla u \cdot \nabla vdA = 0 \quad \forall \ v|_{\Gamma_1} = 0. \tag{3.9}
\]

Since in (3.6) it is required that \( \frac{\partial u}{\partial n}|_{\Gamma_2} = 0 \), we obtain

\[
\int_{\Gamma_2} v \frac{\partial u}{\partial n} \, ds = 0, \tag{3.10}
\]

and hence,

\[
\int_{\Omega} \nabla u \cdot \nabla vdA = 0 \quad \forall \ v|_{\Gamma_1} = 0 \quad \text{smooth.} \tag{3.11}
\]

Hence (3.6) is equivalent to (3.2), if we are not bothered by the smoothness considerations:

\[
\begin{cases}
\text{Find } \hat{u} \text{ smooth, such that } \hat{u}|_{\Gamma_1} = g, \text{ and } \\
\int_{\Omega} \nabla u \cdot \nabla vdA = 0 \quad \text{for all smooth } v \text{ with } v|_{\Gamma_1} = 0. \tag{3.12}
\end{cases}
\]

Here (3.2) is also sometimes referred to as the 'Finite Element' formulation of (3.6). The same principle may be applied to

\[
\begin{cases}
\frac{\partial c}{\partial t} = \Delta c + f, \\
c|_{\Gamma_1} = g, \\
\frac{\partial c}{\partial n}|_{\Gamma_2} = h, \\
c(x, y, 0) = 0, \quad t = 0 \quad (x, y) \quad \Omega. \tag{3.13}
\end{cases}
\]

In the above problem the boundary of the domain of computation \( \Omega \) is given by \( \Gamma_1 \cup \Gamma_2 \). The question is now to find a Finite Element formulation for (3.13). We
multiply the PDE with a testfunction \( v \), that satisfies \( v|_{\Gamma_1} = 0 \), since \( c|_{\Gamma_1} = g \) is prescribed, to obtain, after integration over \( \Omega \),

\[
\int_{\Omega} \frac{\partial c}{\partial t} v dA = \int_{\Omega} v \Delta c dA + \int_{\Omega} f v dA \quad \forall v |_{\Gamma_1} = 0, \tag{3.14}
\]

(\( v \) smooth). By use of the productrule for differentiation, we obtain

\[
\int_{\Omega} \frac{\partial c}{\partial t} v dA = \int_{\Omega} \nabla \cdot \left[ v \nabla c \right] dA - \int_{\Omega} \nabla c \cdot \nabla v dA + \int_{\Omega} f v dA, \quad \forall v |_{\Gamma_1} = 0. \tag{3.15}
\]

The Divergence Theorem implies:

\[
\int_{\Omega} \frac{\partial c}{\partial t} v dA = \int_{\Gamma_2} \frac{\partial c}{\partial n} v ds + \int_{\Gamma_2} \frac{\partial c}{\partial n} v ds - \int_{\Omega} \nabla c \cdot \nabla v dA + \int_{\Omega} f v dA, \quad \forall v |_{\Gamma_1} = 0. \tag{3.16}
\]

Since \( \frac{\partial c}{\partial n} = h \) on \( \Gamma_2 \) and \( v |_{\Gamma_1} = 0 \), we obtain: Find \( c \) with \( c |_{t=0} = 0 \), \( c |_{\Gamma_1} = g \) such that

\[
\int_{\Omega} \frac{\partial c}{\partial t} v dA = \int_{\Gamma_2} h v ds - \int_{\Omega} \nabla c \cdot \nabla v dA + \int_{\Omega} f v dA, \quad \forall v |_{\Gamma_1} = 0. \tag{3.17}
\]

Equation (3.17) is the variational form or Finite Element Form of (3.13). Note that the Neumann BC is changed into a line-integral over \( \Gamma_2 \). Of course, it is easy to show that (3.13) can be derived, once only (3.17) is given:

\[
\int_{\Omega} \left[ \frac{\partial c}{\partial t} - f \right] v dA = \int_{\Gamma_2} h v ds - \int_{\Omega} \nabla \cdot \left[ v \nabla c \right] dA + \int_{\Omega} v \Delta c dA, \quad \forall v |_{\Gamma_1} = 0. \tag{3.18}
\]

Using \( v |_{\Gamma_1} = 0 \), this gives

\[
\int_{\Omega} \left[ \frac{\partial c}{\partial t} - \Delta c - f \right] v dA = \int_{\Gamma_2} \left[ h - \frac{\partial c}{\partial n} \right] v ds \quad \forall v |_{\Gamma_1} = 0. \tag{3.19}
\]

If we set \( v |_{\Gamma_2} = 0 \) besides \( v |_{\Gamma_1} = 0 \), we obtain from DuBois

\[
\int_{\Omega} \left[ \frac{\partial c}{\partial t} - \Delta c - f \right] v dA = 0 \Rightarrow \frac{\partial c}{\partial t} - \Delta c - f = 0 \quad \text{on} \ \Omega. \tag{3.20}
\]

This implies after releasing \( v |_{\Gamma_2} = 0 \) that \( h - \frac{\partial c}{\partial n} = 0 \) on \( \Gamma_2 \) (again from DuBois). We see that (3.17) corresponds with (3.13), since we require \( c |_{\Gamma_1} = g \) and \( c |_{t=0} = 0 \) for both (3.13) and (3.17). Note that for the derivation of the weak form, we always multiply the PDE with a test-function \( v \), which must satisfy \( v = 0 \) on a boundary with a Dirichlet condition (i.e. an essential condition). Subsequently we integrate over the domain of computation.
Exercise 3.1.1. Suppose that we have been given the following problem:

\[
\begin{aligned}
\frac{\partial c}{\partial t} &= \Delta c \quad \text{on } \Omega, \\
c \big|_{\Gamma_1} &= f \quad \text{on } \Gamma_1, \\
c \big|_{\Gamma_2} + \frac{\partial c}{\partial n} \big|_{\Gamma_2} &= g \quad \text{on } \Gamma_2, \\
c(x,y,0) &= 0 \quad \text{for } t = 0 \text{ on } \Omega.
\end{aligned}
\]  

(3.21)

Show that a weak form of the above problem is given by:

Find \( c \) smooth, subject to \( c \big|_{\Gamma_1} = f \) and \( c \big|_{t=0} = 0 \), such that

\[
\int_{\Omega} \frac{\partial c}{\partial t} \, vdA = - \int_{\Omega} \nabla c \cdot \nabla v \, dA + \int_{\Gamma_2} (g - c) \, vds \quad \forall v \big|_{\Gamma_1} = 0.
\]  

(3.22)

\( v \) smooth.

The above weak form (3.22) is used to solve (3.21) by the use of Finite Elements. Note that the Robin-condition is a natural boundary condition, which is contained in the weak form in the second term of the right-hand of the (3.22).

### 3.2 Which weak formulation?

When we considered

\[
\begin{aligned}
\Delta u &= f \quad \text{on } \Omega, \\
u \big|_{\Gamma} &= 0,
\end{aligned}
\]  

(3.23)

then we saw that a weak form is given by (2.30)

\[
\begin{aligned}
\text{Find } u \big|_{\Gamma} &= 0 \text{ such that,} \\
- \int_{\Omega} \nabla u \cdot \nabla v \, dA &= \int_{\Omega} fv \, dA \quad \forall v \big|_{\Gamma} = 0.
\end{aligned}
\]  

(3.24)

The above problem is a weak form, but the following problem is also a weak form:

\[
\begin{aligned}
\text{Find } u \big|_{\Gamma} &= 0 \text{ such that,} \\
\int_{\Omega} v \Delta udA &= \int_{\Omega} fv \, dA \quad \forall v \big|_{\Gamma} = 0,
\end{aligned}
\]  

(3.25)

or even

\[
\begin{aligned}
\text{Find } u \big|_{\Gamma} &= 0 \text{ such that,} \\
- \int_{\Omega} u \Delta v \, dA &= \int_{\Omega} fv \, dA \quad \forall v \big|_{\Gamma} = 0.
\end{aligned}
\]  

(3.26)

Forms (3.24), (3.25) and (3.26) are all possible weak forms of (3.23). However, in the Finite Element calculations, (3.25) and (3.26) are not common. This is due to the reduction of order of the derivatives in the first form (3.24). Here only the first derivatives are used and this will give an advantage for the implementation.
3.3 Mathematical considerations: existence and uniqueness

of the FEM, which we will see later. A more important advantage is that for a minimized order of derivatives in the weak form, the class of allowable solutions is largest, in the sense that solutions that are less smooth are allowable. As a rule of thumb, now, we mention that when we derive a weak form, then we should try to minimize the highest order of the derivatives that occur in the integrals.

Example:

\[
\begin{cases}
u'''' = f & \text{on } x \in (0, 1), \\
u(0) = 0, \\
u'(0) = 0, \\
u(1) = 0, \\
u'(1) = 0.
\end{cases}
\] (3.27)

We derive a weak form with the lowest order for the derivatives.

\[
\int_0^1 u''''v \, dx = \int_0^1 f \, v \, dx \quad \text{for all } v(0) = 0 = v'(0) = v(1) = v'(1),
\] (3.28)

partial integration gives

\[
[u'''v]_0^1 - \int_0^1 u''v' \, dx = \int_0^1 f \, v \, dx \quad \text{for all } v(0) = 0 = v'(0) = v(1) = v'(1),
\] (3.29)

with the condition \( v(0) = 0 = v(1) \) follows

\[
-\int_0^1 u'''v' \, dx = \int_0^1 f \, v \, dx \quad \text{for all } v(0) = 0 = v'(0) = v(1) = v'(1).
\] (3.30)

Partial integration, again, gives (with \( v'(1) = 0 = v'(0) \))

\[
\int_0^1 u''v'' \, dx = \int_0^1 f \, v \, dx \quad \text{for all } v(0) = 0 = v'(0) = v(1) = v'(1).
\] (3.31)

Now we stop, because, another partial integration would increase the maximum order of the derivatives again to obtain

\[
\int_0^1 u'v''' \, dx = \int_0^1 f \, v \, dx \quad \text{for all } v(0) = 0 = v'(0) = v(1) = v'(1).
\] (3.32)

We do not use (3.32) but (3.31) as the weak form for the Finite Element Method.

3.3 Mathematical considerations: existence and uniqueness

This section is intended for the interested reader and it is not necessary for the understanding of the finite element method. Consider the following Poisson
problem
\[-\Delta u = f(x), \quad \text{for } x \in \Omega,\]
\[u = g(x), \quad \text{for } x \in \partial \Omega.\]  
(3.33)

Here we assume that \(f(x)\) and \(g(x)\) are given continuous functions. Let \(\bar{\Omega} = \Omega \cup \partial \Omega\) be the closure of \(\Omega\), then the following assertion can be demonstrated

**Theorem 3.3.1.** If \(f(x)\) and \(g(x)\) are continuous and if the boundary curve is (piecewise) smooth, then problem (3.33) has one and only one solution such that \(u \in C^2(\Omega) \cap C^1(\bar{\Omega})\) (that is the solution has continuous partial derivatives up to at least the second order over the open domain \(\Omega\) and at least continuous first order partial derivatives on the boundary).

We will not prove this result for the existence and uniqueness of a classical solution to problem (3.33). The proof of the above theorem is far from trivial, the interested reader is referred to the monograph of Evans for instance. The fact that the second order partial derivatives need to be continuous is a rather strong requirement.

The finite element representation of the above problem is given by

Find \(u \in H^1(\Omega)\), subject to \(u = g\) on \(\partial \Omega\), such that
\[
\int_{\Omega} \nabla u \cdot \nabla \phi d\Omega = \int_{\Omega} \phi f d\Omega, \quad \text{for all } \phi \in H^1(\Omega).
\]  
(3.34)

In the above problem (4.86), the notation \(H^1(\Omega)\) has been used, this concerns the set of functions for which the integral over \(\Omega\) of the square of the function and its gradient is finite. Informally speaking, this is

\[u \in H^1(\Omega) \iff \int_{\Omega} u^2 d\Omega < \infty \quad \text{and} \quad \int_{\Omega} ||\nabla u||^2 d\Omega < \infty.\]  
(3.35)

This set of functions represents a *Hilbert space* and is commonly referred to as a *Sobolev space*. Using the fact that each function that is in \(H^1(\Omega)\) is also continuous on \(\Omega\), that is \(H^1(\Omega) \subset C^0(\Omega)\), the following claim can be proved

**Theorem 3.3.2.** The problem (4.86) has one and only one solution \(u\), such that \(u \in H^1(\Omega)\).

The proof of the above theorem resides on the Lax-Milgram Theorem (see for instance the book by Kreyszig [1989]):

**Theorem 3.3.3.** Let \(V\) be a Hilbert space and \(a(\cdot, \cdot)\) a bilinear form on \(V\), which is

1. bounded: \(|a(u, v)| \leq C||u||||v||\) and
2. coercive: \(a(u, u) \geq c||u||^2\).

Then, for any linear bounded functional \(f \in V'\), there is a unique solution \(u \in V\) to the equation
\[a(u, v) = f(v), \quad \text{for all } v \in V.\]  
(3.36)
The proof takes into account the fact that the linear operator is positive (more exactly speaking coercive, which is $\int_{\Omega} \|\nabla u\|^2 d\Omega \geq \alpha \int_{\Omega} u^2 d\Omega$ for some $\alpha > 0$) and continuous. Further, the right hand side represents a bounded linear functional. These issues constitute the hypotheses under which the Lax-Milgram theorem holds and hence have to be demonstrated. In this text the (straightforward) proof is omitted and a full proof of the above theorem can be found in van Kan [2006] for instance.

The most important lesson that we learn here, is that the solution to the weak form exists and that it is uniquely defined. Further, the weak form allows a larger class of functions as solutions than the PDE does.
Galerkin’s Finite Element method

In this chapter we treat the Finite Element Method, which was proposed by Galerkin. The method is based on the weak form of the PDE. This Galerkin method is more general than the Ritz’ method, which is based on the solution of a minimization problem and, hence, is only suitable whenever the differential operator is positive and self-adjoint.

4.1 The principle of Galerkin’s method

Given the following weak form:

\[
\begin{align*}
\text{Find } u \mid_{\Gamma} = 0 \text{ such that,} \\
\int_{\Omega} \nabla u \cdot \nabla v dA = 0 \text{ for all } v \mid_{\Gamma} = 0.
\end{align*}
\]

(4.1)

Here \( \Omega \) is a general simply connected domain in \( \mathbb{R}^1 \) or \( \mathbb{R}^2 \) or \( \mathbb{R}^3 \). A crucial principle for the FEM is that we write \( u \) as a sum of “basis-functions” \( \varphi_i(x, y) \), which satisfy

\[
\varphi_i(x, y) \mid_{\Gamma} = 0,
\]

(4.2)

\text{i.e. hence}

\[
u(x, y) = \sum_{j=1}^{\infty} c_j \varphi_j(x, y).
\]

(4.3)

Since, we cannot do calculations with an infinite number of terms, we truncate this series such that we only take the first \( n \) terms into account, then

\[
\hat{u}(x, y) = \sum_{j=1}^{n} c_j \varphi_j(x, y),
\]

(4.4)
where \( \hat{u} \) denotes the approximation of the solution of (4.1). As an example for \( \varphi_i \), one might take powers, sines, cosines (viz. Fourier) and so on. We will assume here that

\[
\hat{u}(x, y) = \sum_{j=1}^{n} c_j \varphi_j(x, y) \rightarrow u(x, y) \quad \text{as} \quad n \rightarrow \infty,
\]

(4.5)

note that \( \hat{u}(x, y) \) represents the approximated solution of (4.1) and \( u(x, y) \) the exact solution of (4.1) respectively. There is a lot of mathematical theory needed to prove that \( \hat{u} \rightarrow u \) as \( n \rightarrow \infty \) for a specific set of basis-functions \( \varphi_i(x, y) \). The Finite Element representation of weak form (4.1) is:

\[
\begin{cases}
\text{Find the set of constants } \{c_1, \ldots, c_n\} \text{ such that,} \\
\sum_{j=1}^{n} \int_{\Omega} c_j \nabla \varphi_j(x, y) \cdot \nabla \varphi_i(x, y) dA = 0 \quad \text{for all } i \in \{1, \ldots, n\}.
\end{cases}
\]

(4.6)

Note that we assume here that all functions \( v \mid_{\Gamma} = 0 \) are represented by (linear combinations of) the set \( \varphi_i(x, y), i \in \{1, \ldots, n\} \). We will use this assumption and skip the mathematical proof (see Strang and Fix [1973], Cumber et al. [1986] and Braess [1996] for instance for a proof). Further, in (4.6), we will make a choice for the functions \( \{\varphi_i(x, y)\} \) and hence they are known in the Finite Element calculations. It turns out that the choice of the basis-functions \( \{\varphi_i(x, y)\} \) influences the accuracy and speed of computations. The accuracy is a difficult subject, which we will treat without detail. The speed of computation is easier to deal with. Note that \( \hat{u} \mid_{\Gamma} = 0 \) due to \( \varphi_i(x, y) \mid_{\Gamma} = 0, i \in \{1, \ldots, n\} \).

(4.6) implies a set of linear equations of \( \{c_i\} \):

\[
\begin{align*}
c_1 \int_{\Omega} \nabla \varphi_1 \cdot \nabla \varphi_2 dA + c_2 \int_{\Omega} \nabla \varphi_1 \cdot \nabla \varphi_2 dA + \cdots + c_n \int_{\Omega} \nabla \varphi_1 \cdot \nabla \varphi_n dA &= 0, \\
c_1 \int_{\Omega} \nabla \varphi_1 \cdot \nabla \varphi_2 dA + c_2 \int_{\Omega} \nabla \varphi_2 \cdot \nabla \varphi_2 dA + \cdots + c_n \int_{\Omega} \nabla \varphi_2 \cdot \nabla \varphi_n dA &= 0, \\
&\vdots \\
c_1 \int_{\Omega} \nabla \varphi_1 \cdot \nabla \varphi_n dA + c_2 \int_{\Omega} \nabla \varphi_2 \cdot \nabla \varphi_n dA + \cdots + c_n \int_{\Omega} \nabla \varphi_n \cdot \nabla \varphi_n dA &= 0,
\end{align*}
\]

(4.7)

The 'discretization' matrix here is referred to as the stiffness matrix, its elements are

\[
A_{ij} = \int_{\Omega} \nabla \varphi_i \cdot \nabla \varphi_j dA
\]

(4.8)

**Exercise 4.1.1.** Show that \( A \) is symmetric, i.e. \( a_{ij} = a_{ji} \).

For a fast solution of the system of linear equations, one would like \( A \) to be as sparse as possible (i.e. \( A \) should contain as many zeros as possible). When \( \{\varphi_i\} \) are orthogonal over the \( \nabla \), then

\[
\int_{\Omega} \nabla \varphi_i \cdot \nabla \varphi_j dA = 0 \quad \text{when } i \neq j.
\]

(4.9)
4.1 The principle of Galerkin’s method

This would be ideal: the $c_j$ then follows very easily:

$$ c_j = \frac{0}{\Omega} \nabla \varphi_j \nabla \varphi_j = 0, \quad (4.10) $$

(Of course when $\mathbf{A}$ is not singular the solution of the above system of equation is given by $c_i = 0$, we here only illustrate the working of the FEM)

In practice it is not always possible to choose a set of orthogonal basis-functions, but we try to choose a set that is almost orthogonal. This means that $\mathbf{A}$ consists of zeroes mainly (i.e. $\mathbf{A}$ is a sparse matrix). We will choose basis-functions $\{\varphi_i(x,y)\}$ that are piecewise linear. Suppose that the domain of computation is divided into a set of gridnodes, see below, with numbers for the unknowns (figure 4.1).

![Finite Element mesh diagram](image)

**Figure 4.1:** An example of a domain divided into a Finite Element mesh

Then, we will choose $\varphi_i(x,y)$ to be piecewise (bi-)linear, such that

$$ \varphi_i(x_j, y_j) = \begin{cases} 1, & \text{for } (x_j, y_j) = (x_i, y_i), \\ 0, & \text{for } (x_j, y_j) \neq (x_i, y_i). \end{cases} \quad (4.11) $$

The reason for this choice will be motivated by the use of a one-dimensional example. It is clear that for this case the integrals of,

$$ \int_{\Omega} \nabla \varphi_i \cdot \nabla \varphi_j dA, \quad (4.12) $$

only do not vanish when $i$ and $j$ are equal or when $i$ and $j$ are neighboring gridpoints, due to piecewise linearity of the basis-functions $\{\varphi_i\}$. This implies that the basis-functions have a compact support and hence, the stiffness-matrix will be sparse. First we motivate the choice of piecewise linear basis-functions by the use of a 1-Dimensional example.
4.2 Motivation of piecewise linear basis-functions

Given any function $u = u(x)$ and a division of gridnodes, see figure 4.2. For $[x_{i-1}, x_i]$, we approximate $u(x)$ by the use of linear interpolation:

$$\bar{u}(x) = u(x_{i-1}) + \frac{u(x_i) - u(x_{i-1})}{x_i - x_{i-1}} (x - x_{i-1}) \quad \text{for} \; x \in [x_{i-1}, x_i], \quad (4.13)$$

or with $u_i = u(x_i)$, we obtain (figure 4.3)

$$\bar{u}(x) = u_{i-1} + \frac{u_i - u_{i-1}}{x_i - x_{i-1}} (x - x_{i-1}),$$

$$= u_{i-1} \left\{ 1 + \frac{x_{i-1} - x}{x_i - x_{i-1}} \right\} + u_i \left\{ \frac{x - x_{i-1}}{x_i - x_{i-1}} \right\},$$

$$= u_{i-1} \frac{x_i - x}{x_i - x_{i-1}} + u_i \frac{x - x_{i-1}}{x_i - x_{i-1}} =: u_{i-1} l_{i-1}(x) + u_i l_i(x). \quad (4.14)$$

Hence

$$\bar{u}(x) = u_{i-1} l_{i-1}(x) + u_i l_i(x) \quad \text{for} \; x \in [x_{i-1}, x_i]. \quad (4.15)$$
4.3 Evaluation of a one-dimensional example

We do the same for \( x \in [x_i, x_{i+1}] \) (see Figure 4.3), to obtain:

\[
 u(x) = u_i l_i(x) + u_{i+1} l_{i+1}(x) \quad \text{on} \quad x \in [x_i, x_{i+1}].
\]  

(4.16)

Thereafter, we write for the approximation \( \bar{u}(x) \) of \( u(x) \):

\[
 \bar{u}(x) = \sum_{j=1}^{n} u_j \phi_j(x),
\]  

(4.17)

where

\[
 \phi_i(x) = \begin{cases} 
 \frac{x - x_{i-1}}{x_i - x_{i-1}}, & \text{for} \quad x \in [x_{i-1}, x_i], \\
 \frac{x - x_{i+1}}{x_i - x_{i+1}}, & \text{for} \quad x \in [x_i, x_{i+1}], \\
 0, & \text{for} \quad x \notin [x_{i-1}, x_{i+1}].
\end{cases}
\]  

(4.18)

Hence \( \phi_i(x) \) is piecewise linear, see Figure 4.4, where

\[
 \phi_i(x_j) = \begin{cases} 
 0, & i \neq j, \\
 1, & i = j.
\end{cases}
\]  

(4.19)

In the Finite Element Method we put in the functions \( \{\phi_i(x)\} \) and we determine \( \{u_i\} \). Thereafter, we obtain the solution.

4.3 Evaluation of a one-dimensional example

Now we treat procedures to approximate the integrals. For this set of basis-functions we solve:

Find \( u(x) \) such that \( u(0) = 0 \) and

\[
 - \int_{0}^{1} u'(x)v'(x)dx = \int_{0}^{1} f(x)v(x)dx, \quad \text{for all} \quad v(0) = 0.
\]  

(4.20)
Exercise 4.3.1. Find the differential equation for $u$ with all boundary conditions that corresponds to (4.20) with $u(0) = 0$.

The approximation $\hat{u}(x) = \sum_{j=1}^{n} u_j \phi_j(x)$ (where $\hat{u} \to u$ as $n \to \infty$ is assumed) leads to the following Finite Element formation: Find coefficient $\{u_i\}$ in $\hat{u}(x) = \sum_{j=1}^{n} u_j \phi_j(x)$, with

$\phi_i(x_j) = \begin{cases} 0, & i \neq j, \\ 1, & i = j, \end{cases}$

(4.21)

such that

$$- \int_0^1 \sum_{j=1}^{n} u_j \phi_j'(x) \phi_i'(x) dx = \int_0^1 f(x) \phi_i(x) dx \quad \forall i \in \{1, \ldots, n\}.$$ 

(4.22)

The computation of $\int_0^1 \phi_i'(x) \phi_j'(x) dx$ is simple, since $\{\phi_i\}$ are piecewise linear functions

$$\int_0^{x_i} \phi_i'(x) \phi_i'(x) dx = \int_{x_{i-1}}^{x_i} \frac{1}{x_{i-1} - x_i} \frac{1}{x_i - x_{i-1}} dx = - \frac{1}{x_i - x_{i-1}},$$

(4.23)

$$\int_0^{x_i} \phi_i'(x) \phi_i'(x) dx = \frac{1}{x_i - x_{i-1}} + \frac{1}{x_{i+1} - x_i},$$

(4.24)

$$\int_0^{x_i} \phi_i'(x) \phi_{i+2}'(x) dx = 0,$$

(4.25)

(why?). Of course, given any function $f(x)$ whose product with $\phi_i(x)$ is integrable, the integral $\int_0^1 f(x) \phi_i(x) dx$ may be evaluated. For some cases it is even possible to find an anti-derivative for $\phi_i(x) f(x)$. However, for many cases this isn’t possible. We then evaluate the integral numerically. Standard FE-
packages use the Newton-Cotes formulas, which read as

\[
\int_{0}^{1} f(x) \phi_i(x) \, dx = f(x_i) \int_{0}^{1} \phi_i(x) \, dx \\
= f(x_i) \int_{x_{i-1}}^{x_{i+1}} \phi_i(x) \, dx \\
= f(x_i) \frac{1}{2} (x_{i+1} - x_{i-1}) \quad \text{for } i \in \{1, 2, \ldots, n-1\}. \tag{4.26}
\]

For \( i = n \) we use

\[
\int_{0}^{1} f(x) \phi_n(x) \, dx = \int_{x_{n-1}}^{x_n} f(x_n) \phi_n(x) \, dx = f(x_n) \frac{1}{2} (x_n - x_{n-1}), \tag{4.27}
\]

or better

\[
\frac{1}{2} f \left( \frac{x_n + x_{n-1}}{2} \right) (x_n - x_{n-1}). \tag{4.28}
\]

Evaluation of all integrals in (4.22) gives a system of linear equations in \( \{u_i\} \).
We also know that \( u'(1) = 0 \) results from (4.20) as a natural (BC). This condition should also follow from (4.22):

\[
\int_{0}^{1} \phi'_{n-1} \phi'_n \, dx = - \frac{1}{x_n - x_{n-1}}, \tag{4.29}
\]

\[
\int_{0}^{1} \phi'_n \phi'_n \, dx = \frac{1}{x_n - x_{n-1}}, \tag{4.30}
\]

\[
\Rightarrow - \frac{1}{x_n - x_{n-1}} u_{n-1} + \frac{1}{x_n - x_{n-1}} u_n = f(\bar{x}_n) \frac{1}{2} (x_n - x_{n-1}), \tag{4.31}
\]

where \( \bar{x}_n = \frac{x_n + x_{n-1}}{2} \) can be chosen (or just \( \bar{x}_n = x_n \)). Further \( f(\bar{x}_n) = u''(\bar{x}_n) \) (see exercise 4.3.1). Implies

\[
\frac{u_n - u_{n-1}}{x_n - x_{n-1}} = \frac{1}{2} f(\bar{x}_n)(x_n - x_{n-1}) \\
= \frac{1}{2} u''(\bar{x}_n)(x_n - x_{n-1}). \tag{4.32}
\]

This implies that \( \frac{u_n - u_{n-1}}{x_n - x_{n-1}} \to 0 \) as \( n \to \infty \) \( (x_n - x_{n-1} \to 0) \). Hence the natural (BC) is recovered.

### 4.4 Ritz’ method of Finite Elements by a simple example

Suppose that we have the following minimization problem:

\[
\min_{u \in U} J(u), \quad J(u) = \frac{1}{2} \int_{0}^{1} \left( \frac{du}{dx} \right)^2 \, dx. \tag{4.33}
\]
4.5 The treatment of a non-homogeneous Dirichlet BC

where $U := \{u \text{ smooth: } u(0) = 0\}$. Then this problem corresponds to the solution of

$$\begin{align*}
\frac{d^2 u}{dx^2} &= 0 \text{ on } (0, 1), \\
u(0) &= 0, \\
\frac{du}{dx}(1) &= 0.
\end{align*}$$

(4.34)

Note that the solution is given by $u = 0$. Now we use Ritz’ FEM to solve the minimazation problem. Let the approximate solution be given by (figure 4.4)

$$\hat{u}_n(x) = \sum_{j=1}^{n} u_j \phi_j(x),$$

(4.35)

where we assume that $\hat{u}_n(x) \to u(x)$ on $x \in [0, 1]$ as $n \to \infty$. Then we look for constants \( \{u_1, u_2, \ldots, u_n\} \) such that $J(u)$ is minimal. In other words:

$$\frac{\partial}{\partial u_i} J(u) = 0 \quad \forall i \in \{1, 2, 3, \ldots, n\}. \quad (4.36)$$

Substitution of $\sum_{j=1}^{n} u_j \phi_j(x) = \bar{u}_n(x)$ into $J(u)$ gives:

$$J(u) = \frac{1}{2} \int_0^1 \sum_{i=1}^{n} u_i \phi_i'(x) \cdot \sum_{j=1}^{n} u_j \phi_j'(x) \, dx. \quad (4.37)$$

Hence

$$\frac{\partial}{\partial u_i} J(u) = 0 \Rightarrow \sum_{j=1}^{n} \int_0^1 u_j \phi_j'(x) \phi_i'(x) \, dx = 0 \quad \forall i \in \{1, \ldots, n\}. \quad (4.38)$$

This gives exactly the same equations as in (4.22). Note the similarity with the result obtained by the use of Galerkin’s method. Since Galerkin’s method is applicable for more general cases we do not treat this method further.

4.5 The treatment of a non-homogeneous Dirichlet BC

Suppose that we have to solve the following variational problem:

Find $u(x)$, subject to $u(0) = u_0$, such that

$$-\int_0^1 u'(x)v'(x) \, dx = \int_0^1 f(x)v(x) \, dx \quad \forall v(0) = 0. \quad (4.39)$$

Now the essential condition is non-zero (when $u_0 \neq 0$), where $u_0$ is given. The treatment is similar to the case where $u_0 = 0$, but now we set

$$\bar{u}_n(x) = \sum_{j=0}^{n} u_j \phi_j(x) = u_0 \phi_0(x) + \sum_{j=1}^{n} u_j \phi_j(x). \quad (4.40)$$
4.5 The treatment of a non-homogeneous Dirichlet BC

Figure 4.5: The piecewise linear functions $\phi_i(x)$ and $\phi_0(x)$

Note that $u_0$ is from the Dirichlet BC. Again, we use piecewise linear basis-functions $\{\phi_i(x)\}$, where

$$
\phi_i(x_j) = \begin{cases} 
1 & i = j, \\
0 & i \neq j,
\end{cases}
$$

and $\phi_i(x)$ is piecewise linear. (4.41)

For $\phi_0(x)$ we have: $\phi_0(x_0) = \phi_0(0) = 1$. For the functions $v(x)$ we set $\phi_i(x)$ where $i \in \{1, \ldots, n\}$ (note that $\phi_i(0) = 0$ since $v(0) = 0$). Then, we obtain the following problem:

Find $\{u_1, u_2, \ldots, u_n\}$ such that

$$
- \int_0^1 \sum_{j=0}^n u_j \phi_j'(x) \phi_i'(x) dx = \int_0^1 f(x) \phi_i(x) dx,
$$

for all $\{\phi_i(x)\}_{i=1}^n$, i.e. $\forall i \in \{1, \ldots, n\}$.

For the functions $\phi_i(x)$ we use the sketches from figure 4.5. For $\phi_1(x)$, where $j = 1$, follows (where $u_0$ is prescribed):

$$
- \int_0^1 u_0 \phi_0'(x) \phi_1'(x) dx - \int_0^1 u_1 \phi_1'(x) \phi_1'(x) dx - \int_0^1 u_2 \phi_2'(x) \phi_1'(x) dx = \int_0^1 f(x) \phi_1(x) dx,
$$

for $j = 1, 2, \ldots, n$.

The integrals can also be evaluated for the other values of $j$. The same equations follow as in the example where $u_0 = 0$. Again, we will have that $\bar{u}_n'(1) \to 0$ as $n \to \infty$. For more dimensional problems, the same occurs: Suppose that we search $u$, subject to $u |_{\Gamma_2} = g$, such that

$$
\int_\Omega \nabla u \cdot \nabla v dA = 0 \quad \forall v |_{\Gamma} = 0.
$$
Let \( n \) be the number of gridnodes that are inside \( \Omega \) (figure 4.5) or on boundary \( \Gamma_2 \), (but not on \( \Gamma_1 \)), and let \( \{n+1, \ldots, n+m\} \) be the gridnodes on boundary \( \Gamma_1 \), then we set

\[
\begin{align*}
\quad u(x, y) & \approx \sum_{j=1}^{n} u_j \phi_j(x, y) + \sum_{j=n+1}^{n+m} g(x_j, y_j) \phi_j(x, y) .
\end{align*}
\]  

(4.46)

Here we take

\[
\phi_j(x_i, y_i) = \begin{cases} 
1, & \text{for } j = i, \\
0, & \text{for } j \neq i,
\end{cases}
\]  

(4.47)

where \( \phi_j(x, y) \) are taken to be piecewise linear. Expression (4.46) is substituted into the weak form to obtain a system of linear equations.

### 4.6 A time-dependent example

Consider the following problem:

\[
\begin{align*}
\quad \frac{\partial c}{\partial t} &= \frac{\partial^2 c}{\partial x^2} \quad \text{on } x \in (0, 1), t > 0, \\
\quad c(0, t) &= 1 \quad t > 0, \\
\quad \frac{\partial c}{\partial x}(1, t) &= 0 \quad t > 0, \\
\quad c(x, 0) &= 0 \quad \text{on } x \in (0, 1).
\end{align*}
\]  

(4.48)

First we search a weak form for (4.48): Search \( c, c(0, t) = 1, c(x, 0) = 0, \)

\[
\int_0^1 \frac{\partial c}{\partial t} v dx = \int_0^1 \frac{\partial^2 c}{\partial x^2} v dx \quad \forall v(0, t) = 0.
\]  

(4.49)

To reduce the order of the derivative, we integrate by parts:

\[
\int_0^1 \frac{\partial c}{\partial t} v dx = \left[ \frac{\partial c}{\partial x} v \right]_0^1 - \int_0^1 \frac{\partial c}{\partial x} \frac{\partial v}{\partial x} dx = -\int_0^1 \frac{\partial c}{\partial x} \frac{\partial v}{\partial x} dx,
\]  

(4.50)

since \( v(0, t) = 0 \) and \( \frac{\partial c}{\partial x}(1, t) = 0 \). Then, we obtain the following weak form

\[
\begin{align*}
\quad \text{Find } c, \text{ subject to } c(0, t) &= 1, c(x, 0) = 0, \text{ such that} \\
\quad \int_0^1 \frac{\partial c}{\partial t} v dx &= -\int_0^1 \frac{\partial c}{\partial x} \frac{\partial v}{\partial x} dx \quad \forall v(0) = 0.
\end{align*}
\]  

(4.51)

We solve (4.51) by use of the Galerkin FEM. Again, we use piecewise linear basis-functions \( \{\phi_i(x)\}_{i=0}^n \) as before on the \( n \) gridnodes, with

\[
\phi_i(x_j) = \begin{cases} 
1, & j = i, \\
0, & j \neq i.
\end{cases}
\]  

(4.52)
Then, we approximate \( c(x, t) \) by
\[
\bar{c}_n(x, t) = \sum_{j=0}^{n} c_j(t) \phi_j(x) = \phi_0(x) + \sum_{j=1}^{n} \phi_j(x) c_j(t).
\] (4.53)

Note that \( \{c_j\} \) should be functions of \( x \) and \( t \) since \( c(x, t) \) is a function of \( t \) and \( \phi_j(x) \) is a function of \( x \) only. Substitution into (4.51) gives
\[
\begin{cases}
\text{Find } \{c_i(t)\}_{i=1}^{n} \\
\int_{0}^{1} \sum_{j=1}^{n} c'_j(t) \phi_j(x) \phi_i(x) dx \\
= - \int_{0}^{1} \left \{ \phi'_0(x) + \sum_{j=1}^{n} \phi'_j(x) c_j(t) \right \} \phi'_i(x) dx \quad \forall i \in \{1, \ldots, n\}.
\end{cases}
\] (4.54)

The above problem represents a system of linear ordinary differential equations. Note that \( c'_j(t) \approx \frac{c_j(t + \Delta t) - c_j(t)}{\Delta t} \). We will deal with the solution of the time-dependent problem in the next chapter.

### 4.7 The principle of element matrices and vectors

We briefly treat the concept of element matrices, which is not a mathematical feature but only a convenient programming trick. This trick is used in most of the implementations of Finite Elements software.

For the treatment we treat a simple one-dimensional example:
\[
\begin{cases}
-u'' = f, \\
u(0) = u_0, \\
u'(1) = 0.
\end{cases}
\] (4.55)

A weak form is then obtained by
\[
-u''v = fv \Leftrightarrow -\int_{0}^{1} u''v dx = \int_{0}^{1} f v dx
\]
\[
\Leftrightarrow \left[ -u' \right]_{0}^{1} + \int_{0}^{1} u'v' dx = \int_{0}^{1} f v dx
\]
\[
\Leftrightarrow \int_{0}^{1} u'v' dx = \int_{0}^{1} f v dx.
\] (4.56)

We use Galerkin’s method on \( n \) gridpoints for the unknowns:
\[
u(x) = \sum_{j=1}^{n} u_j \phi_j(x) + u_0 \phi_0(x).
\] (4.57)
Substitution into the weak form, with
\[
\phi_j(x_i) = \begin{cases} 
1, & j = i, \\
0, & j \neq i,
\end{cases}
\quad (4.58)
\]
piecewise linear, gives
\[
\int_0^1 \sum_{j=0}^n u_j \phi_j'(x) \phi_i'(x) \, dx = \int_0^1 f(x) \phi_i(x) \, dx, \quad i \in \{1, \ldots, n\},
\quad (4.59)
\]
or
\[
\sum_{j=0}^n u_j \int_0^1 \phi_j'(x) \phi_i'(x) \, dx = \int_0^1 f(x) \phi_i(x) \, dx - u_0 \int_0^1 \phi_0'(x) \phi_i'(x) \, dx \quad i \in \{1, \ldots, n\}.
\quad (4.60)
\]
In other words:
\[
\begin{align*}
&u_1 \int_0^1 \phi_1' \phi_1' \, dx + u_2 \int_0^1 \phi_2' \phi_1' \, dx + \cdots + u_n \int_0^1 \phi_n' \phi_1' \, dx = \int_0^1 f \phi_1 \, dx - u_0 \int_0^1 \phi_0' \phi_1' \, dx, \\
&u_1 \int_0^1 \phi_1' \phi_2' \, dx + u_2 \int_0^1 \phi_2' \phi_2' \, dx + \cdots + u_n \int_0^1 \phi_n' \phi_2' \, dx = \int_0^1 f \phi_2 \, dx - u_0 \int_0^1 \phi_0' \phi_2' \, dx,
\end{align*}
\]
\vdots
For simplicity we take \(n=3\); then,
\[
\begin{align*}
&u_1 \int_0^1 \phi_1' \phi_1' \, dx + u_2 \int_0^1 \phi_2' \phi_1' \, dx + u_3 \int_0^1 \phi_3' \phi_1' \, dx = \int_0^1 f \phi_1 \, dx - u_0 \int_0^1 \phi_0' \phi_1' \, dx, \\
&u_1 \int_0^1 \phi_1' \phi_2' \, dx + u_2 \int_0^1 \phi_2' \phi_2' \, dx + u_3 \int_0^1 \phi_3' \phi_2' \, dx = \int_0^1 f \phi_2 \, dx - u_0 \int_0^1 \phi_0' \phi_2' \, dx, \\
&u_1 \int_0^1 \phi_1' \phi_3' \, dx + u_2 \int_0^1 \phi_2' \phi_3' \, dx + u_3 \int_0^1 \phi_3' \phi_3' \, dx = \int_0^1 f \phi_3 \, dx - u_0 \int_0^1 \phi_0' \phi_3' \, dx.
\end{align*}
\quad (4.61)
\]
We take piecewise linear basis-functions. Note that we have \(\int_0^1 \phi_1' \phi_3' \, dx = 0\)
(the basis-functions are called nearly orthogonal) for instance and that only contributions from neighbouring elements are non-zero. Next, we will introduce the concept element as an interval between adjacent meshpoint: \(e_i := [x_{i-1}, x_i]\).
Hence, using the fact that the basis-functions are nearly orthogonal, we write the above equations as follows:

\[
\begin{align*}
&u_1 \int_{e_1 \cup e_2} \phi_1' \phi_1' dx + u_2 \int_{e_2} \phi_1' \phi_2' dx = \int_{e_1 \cup e_2} f \phi_1 dx - u_0 \int_{e_1} \phi_0' \phi_1' dx, \\
u_1 \int_{e_2} \phi_2' \phi_1' dx + u_2 \int_{e_2 \cup e_3} \phi_2' \phi_2' dx + u_3 \int_{e_3} \phi_2' \phi_3' dx = \int_{e_2 \cup e_3} f \phi_2 dx, \\
u_2 \int_{e_3} \phi_2' \phi_3' dx + u_3 \int_{e_3} \phi_2' \phi_3' dx = \int_{e_3} f \phi_3 dx. \quad (4.62)
\end{align*}
\]

Note further that

\[
\begin{align*}
\left\{ \begin{array}{l}
\int_{e_i \cup e_i+1} \phi_i' \phi_i' dx = \int_{e_i} \phi_i' \phi_i' dx + \int_{e_{i+1}} \phi_i' \phi_i' dx \\
\int_{e_i \cup e_i+1} f \phi_i dx = \int_{e_i} f \phi_i dx + \int_{e_{i+1}} f \phi_i dx
\end{array} \right. \quad (4.63)
\]

Now we show a computer procedure to generate the stiffness matrix (discretization matrix) by use of element-matrices: start with,

\[
A_0 = \begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 
\end{pmatrix}, \quad (4.64)
\]

we finish with

\[
A = \begin{pmatrix}
\int_{e_1} \phi_1' \phi_1' + \int_{e_2} \phi_1' \phi_1' & \int_{e_1} \phi_1' \phi_2' & 0 \\
\int_{e_2} \phi_2' \phi_1' & \int_{e_2} \phi_2' \phi_2' + \int_{e_3} \phi_2' \phi_2' & \int_{e_3} \phi_2' \phi_2' \\
0 & \int_{e_3} \phi_2' \phi_3' & \int_{e_3} \phi_3' \phi_3'
\end{pmatrix}. \quad (4.65)
\]

Now we introduce the concept of element-matrices:

\[
s_{e_i} = \begin{bmatrix}
\int_{e_i} \phi_i' \phi_i' \phi_{i-1}' dx & \int_{e_i} \phi_i' \phi_i' \phi_{i-1}' dx \\
\int_{e_i} \phi_i' \phi_i' \phi_{i-1}' dx & \int_{e_i} \phi_i' \phi_i' \phi_{i-1}' dx
\end{bmatrix}. \quad (4.66)
\]

The known quantities do not occur in the stiffness matrix, hence for \( e_1 \) we obtain:

\[
s_{e_1} = \int_{e_1} \phi_1' \phi_1' dx, \quad (4.67)
\]
and for \( e_2 \) and \( e_3 \):

\[
\mathbf{s}_{e_2} = \begin{bmatrix}
\int_{e_2} \phi'_1 \phi'_1 \, dx & \int_{e_2} \phi'_1 \phi'_2 \, dx \\
\int_{e_2} \phi'_2 \phi'_1 \, dx & \int_{e_2} \phi'_2 \phi'_2 \, dx
\end{bmatrix}, \quad (4.68)
\]

\[
\mathbf{s}_{e_3} = \begin{bmatrix}
\int_{e_3} \phi'_2 \phi'_1 \, dx & \int_{e_3} \phi'_2 \phi'_2 \, dx \\
\int_{e_3} \phi'_3 \phi'_1 \, dx & \int_{e_3} \phi'_3 \phi'_3 \, dx
\end{bmatrix}. \quad (4.69)
\]

The matrices \( s_{e_1}, s_{e_2} \) and \( s_{e_3} \) are the element-matrices of elements \( e_1, e_2 \) and \( e_3 \). We will put these matrices into \( A_0 \) and add them to obtain \( A \). The position of the element-matrices is such that \( s_{e_1}[1, 1] = A_0[1, 1] \), hence:

\[
A_1 = \begin{pmatrix}
\int_{e_1} \phi'_1 \phi'_1 \, dx & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}. \quad (4.70)
\]

Subsequently we add \( s_{e_2} \) to \( A_1 \) such that \( s_{e_2}[2, 2] = A_0[2, 2] \):

\[
A_2 = \begin{pmatrix}
\int_{e_1} \phi'_1 \phi'_1 \, dx & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix} + \begin{pmatrix}
\int_{e_2} \phi'_1 \phi'_1 \, dx & \int_{e_2} \phi'_1 \phi'_2 \, dx & 0 \\
\int_{e_2} \phi'_2 \phi'_1 \, dx & \int_{e_2} \phi'_2 \phi'_2 \, dx & 0 \\
0 & 0 & 0
\end{pmatrix}
\]

\[
= \begin{pmatrix}
\int_{e_1} \phi'_1 \phi'_1 \, dx + \int_{e_2} \phi'_1 \phi'_1 \, dx & \int_{e_2} \phi'_1 \phi'_2 \, dx & 0 \\
\int_{e_2} \phi'_2 \phi'_1 \, dx & \int_{e_2} \phi'_2 \phi'_2 \, dx & 0 \\
0 & 0 & 0
\end{pmatrix}. \quad (4.71)
\]

Subsequently we add \( s_{e_3} \) to \( A_2 \) such that \( s_{e_3}[2, 2] = A_0[3, 3] \) to obtain \( A_3 \):

\[
A_3 = \begin{pmatrix}
\int_{e_1} \phi'_1 \phi'_1 \, dx + \int_{e_2} \phi'_1 \phi'_1 \, dx & \int_{e_2} \phi'_1 \phi'_2 \, dx & 0 \\
\int_{e_2} \phi'_2 \phi'_1 \, dx & \int_{e_2} \phi'_2 \phi'_2 \, dx & 0 \\
0 & 0 & 0
\end{pmatrix} + \begin{pmatrix}
0 & 0 & 0 \\
0 & \int_{e_3} \phi'_2 \phi'_2 \, dx & \int_{e_3} \phi'_2 \phi'_3 \, dx \\
0 & \int_{e_3} \phi'_3 \phi'_2 \, dx & \int_{e_3} \phi'_3 \phi'_3 \, dx
\end{pmatrix}
\]

\[
= \begin{pmatrix}
\int_{e_1} \phi'_1 \phi'_1 \, dx + \int_{e_2} \phi'_1 \phi'_1 \, dx + \int_{e_3} \phi'_1 \phi'_1 \, dx & \int_{e_2} \phi'_1 \phi'_2 \, dx & 0 \\
\int_{e_2} \phi'_2 \phi'_1 \, dx + \int_{e_3} \phi'_2 \phi'_2 \, dx & \int_{e_2} \phi'_2 \phi'_2 \, dx + \int_{e_3} \phi'_2 \phi'_3 \, dx & \int_{e_3} \phi'_2 \phi'_3 \, dx \\
0 & \int_{e_3} \phi'_3 \phi'_2 \, dx & \int_{e_3} \phi'_3 \phi'_3 \, dx
\end{pmatrix}. \quad (4.72)
\]

Now, the stiffness-matrix has been built. The same principle is carried out with the element-vectors to generate the right-hand side vector. The same principle gives

\[
\mathbf{f}_{e_i} = \begin{bmatrix}
\int_{e_i} f(x) \phi_{i-1}(x) \, dx \\
\int_{e_i} f(x) \phi_i(x) \, dx
\end{bmatrix}. \quad (4.73)
\]
This is called the element-vector of element \( e_i \),

\[
f_{e_1} = \begin{bmatrix} \int_{e_1} f(x)\phi_1(x)\,dx \end{bmatrix} \Rightarrow f_1 = \begin{pmatrix} \int_{e_1} f(x)\phi_1(x)\,dx \\ 0 \\ 0 \end{pmatrix}. \tag{4.74}
\]

subsequently

\[
f_{e_2} = \begin{bmatrix} \int_{e_2} f(x)\phi_1(x)\,dx \\ \int_{e_2} f(x)\phi_2(x)\,dx \end{bmatrix}
\Rightarrow f_2 = f_1 + \begin{pmatrix} \int_{e_2} f(x)\phi_1(x)\,dx \\ \int_{e_2} f(x)\phi_2(x)\,dx \end{pmatrix}
\tag{4.75}
\]

\[
= \begin{pmatrix} \int_{e_1} f(x)\phi_1(x)\,dx + \int_{e_2} f(x)\phi_1(x)\,dx \\ \int_{e_2} f(x)\phi_2(x)\,dx \end{pmatrix}. \tag{4.76}
\]

Subsequently

\[
f_{e_3} = \begin{bmatrix} \int_{e_3} f(x)\phi_2(x)\,dx \\ \int_{e_3} f(x)\phi_3(x)\,dx \end{bmatrix} \Rightarrow f_3 = \begin{pmatrix} \int_{e_1 \cup e_2} f(x)\phi_1(x)\,dx \\ \int_{e_2 \cup e_3} f(x)\phi_2(x)\,dx \\ \int_{e_3} f(x)\phi_3(x)\,dx \end{pmatrix}. \tag{4.78}
\]

Now we have the right-hand-side if we take the essential condition into account for element \( e_1 \): \(-u_0 \int_{e_1} \phi_0'\phi'_1\,dx\) at \( f'_3 \), to obtain:

\[
f = \begin{pmatrix} \int_{e_1 \cup e_2} f(x)\phi_1(x)\,dx - u_0 \int_{e_1} \phi_0'\phi'_1\,dx \\ \int_{e_2 \cup e_3} f(x)\phi_2(x)\,dx \\ \int_{e_3} f(x)\phi_3(x)\,dx \end{pmatrix}. \tag{4.79}
\]

The principle of element-matrices and vectors should be seen as a convenient programming trick. To tackle non-standard Finite Element problems by the use of Finite Element software packages, it is sufficient to create the element-matrices and element-vectors.

### 4.8 Numerical integration

Consider the following weak form
Find $u(x)$, subject to $u(0) = u_0$, such that

\[ -\int_0^1 u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx \quad \forall v(0) = 0. \]  

(4.80)

Sometimes, the right-hand side contains a complicated integrand that cannot be integrated in terms of an elementary function. Since one uses the concept of element vectors and element matrices, one has to integrate the functions over an element. Many numerical integration methods are available, such as the midpoint Rule, the Trapezoidal Rule, Simpson Rule and so on. A very common numerical integration rule based on the expression of a function in terms of a linear combination of basis-function is the Newton-Cotes integration Rule.

In this section, we will consider Newton-Cotes integration with piecewise linear basis-functions. This rule is based on the following: Consider a function $g(x)$ to be integrated over an element $e_i := [x_{i-1}, x_i]$:

\[ \int_{e_i} g(x)dx, \]  

(4.81)

then we express $g(x)$ as a linear combination of basis-functions with the characteristics as mentioned before on the nodes $x_{i-1}$ and $x_i$:

\[ g(x) = g(x_{i-1})\phi_{i-1}(x) + g(x_i)\phi_i(x). \]  

(4.82)

Then the integration over the interval zero-one, gives:

\[ \int_{e_i} g(x)dx = g(x_{i-1}) \int_{e_i} \phi_{i-1}(x)dx + g(x_i) \int_{e_i} \phi_i(x)dx = \frac{g(x_{i-1}) + g(x_i)}{2}(x_i - x_{i-1}). \]  

(4.83)

Note that the above integration formula represents the Trapezoidal Rule for numerical integration. The same is done with the right-hand of the above weak form:

\[ \int_{e_i} f(x)\phi_k(x)dx = \sum_{j=i-1}^{i} f(x_j)\phi_k(x_j) \int_{e_i} \phi_j(x)dx = \frac{1}{2}f(x_k)(x_{k+1} - x_{k-1}), \]  

(4.84)

for $k \in \{i - 1, i\}$, note that $\phi_k(x_i) = 1$ if $i = k$ and else $\phi_k(x_i) = 0$. This integration rule is easily extended to more dimensional problems. This is beyond the scope of the present course. As an alternative, Gauss integration formulas are used. This is not treated in this course.

## 4.9 Error considerations

This section is not necessary to understand the finite element method. It is intended for the interested reader. The treatment of the error of the finite element solution will not be mathematical, but it will give an idea of what the
4.9 Error considerations

Error is and what issues are important for the derivation of the error. Suppose that we solve the equation in the following weak form of Poisson’s equation

Find \( u \in H^1(\Omega) \), subject to \( u = g \) on \( \partial \Omega \), such that

\[
\int_{\Omega} \nabla u \cdot \nabla \phi d\Omega = \int_{\Omega} \phi f d\Omega, \text{ for all } \phi \in H^1(\Omega).
\]

(4.85)

To solve this problem using Galerkin’s method, we approximate the solution by

\[
u(x) \approx u_h(x) = \sum_{j=1}^{n} u_j \phi_j(x),\]

where \( \phi_j(x) = 0 \) on \( \partial \Omega \). Here \( u(x) \) represents the exact solution and \( u_j \) represents the approximate solution at the nodal points. For the approximate solution \( u_h \), we have

Find \( u_h \in H^1_h(\Omega) \), subject to \( u_h = g \) on \( \partial \Omega \), such that

\[
\int_{\Omega} \nabla u_h \cdot \nabla \phi_h d\Omega = \int_{\Omega} \phi_h f d\Omega, \text{ for all } \phi_h \in H^1_h(\Omega),
\]

(4.86)

where \( H^1_h(\Omega) \) represents the set of solutions for the approximate solution \( u_h \).

Note that \( H^1_h(\Omega) \subset H^1(\Omega) \). Let \( \tilde{u}(x) \) be the approximate solution with the exact values of the solution \( u(x) \), that is

\[
\tilde{u}(x) = \sum_{j=1}^{n} u(x_j) \phi_j(x).
\]

Note that the difference between the exact solution \( u(x) \) and the above solution \( \tilde{u}(x) \) is only determined by the interpolation method that is used. Then it can be proved (see Braess [1996] for instance) that

\[
\int_{\Omega} ||\nabla (u - u_h)||^2 d\Omega \leq \int_{\Omega} ||\nabla (u - \tilde{u})||^2 d\Omega.
\]

(4.87)

The above integrals represent errors in the energy-norm. The left-hand side of the above inequality gives the total error of the finite element solution with respect to the exact solution in the so-called energy norm. This total error basically has two sources:

1. A finite set of basis functions (based on the finite number of meshpoints) is chosen, and hence the summation only concerns a finite number of terms;

2. Using interpolation functions for the basis functions, this gives an interpolation error, which depends on the order of the interpolation functions \( \phi_j(x) \).

The right-hand side of the above inequality only concerns the interpolation error. The above inequality (4.87) is very convenient since it says in the energy norm that the total error is bounded from above by the interpolation error. This lastmentioned error depends on the polynomial order. If linear elements are used, then it can be demonstrated that the energy norm of the interpolational error is of order \( O(h) \) where \( h \) represents the largest side of the element. The actual error \( u - u_h \) is one order higher, that is \( O(h^2) \). The reader should realize
that many statements in this section have been made while omitting some subtle mathematical issues.

As a rule of thumb, we use that if the interpolational error is of order $O(h^p)$, then the order of the error, that is the difference between the finite element solution and the exact solution is of order $O(h^{p+1})$. It should be mentioned that this fact only holds if the elements are not degenerate (the vertices of the triangle are not located on a line in the case of triangular elements).
Time dependent problems: numerical methods

In many cases the mathematical models are "time-dependent", by which we mean that their solution depends on time. Typical examples are

\[
\begin{align*}
\frac{\partial c}{\partial t} + \vec{q} \cdot \vec{\nabla} c &= \Delta c \quad &\text{(convection-diffusion)} \\
\frac{\partial c}{\partial t} - \Delta c &= 0 \quad &\text{(diffusion)} \\
\frac{\partial^2 c}{\partial t^2} &= \Delta c \quad &\text{(wave transmission)} \\
\frac{\partial c}{\partial t} + \frac{\partial}{\partial x} f(c) &= 0 \quad &\text{(Buckley-Leverett in 2-plane flow)}
\end{align*}
\]

In this chapter we will treat some time-integration methods in relation to Finite Differences and Finite Elements. Important subjects will be consistency (convergence), accuracy and stability.

5.1 Time-integration methods

First we consider an ODE (Ordinary Differential Equation) to illustrate some time-integration methods. Subsequently, we will apply these methods to PDE’s. Consider the following problem:

\[
\begin{align*}
y(0) &= y_0 & t &= 0, \\
y'(t) &= f(y,t) & t &> 0.
\end{align*}
\]

Let \( T_{\text{end}} \) be the end-time of the numerical simulation \( h = \frac{T_{\text{end}}}{m}, m= \text{number of timesteps} \), then we introduce the notation \( y^n = y(t_n) = y(nh) \). We will formulate some classical time-integration methods.
5.2 Accuracy of Time-integration methods

[1] Euler’s forward time integration method (explicit):
\[ y^{n+1} = y^n + hf(y^n, t_n) \quad n \geq 0 \quad n \in \{0, \ldots, m-1\} \quad (5.6) \]

[2] Heun’s (or improved Euler’s) time integration method
\[
\begin{cases}
    y^{n+1} = y^n + hf(y^n, t_n) & \text{(predictor)} \\
    \bar{y}^{n+1} = y^n + \frac{h}{2} \left[ f(y^n, t_n) + f(\bar{y}^{n+1}, t_{n+1}) \right] & \text{(corrector)}
\end{cases} \quad (5.7)
\]

[3] Euler’s backward time integration method (implicit)
\[ y^{n+1} = y^n + hf(y^{n+1}, t^{n+1}) \quad (5.8) \]
(Note that we have to solve a non-linear equation whenever \( f \) is non-linear in \( y \))

\[ y^{n+1} = y^n + \frac{h}{2} \left[ f(y^n, t_n) + f(y^{n+1}, t_{n+1}) \right] \quad (5.9) \]
(Here again a non-linear problem has to be solved when \( f \) is nonlinear in \( y \))

We can sum up many more methods for time-integration (see for instance Burden and Faires [2001]) methods. These four methods are the most common and we will analyse them. We also note that the modified Euler method falls within the class of the so-called multi-step methods (due to Runge-Kutta). These methods can be adjusted such that always the desired accuracy can be obtained.

5.2 Accuracy of Time-integration methods

We will consider local truncation error. Given, again
\[
\begin{cases}
    y(0) = y_0 & t = 0, \\
    y'(t) = f(y, t) & t > 0,
\end{cases} \quad (5.10)
\]
then
\[ y(t + h) = y(t) + hy'(t) + \frac{h^2}{2} y''(t) + \frac{h^3}{6} y'''(t) + O(h^4). \quad (5.11) \]

Further, let \( u \) be the approximated solution and let \( y(t^n) = u^n \), then
\[ u(t + h) = u^{n+1} = y^n + hf(y^n, t^n) \quad \text{(Euler)}. \quad (5.12) \]

Since \( f(y^n, t^n) = y'(t) \). Then it follows from (5.11) and (5.12), that
\[ y(t + h) - u^{n+1} = \frac{h^2}{2} y''(t) = O(h^2). \quad (5.13) \]
Herewith, the local truncation error is given by:

\[
\tau_{n+1}(h) = \frac{y(t+h) - u^{n+1}}{h} = \frac{h}{2} y''(t) = \mathcal{O}(h). \tag{5.14}
\]

Hence the order of the local truncation error is \( h \) for the Euler-method. This is similar for the implicit Euler method. We do the same analysis for the accuracy of the modified Euler method. Consider:

\[
\begin{align*}
\bar{u}^{n+1} &= u^n + hf(u^n, t^n), \\
u^{n+1} &= u^n + \frac{h}{2}(f(u^n, t^n) + f(\bar{u}^{n+1}, t^{n+1})).
\end{align*} \tag{5.15}
\]

Then, assuming \( u^n = y^n \), gives

\[
u^{n+1} = y^n + \frac{h}{2}\left\{f(y^n, t^n) + f(y^n, t^n) + hf(y^n, t^n) \frac{\partial f}{\partial y}(y^n, t^n) + h \frac{\partial f}{\partial t}(y^n, t^n)\right\}, \tag{5.16}
\]

this follows from a Taylor expansion. Further:

\[
\frac{d^2 y}{dt^2} = \frac{d}{dt} f(y(t), t) = \frac{\partial f}{\partial y} y' + \frac{\partial f}{\partial t} = \frac{\partial f}{\partial y} + \frac{\partial f}{\partial t}. \tag{5.17}
\]

From this follows

\[
u^{n+1} = y^n + hf(y^n, t^n) + \frac{h^2}{2} \frac{d^2 y}{dt^2}(t^n) + \mathcal{O}(h^3). \tag{5.18}
\]

Hence

\[
u^{n+1} - y^{n+1} = \mathcal{O}(h^3). \tag{5.19}
\]

Herewith, the local truncation error is given by:

\[
\tau_{n+1}(h) = \frac{y^{n+1} - u^{n+1}}{h} = \mathcal{O}(h^2). \tag{5.20}
\]

Hence the accuracy is of 2nd order. The truncation error is of 2nd order, when the Modified Euler method is used.

To look at the global error, we keep in mind that at each time-step an error of \( \mathcal{O}(h^3) \), i.e. Error = \( kh^3 \) (\( k \in \mathbb{R} \)), is made (for Modified Euler method). Since we integrate over \( n \) time-steps the accumulated error is approximately (Global error):

\[
k_1h^3n = k_1h^3T_{\text{end}} \frac{1}{h} = \bar{k}_1h^2. \tag{5.21}
\]

Similarly for Euler’s method (Global error):

\[
k_2h^2n = k_2h^2T_{\text{end}} \frac{1}{h} = \bar{k}_2h. \tag{5.22}
\]

For Crank-Nicholson’s method, it can be shown by use of a similar method that Truncation error \( \sim h^3 \) and Global error \( \sim h^2 \).
5.3 Time-integration of PDE’s

The time integration method for PDE’s is analogous to ODE’s. Suppose that the following problem is given:

\[
\begin{align*}
\frac{\partial u}{\partial t} &= \mathcal{L}(u), \\
\mathcal{L}(u) &= \Delta u \quad \text{(diffusion), or}, \\
\mathcal{L}(u) &= \Delta u - \bar{q} \cdot \nabla u \quad \text{(conv.-diff.)}, \\
\end{align*}
\]

and so on… (5.26)

The time discretization part \( \frac{\partial u}{\partial t} \) is done (for instance) by:

\[
\frac{\partial u}{\partial t} = \frac{u_{i,j}^{n+1} - u_{i,j}^n}{\Delta t},
\]

(5.27)

where \( u_{i,j}^n \) denotes \( u(x_i, y_j, t^n) \). Then we give the following example:

\[
\begin{align*}
\begin{cases}
    u_{i,j}^{n+1} = u_{i,j}^n + \Delta t \mathcal{L}(u^n) & \text{Forward (Explicit) Euler}, \\
    u_{i,j}^{n+1} = u_{i,j}^n + \Delta t \mathcal{L}(u^{n+1}) & \text{Backward (Implicit) Euler}, \\
    u_{i,j}^{n+1} = u_{i,j}^n + \frac{\Delta t}{2} \left\{ \mathcal{L}(u^{n+1}) + \mathcal{L}(u^n) \right\} & \text{Crank-Nicholson}, \\
    \bar{u}_{i,j}^{n+1} = u_{i,j}^n + \Delta t \mathcal{L}(u^n) & \text{Modified Euler (predictor-corrector)}, \\
    u_{i,j}^{n+1} = u_{i,j}^n + \frac{\Delta t}{2} \left\{ \mathcal{L}(u^n) + \mathcal{L}(\bar{u}^{n+1}) \right\}.
\end{cases}
\end{align*}
\]

(5.28)

The global errors are, respectively \( O(\Delta t) \), \( O(\Delta t) \), \( O(\Delta t^2) \) and \( O(\Delta t^2) \). This can be shown by the use of similar procedures as in section 5.2. Now we consider the application to a diffusion equation as our example. We will illustrate the use of Finite Differences and Finite Elements for a 1-dimensional diffusion problem and a 1-dimensional wave equation.

### 5.3.1 The Heat equation

We consider the discretization of the heat equation

\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad t > 0
\]

(5.29)

+ Dirichlet boundary conditions + initial condition.

[1] Finite Differences

\[
\begin{align*}
\frac{\partial u}{\partial t} &= \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2}, \\
\frac{u_{i+1}^{n+1} - u_i^n}{\Delta t} &= \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2} \quad \text{(Forward Euler)}, \\
\frac{u_{i+1}^{n+1} - u_i^n}{\Delta t} &= \frac{u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}}{\Delta x^2} \quad \text{(Backward Euler)}.
\end{align*}
\]

(5.30, 5.31, 5.32)
5.3 Time-integration of PDE’s

\[ \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}. \quad (5.33) \]

Weak form:
\[ \int_\Omega \frac{\partial u}{\partial t} v dx = \int_\Omega \frac{\partial^2 u}{\partial x^2} v dx = - \int_\Omega \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} dx, \quad (5.34) \]

where we took \( v \big|_{\partial \Omega} = 0 \). Find \( u \), subject to \( u(0) = u_0 \), \( u(1) = u_1 \), such that
\[ \int_0^1 \frac{\partial u}{\partial t} v dx = - \int_0^1 \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} dx \quad \forall v(0) = 0 = v(1). \quad (5.35) \]

Provided that the above integral exist. Further take basis-functions (piecewise linear), \( \phi_i(0) = \phi_i(1) \) for \( i \in \{1, \ldots, m-1\} \),
\[ u(x,t) = \sum_{j=1}^{m-1} [c_j(t) \phi_j(x)] + u_0 \phi_0(x) + u_1 \phi_m(x). \quad (5.36) \]

Take for simplicity \( u_0 = 0 = u_1 \) then
\[ \sum_{j=1}^{m-1} c'_j(t) \int_0^1 \phi_j(x) \phi_i(x) dx = - \sum_{j=1}^{m-1} c_j(t) \int_0^1 \phi'_j(x) \phi'_i(x) dx \quad i \in \{1, \ldots, m-1\}. \quad (5.37) \]

For forward Euler, one obtains:
\[ \sum_{j=1}^{m-1} \frac{c_{j+1}^n - c_j^n}{\Delta t} \int_0^1 \phi_j(x) \phi_i(x) dx = - \sum_{j=1}^{m-1} c_j^n \int_0^1 \phi'_j(x) \phi'_i(x) dx \quad i \in \{1, \ldots, m-1\}, \quad (5.38) \]

where we call \( M_{ij} = \int_0^1 \phi_i(x) \phi_j(x) dx \) the entries of the mass-matrix and \( S_{ij} = \int_0^1 \phi'_i(x) \phi'_j(x) dx \) the entries of the stiffness-matrix. Elementwise, we obtain
\[ \frac{c_{i+1}^{n+1} - c_i^n}{\Delta t} \int_{e_i} \phi_i \phi_{i-1} dx + \frac{c_{i+1}^{n+1} - c_i^n}{\Delta t} \int_{e_i \cup e_{i+1}} (\phi_i)^2 dx + \frac{c_{i+1}^{n+1} - c_i^n}{\Delta t} \int_{e_{i+1}} \phi_i \phi_{i+1} dx \]
\[ = - \frac{c_i^n}{e_i} \int_{e_i} \phi_i \phi_{i-1} dx - \frac{c_i^n}{e_i \cup e_{i+1}} \int_{e_i \cup e_{i+1}} \phi'_i \phi'_i dx - \frac{c_{i+1}^{n+1}}{e_{i+1}} \int_{e_{i+1}} \phi'_i \phi'_{i+1} dx \quad \text{for } i \in \{1, \ldots, m-1\}. \quad (5.39) \]

All the integrals can be computed. The discretization method for FEM looks very different. For the case of an equidistant grid, it can be shown that the
result becomes identical to FVM and FDM. For backward Euler, one obtains similarly:

\[
\frac{c_{i+1}^{n+1} - c_{i}^{n+1}}{\Delta t} \int_{e_i} \phi_i \phi_{i-1} dx + \frac{c_{i+1}^{n+1} - c_{i+1}^{n}}{\Delta t} \int_{e_i \cup e_{i+1}} (\phi_i)^2 dx + \frac{c_{i}^{n+1} - c_{i}^{n}}{\Delta t} \int_{e_{i+1}} \phi_i \phi_{i+1} dx
\]

\[
= -c_{i}^{n+1} \int_{e_i} \phi_i' \phi_{i-1} dx - c_{i+1}^{n} \int_{e_i \cup e_{i+1}} \phi_i' \phi_{i} dx - c_{i+1}^{n+1} \int_{e_{i+1}} \phi_i' \phi_{i+1} dx
\]

for \( i \in \{1, \ldots, m-1\} \). (5.40)

All these expressions can be adjusted to Crank-Nicholson or a two-step method (Runge-Kutta 2 or Modified Euler). Later we will consider the stability of numerical methods.

We remark further that the above Finite Element discretization of the heat problem can be written in matrix-form as

\[
M \frac{dc}{dt} = Sc
\]

(5.41)

where \( M_{ij} = \int_0^1 \phi_i(x) \phi_j(x) dx \), \( S_{ij} = \int_0^1 \phi'_i(x) \phi'_j(x) dx \) and \( c = [c_1 \ldots c_n]^T \). Of course, the Euler, modified Euler etc. can be written similarly as before. As an example the Euler-forward time integration method can be written by:

\[
M \mathbf{c}_{n+1} = M \mathbf{c}_n + \Delta t S \mathbf{c}_n.
\]

(5.42)

For the case that \( M \) is not a diagonal matrix, a linear system of equations has to be solved also for the Euler-forward method.

**Exercise 5.3.1.** Write down the corresponding time-integration methods for the backward Euler, Modified Euler and Crank-Nicholson methods.

### 5.3.2 The wave equation

We consider the discretization of the wave equation

\[
\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad t > 0,
\]

+ Dirichlet boundary conditions + initial conditions for \( u \) and \( u_t \).

Here \( c \) represents the given wave-speed, which is assumed to be constant in the text. The weak form of the above equation is obtained after multiplication by a test-function, to obtain

\[
\int_{\Omega} \frac{\partial^2 u}{\partial t^2} v dx = c^2 \int_{\Omega} \frac{\partial^2 u}{\partial x^2} v dx = -c^2 \int_{\Omega} \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} dx,
\]

(5.44)
where we took \( v \mid_{\partial \Omega} = 0 \). Find \( u \), subject to \( u(0) = u_0, u(1) = u_1 \), such that

\[
\int_0^1 \frac{\partial^2 u}{\partial t^2} v dx = -c^2 \int_0^1 \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} \, dx \quad \forall v(0) = 0 = v(1).
\] (5.45)

Provided that the above integral exist. Further take basis-functions (piecewise linear), \( \phi_i(0) = 0 = \phi_i(1) \) for \( i \in \{1, \ldots, m - 1\} \),

\[
u(x, t) = \sum_{j=1}^{m-1} [c_j(t) \phi_j(x)] + u_0 \phi_0(x) + u_1 \phi_m(x).
\] (5.46)

Take for simplicity \( u_0 = 0 = u_1 \) then

\[
\sum_{j=1}^{m-1} c_j''(t) \int_0^1 \phi_j(x) \phi_i(x) dx = -c^2 \sum_{j=1}^{m-1} c_j(t) \int_0^1 \phi_j'(x) \phi_i'(x) dx \quad i \in \{1, \ldots, m - 1\}.\] (5.47)

where we call \( M_{ij} = \int_0^1 \phi_i(x) \phi_j(x) dx \) the entries of the mass-matrix and \( S_{ij} = \int_0^1 \phi'_i(x) \phi'_j(x) dx \) the entries of the stiffness-matrix. Hence, we obtain

\[
M \frac{d^2 \bar{c}}{dt^2} = c^2 S \bar{c}.
\] (5.48)

One possibility to integrate the above equation in time is to write it as system of first-order differential equations in time for \( c \) and \( w \):

\[
\frac{dc}{dt} = w,
\]

\[
M \frac{dw}{dt} = c^2 S \bar{c}.
\] (5.49)

with initial conditions for \( c \) and \( w \).

Now, we can apply the Forward Euler method to the above problem:

\[
\bar{c}^{n+1} = \bar{c}^n + \Delta t w^n,
\] (5.50)

\[
M \bar{w}^{n+1} = M w^n + \Delta t c^2 S \bar{c}^n.
\]

**Exercise 5.3.2.** Write down the corresponding time-integration methods for the backward Euler, Modified Euler and Crank-Nicholson methods.

It can be shown that the numerical integration of the above equation gives rise to dissipation \( \bar{c}^n \to 0 \) as \( n \to \infty \). The amount of dissipation can be decreased when a Runge-Kutta method is used. Further one can use a direct time-integration of the second-order system of differential equations. This is not treated any further.
5.4 Stability analysis

In general after discretization we obtain a system of ordinary differential equation in the form

\[ M \frac{du}{dt} = Su \iff \frac{du}{dt} = M^{-1}Su. \]  

(5.51)

For the stability the eigenvalues of the above matrix \( M^{-1}S \) are crucial. First, for the analytical asymptotic stability we require

\[ \lambda(M^{-1}S) < 0, \quad \text{and if } \lambda \notin \mathbb{R} \text{ we have } \Re\{\lambda(M^{-1}S)\} < 0. \]

(5.52)

**Exercise 5.4.1.** Derive stability criteria for the eigenvalues of \( M^{-1}S \) for the Forward Euler, Modified Euler, Backward Euler and Trapezoid time integration methods.

Hence, it is crucially important to have some knowledge on the eigenvalues of the matrix \( M^{-1}S \). For large matrices it is not easy to compute the eigenvalues. Fortunately, Gershgorin’s Theorem gives a often very usefull estimate of the eigenvalues. For many cases the mass matrix \( M \) is diagonal as a result of numerical integration by use of the Rule of Newton-Cotes. This is commonly called lumping.

**Exercise 5.4.2.** Show by the numerical integration of Newton-Cotes that

\[ m_{ji} = \int_0^1 \phi_i(x)\phi_j(x)dx = \begin{cases} \frac{x_{i+1} - x_{i-1}}{2}, & \text{if } j = i, \\ 0, & \text{else.} \end{cases} \]  

(5.53)

For this case the Theorem reads as follows:

**Theorem 5.4.1.** Let \( M \) be diagonal, then, for all eigenvalues \( \lambda \) of \( M^{-1}S \) holds:

\[ |\lambda| \leq \sup_k \frac{1}{|m_{kk}|} \sum_{i=1}^n |s_{ki}|. \]  

(5.54)

Note that the eigenvalues may be complex if \( S \) is not symmetric. Then the areas in which the eigenvalues are allowed to be in, consist of circles in the complex plane.

**Proof 5.4.1.** Let \( \lambda \) be an eigenvalue of the generalized eigenvalue problem with corresponding eigenvector \( \underline{v} \), then,

\[ S\underline{v} = \lambda M\underline{v}. \]  

(5.55)
In other words, this becomes for each component $k$:

$$
\sum_{i=1}^{n} s_{ki}v_i = \lambda m_{kk}v_k. \tag{5.56}
$$

Let $v_j$ be the component of $v$ with the largest modulus, then, we have for this index $j$:

$$
\lambda = \frac{1}{m_{jj}} \sum_{i=1}^{n} s_{ji} \frac{v_i}{v_j}, \tag{5.57}
$$

and since $|v_i/v_j| \leq 1$, we get

$$
|\lambda| \leq \frac{1}{|m_{kk}|} \sum_{i=1}^{n} |s_{ki}|. \tag{5.58}
$$

This proves the assertion. \hfill \Box

We illustrate the use of Gershgorin’s Theorem by the following example.

**Example:** Suppose we use Finite Differences, then $M = I$. Let $h$ be the stepsize and $\Delta t$ be the time-step. Further in one dimension, we have $s_{ii} = -2/h^2$ and $s_{i-1} = 1/h^2 = s_{i+1}$. From this, we obtain $\lambda \leq 4/h^2$. Hence we obtain for the Forward Euler method: $\Delta t \leq h^2/2$.

The analysis of stability of time-integration methods for Partial Differential Equations can also be done by the use of the Von Neumann analysis, which is based on Fourier analysis. This is omitted in the present course.
Numerical linear algebra

In this part we will consider some basic methods to solve the matrix-vector equations that arise from the various types of discretizations. For small systems, one preferably uses a so-called “direct” method to solve the equation. For large, sparse (i.e. with mainly zeroes in the matrix and non-zero entries in the vicinity of the main diagonal) systems/matrices, one better relies on iterative methods. The treatment given here is far from complete. Before we treat some methods, we will review some basic linear algebra concepts, which are essential for understanding.

### 6.1 Basics

Given a matrix $A$ and vector $x$:

\[
A = \begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} \\
\vdots \\
a_{n1} & \cdots & \cdots & a_{nn}
\end{pmatrix}, 
\quad x = \begin{pmatrix}
x_1 \\
x_2 \\
\vdots \\
x_n
\end{pmatrix}
\]  

(6.1)

then we can see $A$ as a row of columns: $A = (a_1, a_2, \ldots, a_n)$, where

\[
a_j = \begin{pmatrix}
a_{1j} \\
a_{2j} \\
\vdots \\
a_{nj}
\end{pmatrix}, \quad i \in \{1, \ldots, n\}.
\]  

(6.2)
The product of $A$ and $\mathbf{x}$ can be written by

$$AX = \left( a_1, a_2, \ldots, a_n \right) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} := a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

$$= \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} x_1 + \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{pmatrix} x_2 + \cdots + \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{nn} \end{pmatrix} x_n$$

$$= \begin{pmatrix} a_{11} x_1 + a_{12} x_2 + \cdots + a_{1n} x_n \\ a_{21} x_1 + a_{22} x_2 + \cdots + a_{2n} x_n \\ \vdots \\ a_{n1} x_1 + a_{n2} x_2 + \cdots + a_{nn} x_n \end{pmatrix}.$$  \hspace{1cm} (6.3)

**6.1.1 Eigenvalues and eigenvectors of $A$**

The eigenvalues and eigenvectors are defined by the solution of the following problem:

$$\text{Find } \lambda, \mathbf{v} \text{ such that } AV = \lambda \mathbf{v} \text{ for } \mathbf{v} \neq \mathbf{0}. \hspace{1cm} (6.4)$$

$\lambda$ is called an eigenvalue and $\mathbf{v}$ is called an eigenvector. This implies that $AV - \lambda \mathbf{v} = \mathbf{0}$ for $\mathbf{v} \neq \mathbf{0}$. Let $I = \text{diag}(1 \ldots 1)$ be the identity matrix then it follows that

$$(A - \lambda I)\mathbf{v} = \mathbf{0} \text{ for } \mathbf{v} \neq \mathbf{0}. \hspace{1cm} (6.5)$$

This implies that $A - \lambda I$ is singular and hence that its determinant is zero, \textit{i.e.}

$$\det(A - \lambda I) = 0. \hspace{1cm} (6.6)$$

The above equation is referred to as the characteristic equation, which defines the eigenvalues ($\lambda$).

**Example:** Given

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}. \hspace{1cm} (6.7)$$

Eigenvalues

$$\det(A - \lambda I) = 0 \Rightarrow \begin{vmatrix} 2 - \lambda & -1 \\ -1 & 2 - \lambda \end{vmatrix} = 0 \hspace{1cm} (6.8)$$

$$\Leftrightarrow (2 - \lambda)^2 - 1 = 0 \Leftrightarrow 2 - \lambda = \pm 1 \Rightarrow \lambda = 3 \lor \lambda = 1. \hspace{1cm} (6.9)$$
Eigenvectors:

\[
Av = \lambda v, \quad (6.11)
\]

\[
\lambda = 3 : \begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} v^{(1)} \\ v^{(2)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow v = \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \quad (6.12)
\]

\[
\lambda = 1 : \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} v^{(1)} \\ v^{(2)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow v = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \quad (6.13)
\]

6.1.2 Diagonalization of A

Suppose that \( \lambda_1, \ldots, \lambda_n \) are the eigenvalues of \( A \) with eigenvectors \( v_1, \ldots, v_n \) and let \( P = (v_1 \ v_2 \ \ldots \ v_n) \), then

\[
AP = A(v_1 \ v_2 \ \ldots \ v_n) = (Av_1 \ Av_2 \ \ldots \ Av_n) = (\lambda_1 v_1 \ \lambda_2 v_2 \ \ldots \ \lambda_n v_n). \quad (6.14)
\]

Further let \( D = \text{diag}(\lambda_1 \ \lambda_2 \ldots \lambda_n) \) then

\[
PDP = (v_1 \ v_2 \ \ldots \ v_n) \begin{pmatrix} \lambda_1 & 0 & \cdots \\ 0 & \ddots & \vdots \\ \vdots & \ddots & \lambda_n \end{pmatrix} = (\lambda_1 v_1 \ \lambda_2 v_2 \ \ldots \ \lambda_n v_n). \quad (6.15)
\]

Hence \( AP = PD \). Note that we assumed that for each \( \lambda_i \), there exists a \( v_i \).

Hence \( A = PDP^{-1} \), \( \rightarrow \) diagonalization. Suppose that \( A \) is symmetric, \( a_{ij} = a_{ji}, \ A^T = A \), then

\[
A = PDP^{-1}, \quad (6.16)
\]

\[
A^T = (PDP^{-1})^T = P^{-T}DP^T. \quad (6.17)
\]

Since \( A = A^T \), we obtain that \( P^T = P^{-1} \) satisfies! Hence when \( A = A^T \), then \( A = PDP^T \).

6.1.3 Eigenvalues of a symmetric matrix

**Theorem 6.1.1.** The eigenvalues of a real-valued symmetric matrix are real. Further its eigenvectors of different eigenvalues are orthogonal.

Before we prove this assertion, we introduce the concept of inner-product and orthogonality. Given two vectors in \( \mathbb{C}^n \), then we define its inner product by

\[
\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = \mathbf{v}_1^T \mathbf{v}_2, \quad (6.18)
\]
where \( \bar{v}_1 \) is the complex conjugate of \( v_1 \). Consequently, when \( v_1, v_2 \in \mathbb{R}^n \), \( A \in \mathbb{R}^{n \times n} \), then
\[
\langle A v_1, v_2 \rangle = (A v_1)^T v_2 = v_1^T A^T v_2 = \langle v_1, A^T v_2 \rangle.
\] (6.19)
Hence \( \langle A v_1, v_2 \rangle = \langle v_1, A^T v_2 \rangle \). Now we consider the proof of Theorem 6.1.1:

**Proof 6.1.1.** We compute
\[
\langle A v_1, v_1 \rangle = \langle \lambda_1 v_1, v_1 \rangle = \bar{\lambda}_1 \langle v_1, v_1 \rangle.
\] (6.20)
Further
\[
\langle v, A v_1 \rangle = \langle v_1, \lambda_1 v_1 \rangle = \lambda_1 \langle v, v_1 \rangle.
\] (6.21)
Since \( A = A^T \), it follows that \( \langle A v_1, v_1 \rangle = \langle v_1, A v_1 \rangle \), and hence the expressions (6.20) and (6.21) should be equal. Hence \( \lambda_1 = \bar{\lambda}_1 \) and hence \( \lambda_1 \in \mathbb{R} \). This holds for all eigenvalues analogously. (since \( A v_i = \lambda_i v_i \)). Two vectors are orthogonal if and only if \( \langle v_1, v_2 \rangle = 0 \). Compute
\[
\langle A v_1, v_2 \rangle = \lambda_1 \langle v_1, v_2 \rangle,
\] (6.22)
since \( \lambda_1 \in \mathbb{R} \). Further
\[
\langle v_1, A v_2 \rangle = \lambda_2 \langle v_1, v_2 \rangle.
\] (6.23)
Since \( A = A^T \) we have \( \langle A v_1, v_2 \rangle = \langle v_1, A v_2 \rangle \) and hence
\[
\lambda_1 \langle v_1, v_2 \rangle = \lambda_2 \langle v_1, v_2 \rangle \implies \langle v_1, v_2 \rangle = 0.
\] (6.24)
This implies that \( v_1 \) and \( v_2 \) are orthogonal.

**6.2 Solution of systems of linear equations**

- **6.2.1 The LU-decomposition**

Consider the following system of equations:
\[
\begin{align*}
    a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\
    a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\
    & \vdots \\
    a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n.
\end{align*}
\] (6.25)

We can see the above system as a matrix equation
\[
\begin{pmatrix}
    a_{11} & a_{12} & \cdots & a_{1n} \\
    a_{21} & a_{22} & \cdots & a_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{n1} & a_{n2} & \cdots & a_{nn}
\end{pmatrix}
\begin{pmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_n
\end{pmatrix} =
\begin{pmatrix}
    b_1 \\
    b_2 \\
    \vdots \\
    b_n
\end{pmatrix},
\] (6.26)
\[ Ax = b \] (6.27)

Direct methods are Gauss elimination, \( LU \)-decomposition. \( LU \)-decomposition is used in many practical situations: Therefore, we consider the following example: Let

\[
A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \rightarrow \begin{pmatrix} a_{11} - \frac{a_{21}}{a_{11}} a_{12} \\ 0 \end{pmatrix} \sim \begin{pmatrix} a_{11} & a_{12} - \frac{a_{21}}{a_{11}} a_{12} \\ 0 & a_{22} - \frac{a_{21}}{a_{11}} a_{12} \end{pmatrix} = U. \quad (6.28)
\]

Further let

\[
L = \begin{pmatrix} 1 & 0 \\ \frac{a_{21}}{a_{11}} & 1 \end{pmatrix}. \quad (6.29)
\]

We calculate the product of \( LU \):

\[
LU = \begin{pmatrix} 1 & 0 \\ \frac{a_{21}}{a_{11}} & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ 0 & a_{22} - \frac{a_{21}}{a_{11}} a_{12} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ \frac{a_{21}}{a_{11}} a_{11} & a_{22} - \frac{a_{21}}{a_{11}} a_{12} \end{pmatrix} = A. \quad (6.31)
\]

Hence

\[
A = LU \quad \begin{cases} L : \text{lower matrix (triangular).} \\ U : \text{upper triangular matrix.} \end{cases} \quad (6.32)
\]

\[
Ax = b \iff LUx = b 
\Rightarrow \begin{cases} Ls = b, \\ Ux = s. \end{cases} \quad (6.34)
\]

solve \( s \) and \( x \) by use of these two steps in (6.34). For larger matrices this reads as

\[
L = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \frac{a_{21}}{a_{11}} & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \frac{a_{n-1}}{a_{n}} & \cdots & \cdots & 1 \end{pmatrix}, \quad U = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \bar{a}_{22} & \cdots & \cdots & \bar{a}_{2n} \\ \bar{a}_{33} & \cdots & \cdots & \bar{a}_{3n} \\ \vdots & \ddots & \ddots & \vdots \\ \bar{a}_{nn} & \cdots & \cdots & \bar{a}_{nn} \end{pmatrix} \quad (6.35)
\]

\[
Ls = b \Rightarrow \begin{cases} s_1 = b_1 \\ s_2 = b_2 - \frac{a_{21}}{a_{11}} s_1 = b_2 - \frac{b_1}{a_{11}} \\ \vdots \end{cases} \quad (6.36)
\]

and \( x_n = \frac{s_n}{a_{nn}} \), etc. This gives a gain in the efficiency. For one-dimensional problems, this is a nice method. For sparse matrices for 3D geometries this is not efficient due to "fillin", which is the loss of the sparse band structure of the matrix.
The LU-decomposition $A = LU$ can be obtained by Gauss-elimination:

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & & & \ddots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} \sim \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ \theta_1 & a_{22} & \ldots & a_{2n} \\ \vdots & & & \ddots \\ \theta_n & a_{n2} & \ldots & a_{nn} \end{pmatrix} = U^{(1)},$$

where $a_{22}^{(1)} = a_{22} - \frac{a_{21}}{a_{11}}a_{11}$ etc, and $a_{n2}^{(1)} = a_{n2} - \frac{a_{n1}}{a_{11}}a_{11}$ etc.

$$L^{(1)} = \begin{pmatrix} 1 & \theta_1 & \emptyset \\ \frac{a_{21}}{a_{11}} & 1 & \emptyset \\ \vdots & \emptyset & \ddots \\ \frac{a_{n1}}{a_{11}} & \emptyset & \cdots & 1 \end{pmatrix}.$$  \hfill (6.39)

We can check that $L^{(1)}U^{(1)} = A$. We repeat this on $U^{(1)}$,

$$U^{(1)} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ \theta_1 & a_{22} & \ldots & a_{2n} \\ \vdots & & & \ddots \\ \theta_n & a_{n2} & \ldots & a_{nn} \end{pmatrix} \sim \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ \emptyset & a_{22} & \ldots & a_{2n} \\ \vdots & & & \ddots \\ \emptyset & a_{n2} & \ldots & a_{nn} \end{pmatrix} = U^{(2)},$$

$$L^{(2)} = \begin{pmatrix} 1 & \emptyset \\ \frac{a_{21}}{a_{11}} & \emptyset \\ \emptyset & \ddots \\ \frac{a_{n1}}{a_{11}} & \emptyset & \cdots & 1 \end{pmatrix}.$$  \hfill (6.42)

We can check that $A = L^{(2)}U^{(2)}$. This procedure repeated until $U^{(n)}$ is an upper matrix (triangle). Hence, $n$ times!
An alternative is Crout’s scheme:

$$A = LU,$$

(6.43)

suppose that

$$A = \begin{pmatrix} \tilde{A} & a_c \\ \mathbf{a}_r^T & a_{nn} \end{pmatrix},$$

(6.44)

and that $\tilde{A} = \tilde{L}\tilde{U}$ is known, then

$$L = \begin{pmatrix} \tilde{L} & \emptyset \\ \tilde{a}_r^T & 1 \end{pmatrix} \text{ and } U = \begin{pmatrix} \tilde{U} & u_c \\ \emptyset & u_{nn} \end{pmatrix}.$$  

(6.45)

Then

$$A = LU \Rightarrow \begin{pmatrix} \tilde{A} & a_c \\ \mathbf{a}_r^T & a_{nn} \end{pmatrix} = \begin{pmatrix} \tilde{L} & \emptyset \\ \tilde{a}_r^T & 1 \end{pmatrix} \begin{pmatrix} \tilde{U} & u_c \\ \emptyset & u_{nn} \end{pmatrix}.$$  

(6.46)

Evaluation of the products of these block matrices, gives:

$$\begin{aligned}
\tilde{A} &= \tilde{L}\tilde{U} \\
\mathbf{a}_r^T &= \tilde{a}_r^T \tilde{U} \Rightarrow \tilde{U}\tilde{L}_r = \mathbf{a}_r \\
a_r &= \tilde{L}_r \mathbf{a}_r \\
a_{nn} &= \tilde{U}\tilde{U}_c + u_{nn} \Rightarrow u_{nn} = a_{nn} - \tilde{U}_c \mathbf{a}_r
\end{aligned}$$

(6.47)

Here we have to determine $\tilde{L}_r^T$, $u_c$, and $u_{nn}$.

**Exercise 6.2.1.** Given

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$  

(6.48)

*Give the LU-decomposition of $A$.*

A disadvantage is the fill-in that takes place for more-dimensional problems:

$$A = \begin{pmatrix} \emptyset \\ \emptyset \\ \emptyset \\ \emptyset \\ \emptyset \end{pmatrix} \sim \cdots \sim \begin{pmatrix} \text{extra} & \text{non-} & \emptyset \\ \text{non-} & \text{non-} & \text{zero’s} \\ \emptyset & \emptyset & \emptyset \end{pmatrix} = U$$

(6.49)

and

$$L = \begin{pmatrix} \text{extra} & \emptyset \\ \text{non-} & \emptyset \\ \emptyset & \text{zero’s} \end{pmatrix}$$

(6.50)

Especially when $A$ is a discretization matrix of a 3D-model then this will destroy the efficiency. For these problems one takes refuge to iterative solvers.
6.2 Solution of systems of linear equations

6.2.2 Basic, classical iterative solvers

We consider again

\[
\begin{align*}
 a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\
 a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\
 & \vdots \\
 a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n.
\end{align*}
\]  

(6.51)

\(Ax = b\), consider the sequence of \(x^k = (x_1^k, x_2^k, \ldots, x_n^k)^T\) were \(x^k\) is the solution after \(k\) iterations. Then, Jacobi-iterations work as follows:

\[
\begin{align*}
 a_{11}x_1^{k+1} &= b_1 - a_{12}x_2^k - a_{13}x_3^k - \cdots - a_{1n}x_n^k, \\
 a_{22}x_2^{k+1} &= b_2 - a_{21}x_1^k - a_{23}x_3^k - \cdots - a_{2n}x_n^k, \\
 & \vdots \\
 a_{nn}x_n^{k+1} &= b_n - a_{n1}x_1^k - a_{n2}x_2^k - \cdots - a_{nn}x_n^k.
\end{align*}
\]

(6.52)

In other words, when we introduce the following notation:

- \(D = \text{diag}(a_{11}, a_{22}, \ldots, a_{nn})\),
- \(L_{ij} = -A_{ij}\) when \(i > j\), \(L_{ij} = 0\) when \(i \leq j\) (lower),
- \(U_{ij} = -A_{ij}\) when \(i < j\), \(U_{ij} = 0\) when \(i \geq j\) (upper),

then \(A = D - L - U\). Jacobi-iteration then works as follows

\[
\begin{align*}
 \text{Find } x^0 \text{ (initial choice/estimate)} \\
 D x^{k+1} &= b + (L + U)x^k, \quad k \in \{0, 1, 2, \ldots\}.
\end{align*}
\]

(6.53)

The Jacobi-iterations represent a classical iterative solution method. It converges however very slowly.

Looking at equations (system (6.52)), we see that after \(x_1^{k+1}\) is determined, it could be substituted into the equation for \(x_2^{k+1}\). Subsequently, the values of \(x_1^{k+1}\) and \(x_2^{k+1}\) can be substituted into the equation for \(x_3^{k+1}\), and so on. This will give, on intuitive grounds, a faster convergence behaviour. This is the iteration method due to Gauss-Seidel. It runs as follows:

\[
\begin{align*}
 a_{11}x_1^{k+1} &= b_1 - a_{12}x_2^k - a_{13}x_3^k - \cdots - a_{1n}x_n^k, \\
 a_{22}x_2^{k+1} &= b_2 - a_{21}x_1^k - a_{23}x_3^k - \cdots - a_{2n}x_n^k, \\
 a_{33}x_3^{k+1} &= b_3 - a_{31}x_1^k - a_{32}x_2^k - \cdots - a_{3n}x_n^k, \\
 & \vdots \\
 a_{nn}x_n^{k+1} &= b_n - a_{n1}x_1^k - a_{n2}x_2^k - \cdots - a_{nn}x_n^k.
\end{align*}
\]

(6.54)

This implies in vector-form:

\[
\begin{align*}
 \text{Choose } x^0 \\
 (D - L)x^{k+1} &= Ux^k + b.
\end{align*}
\]

(6.55)
This is known as Gauss-Seidel’s iteration method. Convergence is faster than the convergence of Jacobi’s method.

Exercise 6.2.2. Given

\[
\mathbf{b} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}^T \tag{6.56}
\]

and

\[
A = \begin{pmatrix}
4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
-1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0 \\
-1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\
0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\
0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\
0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\
0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4
\end{pmatrix} \tag{6.57}
\]

Give the matrices \( D \), \( L \) and \( U \) for the Gauss-Seidel method.

Now we will analyze some convergence properties.

### 6.3 Convergence properties of iterative methods

The Jacobi and Gauss-Seidel iterations can be written in the following form:

\[
(D - L)x^{k+1} = kx^k + \mathbf{b} \Rightarrow x^{k+1} = (D - L)^{-1}Ux^k + (D - L)^{-1}, \tag{6.58}
\]

provided that \( \det(D - L) \neq 0 \). In other words when we put \( M := (D - L)^{-1}U \) and \( f := (D - U)^{-1}b \), then

\[
x^{k+1} = Mx^k + f. \tag{6.59}
\]

This is the most basic iterative scheme. We will analyze its convergence. Therefore, we need to prove the following theorem:

**Theorem 6.3.1.** Let \( x^k \in \mathbb{R}^n \), \( M \in \mathbb{R}^{n \times n} \) and \( f \in \mathbb{R}^n \).

The sequence \( x^{k+1} = Mx^k + f \) converges to a limit if and only if \( | \lambda_1 | < 1 \), where \( \lambda_1 \) is the eigenvalue of matrix \( M \) with the largest absolute value.

Before we "prove" this assertion we remind that we mean convergence if there exists a vector \( \mathbf{x} \) for which \( x^k \to \mathbf{x} \) as \( k \to \infty \). In other words \( x^k \) has a limit. Mathematicians are rather formal about convergence (and that’s their problem and right), and they will use the following definition in general:
Definition 6.3.2. A sequence of vectors is called convergent to \( x \) if there exists \( \varepsilon > 0 \) and \( N \in \mathbb{N} \) such that
\[
| x^k - x | < \varepsilon \quad \text{for all} \quad n > N (= N(\varepsilon)).
\] (6.60)

We then have \( x^k \to x \) as \( n \to \infty \).

The proof that we present here, will not be very formal. We sketch the proof for diagonalizable matrices only, i.e. There exists \( P, \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n) \) such that \( A = PDP^{-1} \), where the columns of \( P \) represent eigenvectors, see Section 6.1. The proof can be generalized using Jordan decompositions with generalized eigenvectors for non-diagonalizable matrices. An alternative proof is based on writing the start-vector as a linear combination of the eigenvectors of the matrix \( M \). These approaches will be omitted.

Proof 6.3.1. (Proof of Theorem 6.3.1): We assume that \( M = P\Lambda P^{-1} \), where \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n) \) with \( \{\lambda_i\} \) representing the eigenvalue of \( M \) and \( P = (\mathbf{v}_1, \ldots, \mathbf{v}_n) \) is produced by the use of eigenvectors of \( M \) (see section 6.1). Then,
\[
x^{k+1} = P\Lambda P^{-1}x^k + f,
\] (6.61)
or equivalently
\[
P^{-1}x^{k+1} = \Lambda P^{-1}x^k + P^{-1}f,
\] (6.62)
where \( P^{-1}x^{k+1} = y^{k+1}, \Lambda P^{-1}x^k = \Lambda y^k \) and \( P^{-1}f = \tilde{f} \). Introduction of \( y^k := P^{-1}x^k \) and \( \tilde{f} = P^{-1}f \), then gives:
\[
y^{k+1} = \Lambda y^k + \tilde{f}.
\] (6.63)

Suppose that this limit exists, i.e. \( y^k \to y \) as \( k \to \infty \), then
\[
\begin{cases}
  y = \Lambda y + \tilde{f} \\
  y^{k+1} = \Lambda y^k + \tilde{f}
\end{cases}
\]
holds for the exact limit, and
\[
y - y^{k+1} = \Lambda (y - y^k)
\] (6.64)
where \( y - y^{k+1} \) is the error after \( k+1 \) iterations and \( y - y^k \) is the error after \( k \) iterations. Component-wise we see \( y_i - y_i^{k+1} = \lambda_i(y_i - y_i^k) \), hence
\[
| y_i - y_i^{k+1} | \leq | \lambda_i | \cdot | y_i - y_i^k | .
\] (6.65)
We see that \( | \lambda_i | < 1 \) implies \( | y - y^{k+1} | \leq | y - y^k | \Leftrightarrow y^k \to y \). Further, we see that if \( y^k \to y \) then
\[
| y_i - y_i^{k+1} | \leq | y_i - y_i^k | \quad \text{(i.e. convergence),}
\] (6.66)
implies \( | \lambda_i | < 1 \). This proves the Theorem. □
Consequence: The iterative methods of Jacobi and Gauss-Seidel converge if and only if
\[
\left\{ \begin{array}{l}
| \lambda(D^{-1}(L + U)) | < 1, \\
| \lambda((D - L)^{-1}U) | < 1.
\end{array} \right.
\] (6.67)

We see from the above that it is important to know the eigenvalues. However, for large matrices the computation of the eigenvalues is expensive. We will provide some Theorems on the estimation of the eigenvalues. First we will show that Gauss-Seidel converges if \( A \) is a symmetric and positive definite matrix. First we need the definition of positive definitity:

**Definition 6.3.3.** A real-valued symmetric matrix \( A = A^T \) is positive definite if and only if
\[
\langle \mathbf{x}, A\mathbf{x} \rangle = \mathbf{x}^T A \mathbf{x} > 0 \quad \forall \mathbf{x} \in \mathbb{R}^n, \quad \mathbf{x} \neq \mathbf{0}.
\] (6.68)

We remark that \( A \) is positive definite \( (A = A^T) \) if and only if \( \lambda(A) > 0 \) for all its eigenvalues. Now we will state a theorem for convergence of Gauss-Seidel for a symmetric positive definite matrix \( A \).

**Theorem 6.3.4.** If the matrix \( A \) is symmetric positive definite (SPD) then, the Gauss-Seidel process converges.

**Proof 6.3.2.** (Proof of Theorem 6.3.4): Let \( A \) be symmetric \( A = A^T \), then
\[
A = D - L - L^T \quad (L^T = U).
\] (6.69)

Further we have to show that \( | \lambda((D - L)^{-1}L^T) | < 1 \). Let \( \lambda \) and \( \mathbf{v} \) be an eigenvalue and an eigenvector of \( (D - L)^{-1}L^T \), then
\[
(D - L)^{-1}L^T \mathbf{v} = \lambda \mathbf{v},
\] (6.70)

and
\[
(D - L)^{-1}L^T \bar{\mathbf{v}} = \bar{\lambda} \bar{\mathbf{v}}.
\] (6.71)

This implies
\[
L^T \mathbf{v} = \lambda(D - L) \mathbf{v},
\] (6.72)

and
\[
L^T \bar{\mathbf{v}} = \bar{\lambda}(D - L) \bar{\mathbf{v}}.
\] (6.73)

Further we will compute \( \langle \mathbf{v}, A \mathbf{v} \rangle \) by the use of the two following possibilities:

[1]
\[
\langle \mathbf{v}, A \mathbf{v} \rangle = \mathbf{v}^T A \mathbf{v} = \mathbf{v}^T (D - L - L^T) \mathbf{v} = \mathbf{v}^T (D - L) \mathbf{v} - \mathbf{v}^T L^T \mathbf{v}
= \mathbf{v}^T (D - L) \mathbf{v} - \bar{\mathbf{v}}^T \lambda(D - L) \mathbf{v}
= (1 - \lambda) \mathbf{v}^T (D - L) \mathbf{v}.
\] (6.74)
The above two relations imply:

\[
\frac{1}{1-\lambda} \langle \psi, A\psi \rangle = \bar{\psi}^T (D - L)^T \psi
\]

\[
\frac{1}{1-\lambda} \langle \psi, A\psi \rangle = \bar{\psi}^T (D - L^T)^T \psi
\]

\[
\frac{1}{1-\lambda} + \frac{1}{1-\lambda} \langle \psi, A\psi \rangle = \langle \psi, D\psi \rangle + \langle \psi, A\psi \rangle
\]

(6.76)

where \( \langle \psi, A\psi \rangle > 0 \), \( \langle \psi, D\psi \rangle > 0 \) (this can be shown by use of Gershgorin's Theorem, to be explained later) and \( \langle \psi, A\psi \rangle > 0 \). From this follows

\[
\frac{1}{1-\lambda} + \frac{1}{1-\lambda} > 1.
\]

(6.77)

Hence

\[
\frac{1-\lambda+1-\lambda}{(1-\lambda)(1-\lambda)} > 1
\]

\[
\iff 2 - 2\Re\lambda > 1 - (\bar{\lambda} + \lambda) + |\lambda|^2
\]

\[
\iff 2 - 2\Re\lambda > 1 - 2\Re\lambda + |\lambda|^2
\]

\[
\iff |\lambda|^2 < 1.
\]

(6.78)

This implies that \( |\lambda((D - L)^{-1}L^T)| < 1 \) and hence due to Theorem 6.3.1 we have convergence for Gauss-Seidel.

Now we will derive a theorem, due to the Russian mathematician Gershgorin, for the estimation of the eigenvalues of \( A \).

\[\textbf{Theorem 6.3.5.} \text{ Let } A \in \mathbb{R}^{n \times n} \text{ with eigenvalues } \{\lambda_1, \lambda_2, \ldots, \lambda_n\} \text{ then}
\]

\[
|\lambda - a_{ii}| \leq \sum_{\substack{j = 1 \\ j \neq i}}^{n} |a_{ij}| \quad \text{for all } i \in \{1, \ldots, n\}.
\]

(6.79)

The above Theorem provides a bound in which the eigenvalues should be located. It can be used for the derivation of stability of a numerical method. Let's prove the theorem!
Proof 6.3.3. (Proof of Theorem 6.3.5): Let $\lambda$ be an eigenvalue with eigenvector $v$, $v = (v^1, v^2, \ldots, v^n)^T$, $Av = \lambda v$, note $A = (a_{ij})$, then $Av = (a_1 \ldots a_n)^T v = a_1 v^1 + a_2 v^2 + \cdots + a_n v^n = \lambda v$. Consider the $i^{th}$ component, which is the largest component of $v$:

$$a_{i1} v^1 + a_{i2} v^2 + \cdots + a_{in} v^n = \lambda v^i$$

$$\Rightarrow (\lambda - a_{ii}) v^i = a_{i1} v^1 + a_{i2} v^2 + \cdots + a_{ii-1} v^{i-1} + a_{ii+1} v^{i+1} + \cdots + a_{in} v^n. \tag{6.80}$$

Hence:

$$(\lambda - a_{ii}) v^i = \sum_{j=i}^{n} a_{ij} v^j$$

$$\Rightarrow |\lambda - a_{ii}| = |\sum_{j=i}^{n} a_{ij} \frac{v^j}{v^i}| \leq \sum_{j=i}^{n} |a_{ij} \frac{v^j}{v^i}|$$

$$\leq \sum_{j=i}^{n} |a_{ij}| \quad (since \ |v^j| \leq |v^i|). \tag{6.81}$$

This holds for all eigenvalues and hence the theorem is proven. \qed

A consequence of the above theorem is that if a symmetric matrix with positive elements on the main diagonal is diagonally dominant, i.e.

$$a_{ii} \geq \sum_{j=i}^{n} |a_{ij}| > 0 \quad \text{for } j \neq i, \tag{6.82}$$

then the matrix is positive (semi) definite. This is due to

$$|\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \leq \sum_{j \neq i} |a_{ij}| \quad \text{for } j \neq i. \tag{6.83}$$

$$-\sum_{j \neq i} |a_{ij}| \leq \lambda - a_{ii} \leq \sum_{j \neq i} |a_{ij}| \quad \text{for } j \neq i. \tag{6.84}$$

$$0 \leq a_{ii} - \sum_{j \neq i} |a_{ij}| \leq \lambda \leq a_{ii} + \sum_{j \neq i} |a_{ij}| \quad \text{for } j \neq i. \tag{6.85}$$

Hence $\lambda > 0$ and hence $A$ is positive (semi) definite. We, then, proved the following Theorem:

**Theorem 6.3.6.** Let $A = A^T \in \mathbb{R}^{n \times n}$ and $a_{ii} > 0$, $i \in \{1, \ldots, n\}$ and let

$$a_{ii} > \sum_{j=i}^{n} |a_{ij}|, \text{ then } A \text{ is positive definite.}$$
6.4 Gradient methods for $Ax = b$

The schemes of Gauss-Seidel and Jacobi are classical schemes. A completely different, but fast converging, class is the so-called Krylov-subspace class of iterative solution methods for linear equations. The mathematical theory is too complicated for treatment here. Further, there are numerous choices for Krylov subspace methods, which will not be treated here. A nice feature of the Krylov methods is their suitability to large systems and the fact that they only consist of matrix-vector multiplications, which is efficient when the matrix $A$ is sparse.

In this section we will only consider a gradient method which is one of the most important principles on which the conjugate gradient method is based. The class of gradient methods is based on the minimization of a functional in which the matrix $A$ is included. An important condition here is that $A$ is symmetric and positive definite ($A$ is SPD). Consider the following functional $J(x)$:

$$J(x) = \frac{1}{2} \langle x, Ax \rangle - \langle x, b \rangle; \quad x \in \mathbb{R}^n.$$  \hfill (6.86)

Now, we consider minimization of the above functional:

<table>
<thead>
<tr>
<th>Find $y \in \mathbb{R}^n$ such that</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J(y) \leq J(x)$ for all $x \in \mathbb{R}^n.$</td>
</tr>
<tr>
<td>-------------------------------------</td>
</tr>
</tbody>
</table>

(6.87)

Now we will show that this minimum coincides with the solution of $Ay = b$, where $y$ should be determined. Let $x = y + \varepsilon v$, then

$$J(y) \leq J(y + \varepsilon v) \quad \forall v \in \mathbb{R}^n.$$  \hfill (6.88)

In other words

$$\frac{d}{d\varepsilon} J(y + \varepsilon v) |_{\varepsilon=0} = 0 \quad \forall v \in \mathbb{R}^n.$$  \hfill (6.89)

$$J(y + \varepsilon v) = \frac{1}{2} \langle y + \varepsilon v, Ay + \varepsilon v \rangle - \langle y + \varepsilon v, b \rangle$$

$$= \frac{1}{2} \langle y, Ay \rangle + \frac{\varepsilon}{2} \langle y, Av \rangle + \frac{\varepsilon}{2} \langle v, Ay \rangle$$

$$+ \frac{\varepsilon^2}{2} \langle v, Av \rangle - \langle y, b \rangle - \varepsilon \langle v, b \rangle.$$  \hfill (6.90)

Then,

$$\frac{d}{d\varepsilon} J(y + \varepsilon v) |_{\varepsilon=0} = \langle v, Ay \rangle - \langle v, b \rangle = 0,$$  \hfill (6.91)

note that

$$\langle y, Av \rangle = \langle Ay, v \rangle = \langle v, Ay \rangle,$$  \hfill (6.92)

because $A^T = A$ and $v, y \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$. Hence the minimization problem transforms into:
Find $y \in \mathbb{R}^n$ such that

$$\langle v, Ay - b \rangle = 0 \quad \forall v \in \mathbb{R}^n. \quad (6.93)$$

This implies that $Ay - b = 0 \iff Ay = b$. So the solution of (6.87) co-incides with the solution of (6.93) and of:

$$Ay = b \quad \text{with A SPD, } y \in \mathbb{R}^n. \quad (6.94)$$

This is the principle on which (conjugate) gradient method are based. Now we treat the simplest, most basic gradient method that is used to solve (6.87). We, therefore, keep in mind that $\nabla J(x)$ gives the direction of the steepest incline and hence $-\nabla J(x)$ gives the steepest descent. Now the gradient of $J(x)$ can be computed by evaluation of all the partial derivatives:

$$-\nabla J(x) = \begin{pmatrix}
-\frac{\partial J}{\partial x_1}(x) \\
-\frac{\partial J}{\partial x_2}(x) \\
\vdots \\
-\frac{\partial J}{\partial x_n}(x)
\end{pmatrix}. \quad (6.95)$$

This will give $-\nabla J(x) = b - Ax =: r_k$. $b - Ax$ is commonly referred to as the residual. Suppose that $\{x^k\}$ represent the set of iterates, then we define:

$$r_k = b - Ax_k = -\nabla J(x_k). \quad (6.96)$$

The vector $r_k$ gives the direction of the steepest descent. Now we try to walk from $x_k$ over the direction of $r_k$ such that we find $x_{k+1}$ which minimizes $J(x)$ over this direction: in other words

Find $\alpha_k \in \mathbb{R}$ such that

$$J(x_k + \alpha_k r_k) \leq J(x_k + \beta_k r_k) \quad \forall \beta \in \mathbb{R}. \quad (6.97)$$

Then, we set $x_{k+1} = x_k + \alpha_k r_k$. This $\alpha_k$ should be determined and then this process is executed recursively. This gives find $\alpha_k$ from

$$\frac{d}{d\alpha_k} J(x_k + \alpha_k r_k) = 0 \Rightarrow \alpha_k = \frac{\langle r_k, r_k \rangle}{\langle r_k, A r_k \rangle}. \quad (6.98)$$

Now we are able to construct the most basic gradient method. Choose $x_0$, $r_0 = b - Ax_0$,

$$\text{do } k = 1 : n$$

$$\alpha_{k-1} = \frac{\langle r_{k-1}, r_{k-1} \rangle}{\langle r_{k-1}, A r_{k-1} \rangle}$$

$$x_k = x_{k-1} + \alpha_{k-1} r_{k-1}$$

$$r_k = b - A x_k = r_{k-1} - \alpha_{k-1} A r_{k-1}$$

end do
Of course, in practice the do-loop is replaced by a while loop until convergence is obtained, \textit{i.e.}

$$\| \mathbf{z}_k \| < \varepsilon. \quad (6.99)$$

The above algorithm is the basis for the more commonly used conjugate gradient method. The CG-method has the important different (from the gradient method) characteristic that minimization takes place in the so-called $A$-norm:

$$\| \mathbf{v} \|_A := \sqrt{\langle \mathbf{v}, A\mathbf{v} \rangle}. \quad (6.100)$$

Minimization in the $A$-norm is natural since the solution $A\mathbf{y} = \mathbf{b}$ has to be found where $\mathbf{b}$ (\textit{i.e.} $A\mathbf{y}$) is known. CG tries to find

$$\min_{\mathbf{x}_k \in K} \| \mathbf{y} - \mathbf{x}_k \|_A, \quad (6.101)$$

where $\mathbf{y}$ is exact and $\mathbf{x}_k$ is the numerical solution in a subspace of $\mathbb{R}^n$. We refer to Golub and Van Loan [1996] for a further (algorithmic and mathematical) treatment of the CG-method.


