

Advanced Numerical Methods and Their Applications to Industrial Problems

Adaptive Finite Element Methods

Lecture Notes

Summer School

Yerevan State University

Yerevan, Armenia

2004



Alfred Schmidt, Arsen Narimanyan

Center for Industrial Mathematics

University of Bremen

Bremen, Germany

www.math.uni-bremen.de/zetem/



Contents

1	Introduction, motivation	1
1.1	Introduction	1
1.2	Multiple scales in the modelling of real world problems	2
2	Mathematical modeling	5
2.1	Density, flux, and conservation	5
2.2	PDEs as a modeling tool	11
3	Functional analysis background	15
3.1	Banach spaces and Hilbert spaces	15
3.2	Basic concepts of Lebesgue spaces	16
3.3	Weak derivatives	17
3.4	Introduction to Sobolev spaces	18
3.5	Some useful properties of Sobolev spaces	19
4	Variational formulation of elliptic problems	21
4.1	Variational formulation of Poisson problem	21
4.2	Existence and uniqueness of weak solution	22
5	Finite element approximation	25
5.1	Galerkin discretization	25
5.2	Finite element method	26
5.3	Discretisation of 2nd order equation	31
5.4	Simplices of arbitrary dimension	34
6	A priori error estimates for elliptic problems	37
6.1	Abstract error estimates: Céa's lemma	37
6.2	Interpolation estimates	38
6.2.1	Clement interpolation	39
6.2.2	Lagrange interpolation	42
6.3	A priori error estimate	45
7	A posteriori error estimation for elliptic problems	46
7.1	A posteriori error estimation in the energy norm	46
8	Mesh refinement and coarsening	51
8.1	Refinement algorithms for simplicial meshes	52
8.2	Prolongation of data during refinement	58
8.3	Coarsening algorithms	58
8.4	Restriction of data during coarsening	60
8.5	Storage methods for hierarchical meshes	61

9	Adaptive strategies for elliptic problems	63
9.1	Quasi-optimal meshes	63
9.2	Mesh refinement strategies	64
9.3	Coarsening strategies	68
10	Aspects of efficient implementation	70
10.1	Numerical integration (quadrature schemes)	70
10.2	Efficient solvers for linear systems	71
10.2.1	Methods of Jacobi, Gauss-Seidel and Relaxation	71
10.2.2	Conjugate gradient method and other Krylov subspace iterations	75
10.2.3	Multilevel preconditioners and solvers	76
11	Error estimates via dual techniques	77
11.1	Estimates for the L_2 norm of the error	77
11.2	A priori error estimation in the L_2 norm	78
11.3	A posteriori error estimation in the L_2 norm	78
12	Parabolic problems - heat equation	81
12.1	Weak solutions of heat equation	81
12.2	Discretization of heat equation	83
12.3	A priori error estimates	86
13	A posteriori error estimates for parabolic problems	87
13.1	Abstract error estimate for ordinary differential equations	87
13.2	Weak formulation of the heat equation	89
13.3	Discretization	89
13.4	Error representation and dual problem	90
13.5	A posteriori error estimate	92
14	Adaptive methods for parabolic problems	93
14.1	Adaptive control of the time step size	94
15	The Stefan problem of phase transition	98
15.1	Problem setting:	99
15.2	Discretization	99
15.3	Error control for Stefan problem	100
15.4	Error Representation Formula	100
15.5	A Posteriori Error Estimators	101
15.6	Adaptive Algorithm	102
15.7	Numerical Experiments	103
15.7.1	Example 1: Oscillating Circle	103
15.7.2	Example 2: Oscillating Source	103
15.8	Nonlinear solver	106

16 The continuous casting problem	107
16.1 Setting	109
16.2 Discretization	111
16.3 Parabolic Duality	113
16.4 Robin Inflow Condition	114
16.5 Dirichlet Inflow Condition	115
16.6 Discontinuous p	116
16.7 Error Representation Formula	117
16.8 A Posteriori Error Analysis	118
16.9 Residuals	121
16.10 Proof of Theorem 16.1	125
16.11 Proof of Theorem 16.2	125
16.12 Discontinuous p	125
16.13 Performance	126
16.14 Localization and adaption	126
16.15 Example: Traveling wave	127
16.16 Applications to Casting of Steel	129
16.17 Scaling	129
16.18 Example: Oscillating Velocity	132
16.19 Example: Oscillating Cooling	133
17 Mathematical modeling of thermal cutting	135
17.1 Introduction	135
17.2 Problem description and physical modeling	135
17.3 Mathematical formulation of the problem	137
17.4 Variational inequalities and the weak formulation of the problem	139
17.4.1 Notation and Functional Spaces	139
17.4.2 A VI equivalent of Stefan-Signorini problem	140
17.5 Level set formulation	141
17.5.1 Stefan condition as level-set equation	141
17.6 Weak formulation of Stefan-Signorini problem	144
17.7 Heat Flux Density	144
17.8 Solution algorithm	148
References	150

1 Introduction, motivation

1.1 Introduction

When we talk about the use of mathematics or mathematical modeling for the industry, we mean the transformation of real world problems into mathematics. Often, this means neglecting some details, which are unimportant with respect to the posed questions. While experiments reveal the particular features of any process, the mathematical model permits the establishment of the general laws and thus contributes to the fundamental knowledge of the process.

The subject of partial differential equations (PDEs) holds a special and an important position in mathematical modeling. Partial differential equations describe a huge range of physical principles and they have been becoming increasingly powerful since the times of Euler and Lagrange.

The study of PDEs contains two main aspects:

1. Analytic methods for PDEs which involves the issues concerning the existence and uniqueness of solutions,
2. Numerical approximation of PDEs.

Both the mathematical analysis of the PDEs and the numerical analysis of methods rely heavily on the strong tools of functional analysis.

Numerical approximation of PDEs is a cornerstone of the mathematical modeling since almost all modeled real world problems fail to have analytic solutions or they are not known in the scope of pure mathematics because of their complexity.

The history of numerical solution of PDEs is much younger than that of analytic methods, but the development of high speed computers nowadays makes the advent of numerical methods very fast and productive.

On the other hand, the numerical approximation of PDEs often demands a knowledge of several aspects of the problem, such as the physical background of the problem, in order to understand and interpret the behavior of expected solutions, or the algorithmic aspects concerned with the choice of the numerical method and the accuracy that can be achieved.

The aim of the lecture is to discuss some modeling problems and provide the students with the knowledge of Finite Element techniques for the numerical approximation of the model equations.

Especially the theory and application of finite element methods is a very nice combination of mathematical theory with aspects of implementation, modelling, and applications. So-called “adaptive” methods enable on one hand the prescription of a tolerance for the approximation error, while on the other hand they make computations possible in

cases where, for example, a uniformly refined mesh would be prohibitively costly even on nowadays' computers, especially in three space dimensions or for problems that need the resolution of different scales.

1.2 Multiple scales in the modelling of real world problems

Most of the phenomena in nature are concerned with the behavior of a big number of *individual objects* which are always in a close interaction with each other. On the other hand, the important features are visible on a much coarser *macro scale*, where a *mean behaviour* of objects is observable.

Let us consider a very short list of fields where such behaviour arises in real life problems and where the mathematical investigation is needed to answer the questions stated by the problem.

1. Meteorology: In this field we are mainly interested in the interaction of air (oxygen, nitrogen, ozone, etc.) and water molecules, because those interactions are the ones who are responsible for the behaviour of macroscopic variables like temperature, pressure, humidity, and wind. The main objective of meteorological stations is to develop a system which permits reliable monitoring of climate changes. The monitoring is of high importance for companies like airports, off-shore wind parks, etc.

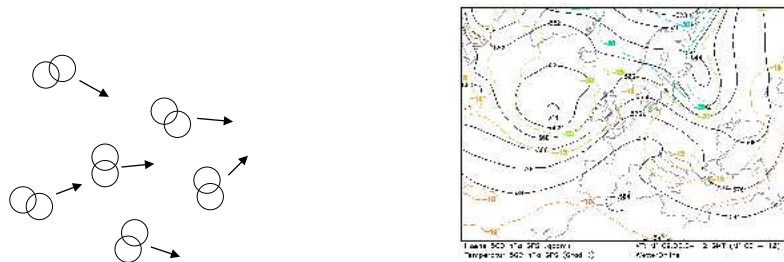


Figure 1.1: Air molecules — Weather map with isobars

2. Civil Engineering: In many aspects of our life a huge amount of different materials are used. Glass, wood, metals, or concrete, those are some instances of materials which we directly use almost every minute in our everyday life. Thus, the modification of materials and prediction of their properties are very important objectives for the manufacturers. In order to produce high quality materials the engineers in industry, among other problems, are very much interested in the elastic behavior or loading capacity of the material. While it is known that the bonding forces between the atoms of the material are responsible for the above mentioned properties, the averaged behaviour can be modelled using *continuum mechanics*.

So, to manufacture a new product with higher quality, a detailed investigation of the material on the atomic level is *not* required in most cases. A mathematical model is needed for the quantitative description of the change of material properties under external influences such as melting or cooling. The theory of differential equations comes to

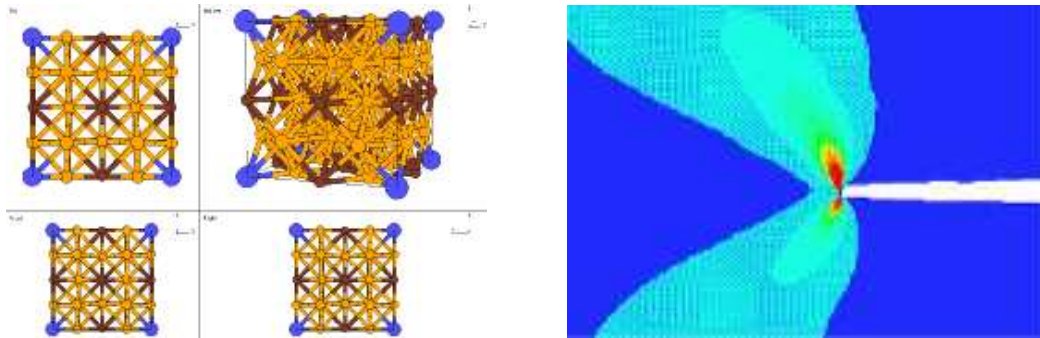


Figure 1.2: Crystal lattice of steel — Crack with inner stresses

help us as an excellent tool for the development of such a model.

3. Biology: Today the connection between biology and mathematics is a vast growing area of research. One of the target objects are bacteria. They are unicellular organisms growing on different substrates, and they gain energy from degradation of substrate components. Important is that as a result of degradation of substrate, bacteria produce other materials which are of high industrial interest, because they are used e.g. in medical care.

Another research area is the population growth problem. The problem aims in forecasting the change in a given population. The population can consist of cells in a tumor, or insects, or petri dishes.

While the behaviour of *single individuals* is the underlying principle, models for the behaviour of *the whole population* or *local parts of the population* can be derived by averaging over individuals.

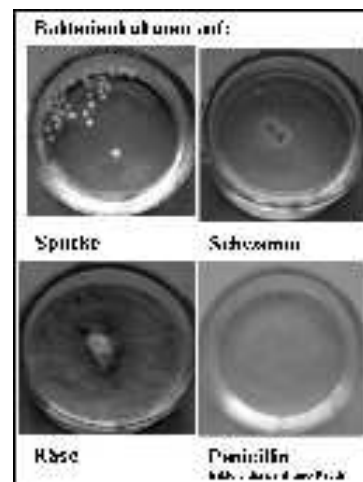


Figure 1.3: Single bacterias — Colonies in Petri dishes

The problem is very complex and old. For simplified cases, ordinary differential equations

are a wide spread method to model the population growth, but localized phenomena need the treatment with PDEs.

4. Traffic Flow: After a usual working day, many people in many countries of the world spend several hours on their way back home because of the traffic jams on the roads. During the driving process every driver has its own behavior which depends on the objectives of being fast and avoiding accidents. So, in this way a driver (with his car) interacts with other cars. But people are not able intuitively to drive in such a manner as to avoid the traffic jams on the roads.



Figure 1.4: Traffic flow

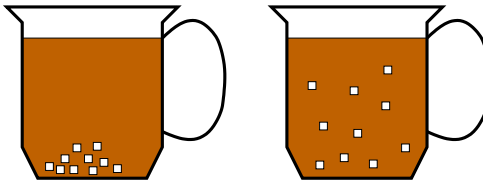
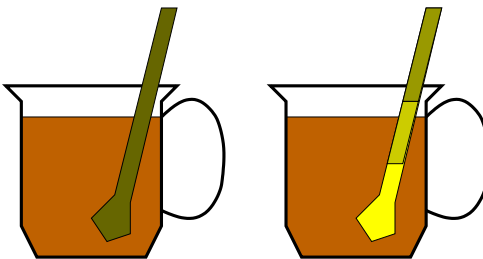
Again we need the help of mathematical model which can provide the understanding necessary to make the life of drivers more pleasant. The developed model can serve as an efficient model also for other application problems. For example, the traffic jam model is similar to gas flow models which allow for the appearance of shock waves. In aircraft traffic, analogous problems cause noise pollution near airports.

2 Mathematical modeling

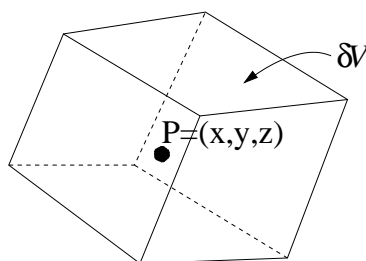
2.1 Density, flux, and conservation

The simplest mathematical models can be developed with the help of *density*, *flux* and a *conservation law*.

Density. As examples of a density we can consider some space quantities which can vary in time. Quantities like concentration of a substance or the heat density in a body are two simple examples.

- Sugar in Coffee (Concentration)

- Temperature in a spoon or in a pot (Heat Density)


The mass density is, for sure, the simplest example of density. To define the mass density (density of a material), we consider a point $P = (x, y, z)$ in the space, and let δV be a small volume element containing P .

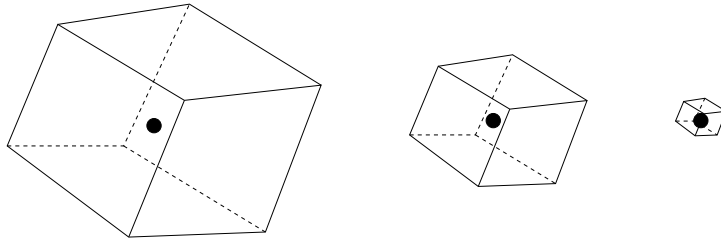


Volume Element δV Containing P

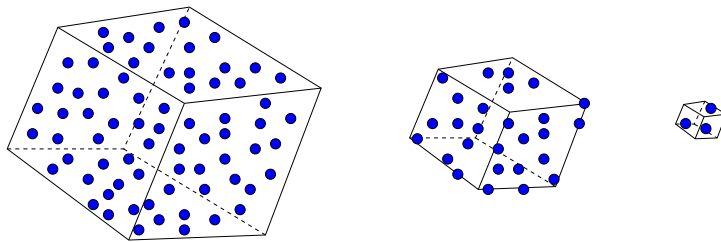
The average mass density ρ in δV at time t is equal to the mass contained in δV (which is proportional to the number of molecules), divided by the volume of $|\delta V|$:

$$\rho(\delta V, t) = \frac{\text{Mass in } \delta V \text{ at time } t}{|\delta V|}$$

In order to determine the mass density $\rho(P, t)$ in the point P at time t , we should allow δV to become smaller and smaller.



In reality we can not make the volume element arbitrarily small, as for a very small element δV a few number of molecules might remain inside δV (it can become even empty). Thus, the mass density in a point is kind of theoretical property, it is just an idea.

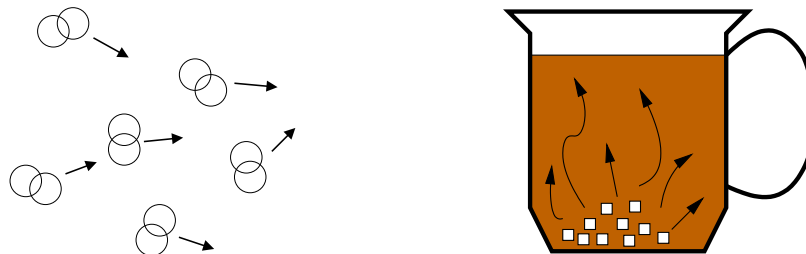


On the other hand, we know that there are 10^{23} molecules in 1 cm^3 of water, therefore, it is possible to make δV small enough to define reasonable average values of the density.

Analogous to mass density one can define

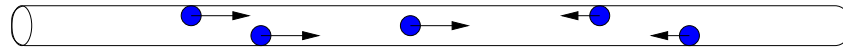
- Electrical charge density,
- Population density (biological organisms),
- Chemical concentration (e.g. of components in a mixture),
- Energy density $e(P, t)$ (for example thermal energy / temperature).

Flux. It is known that single objects like molecules or organisms are in continuous movement.



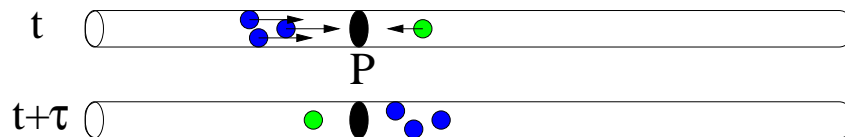
So, we want to define the flow vector (flux) $\mathbf{q}(P, t)$ in a point P at time t to be the rate and direction of average movement of the objects. Like in the case of density, the flux can be also defined through a limiting process.

For the simplicity, first we discuss the one dimensional case. Suppose we want to consider the heat conduction in a rod. Similarly you can imagine a fluid flow in a tube.



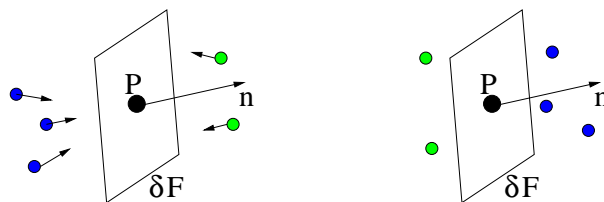
Let $\delta I = [t, t + \Delta t)$ a (small) time interval starting at t . We define the mass flux (or the fluid flow vector) $q(P, \delta I)$ to be the total mass, which moves from left to right through the point P in the time interval δI , divided by the length $|\delta I| = \Delta t$ of the time interval:

$$q(P, \delta t) = \frac{\text{Mass moved through } P \text{ from left in time interval } \delta I}{\Delta t}$$



Letting Δt get smaller and smaller, we arrive to the limiting value of mass flux rate $q(P, t)$ through P at time t . $q(P, t)$ is negative, when the amount of mass which passed from right to left is bigger than from left to right. Thus, the sign of the flux vector in one dimensional case shows the direction of the movement of the objects.

Now let us consider the **higher dimensional case**. Consider a surface element δF containing the point P with a unit normal vector $\mathbf{n}_{\delta F}$ at the point P . Analog to 1D case, we consider again the number of objects which move from one side of δF to another through the surface element δF in a time interval δI .

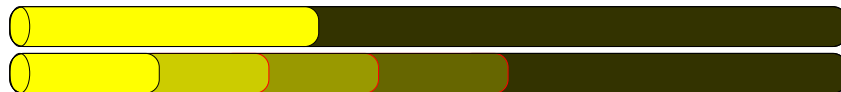


Then the flux in the direction $\mathbf{n}_{\delta F}$ is defined as

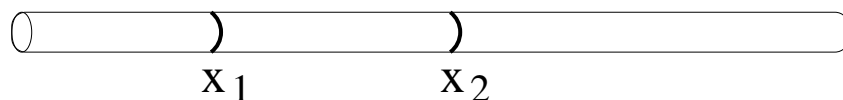
$$\mathbf{q}(\delta F, \delta I) \cdot \mathbf{n}_{\delta F} = \frac{\text{Number of objects passing across } \delta F \text{ in } \delta I}{|\delta F| \Delta t}$$

Again tending the size of the element surface and Δt to zero, we get in limit the flux $\mathbf{q}(P, t) \cdot \mathbf{n}$. Here the flow vector $\mathbf{q}(P, t)$ determines the direction and strength of the flow in the point P at time t .

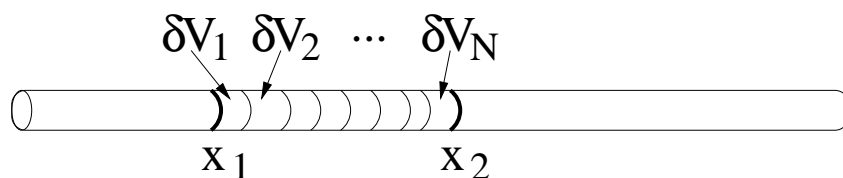
Conservation Laws. We start with one dimensional case. Suppose we want to derive a conservation of energy (e.g. heat) in a rod.



For this purpose, we consider an arbitrary section $[x_1, x_2]$ of the rod.



Next we divide the section $[x_1, x_2]$ into volume elements $\delta V_1, \delta V_2, \dots, \delta V_N$.



Note that the total energy at time instant t is equal to the sum of energies in the volume elements $\delta V_1, \delta V_2, \dots, \delta V_N$. Thus

$$E_{[x_1, x_2]}(t) = \sum_{i=1}^N E(\delta V_i, t) = \sum_{i=1}^N |\delta V_i| \underbrace{\frac{E(\delta V_i, t)}{|\delta V_i|}}_{\text{Energy density } e(\delta V_i, t)}$$

For $|\delta V| \rightarrow 0$ we get

$$E_{[x_1, x_2]}(t) = \int_{x_1}^{x_2} e(x, t) dx$$

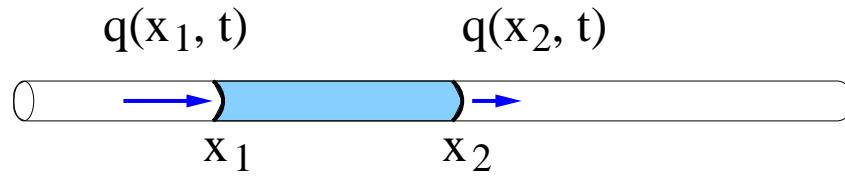
We state now the general conservation law for the case when no source and sink terms are present. The basic law of conservation of heat energy for the section $[x_1, x_2]$ can be expressed in the following way

Conservation Law:

The net accumulation of the energy in the section is equal to the input across the end points of the section minus the output across the end points of the section

Energy flow from the section $[x_1, x_2]$ through the end points is the following

- in x_1 flows $q(x_1, t)$ to right (inside of the section, input),
- in x_2 flows $q(x_2, t)$ to right (outside of the section, output).



Thus

$$\frac{d}{dt}E_{[x_1, x_2]}(t) = q(x_1, t) - q(x_2, t)$$

If we assume that q is a differentiable function with respect to x , we obtain

$$q(x_2, t) - q(x_1, t) = \int_{x_1}^{x_2} \frac{\partial}{\partial x} q(x, t) dx$$

and from equation

$$\frac{d}{dt}E_{[x_1, x_2]}(t) = \frac{\partial}{\partial t} \int_{x_1}^{x_2} e(x, t) dx = \int_{x_1}^{x_2} \frac{\partial}{\partial t} e(x, t) dx$$

follows

$$\int_{x_1}^{x_2} \frac{\partial}{\partial t} e(x, t) dx = - \int_{x_1}^{x_2} \frac{\partial}{\partial x} q(x, t) dx$$

or

$$\int_{x_1}^{x_2} \left(\frac{\partial}{\partial t} e(x, t) + \frac{\partial}{\partial x} q(x, t) \right) dx = 0$$

The last equation is valid for all sections $[x_1, x_2]$, therefore, due to the result from the calculus course, the integrand must vanish:

$$\frac{\partial}{\partial t} e(x, t) + \frac{\partial}{\partial x} q(x, t) = 0$$

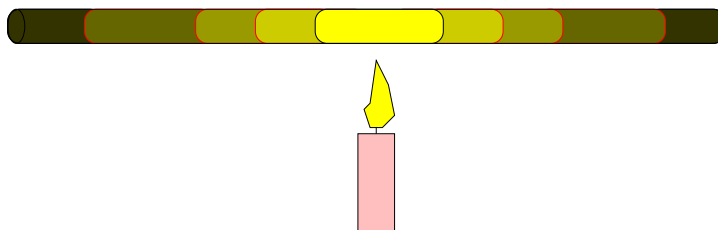
for all points x and for any time t .

This is a partial differential equation, named the

Conservation Equation:

$$\frac{\partial}{\partial t} e(x, t) = - \frac{\partial}{\partial x} q(x, t)$$

All above considerations were done when no sources or sinks are present. Let assume that the 1D rod is additionally heated with a candle from outside.



What he have to do is to add a *source* term f (function describing the heating by the candle) to the right hand side of the conservation equation:

<p>rate of change of energy = space change of the flux + Production</p> $\frac{\partial}{\partial t}e(x, t) = -\frac{\partial}{\partial x}q(x, t) + f(x, t)$

It is obvious that the source may be negative (a *sink*), $f < 0$, when we cool down the heated rod (e.g. by spraying water onto the rod).

Some examples of sources and sinks in other models are:

- Chemistry:
Production and consumption of substrates through chemical reactions
- Biology:
Creation of new biological individuals (e.g. reproduction of bacterias or their die off),
- Traffic Flow:
Driving in and out of the road,
- ...

Conservation law in n dimensions:

Consider a n -dimensional cube with faces which are parallel to the coordinate planes. Then the accumulation in the cube is given by the in- and outflow over all faces. Using the fact that the unit normal vectors of the faces are the unit vectors $\pm e_1, \dots, \pm e_n$, the flux in direction e_i is just the i -th component of the flux vector, q_i .

With similar arguments as above, we can model the accumulation by the sum of partial derivatives of the flux vector components, its *divergence*: Let $x = (x_1, \dots, x_n)$, then

$$\frac{\partial}{\partial t}e(x, t) = -\sum_{i=1}^n \frac{\partial}{\partial x_i}q_i(x, t) + f(x, t),$$

which can equivalently be written, using the notion of *divergence*, as

$$\frac{\partial}{\partial t}e(x, t) = -\operatorname{div} q(x, t) + f(x, t)$$

2.2 PDEs as a modeling tool

Heat conduction equation (1D). If the left end of the rod is at a higher temperature, then heat energy will be transferred from left to right across the rod toward the colder part of the rod. Note that in conduction process we do not consider any any motion of the material as a whole. The quantity that is conserved in the theory of heat conduction is the thermal energy. The energy density $e(x, t)$ per unit mass of a material in the interval $[x_1, x_2]$ of density ρ and specific heat c_p depends on the temperature T via

$$e(x, t) = \rho c_p T(x, t)$$

The heat flux tries to equidistribute heat over a piece of material. Thus, heat flows from warmer parts to cooler ones. This mechanism can be modelled by the empirical Fourier's law for heat conduction, which states that the rate of heat flow through a homogeneous medium is directly proportional to the temperature difference along the path of heat flow, $\partial T/\partial x$ i.e.

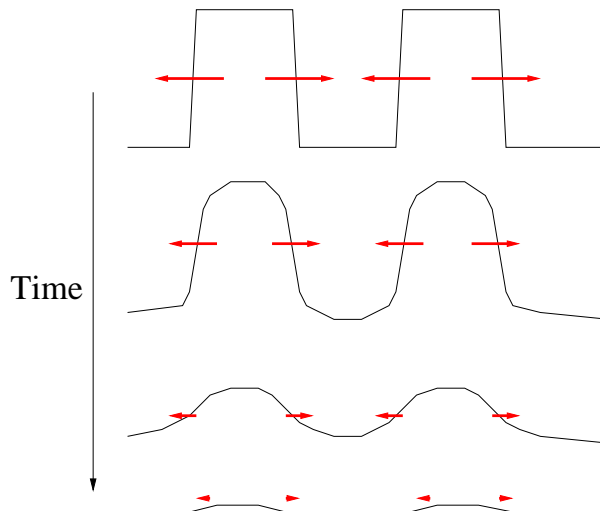
$$q(x, t) = -k \frac{\partial}{\partial x} T(x, t)$$

where $k > 0$ is a material parameter and is called the *thermal conductivity*.

Finally, after some calculations, we arrive to the one-dimensional heat conduction equation

$$\rho c_p \frac{\partial T}{\partial t} = k \frac{\partial^2 T}{\partial x^2} + f$$

The temperature balancing through the heat flow without sources is illustrated in the next figure.



Heat conduction equation (3D): The derivation of the heat conduction equation in 3D case follows from similar steps. We should only substitute the energy density $e(x, t)$ and heat flow vector $q(x, t)$ in the conservation equation with their corresponding expressions in three-dimensional space. The Fourier's law now can be written as

$$\begin{aligned} \mathbf{q}(x, t) &= \begin{pmatrix} -k \frac{\partial}{\partial x_1} T(x, t) \\ -k \frac{\partial}{\partial x_2} T(x, t) \\ -k \frac{\partial}{\partial x_3} T(x, t) \end{pmatrix} \\ &= -k \nabla T(x, t) \end{aligned}$$

The thermal energy density $e(x, t)$, analog to one-dimensional case, for domain Ω is equal to

$$e(x, t) = c_p \rho T(x, t)$$

Substituting both quantities in the conservation equation and assuming the material parameters to be constant, we obtain the three-dimensional heat conduction equation

$$\rho c_p \frac{\partial T}{\partial t} = \operatorname{div}(k \nabla T) + f$$

Since we assume that k is a constant, we can write this also using the Laplace operator,

$$\rho c_p \frac{\partial T}{\partial t} = k \Delta T + f$$

Poisson equation: Imagine we have some fluid which is incompressible and irrotational in a domain $\Omega \subset \mathbb{R}^3$ and we would like to follow the flow of the fluid. This is an often met problem in several areas of fluid mechanics. Assume that the fluid has a density $\rho(x, t)$ and the flow is defined by the vector velocity field $v(x, t)$. Further we assume that there are some mass sources per unit volume equal to $Q(x, t)$. Following the mass conservation law introduced in previous section we can write the differential equation

$$(2.1) \quad \rho_t - \operatorname{div}(\rho v) = Q$$

The definition of an incompressible flow yields that the density is constant, so we get

$$(2.2) \quad \operatorname{div} v = Q/\rho := \tilde{Q}$$

Now we use the fact that the fluid is irrotational, i.e. $\operatorname{curl} v = 0$. Then the velocity vector v is the gradient of a scalar potential $u(x)$, that is

$$(2.3) \quad v = \nabla u.$$

Finally, from (2.2) and (2.3) we obtain the Poisson equation

$$(2.4) \quad \Delta u = \tilde{Q},$$

a scalar equation for the potential of the flow. Once u is determined by solving the Poisson problem, then v can be computed by (2.3).

Remark 2.1. *Mathematically speaking, the conservation equation must be supplemented with some additional conditions, i.e. initial conditions (e.g. the temperature distribution in the material at time $t = 0$) and boundary conditions (e.g. the temperature on the boundary or the flux across the boundary, etc.).*

Without these supplemental conditions, we cannot expect to have a unique solution to the problem.

Stefan problem The classical Stefan problem is a typical free boundary problem which describes the melting of ice surrounded by water. This problem admits the following mathematical formulation. Given a container $\Omega \subset \mathbb{R}^n$ with liquid (water) with a temperature distribution $T_0 = T_0(x)$. An initial piece of ice inside the water is described by the curve (surface) Γ_0 (see figure 2.2). A model for the melting of ice in the water can

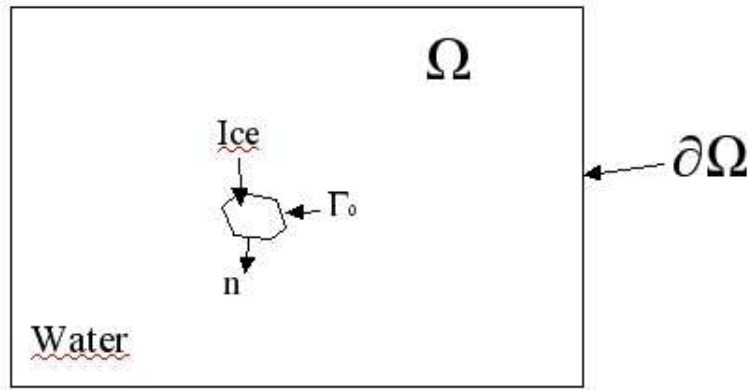


Figure 2.1: Melting of ice in the Water

be formulated in the following way:

given: $\Omega \subset \mathbb{R}^n$, Γ_0 and $T_0 = T_0(x)$ ($x \in \bar{\Omega}$). At a time $t > 0$, Ω is composed of two subdomains Ω^s and Ω^l , appropriately occupied by the solid (ice) and liquid (water) phases. The subdomains are separated by a regular interface Γ_t .

compute: $T(x, t)$ and $\Gamma(t)$, $t > 0$, such that in each of the subdomains $\Omega^{s,l}$ the heat equation is fulfilled

$$(2.5) \quad \rho c \frac{\partial T}{\partial t} - \operatorname{div}(k \nabla T) = f$$

where $f(x, t)$ is a given function describing heat sources, ρ denotes the density, k the heat conductivity, and c the heat capacity. All these coefficients are assumed to be constant in each of the phases, but may differ between them.

In addition to heat equation we must impose boundary and initial conditions. On the unknown interface Γ_t between phases two conditions are prescribed. First,

$$T = 0 \quad \text{on } \Gamma_t,$$

the temperature is equal to the melting temperature of the ice. The second condition follows from the energy conservation law by its application to elementary volumes that contain both phases at the same time. Let us consider an element $d\gamma$ of interface that moves with velocity v , and denote by q_l the heat flux (per unit surface) contributed by the liquid phase and by q_s the heat flux (again per unit surface) absorbed by the solid phase, and L being the constant latent heat of melting. Latent heat is either absorbed or released at a rate $Lv \cdot nd\gamma$. The heat exchanged by the interface Γ_t itself through $d\gamma$ is equal to $(q_l \cdot n - q_s \cdot n)d\gamma$. Applying the energy conservation law to the elementary surface $d\gamma$, we obtain

$$(q_l \cdot n - q_s \cdot n) d\gamma = Lv \cdot nd\gamma$$

This yields (dividing both sides by $d\gamma$ and using the Fourier law for the heat flux) the classical *Stefan condition* on the moving interface.

$$\rho Lv \cdot n = k \frac{\partial T}{\partial n}|_l - k \frac{\partial T}{\partial n}|_s \quad \text{on } \Gamma_t.$$

On the fixed boundary $\partial\Omega$ we impose a Dirichlet boundary condition, that is

$$(2.6) \quad T = T_D$$

As for initial conditions, we have

$$(2.7) \quad T(x, 0) = T_0(x) \quad x \in \Omega$$

$$(2.8) \quad \Gamma(0) = \Gamma_0$$

The above formulation is called two-phase one-front Stefan problem and represents one of the simplest problem settings. Modifications are possible in different directions (one-phase Stefan problem, multi-phase multi-front Stefan problem, etc.).

Remark 2.2. *We have so far introduced some models which are derived with the help of mass conservation law. After the mathematical model is complete, several questions arise which can be answered only after a detailed investigation and mathematical analysis of the model. Typical questions are:*

- *Existence: Does a solution to the problem exist?*
- *Uniqueness: Can there be two or more (different) solutions?*
- *Regularity: How smooth are those solutions? How does the smoothness of the solution depend on the source term or the smoothness of the boundary of the domain?*
- ...

3 Functional analysis background

In the following chapter we are going to develop function spaces that are used in the weak formulation of partial differential equations. Using the main concepts of Lebesgue functional spaces we will define spaces commonly referred to as Sobolev spaces. The chapter includes only a small part of functional analysis and the theory for Sobolev spaces – just enough to be able to establish the foundation for the finite element method.

3.1 Banach spaces and Hilbert spaces

Let (X, d) be a metrical space.

Definition 3.1. A sequence $(x_k)_{k \in \mathbb{N}}$ in X is called a Cauchy sequence if and only if

$$d(x_k, x_l) \rightarrow 0 \text{ for } k, l \rightarrow \infty$$

We say that (X, d) is a complete metrical space if every Cauchy sequence in X converges to a limit in X .

Definition 3.2. (Banach Space) A normed space which is complete with respect to the induced metric is called a Banach space.

Let $\Omega \in \mathbb{R}^n$ be an open and bounded set.

Definition 3.3. (Spaces of Hölder Continuous Functions) For $0 < \lambda \leq 1$ and we define $\mathbb{C}^{m, \lambda}(\overline{\Omega})$ to be the subspace of $\mathbb{C}^m(\overline{\Omega})$ consisting of those functions f for which there exists a constant h such that

$$|D^\alpha f(x) - D^\alpha f(y)| \leq h|x - y|^\lambda, \quad x, y \in \Omega$$

for $0 \leq \alpha \leq m$. The functions from the space $\mathbb{C}^{m, \lambda}(\overline{\Omega})$ are called Hölder continuous and Lipschitz continuous for the case $\lambda = 1$.

The constant h is called the Hölder constant. The space $\mathbb{C}^{m, \lambda}(\overline{\Omega})$ is then a Banach space with norm given by

$$\|f\|_{\mathbb{C}^{m, \lambda}(\overline{\Omega})} = \|f\|_{\mathbb{C}^m(\overline{\Omega})} + \max_{0 \leq |\alpha| \leq m} \sup_{x, y \in \Omega, x \neq y} \frac{|D^\alpha f(x) - D^\alpha f(y)|}{|x - y|^\lambda}$$

Definition 3.4. (Inner Product Spaces and Hilbert Spaces) Let X be a vector space and (\cdot, \cdot) be a symmetric positive definite bilinear form (an inner product) on $X \times X$. Then X is called an inner product space and the norm on this space may be defined as

$$\|x\|_X = \sqrt{(x, x)_X}, \quad x \in X$$

If X is complete under this norm, then it is called a Hilbert space.

Denote by X' the normed dual of the space X with the norm

$$\|x'\|_{X'} = \sup \{|x'(x)| : \|x\|_X \leq 1\}.$$

The following theorem shows that there exists an isometry between a Hilbert spaces X and X' .

Theorem 3.1. (*Riesz Representation Theorem*) *Let X be a Hilbert space. Then for any continuous linear functional x' from the space X' there exists exactly one $x \in X$ such that x' can be represented as*

$$x'(y) = (x, y) \text{ for all } y \in X.$$

In this case

$$\|x'\|_{X'} = \|x\|_X$$

According to the Riesz Representation Theorem, we can identify any Hilbert space with its normed dual.

3.2 Basic concepts of Lebesgue spaces

Let Ω be a Lebesgue-measurable domain in R^n and let p be a positive real number. We denote by $\mathbb{L}^p(\Omega)$ the class of all measurable functions, defined on Ω :

$$(3.1) \quad \mathbb{L}^p(\Omega) := \{u : \|u\|_{L^p(\Omega)} < \infty\},$$

where the norm $\|u\|_{L^p(\Omega)}$ is defined in the following way: for $1 \leq p < \infty$

$$(3.2) \quad \|u\|_{L^p(\Omega)} := \left(\int_{\Omega} |u(x)|^p dx \right)^{\frac{1}{p}},$$

and for $p = \infty$ we set

$$(3.3) \quad \|u\|_{L^\infty(\Omega)} := \text{ess sup}_{x \in \Omega} \{|u(x)|\}.$$

The elements of $\mathbb{L}^p(\Omega)$ are actually equivalence classes of measurable functions satisfying (3.2) or (3.3), because we can identify all functions in $\mathbb{L}^p(\Omega)$ which are equal almost everywhere on Ω .

Theorem 3.2. *For $1 \leq p, q \leq \infty$, $\frac{1}{p} + \frac{1}{q} = 1$ and $u, v \in \mathbb{L}^p(\Omega)$, $w \in \mathbb{L}^q(\Omega)$ we have*

$$\|u + v\|_{L^p(\Omega)} \leq \|u\|_{L^p(\Omega)} + \|v\|_{L^p(\Omega)}, \quad \text{Minkowski's inequality,}$$

$$\int_{\Omega} |u(x)w(x)| dx \leq \|u\|_{L^p(\Omega)} \|w\|_{L^q(\Omega)} \quad \text{Hölder's inequality,}$$

$$\int_{\Omega} |u(x)w(x)| dx \leq \|u\|_{L^2(\Omega)} \|w\|_{L^2(\Omega)} \quad \text{Schwarz' inequality.}$$

Note that Schwarz' inequality is just the Hölder's inequality in the special case $p = q = 2$.

The following theorem gives some useful properties of \mathbb{L}^p -spaces over domains with finite volume. The proof of the theorem may be found in the references given in the bibliography.

Theorem 3.3. *Assume $1 \leq p \leq q \leq \infty$. Then*

1. $\mathbb{L}^p(\Omega)$ is a Banach space.
2. if $u \in \mathbb{L}^q(\Omega)$, then $u \in \mathbb{L}^p(\Omega)$ and

$$(3.4) \quad \|u\|_{L^p(\Omega)} \leq (\text{volume}\Omega)^{\left(\frac{1}{p}-\frac{1}{q}\right)} \|u\|_{L^q(\Omega)}.$$

3. as a consequence of (3.4) we get a useful embedding result for \mathbb{L}^p -spaces, namely

$$(3.5) \quad \mathbb{L}^q(\Omega) \hookrightarrow \mathbb{L}^p(\Omega).$$

4. if $p < \infty$ then $\mathbb{L}^p(\Omega)$ is separable.
5. $\mathbb{L}^p(\Omega)$ is reflexive if and only if $1 < p < \infty$

3.3 Weak derivatives

The classical definition of derivative contains information about the function only near the given point. Of special importance is the notion of weak or distributional derivatives which does not care about the pointwise values. Therefore, we will consider derivatives that can be interpreted as functions in the Lebesgue spaces. We know that pointwise values of functions in Lebesgue spaces are irrelevant and these functions are determined only by their global behavior. The weak derivative will be used in the development of the variational formulation of partial differential equations.

Let Ω be a domain in R^n . Denote by $\mathbb{C}_0^\infty(\Omega)$ the subset of \mathbb{C}^∞ functions with compact support in Ω . By support of the function u defined on a compact $K \subset \subset \Omega$ we mean

$$\text{supp } u = \overline{\{x \in K : u(x) \neq 0\}},$$

and if $\text{supp } u \subset \subset \Omega$ then we say that u has a compact support.

Definition 3.5. *A function u is said to be locally integrable on Ω , if it is defined on Ω almost everywhere and $u \in \mathbb{L}^1(K)$ for every compact K lying in the interior of Ω . The locally integrable function space is denoted by $\mathbb{L}_{loc}^1(\Omega)$.*

Now we are ready to define the notion of weak derivative.

Definition 3.6. The function $u \in \mathbb{L}_{loc}^1(\Omega)$ possesses a weak derivative, if there exists a function $v \in \mathbb{L}_{loc}^1(\Omega)$ such that

$$\int_{\Omega} v(x)\phi(x) dx = (-1)^{|\alpha|} \int_{\Omega} u(x)D^{\alpha}\phi(x) dx \quad \text{for all } \phi \in \mathbb{C}_0^{\infty}(\Omega).$$

where

$$D^{\alpha} = \frac{\partial^{|\alpha|} v}{\partial^{\alpha_1} x_1 \partial^{\alpha_2} x_2 \dots \partial^{\alpha_n} x_n}$$

with a multi index $\alpha = (\alpha_1, \dots, \alpha_n)$, $\alpha_i \in \mathbf{Z}$, $\alpha_i \geq 0$, $|\alpha| = \alpha_1 + \alpha_2 + \dots + \alpha_n$.

We denote the weak derivative of u by $D_w^{\alpha}u$ and define $D_w^{\alpha}u=v$ (if such a v exists, of course).

3.4 Introduction to Sobolev spaces

The Sobolev spaces which will play an important role in the variational formulation of partial differential equations are built on the function spaces $\mathbb{L}^p(\Omega)$ introduced in the previous chapter. The idea is to generalize the Lebesgue norms and spaces to include weak derivatives.

Let again Ω be an open subset of R^n .

Definition 3.7. Let m be a non-negative integer and $1 \leq p \leq \infty$. We define the Sobolev norm $\|\cdot\|_{m,p}$ for any function $u \in \mathbb{L}_{loc}^1(\Omega)$ in the following form

$$(3.6) \quad \|u\|_{m,p} = \left(\sum_{0 \leq |\alpha| \leq m} \|D_w^{\alpha}u\|_{L^p(\Omega)}^p \right)^{\frac{1}{p}} \quad \text{if } 1 \leq p < \infty,$$

$$(3.7) \quad \|u\|_{m,\infty} = \max_{0 \leq |\alpha| \leq m} \|D_w^{\alpha}u\|_{L^{\infty}(\Omega)} \quad \text{if } p = \infty$$

where we assume that the weak derivatives $D_w^{\alpha}u$ of u exist for all $|\alpha| \leq m$.

The Sobolev norm defines a norm on any vector space of functions provided we identify all functions in the case they are equal almost everywhere in Ω .

Definition 3.8. For any positive integer m and $1 \leq p \leq \infty$ we define the Sobolev spaces

$$\mathbb{W}^{m,p}(\Omega) := \{u \in \mathbb{L}_{loc}^1 : \|u\|_{m,p} < \infty\}$$

Clearly $\mathbb{W}^{0,p}(\Omega) = \mathbb{L}^p(\Omega)$. For the finite element approximation of differential equations the following space is of great importance

$$\mathbb{W}_0^{m,p}(\Omega) \equiv \text{the closure of } \mathbb{C}_0^{\infty}(\Omega) \text{ in the space } \mathbb{W}^{m,p}(\Omega)$$

For an arbitrary integer m we get the following obvious chain of embeddings

$$\mathbb{W}_0^{m,p}(\Omega) \hookrightarrow \mathbb{W}^{m,p}(\Omega) \hookrightarrow \mathbb{L}^p(\Omega)$$

In a special cases of Sobolev spaces $\mathbb{W}^{m,p}(\Omega)$, i.e. when $p = 2$, we will use the notation $\mathbb{H}^m(\Omega)$ instead of $\mathbb{W}^{m,2}\Omega$ and $\mathbb{H}_0^m(\Omega)$ instead of $\mathbb{W}_0^{m,2}\Omega$.

3.5 Some useful properties of Sobolev spaces

In this section we will present, mainly without proofs, some useful properties (useful for our further considerations) enjoyed by functions from Sobolev spaces. We will provide results in their general formulations. The special cases used in Finite Element formulations can be easily obtained with very simple calculations.

Let again $\Omega \subset R^d$ be an open, bounded domain with $\partial\Omega \in \mathbb{C}^{0,1}$.

Theorem 3.4. *The Sobolev space $\mathbb{W}^{m,p}$ is a Banach space.*

Proof. Let $\{u_n\}$ be a Cauchy sequence in $\mathbb{W}^{m,p}(\Omega)$. Then for all $|\alpha| \leq m$, $\{D_w^\alpha u_n\}$ is a Cauchy sequence with respect to the norm $\|\cdot\|_{L^p(\Omega)}$, since the $\|\cdot\|_{\mathbb{W}^{m,p}(\Omega)}$ norm is a combination of $\|\cdot\|_{L^p(\Omega)}$ norms of weak derivatives. Because of the completeness of $\mathbb{L}^p(\Omega)$ (the space $\mathbb{L}^p(\Omega)$ is a Banach space), there exists a $u^\alpha \in \mathbb{L}^p(\Omega)$ such that

$$\|D_w^\alpha u_n - u^\alpha\|_{L^p(\Omega)} \rightarrow 0 \text{ for } n \rightarrow \infty$$

Particularly, $u_n \rightarrow u^{(0,\dots,0)} =: u$ in $\mathbb{L}^p(\Omega)$. To end the proof of the theorem, it remains to show that $D_w^\alpha u$ exists and is equal to u^α .

First, note that if $v_n \rightarrow v$ in $\mathbb{L}^p(\Omega)$, then for all $\phi \in \mathbb{C}_0^\infty(\Omega)$ (using the Hölder's inequality)

$$(3.8) \quad \|v_n \phi - v \phi\|_{L^p(\Omega)} \leq \|v_n - v\|_{L^p(\Omega)} \|\phi\|_{L^\infty(\Omega)} \rightarrow 0 \text{ for } n \rightarrow \infty$$

Thus

$$(3.9) \quad \int_{\Omega} v_n(x) \phi(x) dx \rightarrow \int_{\Omega} v(x) \phi(x) dx$$

Now having in hand the definition of the weak derivative and two times applying (3.9), we obtain

$$\int_{\Omega} u^\alpha \phi dx = \lim_{n \rightarrow \infty} \int_{\Omega} (D_w^\alpha u_n) \phi dx = \lim_{n \rightarrow \infty} (-1)^{|\alpha|} \int_{\Omega} u_n \partial^\alpha \phi dx = (-1)^{|\alpha|} \int_{\Omega} u \partial^\alpha \phi dx$$

Thus

$$\int_{\Omega} u^\alpha \phi dx = (-1)^{|\alpha|} \int_{\Omega} u \partial^\alpha \phi dx$$

and the proof of the theorem is complete.

Theorem 3.5. (Sobolev Embedding Theorem). Assume $m, l \in \mathbb{N}$ and $p, q \in [1, \infty]$. Then the following statements hold:

1. if $m \geq l$ and $m - \frac{d}{p} > l - \frac{d}{q}$, then $W^{m,p}(\Omega)$ is continuously embedded in $W^{l,q}(\Omega)$, i.e. there exists a constant c , such that

$$\|u\|_{l,q} \leq c \cdot \|u\|_{m,p}$$

The number $m - \frac{d}{p}$ is called the Sobolev number.

2. if $m > l$ and $m - \frac{d}{p} > l - \frac{d}{q}$, then the embedding is compact.
3. if $m - \frac{d}{p} > k + \alpha$, then

$$\|u\|_{C^{k,\alpha}(\overline{\Omega})} \leq c \cdot \|u\|_{m,p} \quad \forall u \in W^{m,p}(\Omega)$$

i.e. the space $W^{m,p}(\Omega)$ is continuously embedded in $C^{k,\alpha}(\overline{\Omega})$.

Recall, that Sobolev functions can not be in general evaluated over lower-dimensional subsets, i.e. subsets having measure equal zero.

Theorem 3.6. (Trace Theorem). Using the notations of the previous theorem, we assume that $m > l$ and $m - \frac{d}{p} > l - \frac{d-r}{q}$. Then there exists a continuous linear embedding $\gamma : W^{m,p}(\Omega) \rightarrow W^{l,q}(S)$, where S is a smooth $(d-r)$ dimensional sub-manifold of Ω . Thus the estimate holds

$$\|\gamma(u)\|_{l,q,S} \leq c \cdot \|u\|_{m,p} \quad \forall u \in W^{m,p}(\Omega)$$

The embedding operator γ is then called the trace operator. For example, if $u \in C^\infty(\overline{\Omega})$, then the trace operator γ is determined as $\gamma(u) = u$.

For our future discussions it is useful to introduce the notation of Sobolev semi-norms.

Definition 3.9. Let m be a non-negative integer and $u \in W^{m,p}(\Omega)$. We define the Sobolev semi-norm $|\cdot|_{m,p}$

$$(3.10) \quad |u|_{m,p} = \left(\sum_{|\alpha|=m} \|D_w^\alpha u\|_{L^p(\Omega)}^p \right)^{\frac{1}{p}} \quad \text{if } 1 \leq p < \infty,$$

$$(3.11) \quad |u|_{m,\infty} = \max_{|\alpha|=m} \|D_w^\alpha u\|_{L^\infty(\Omega)} \quad \text{if } p = \infty$$

4 Variational formulation of elliptic problems

In most mathematical models of real world problems one has the difficulty with analytical investigation of the problem. The difficulties can be caused by several factors, e.g. complicated geometry, etc. Therefore, more and more often we have to apply the capacity of high performance computers for getting adequate solutions of developed mathematical models.

4.1 Variational formulation of Poisson problem

In this chapter we will discuss the weak or variational formulation of boundary value problems. Since finite element approximation methods are most naturally defined in terms of weak formulations, we briefly indicate how elliptic problems can be cast in weak form. Moreover, the weak formulation provides a relatively simple way to develop the existence and uniqueness of so-called “weak solutions”.

Let $\Omega \subset R^d$ be a bounded domain with Lipschitz boundary and $f \in C(\overline{\Omega})$. Consider the Poisson problem for the unknown function $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$:

$$(4.1) \quad -\Delta u = f \quad \text{in } \Omega,$$

$$(4.2) \quad u = 0 \quad \text{on } \partial\Omega.$$

(We impose here Dirichlet boundary condition on $\partial\Omega$, in the case of Neumann or Robin boundary condition the procedure looks very similar.)

Now let us replace the classical representation (4.1)+(4.2) by a weak or variational formulation. The idea is to multiply the equation (4.1) with a so called test function $v \in \mathbb{H}_0^1(\Omega)$ and then integrate both sides of the new equation over the domain Ω .

$$(4.3) \quad - \int_{\Omega} \Delta u v \, dx = \int_{\Omega} f v \, dx$$

Using the known rules of integration by parts and then employing the Dirichlet boundary condition (4.2), equivalently it holds

$$\int_{\Omega} \nabla u \nabla v \, dx = \int_{\Omega} f v \, dx$$

Now we can ask a solution just to fulfill this *weak* form of the poisson problem. For such a weak solution is is not longer necessary that second derivatives exist, as the equation above contains only integrals over first derivatives. Additionally, it is sufficient that the first derivatives exist only in a weak sense. Thus, we can state the *weak formulation of the Poisson problem* in the following way:

Find $u \in \mathbb{H}_0^1(\Omega)$, such that for all test functions $v \in \mathbb{H}_0^1(\Omega)$ holds

$$(4.4) \quad \int_{\Omega} \nabla u \nabla v \, dx = \int_{\Omega} f v \, dx$$

We call the solution of (4.4) the *weak (variational) solution*.

Remark 4.1. *Why we are replacing the classical formulation?*

The answer is that by writing the problem in its weak formulation it is possible to treat a larger class of problems. It is easy to see that less regularity is required for the weak solution, thus the existence proofs are simplified. The main advantage of such replacement is that for solving the problems, numerical methods can be constructed which converge with less assumptions on the regularity of the solution, for example, finite element method. Finally, it is well known that if the classical solutions exist, then the weak solutions coincide with them. However, there are a lot of cases, when the weak solution fails to be classical.

Another important reason of the replacement is that the weak formulation allows us to use the nice theory of functional analysis and easily arrive at existence results and to error estimates for approximations in different functional spaces.

4.2 Existence and uniqueness of weak solution

Before we establish the existence result of weak formulation (4.4), let us first recall some definitions and state an abstract theorem.

Let X be a Hilbert space with a scalar product $(\cdot, \cdot)_X$ and associated norm $\|\cdot\|$, namely

$$\|v\|_X := \sqrt{(v, v)_X}.$$

Further we define the dual space, X^* , to the Hilbert space X as a set of all linear functionals on X . We introduce the dual pairing $\langle \cdot, \cdot \rangle_{X^* \times X}$ in the form

$$\langle \cdot, \cdot \rangle_{X^* \times X} : X^* \times X \rightarrow \mathbb{R},$$

$$\langle l, v \rangle_{X^* \times X} := l(v) \quad \forall l \in X^*, v \in X.$$

Definition 4.1. *A mapping $a(\cdot, \cdot) : X \times X \rightarrow R$ is said to be a **bilinear form**, if for any fixed $v \in X$ each of the maps $a(v, \cdot) : X \times X \rightarrow R$ and $a(\cdot, v) : X \times X \rightarrow R$ is a linear form on X . If*

$$a(w, v) = a(v, w) \quad \forall u, w \in X,$$

*then the bilinear form $a(\cdot, \cdot)$ is called **symmetric**.*

We define also the mapping $L : X \rightarrow X^*$

$$\langle Lw, v \rangle_{X^* \times X} := a(w, v) \quad \forall w, v \in X.$$

Assume that we are given the bilinear form $a(\cdot, \cdot)$ (not necessarily symmetric) and the norm $\|\cdot\| : X \rightarrow R$ defined on the space X with the property

$$\|v\| \leq a(v, v), \quad \forall v \in X$$

In the case when the bilinear form $a(\cdot, \cdot)$ is symmetric, then

$$\|v\| := \sqrt{a(v, v)}$$

Remark 4.2. *It is easy to check that $\|v\| := \sqrt{a(v, v)}$ defines a norm in the Hilbert space X . This norm is usually called the **energy norm**.*

The problem in the solution of which we are mainly interested is the following:

Problem 4.1. *For a given $f \in X^*$, find $u \in X$ such that*

$$a(u, v) = \langle f, v \rangle \text{ for all } v \in X.$$

Equivalently we can rewrite the problem as: find $u \in X$ such that

$$Lu = f \text{ in } X^*.$$

Let the bilinear form $a(\cdot, \cdot)$ be coercive, it means there exists $C_a > 0$ such that

$$\|v\|^2 \geq C_a \|v\|_X^2 \quad \forall v \in X,$$

or in terms of the bilinear form

$$a(v, v) \geq C_a \|v\|_X^2$$

Suppose also that $a(\cdot, \cdot)$ is continuous, meaning that there exist $C_a > 0$ and $C_l > 0$ such that

$$\begin{aligned} a(v, w) &\leq C_a \|v\|_X \|w\|_X, \\ a(v, w) &\leq C_l \|v\| \cdot \|w\|_X. \end{aligned}$$

Theorem 4.1. *(Lax-Milgram lemma). Let X be a Hilbert space, $a(\cdot, \cdot) : X \times X \rightarrow R$ be a continuous coercive bilinear form and let $f : X \rightarrow R$ be a continuous linear form. Then the Problem 4.1 has one and only one solution.*

Now let apply the Lax-Milgram lemma to the weak formulation (4.4). We set

$$X = \mathbb{W}_0^1(\Omega),$$

$$a(u, v) = \int_{\Omega} \nabla u \nabla v \, dx \quad \text{for } u, v \in X,$$

and

$$\langle f, v \rangle = \int_{\Omega} f v \, dx$$

and it is up to the reader to verify that all assumptions of Lax-Milgram lemma are satisfied.

Remark 4.3. Above, the Sobolev spaces appear without much motivation, just saying that the regularity of Sobolev functions is enough to write down the formulation. You can introduce them also in a different way: The weak formulation appears naturally as the condition for the minimizer u of the energy functional

$$E(v) = \int_{\Omega} \frac{1}{2} |\nabla v|^2 - f v \, dx, \quad v \in \mathbb{C}_0^1(\Omega).$$

Trying to prove existence of a (unique) minimizer, you need completeness of the underlying space with respect to a norm which is compatible with the energy functional. This leads naturally to the completion

$$\mathbb{H}_0^1 = \overline{\mathbb{C}_0^1(\Omega)}^{\|\cdot\|_{H^1}}.$$

5 Finite element approximation

The finite element method was first conceived in a paper by Courant in 1943, but the importance of this contribution was ignored at the time. Then the engineers independently re-invented the method in the early fifties. Nowadays the whole procedure of the finite element method is a field of mathematical research since many years and it has become one of the most popular techniques for obtaining numerical solutions of differential equations arising from engineering problems.

In this chapter we will briefly introduce the main aspects of the finite element method.

5.1 Galerkin discretization

In the theory of classical solutions it is natural to use approximation procedures which are based on a pointwise evaluation of functions and differential operators. When dealing with weak solutions this approach cannot be taken over, because point values of functions in $\mathbb{H}^m(\Omega)$ are in general not defined, if $m - \frac{n}{2} \leq 0$ (m and n are the corresponding Sobolev numbers). The formulation of the original problem as weak problem suggests a different strategy to convert the infinite dimensional space into a finite dimensional one which then allows a numerical treatment.

Having in mind the old notations, let us again consider the following problem:

Problem 5.1. For a given $f \in X^*$, find $u \in X$ such that

$$a(u, v) = \langle f, v \rangle \text{ for all } v \in X.$$

Then discretization is obtained by replacing X with a *finite dimensional subspace* $X_h \subset X$. To get a numerical approximation to the unknown function u , the idea of *Galerkin method* is

Problem 5.2. For a given $f \in X^*$, find $u_h \in X_h$ such that

$$(5.1) \quad a(u_h, v_h) = \langle f, v_h \rangle \text{ for all } v_h \in X_h.$$

Existence of a *discrete solution* $u_h \in X_h$ follows directly by applying the above theory on the subspace, or by looking for a minimizer of the energy functional in the (finite-dimensional, so complete) subspace.

We introduce a basis $\{\varphi_1, \varphi_2, \dots, \varphi_n\}$ of V_h and taking into consideration the fact that (5.1) is satisfied for any $v_h \in V_h$, we replace v_h by basis functions.

$$(5.2) \quad a(u_h, \varphi_i) = \langle f, \varphi_i \rangle.$$

Now the desired approximate solution u_h is represented by means of chosen basis functions:

$$(5.3) \quad u_h = \sum_{j=1}^n U_j \varphi_j.$$

If this expression for u_h is now substituted into (5.2), we obtain the following system of equations:

$$(5.4) \quad \sum_{j=1}^n a(\varphi_j, \varphi_i) U_j = \langle f, \varphi_i \rangle \text{ for } i = 1, \dots, n,$$

which must be solved for the unknowns U_1, U_2, \dots, U_n .

The equation (5.4) permits us to write it in the form:

$$(5.5) \quad SU = B$$

with the matrix $S_{ij} = a(\varphi_j, \varphi_i)$, vector $B_i = \langle f, \varphi_i \rangle$ and U being the column vector of coefficients U_j .

S is called the *stiffness matrix* and B is called the *load vector*.

For a given space X_h , solving the corresponding discrete problem (5.1) amounts to finding the coefficients U_j of the expansion (5.3) over the basis functions φ_j , $j = 1, 2, \dots, n$. Thus, in order to obtain the numerical solution of any second order elliptic problem one has first, to compute the stiffness matrix S and load vector B for the specific problem, and second, solve the algebraic system (5.5).

Some details about aspects of efficient implementations are given in Section 10

5.2 Finite element method

The finite element method can be described in a few words. Suppose that the problem to be solved is in weak formulation. The idea of finite element method is simple. It starts by a subdivision of the structure, or the region of physical interest, into smaller pieces. These pieces must be easy for the computer to record and identify: they may be triangles or rectangles.

Then within each piece the trial functions are given an extremely simple form—normally they are polynomials of arbitrary degree. Boundary conditions are easier to impose locally along the edge of a triangle or rectangle, than globally along a more complicated boundary.

So, let us start from the first step: divide the domain into finitely many smaller pieces. These small pieces are called elements. There are several kinds of elements, which can be used for the decomposition of the domain and it is not clear whether to subdivide the region into triangles, rectangles or other types of elements. We will not discuss the advantages and disadvantages of each type of elements and will subdivide the region of interest into triangles.

If we decompose the given domain Ω by triangles (see Figure 5.1), we will see that the union of these triangles will be a polygon Ω_h and in general - if $\partial\Omega$ is a curved boundary

- there will be a nonempty region $\Omega \setminus \Omega_h$, which will later, of course, contribute to the error. So one of the main tasks when considering curved boundary, will be to make the nonempty region as small in area as possible. For simplicity we will consider only polygonal domains, i.e. the case when $\Omega_h = \Omega$.

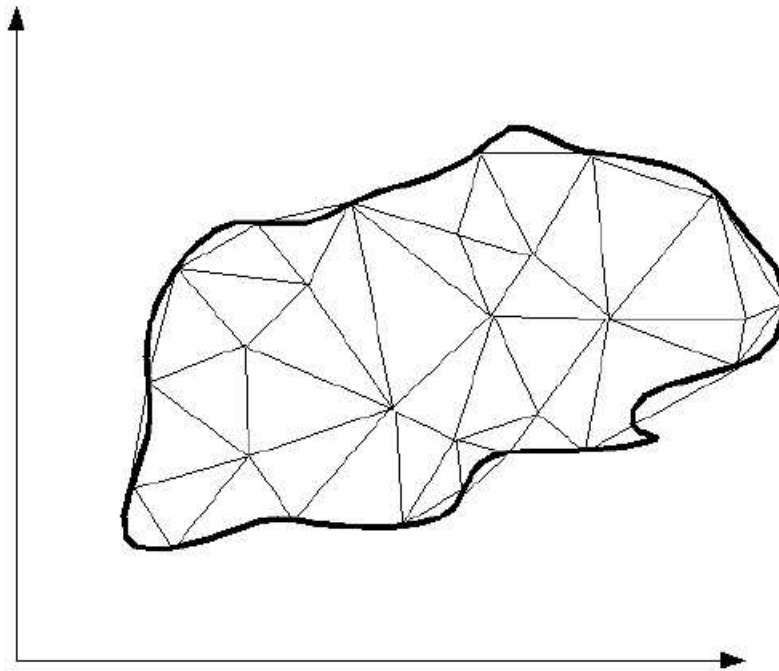


Figure 5.1: A triangulation of a domain

Definition 5.1. $\tau := \{T_1, \dots, T_{N_\tau}\}$ is called a (conforming) triangulation of Ω , if the following conditions are fulfilled: (see [21])

1. T_i are open triangles (elements) for $1 \leq i \leq N_\tau$;
2. T_i are disjoint, i.e. $T_i \cap T_j = \emptyset$ for $i \neq j$;
3. $\bigcup_{1 \leq i \leq N_\tau} \overline{T_i} = \overline{\Omega}$;
4. for $i \neq j$ the set $\overline{T_i} \cap \overline{T_j}$ is either
 - i. empty, or
 - ii. a common edge of T_i and T_j , or
 - iii. a common vertex of T_i and T_j .

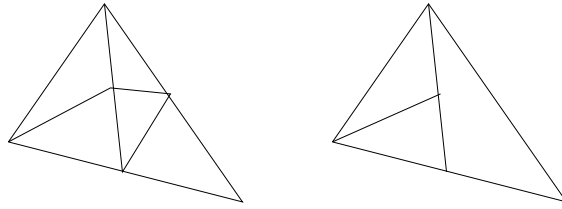


Figure 5.2: Examples of conforming and non-conforming triangulation

The examples of conforming and non-conforming triangles are given in Figure 5.2.

Let τ_0 be a triangulation of Ω . If we subdivide a subset of triangles of τ_0 into sub-triangles such that the resulting set of triangles is again a triangulation of Ω , then we call this a refinement of τ_0 . Let the new triangulation be τ_1 . If we proceed in this way, we can construct a sequence of triangulations $\{\tau_k\}_{k \geq 0}$ such that τ_{k+1} is a refinement of τ_k .

Let now τ be a *conforming triangulation*. Our next task is to define a *finite element space* X_h . For the moment we know only that X_h is a finite dimensional space of functions defined over the domain $\overline{\Omega}$. With the help of X_h we define the space

$$P_T = \{v_h|_T; v_h \in X_h\}.$$

The members of this space are the restrictions of the functions $v_h \in X_h$ to the elements (triangles) $T \in \tau$. It is natural now to obtain some conditions guaranteeing that the inclusion $X_h \subset \mathbb{H}^1(\Omega)$ holds (if you remember, our goal is to approximate solutions of problems belonging to the space $\mathbb{H}^1(\Omega)$).

Theorem 5.1. *Assume that $X_h \subset \mathbb{C}(\overline{\Omega})$ and $P_T \subset \mathbb{H}^1(T)$ for all $T \in \tau$. Then*

$$X_h \subset \mathbb{H}^1(\Omega),$$

$$X_{h0} := \{v_h \in X_h; v_h = 0 \text{ on } \partial\Omega\} \subset \mathbb{H}_0^1(\Omega)$$

Having in mind all previous considerations, we summarize the properties of a finite element space.

1. A finite element space is described by the underlying triangulation τ of the domain $\overline{\Omega}$
2. For each element T of the triangulation τ the space

$$P_T = \{v_h|_T; v_h \in X_h\}.$$

contains polynomials of certain degree.

3. there exist a canonical basis in the space X_h , whose functions are easy to describe using the information on local elements.

Example 5.1. (*Linear finite elements*).

Let assume that we are given a domain $\Omega \in R^2$ with polygonal boundary $\partial\Omega$. Let τ be some triangulation of Ω into triangles T . For an arbitrary positive integer r we define the space

$$P_r(T) := \{v; v \text{ is a polynomial of degree } \leq r \text{ on } T\}.$$

Thus for $r = 1$, $P_1(T)$ is the space of linear functions defined on T . These linear functions can be represented as

$$v(x, y) = a + bx + cy, \quad x \in T$$

where $a, b, c \in R$. Here we immediately see how the basis $\{\varphi_1, \varphi_2, \varphi_3\}$ for $P_1(T)$ looks like, compare Figures 5.3 and 5.4

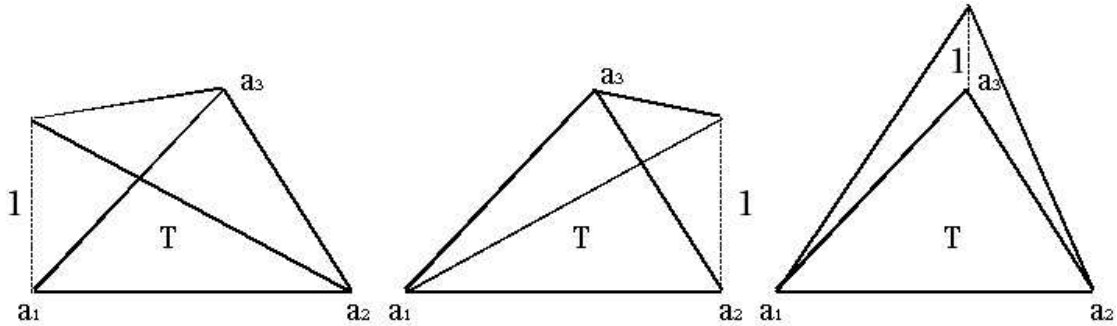


Figure 5.3: Linear basis functions for the triangle T

Note that $\dim P_1(T) = 3$. Let the finite dimensional space X_h be the space of piecewise linear functions, i.e.

$$X_h = \{v \in C(\overline{\Omega}); v|_T \in P_1(T), \forall T \in \tau\}$$

It is clear that any function $v \in X_h$ is uniquely determined by the values (called also degrees of freedom) at the vertices of T (called also the nodes). Indeed, let $T \in \tau$ be a triangle with vertices $a_1 = (x_1, y_1), a_2 = (x_2, y_2), a_3 = (x_3, y_3)$. Since the function $v \in X_h$ is defined on the arbitrary triangle T , the vertices of T must satisfy the equation for v , i.e. $v(x, y) = a + bx + cy$. If we denote the values of $v(x, y)$ at the vertices by $\alpha_i, i = 1, 2, 3$, we obtain the linear system of equations

$$a + bx_i + cy_i = \alpha_i, \quad i = 1, 2, 3,$$

for the unknowns a, b, c . From the basics of linear algebra we know that for given α_i the system has a unique solution if and only if the determinant of coefficient matrix does not vanish, i.e.

$$\det \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{pmatrix} \neq 0.$$

On the other hand it is well known that the above mentioned determinant is equal to twice the area of triangle T , thus it can not be equal to zero, which means that the unknowns a, b, c are uniquely determined and therefore, the function $v(x, y)$ is also uniquely determined by its given degrees of freedom (values at the vertices of T).

Remark 5.1. *In general, for our approximations we will use the finite element space X_{h0} for solving the second-order problems with homogeneous Dirichlet boundary and the space X_h if we are solving a second order Neumann problem.*

Remark 5.2. *A common way to define the basis functions associated with the degrees of freedom is to take functions $\varphi_i \in P_1(T)$, $i = 1, 2, 3$, such that*

$$\varphi_i(a_j) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

for $i, j = 1, 2, 3$ (see Figure 5.3).

Note, that we can analogously define other finite element spaces using the spaces of higher degree polynomials. Here we state only a theorem on general Lagrange elements.

Definition 5.2. *The Lagrange grid $G_k(T)$ on a triangle T with vertices a_0, a_1, a_2 is given by the set of points*

$$G_k(T) = \left\{ x = \sum_{j=0}^d \lambda_j a_j : \lambda_j \in \left\{ \frac{m}{k}, m = 0, \dots, k \right\}, \lambda_j \geq 0, \sum_{j=0}^d \lambda_j = 1 \right\}$$

On each triangle T , each polynomial $p \in \mathbb{P}_k$ of degree k is defined uniquely by its values on the Lagrange grid $G_k(T)$. It holds $\dim \mathbb{P}_k = \binom{k+2}{k} = \#G_k(T)$.

Theorem 5.2. *Let the domain $\Omega \in \mathbb{R}^d$ is decomposed into triangles through the triangulation \mathcal{T} . Assume that the grid G_k is of order k , it means*

$$G_k := \cup_{T \in \mathcal{T}} G_k(T) = \{a_j, j = 1, 2, \dots, N\}.$$

If the values of u_h on the grid G_k are known, then using these values we can uniquely determine a function $u_h \in X_h \subset \mathbb{H}^1(\Omega)$ with

$$X_h = \{u_h \in \mathbb{C}^0(\overline{\Omega}); u_h|_T \in \mathbb{P}_k(T), T \in \mathcal{T}\},$$

A basis of X_h is given as a collection of functions $\varphi_j \in X_h$ such that

$$\varphi_j(a_i) = \delta_{ji} \quad i, j = 1, 2, \dots, N.$$

where δ_{ji} is the well know Kronecker delta function.

The basis functions on the given triangulation for linear, quadratic and 4th order finite elements are visualized in figures 5.4, 5.5 and 5.6.

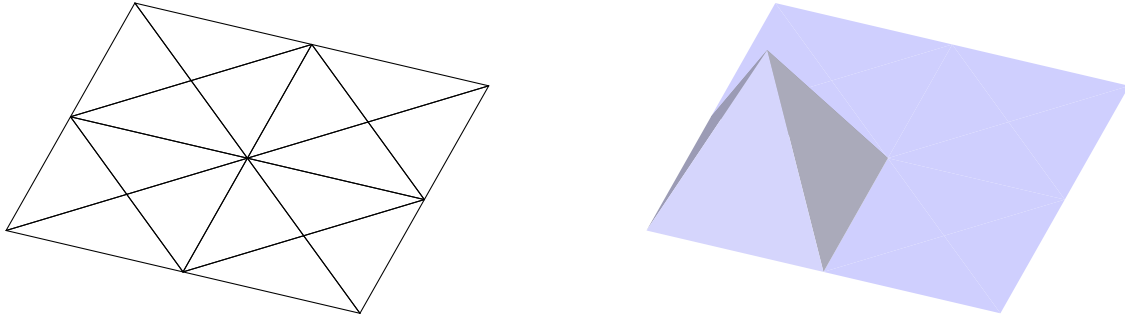


Figure 5.4: Mesh and linear basis function

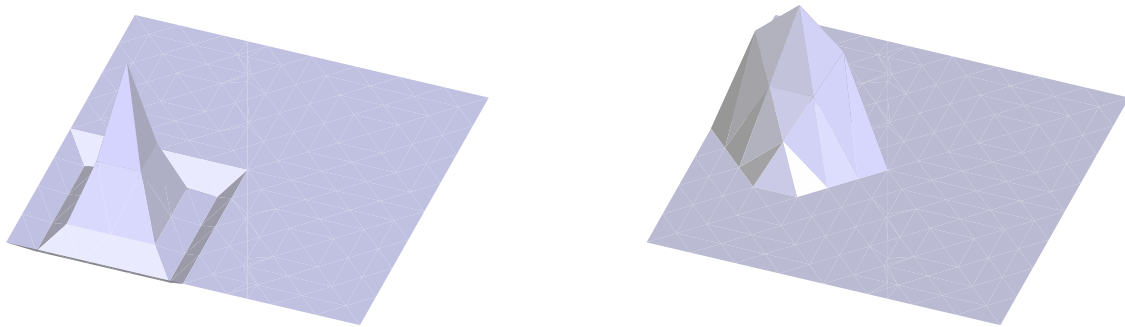


Figure 5.5: Quadratic basis functions

5.3 Discretisation of 2nd order equation

In this section we describe the assembling of the discrete system in detail.

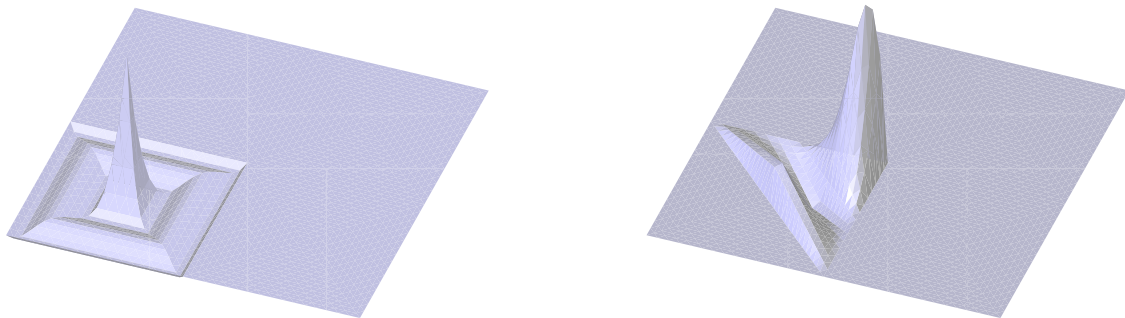
We consider the following second order differential equation in divergence form:

$$(5.6a) \quad Lu := -\nabla \cdot A \nabla u + b \cdot \nabla u + c u = f \quad \text{in } \Omega,$$

$$(5.6b) \quad u = g \quad \text{on } \Gamma_D,$$

$$(5.6c) \quad \nu_\Omega \cdot A \nabla u = 0 \quad \text{on } \Gamma_N,$$

where $A \in L^\infty(\Omega; \mathbb{R}^{d \times d})$, $b \in L^\infty(\Omega; \mathbb{R}^d)$, $c \in L^\infty(\Omega)$, and $f \in L^2(\Omega)$. By $\Gamma_D \subset \partial\Omega$ (with $|\Gamma_D| \neq 0$) we denote the Dirichlet boundary and we assume that the Dirichlet boundary values $g: \Gamma_D \rightarrow \mathbb{R}$ have an extension to some function $g \in \mathbb{H}^1(\Omega)$.



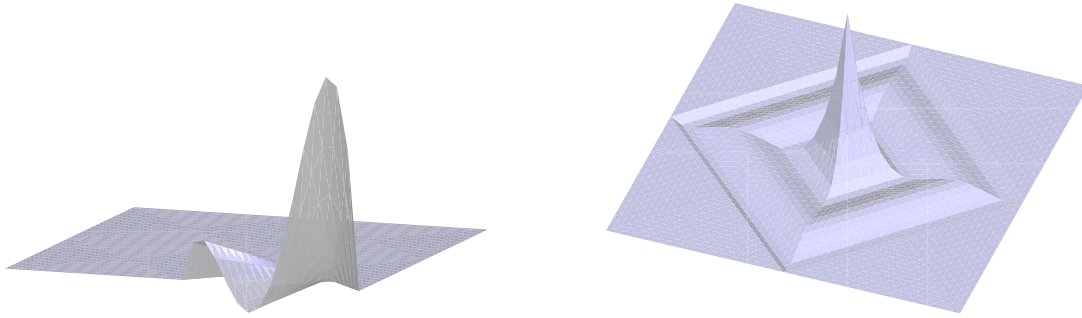


Figure 5.6: 4th order basis functions

By $\Gamma_N = \partial\Omega \setminus \Gamma_D$ we denote the Neumann boundary, and by ν_Ω we denote the outer normal vector on $\partial\Omega$. The boundary condition (5.6c) is a so called *natural* Neumann condition.

Equations (5.6) describe not only a simple model problem. The same kind of equations result from a linearization of nonlinear elliptic problems (for example by a Newton method) as well as from a time discretization scheme for (non-) linear parabolic problems.

Setting

$$(5.7) \quad X = \mathbb{H}^1(\Omega) \quad \text{and} \quad \mathring{X} = \mathbb{H}_0^1(\Omega) = \{v \in \mathbb{H}^1(\Omega); v = 0 \text{ on } \Gamma_D\}$$

this equation has the following weak formulation: We are looking for a solution $u \in X$, such that $u \in g + \mathring{X}$ and

$$(5.8) \quad \int_{\Omega} (\nabla\varphi(x)) \cdot A(x)\nabla u(x) + \varphi(x) b(x) \cdot \nabla u(x) + c(x) \varphi(x) u(x) dx = \int_{\Omega} f(x) \varphi(x) dx$$

for all $\varphi \in \mathring{X}$

Denoting by \mathring{X}^* the dual space of \mathring{X} we identify the differential operator L with the linear operator $L \in \mathcal{L}(X, \mathring{X}^*)$ defined by

$$(5.9) \quad \langle Lv, \varphi \rangle_{\mathring{X}^* \times \mathring{X}} := \int_{\Omega} \nabla\varphi \cdot A\nabla v + \int_{\Omega} \varphi b \cdot \nabla v + \int_{\Omega} c \varphi v \quad \text{for all } v, \varphi \in \mathring{X}$$

and the right hand side f with the linear functional $f \in \mathring{X}^*$ defined by

$$(5.10) \quad \langle F, \varphi \rangle_{\mathring{X}^* \times \mathring{X}} := \int_{\Omega} f \varphi \quad \text{for all } \varphi \in \mathring{X}.$$

With these identifications we use the following reformulation of (5.8): Find $u \in X$ such that

$$(5.11) \quad u \in g + \mathring{X} : \quad Lu = f \quad \text{in } \mathring{X}^*$$

holds.

Suitable assumptions on the coefficients imply that L is elliptic, i.e. there is a constant $C = C_{A,b,c,\Omega}$ such that

$$\langle L\varphi, \varphi \rangle_{\mathring{X}^* \times \mathring{X}} \geq C \|\varphi\|_{\mathring{X}}^2 \quad \text{for all } \varphi \in \mathring{X}.$$

The existence of a unique solution $u \in X$ of (5.11) is then a direct consequence of the Lax–Milgram–Theorem.

We consider a finite dimensional subspace $X_h \subset X$ for the discretization of (5.11) with $N = \dim X_h$. We set $\mathring{X}_h = X_h \cap \mathring{X}$ with $\mathring{N} = \dim \mathring{X}_h$. Let $g_h \in X_h$ be an approximation of $g \in X$. A discrete solution of (5.11) is then given by: Find $u_h \in X_h$ such that

$$(5.12) \quad u_h \in g_h + \mathring{X}_h : \quad Lu_h = f \quad \text{in } \mathring{X}_h^*,$$

i.e.

$$u_h \in g_h + \mathring{X}_h : \quad \langle Lu_h, \varphi_h \rangle_{\mathring{X}_h^* \times \mathring{X}_h} = \langle f, \varphi_h \rangle_{\mathring{X}_h^* \times \mathring{X}_h} \quad \text{for all } \varphi_h \in \mathring{X}_h$$

holds. If L is elliptic, we have a unique discrete solution $u_h \in X_h$ of (5.12), again using the Lax–Milgram–Theorem.

Choose a basis $\{\varphi_1, \dots, \varphi_N\}$ of X_h such that $\{\varphi_1, \dots, \varphi_{\mathring{N}}\}$ is a basis of \mathring{X}_h . For a function $v_h \in X_h$ we denote by $\mathbf{v} = (v_1, \dots, v_N)$ the coefficient vector of v_h with respect to the basis $\{\varphi_1, \dots, \varphi_N\}$, i.e.

$$v_h = \sum_{j=1}^N v_j \varphi_j.$$

Using (5.12) with test functions φ_i , $i = 1, \dots, \mathring{N}$, we get the following N equations for the coefficient vector $\mathbf{u} = (u_1, \dots, u_N)$ of u_h :

$$(5.13a) \quad \sum_{j=1}^N u_j \langle L\varphi_j, \varphi_i \rangle_{\mathring{X}_h^* \times \mathring{X}_h} = \langle f, \varphi_i \rangle_{\mathring{X}_h^* \times \mathring{X}_h} \quad \text{for } i = 1, \dots, \mathring{N},$$

$$(5.13b) \quad u_i = g_i \quad \text{for } i = \mathring{N} + 1, \dots, N.$$

Defining the *system matrix* \mathbf{L} by

$$(5.14) \quad \mathbf{L} := \begin{bmatrix} \langle L\varphi_1, \varphi_1 \rangle & \dots & \langle L\varphi_{\mathring{N}}, \varphi_1 \rangle & \langle L\varphi_{\mathring{N}+1}, \varphi_1 \rangle & \dots & \langle L\varphi_N, \varphi_1 \rangle \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \langle L\varphi_1, \varphi_{\mathring{N}} \rangle & \dots & \langle L\varphi_{\mathring{N}}, \varphi_{\mathring{N}} \rangle & \langle L\varphi_{\mathring{N}+1}, \varphi_{\mathring{N}} \rangle & \dots & \langle L\varphi_N, \varphi_{\mathring{N}} \rangle \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \vdots & \ddots & 0 & 0 & 0 & \ddots & \vdots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

and the *right hand side vector* or *load vector* \mathbf{f} by

$$(5.15) \quad \mathbf{f} := \begin{bmatrix} \langle f, \varphi_1 \rangle \\ \vdots \\ \langle f, \varphi_N \rangle \\ g_{N+1} \\ \vdots \\ g_N \end{bmatrix},$$

we can write (5.13) as the linear $N \times N$ system

$$(5.16) \quad \mathbf{L} \mathbf{u} = \mathbf{f},$$

which has to be assembled and solved numerically.

5.4 Simplices of arbitrary dimension

Above, we considered for simplicity only triangulations of 2-dimensional domains, built out of triangles or 2-simplices. But the concept of finite elements and triangulations can be used in any space dimension.

Definition 5.3 (Simplex). *a) Let $a_0, \dots, a_d \in \mathbb{R}^n$ be given such that $a_1 - a_0, \dots, a_d - a_0$ are linear independent vectors in \mathbb{R}^n . The convex set*

$$(5.17) \quad T = \text{conv hull}\{a_0, \dots, a_d\}$$

is called a d -simplex in \mathbb{R}^n . For $k < d$ let

$$(5.18) \quad T' = \text{conv hull}\{a'_0, \dots, a'_k\} \subset \partial T$$

be a k -simplex with $a'_0, \dots, a'_k \in \{a_0, \dots, a_d\}$. Then T' is called a k -sub-simplex of T . A 0-sub-simplex is called vertex, a 1-sub-simplex edge and a 2-sub-simplex face.

b) The standard simplex in \mathbb{R}^d is defined by

$$(5.19) \quad \hat{T} = \text{conv hull}\{\hat{a}_0 = 0, \hat{a}_1 = e_1, \dots, \hat{a}_d = e_d\},$$

where e_i are the unit vectors in \mathbb{R}^d .

c) Let $F_T: \hat{T} \rightarrow T \subset \mathbb{R}^n$ be an invertible, differentiable mapping. Then T is called a parametric d -simplex in \mathbb{R}^n . The k -sub-simplices T' of T are given by the images of the k -sub-simplices \hat{T}' of \hat{T} . Thus, the vertices a_0, \dots, a_d of T are the points $F_T(\hat{a}_0), \dots, F_T(\hat{a}_d)$.

d) For a d -simplex T , we define

$$(5.20) \quad h_T := \text{diam}(T) \quad \text{and} \quad \rho_T := \sup\{2r; B_r \subset T \text{ is a } d\text{-ball of radius } r\},$$

the diameter and in-ball-diameter of T .

Let T be an element of the triangulation with vertices $\{a_0, \dots, a_d\}$; let $F_T: \hat{T} \rightarrow T$ be the diffeomorphic parameterization of T over \hat{T} with regular Jacobian DF_T , such that

$$F(\hat{a}_k) = a_k, \quad k = 0, \dots, d$$

holds. For a point $x \in T$ we set

$$(5.21) \quad \hat{x} = F_T^{-1}(x) \in \hat{T}.$$

For a simplex T the easiest choice for F_T is the unique affine mapping

$$F_T(\hat{x}) = A_T \hat{x} + a_0$$

where the matrix $A_T \in \mathbb{R}^{n \times d}$ is defined for any d -simplex T in \mathbb{R}^n as

$$A_T = \begin{bmatrix} \vdots & & \vdots \\ a_1 - a_0 & \cdots & a_d - a_0 \\ \vdots & & \vdots \end{bmatrix},$$

Since F_T is affine linear it is differentiable. It is easy to check that $F_T: \hat{T} \rightarrow T$ is invertible and that $F_T(\hat{a}_i) = a_i$, $i = 0, \dots, d$ holds. For an affine mapping, DF_T is constant. In the following, we assume that the parameterization F_T of a simplex T is affine.

For a simplex T the barycentric coordinates

$$\lambda^T(x) = (\lambda_0^T, \dots, \lambda_d^T)(x) \in \mathbb{R}^{d+1}$$

of some point $x \in \mathbb{R}^d$ are (uniquely) determined by the $(d+1)$ equations

$$\sum_{k=0}^d \lambda_k^T(x) a_k = x \quad \text{and} \quad \sum_{k=0}^d \lambda_k^T(x) = 1.$$

The following relation holds:

$$x \in T \quad \text{iff} \quad \lambda_k^T(x) \in [0, 1] \text{ for all } k = 0, \dots, d \quad \text{iff} \quad \lambda^T \in \bar{T}.$$

On the other hand, each $\lambda \in \bar{T}$ defines a unique point $x^T \in T$ by

$$x^T(\lambda) = \sum_{k=0}^d \lambda_k a_k.$$

Thus, $x^T: \bar{T} \rightarrow T$ is an invertible mapping with inverse $\lambda^T: T \rightarrow \bar{T}$. The barycentric coordinates of x on T are the same as those of \hat{x} on \hat{T} , i.e. $\lambda^T(x) = \lambda^{\hat{T}}(\hat{x})$.

In the general situation, when F_T may not be affine, i.e. we have a parametric element, the barycentric coordinates λ^T are given by the inverse of the parameterization F_T and the barycentric coordinates on \hat{T} :

$$(5.22) \quad \lambda^T(x) = \lambda^{\hat{T}}(\hat{x}) = \lambda^{\hat{T}}(F_T^{-1}(x))$$

and the *world coordinates* of a point $x^T \in T$ with barycentric coordinates λ are given by

$$(5.23) \quad x^T(\lambda) = F_T \left(\sum_{k=0}^d \lambda_k \hat{a}_k \right) = F_T(x^{\hat{T}}(\lambda))$$

(see also Figure 5.7).

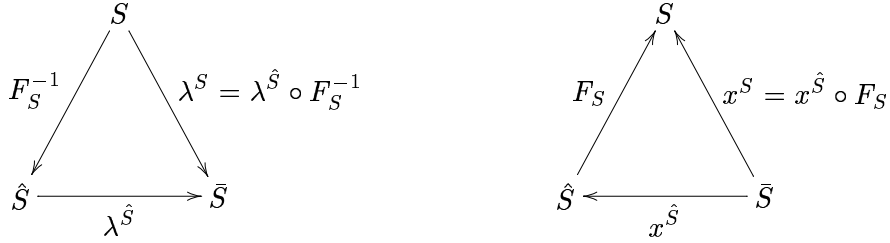


Figure 5.7: Definition of $\lambda^T: T \rightarrow \bar{T}$ via F_T^{-1} and $\lambda^{\hat{T}}$, and $x^T: \bar{T} \rightarrow T$ via $x^{\hat{T}}$ and F_T

Every function $f: T \rightarrow V$ defines (uniquely) two functions

$$\begin{array}{ccc} \bar{f}: \bar{T} & \rightarrow & V \\ \lambda & \mapsto & f(x^T(\lambda)) \end{array} \quad \text{and} \quad \begin{array}{ccc} \hat{f}: \hat{T} & \rightarrow & V \\ \hat{x} & \mapsto & f(F_T(\hat{x})). \end{array}$$

Accordingly, $\hat{f}: \hat{T} \rightarrow V$ defines two functions $f: T \rightarrow V$ and $\bar{f}: \bar{T} \rightarrow V$, and $\bar{f}: \bar{T} \rightarrow V$ defines $f: T \rightarrow V$ and $\hat{f}: \hat{T} \rightarrow V$.

Assuming that a function space $\bar{P} \subset \mathbb{C}^0(\bar{T})$ is given, it uniquely defines function spaces \hat{P} and P_T by

$$(5.24) \quad \hat{P} = \left\{ \hat{p} \in \mathbb{C}^0(\hat{T}); \bar{p} \in \bar{P} \right\} \quad \text{and} \quad P_T = \left\{ \varphi \in \mathbb{C}^0(T); \bar{p} \in \bar{P} \right\}.$$

We can also assume that the function space \hat{P} is given and this space uniquely defines \bar{P} and P_T in the same manner. In numerical implementation it makes sense to use the function space \bar{P} on \bar{T} ; the implementation of a basis $\{\bar{p}^1, \dots, \bar{p}^m\}$ of \bar{P} is much simpler than the implementation of a basis $\{\hat{p}^1, \dots, \hat{p}^m\}$ of \hat{P} as we are able to use symmetry properties of the barycentric coordinates λ .

6 A priori error estimates for elliptic problems

In the following chapter we will consider general linear elliptic partial differential equations and introduce an abstract a priori error estimate for those problems. A priori estimate means to find some connection relating the error between the exact solution of the problem and its approximation to the regularity properties of the exact solution itself. Note that in most cases the exact solution is not available.

The problem in the solution of which we are mainly interested is the following:

Problem 6.1. For a given $f \in X^*$, find $u \in X$ such that

$$a(u, v) = \langle f, v \rangle \text{ for all } v \in X.$$

Equivalently we can rewrite the problem as: find $u \in X$ such that

$$Lu = f \text{ in } X^*.$$

For the existence and uniqueness result of the solution of Problem 6.1 we refer to Lax-Milgram lemma, which has been the subject of Chapter 2.

After formulation of the continuous problem, we introduce now the discretized form of Problem 6.1. For this purpose assume that $\{X_h\}$ is the family of finite dimensional subspaces of the space X . The subscript h must be understood as a parameter, which defines the family and it will tend to zero.

Problem 6.2. for a given $f \in X^*$, find the discrete solution u_h associated with each finite dimensional space X_h such that

$$a(u_h, v_h) = \langle f, v_h \rangle \text{ for all } v_h \in X_h.$$

6.1 Abstract error estimates: Céa's lemma

Let u be the solution of problem 6.1 and u_h be the solution to discretized problem 6.2. What we want to do is to estimate the error between the exact solution u and discrete solution u_h .

Theorem 6.1. (Céa's Lemma). Let $f \in X^*$. Then the following inequalities hold

$$\sqrt{C_a} \|u - u_h\|_X \leq \|u - u_h\| \leq C_l \inf_{v_h \in X_h} \|u - v_h\|$$

Proof: Let v_h be an arbitrary element in X_h . Then using the coercivity and continuity of the bilinear form $a(\cdot, \cdot)$ we obtain

$$\begin{aligned} \|u - u_h\|^2 &\leq a(u - u_h, u - u_h) = a(u - u_h, u - v_h) + \underbrace{a(u - u_h, v_h - u_h)}_{=0} \\ &\leq C_l \|u - u_h\| \cdot \|u - v_h\| \end{aligned}$$

$$\begin{aligned} \Rightarrow \| |u - u_h| \| &\leq C_l \| |u - v_h| \|, \quad \forall v_h \in X_h \\ \Rightarrow \| |u - u_h| \| &\leq C_l \inf_{v_h \in X_h} \| |u - v_h| \| \end{aligned}$$

Remark 6.1. *The problem of estimating the error $\| |u - u_h| \|$ is now reduced to a problem of evaluating the term $\inf_{v_h \in X_h} \| |u - v_h| \|$. Usually instead of v_h one takes the interpolant of u on the subspace X_h , i.e.*

$$v_h := \mathbb{I}_h u \in X_h.$$

What we need now is the local interpolation estimates between u and its interpolant $\mathbb{I}_h u$, more precisely, we need an estimate for the norm $\| |u - \mathbb{I}_h u| \|_X$.

6.2 Interpolation estimates

One of the most used inequality in the process of derivation of any error estimate is the Poincaré's inequality, which we introduce in the following theorem.

Theorem 6.2. *(Poincaré's inequality for functions with vanishing mean value). Let $\Omega \subset \mathbb{R}^d$ be any convex bounded domain with Lipschitz boundary $\partial\Omega \in \mathbb{C}^{0,1}$. Then there exists a constant $C(\Omega)$ such that for all $u \in \mathbb{H}^1(\Omega)$ with zero mean value, i.e. $\int_{\Omega} u = 0$ the following inequality holds*

$$(6.1) \quad \| |u| \|_{L^2(\Omega)} \leq C \| |\nabla u| \|_{L^2(\Omega)}.$$

Basically, all estimates will be derived by transformation to the standard element \hat{T} , compare (5.19). Thus, we need estimates for the transformation from T to \hat{T} .

Lemma 6.1. *The transformation $F_T : \hat{T} \rightarrow T$ is affine linear, it can be written as*

$$F_T(\hat{x}) = A_T \hat{x} + a_0.$$

For a regular n -simplex T , the matrix $A_T \in \mathbb{R}^{n \times n}$ is invertible. The following estimates hold:

$$\begin{aligned} |A_T| &\leq \frac{h_T}{\rho_{\hat{T}}}, \quad |A_T^{-1}| \leq \frac{h_{\hat{T}}}{\rho_T}, \\ |\det A_T| &= \frac{|T|}{|\hat{T}|}, \quad c(n) \rho_T^n \leq |\det A_T| \leq c(n) h_T^n. \end{aligned}$$

Here, $|A|$ is the matrix norm corresponding to the Euclidian vector norm.

By the chain rule, now the following estimates hold:

Lemma 6.2. *For $g \in \mathbb{H}^m(T)$ define $\hat{g}(\hat{x}) := g(F_T(\hat{x}))$. Then $\hat{g} \in \mathbb{H}^m(\hat{T})$ and*

$$|\hat{g}|_{\mathbb{H}^m(\hat{T})} \leq c \frac{(h_T)^m}{|\det A_T|^{\frac{1}{2}}} |g|_{\mathbb{H}^m(T)}, \quad |g|_{\mathbb{H}^m(T)} \leq c \frac{|\det A_T|^{\frac{1}{2}}}{(\rho_T)^m} |\hat{g}|_{\mathbb{H}^m(\hat{T})}.$$

For functions in a finite-dimensional subspace of $\mathbb{H}^m(T)$ (for example polynomials of a fixed degree), we obtain the following *inverse estimates*.

Lemma 6.3. *Let $\hat{X}_T(\hat{T}) \subset \mathbb{H}^m(\hat{T})$ a finite dimensional subspace, $\hat{q} \in \hat{X}_T(\hat{T})$, and define $q(F_T(\hat{x})) := \hat{q}(\hat{x})$. Then it holds for all k with $0 \leq k < m$*

$$|q|_{\mathbb{H}^m(T)} \leq c(\sigma_T)^m \frac{1}{(h_T)^{m-k}} |q|_{\mathbb{H}^k(T)}.$$

Here, $\sigma_T = \frac{h_T}{\rho_T}$ is a measure for the aspect ratio of T , thus a geometrical parameter of the triangulation.

Notice that we can estimate higher derivatives of q by lower ones, loosing powers of h_T .

6.2.1 Clement interpolation

To describe the procedure of obtaining some interpolation estimates we assume that a conforming triangulation τ of $\Omega \in \mathbb{R}^d$ is given and we set $X = \mathbb{H}^1(\Omega)$. Let $X_h := \{v_h \in \mathbb{C}(\bar{\Omega}); v_h|_T \in P_k, \forall T \in \tau\}$ be some finite element space of X . Let further ω be the support of some basis function φ_h of X_h . For any $T \in \tau$ we define

$$h(\omega) := \text{diam}(\omega)$$

$$X_h(\omega) := \{v|_\omega; v \in X_h\}$$

Lemma 6.4. *There exists a mapping*

$$P_\omega : \mathbb{L}^2(\omega) \rightarrow X_h(\omega)$$

such that for all $v \in \mathbb{L}^2(\omega)$ the following equation holds

$$\|v - P_\omega v\|_{L^2(\omega)} = \inf_{\varphi_h \in X_h(\omega)} \|v - \varphi_h\|_{L^2(\omega)}$$

Corollary 6.1. *Let ω , $X_h(\omega)$ and P_ω are the same as above. Then for any $v \in \mathbb{H}^1(\omega)$ the estimates are true*

$$\begin{aligned} \|v - P_\omega v\|_{L^2(\omega)} &\leq Ch(\omega) \|\nabla v\|_{L^2(\omega)} \\ \|\nabla(v - P_\omega v)\|_{L^2(\omega)} &\leq C \|\nabla v\|_{L^2(\omega)} \end{aligned}$$

Proof: We first note that P_ω is a \mathbb{L}^2 -projection. Thus

$$\|v - P_\omega v\|_{L^2(\omega)} \leq \|v - \int_\omega v\|_{L^2(\omega)} \leq C \cdot \text{diam}(\omega) \cdot \|\nabla v\|_{L^2(\omega)}$$

In the last inequality we have used the Poincaré's inequality.

$$\begin{aligned}
\|\nabla(v - P_\omega v)\|_{L^2(\omega)} &\leq \|\nabla\left(v - \int_\omega v\right)\|_{L^2(\omega)} + \|\nabla\left(\int_\omega v - P_\omega v\right)\|_{L^2(\omega)} \\
&\leq \|\nabla v\|_{L^2(\omega)} + \frac{C}{h(\omega)} \left\| \int_\omega v - P_\omega v \right\|_{L^2(\omega)} \\
&\leq \|\nabla v\|_{L^2(\omega)} + \frac{C}{h(\omega)} \left(\left\| \int_\omega v - v \right\|_{L^2(\omega)} + \|v - P_\omega v\|_{L^2(\omega)} \right) \\
&\leq C\|\nabla v\|_{L^2(\omega)}^2
\end{aligned}$$

Theorem 6.3. For all $v \in \mathbb{H}^1(\Omega)$ there exists a linear mapping $\mathbb{I}_h \in L(\mathbb{H}^1(\Omega), X_h)$, such that

$$\|v - \mathbb{I}_h v\|_{L^2(\Omega)} \leq C \left(\sum_{T \in \tau} h(T)^2 \|\nabla v\|_{L^2(T)}^2 \right)^{\frac{1}{2}}$$

and respectively

$$\left(\sum_{T \in \tau} h^{-2}(T) \|v - \mathbb{I}_h v\|_{L^2(T)}^2 \right)^{\frac{1}{2}} \leq C \|\nabla v\|_{L^2(\Omega)}$$

as well as

$$\|\nabla(v - \mathbb{I}_h v)\|_{L^2(\Omega)} \leq C \|\nabla v\|_{L^2(\Omega)}$$

Proof: Let $a_j, j = 1, \dots, N$ be the vertices of the triangulation and φ_j the piecewise linear basis functions, i.e.

$$\varphi_i(a_j) = \delta_{ij}$$

and let $\omega_j = \text{supp}(\varphi_j)$. We define \mathbb{I}_h as

$$(\mathbb{I}_h v)(x) := \sum_{j=1}^N (P_{\omega_j} v)(a_j) \cdot \varphi_j$$

The defined interpolation operator \mathbb{I}_h is called *Clement Interpolation*. We will prove the estimates element wise. For that let $T \in \tau$ be an element and ω_T the support of T , i.e.

$$\omega_T := \cup_{T' \in \tau} \{T \cap T' \neq \emptyset\}$$

We have

$$\frac{1}{C} h(T') \leq h(T) \leq C h(T') \quad \forall T' \subset \omega_T$$

with

$$\text{number of}(T' \subset \omega_T) \leq C < \infty$$

Let a_0, \dots, a_d be the vertices of $T \in \tau$, then

$$\begin{aligned}
\mathbb{I}_h v|_T &= \sum_{j=1}^d (P_{\omega_j} v)(a_j) \cdot \varphi_j = P_{\omega_0} v + \sum_{j=1}^d (P_{\omega_j} v - P_{\omega_0} v)(a_j) \cdot \varphi_j \\
\| (P_{\omega_j} v - P_{\omega_0} v)(a_j) \cdot \varphi_j \|_{L^2(T)} &\leq \| P_{\omega_j} v - P_{\omega_0} v \|_{L^\infty(T)} \| \varphi_j \|_{L^2(T)} \\
&\leq \frac{C}{\sqrt{|T|}} \| P_{\omega_j} v - P_{\omega_0} v \|_{L^2(T)} \sqrt{|T|} \\
&\leq C (\| P_{\omega_j} v - v \|_{L^2(T)} + \| v - P_{\omega_0} v \|_{L^2(T)})
\end{aligned}$$

Thus

$$\begin{aligned}
\| v - \mathbb{I}_h v \|_{L^2(T)} &\leq C \sum_{j=0}^d \| v - P_{\omega_j} v \|_{L^2(T)} \\
&\leq C \sum_{j=0}^d \| v - P_{\omega_j} v \|_{L^2(\omega_j)} \\
&\leq C \sum_{j=0}^d h(\omega_j) \| \nabla v \|_{L^2(\omega_j)} \\
\Rightarrow \| v - \mathbb{I}_h v \|_{L^2(\Omega)} &= \left(\sum_{T \in \tau} \| v - \mathbb{I}_h v \|_{L^2(T)}^2 \right)^{\frac{1}{2}} \\
&\leq \left(\sum_{T \in \tau} C \sum_{j=0}^d h(\omega_{T,j})^2 \| \nabla v \|_{L^2(\omega_{T,j})}^2 \right)^{\frac{1}{2}} \\
&\leq C \left(\sum_{T \in \tau} h(T)^2 \| \nabla v \|_T^2 \right)^{\frac{1}{2}}
\end{aligned}$$

which is the required result. The other two inequalities are proved in a analog way.

Lemma 6.5. *(The Scaled Trace Theorem) Let T be a d -simplex and Γ a $(d-1)$ -dimensional sub-simplex of T . Then there exist a $C > 0$ such that for all $v \in \mathbb{H}^1(T)$ the following inequality is valid*

$$\| v \|_{L^2(\Gamma)} \leq C \left(h(T)^{-\frac{1}{2}} \| v \|_{L^2(T)} + h(T)^{\frac{1}{2}} \| \nabla v \|_{L^2(T)} \right)$$

Corollary 6.2. *Let T be a d -simplex and Γ a $(d-1)$ -dimensional sub-simplex of T . Then for all $v \in \mathbb{H}^1(\Omega)$ the following inequality is satisfied*

$$\| v - \mathbb{I}_h v \|_{L^2(\Gamma)} \leq C h(\Gamma)^{\frac{1}{2}} \| \nabla v \|_{L^2(\omega_T)}$$

6.2.2 Lagrange interpolation

To derive Lagrange interpolation estimates we state some results first.

Theorem 6.4. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary $\partial\Omega \in \mathbb{C}^{0,1}$, u be a function in the space $\mathbb{H}^l(\Omega)$ and*

$$\int_{\Omega} D^{\alpha} u = 0 \text{ for a multi-index } \alpha, \quad |\alpha| = 0, 1, \dots, l-1.$$

Then

$$\|u\|_{H^1(\Omega)} \leq C(\Omega, l) |u|_{H^1(\Omega)}$$

where $|\cdot|$ is the earlier defined semi-norm and C is a constant depending on Ω and l .

For the proof of this theorem one should successively apply the Poincaré's inequality.

Lemma 6.6. *For a function $u \in \mathbb{H}^{l+1}(\Omega)$ there exists exactly one polynomial $q \in P_l(\Omega)$ such that*

$$\int_{\Omega} D^{\alpha} (u - q) = 0, \quad |\alpha| = 0, 1, \dots, l.$$

Proof: The polynomial q in general has the following form

$$q(x) = \sum_{|\beta|=0}^l c_{\beta} x^{\beta}.$$

With this representation we arrive to a linear system of equations

$$\sum_{|\beta|=0}^l c_{\beta} \int_{\Omega} D^{\alpha} x^{\beta} dx = \int_{\Omega} D^{\alpha} u(x) dx, \quad |\alpha| = 0, 1, \dots, l.$$

The system has a unique solution. Indeed, from

$$\sum_{|\beta|=0}^l c_{\beta} \int_{\Omega} D^{\alpha} x^{\beta} dx = 0 \quad |\alpha| = 0, 1, \dots, l,$$

follows that

$$\int_{\Omega} D^{\alpha} q dx = 0 \quad |\alpha| = 0, 1, \dots, l,$$

which is equivalent to $q = 0$.

As a consequence of these two results we obtain

Theorem 6.5. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary $\partial\Omega \in \mathbb{C}^{0,1}$ and $k \in \mathbb{N}$. Then there exist a constant $C = C(\Omega, k)$ such that for all functions u from the quotient space $\mathbb{H}^{k+1}(\Omega)/\mathbb{P}_k(\Omega)$ the following inequality is satisfied*

$$\|u\|_{\mathbb{H}^{k+1}(\Omega)/\mathbb{P}_k(\Omega)} \leq C|u|_{\mathbb{H}^{k+1}(\Omega)}.$$

We recall her the definition of the norm on the quotient functional space

$$\|u\|_{\mathbb{H}^{k+1}(\Omega)/\mathbb{P}_k(\Omega)} = \inf_{q \in \mathbb{P}_k(\Omega)} \|u - q\|_{\mathbb{H}^{k+1}(\Omega)}.$$

Corollary 6.3. *Let again $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary $\partial\Omega \in \mathbb{C}^{0,1}$ and $k, m \in \mathbb{N}_0$, $m \leq k + 1$ and $\mathbb{I} \in L(\mathbb{H}^{k+1}(\Omega), \mathbb{H}^m(\Omega))$ be a interpolation operator which leaves the polynomial space $\mathbb{P}_k(\Omega)$ invariant, i.e. $\mathbb{I}q = q$ for all $q \in \mathbb{P}_k(\Omega)$. Then there exists a constant $C = C(k, m, \Omega, \|\mathbb{I}\|)$ such that for all functions $u \in \mathbb{H}^{k+1}(\Omega)$ we have*

$$\|u - \mathbb{I}u\|_{\mathbb{H}^m(\Omega)} \leq C|u|_{\mathbb{H}^{k+1}(\Omega)}.$$

Proof: We recall one embedding result of Sobolev spaces.

$$\mathbb{H}^m(\Omega) \hookrightarrow \mathbb{H}^{k+1}(\Omega),$$

which yields

$$\|u\|_{\mathbb{H}^m(\Omega)} \leq \|u\|_{\mathbb{H}^{k+1}(\Omega)}.$$

Then for every $q \in \mathbb{P}_k(\Omega)$ we have

$$\|u - \mathbb{I}u\|_{\mathbb{H}^m(\Omega)} = \|(u - q) - \mathbb{I}(u - q)\|_{\mathbb{H}^m(\Omega)} \leq (1 + \|\mathbb{I}\|)\|u - q\|_{\mathbb{H}^{k+1}(\Omega)}$$

Finally, applying the Theorem 6.5 we obtain

$$\|u - \mathbb{I}u\|_{\mathbb{H}^m(\Omega)} \leq (1 + \|\mathbb{I}\|)\|u\|_{\mathbb{H}^{k+1}(\Omega)/\mathbb{P}_k(\Omega)} \leq C(1 + \|\mathbb{I}\|)|u|_{\mathbb{H}^{k+1}(\Omega)}$$

The proof is complete.

Now we apply the interpolation operator locally on each simplex of the triangulation of the domain Ω .

Theorem 6.6. *Let T be a non-degenerate d -simplex, \hat{T} a reference simplex and $F_T : \hat{T} \rightarrow T$ the corresponding affine mapping. Let $k, m \in \mathbb{N}_0$, $m \leq k + 1$ and let $\hat{\mathbb{I}} \in L(\mathbb{H}^{k+1}(\hat{T}), \mathbb{H}^m(\hat{T}))$ be a interpolation operator which leaves the polynomial space $\mathbb{P}_k(\hat{T})$ invariant.*

Then for an interpolation operator $\mathbb{I} \in L(\mathbb{H}^{k+1}(T), \mathbb{H}^m(T))$ defined as

$$(\mathbb{I}u) \circ F_T = \hat{\mathbb{I}}(u \circ F),$$

we get the following interpolation estimate for each $u \in \mathbb{H}^{k+1}(T)$

$$|u - \mathbb{I}u|_{\mathbb{H}^m(T)} \leq C\sigma(T)^m h(T)^{k+1-m} |u|_{\mathbb{H}^{k+1}(T)}$$

where $C = C(k, m, \hat{T}, \|\hat{\mathbb{I}}\|)$.

Proof: The proof follows directly from the result of corollary 6.3.

For our purposes we need an interpolation operator \mathbb{I} (respectively $\hat{\mathbb{I}}$) such that $\hat{\mathbb{I}}(\hat{u}) \in \mathbb{P}_k(\hat{T})$, because what we would like to estimate is the finite element error, i.e.

$$\|u - u_h\|_{H^1(\Omega)} \leq C \inf_{v_h \in X_h} \|u - v_h\|_{H^1(\Omega)} \leq C \|u - \mathbb{I}_h u\|_{H^1(\Omega)}.$$

The interpolation operator \mathbb{I}_h must fulfill the following conditions:

$$\mathbb{I}_h u|_T \in \mathbb{P}_k(T),$$

$$\mathbb{I}_h q|_T = q \quad \text{for all } q \in \mathbb{P}_k(T),$$

and

$$\mathbb{I}_h u \in X_h.$$

We have already get acquainted with such interpolation operators in section 5 when we discussed the linear finite elements. We know that the values of u in Lagrange nodes (in the case of linear elements Lagrange nodes are the vertices of elements) uniquely determine $u_h \in X_h$ with $u_h|_T \in \mathbb{P}_k$ and because of the unique representation the space $P_k(T)$ remains invariant. The problem that arises here is the so called ‘‘point values’’ of u , as u is in general a function from a Sobolev space ($u \in \mathbb{H}^{k+1}(\Omega)$) and the point values are only defined for continuous functions.

Theorem 6.7. *Let $\Omega \in \mathbb{R}^d$, $d < 4$, be an open and bounded domain with a conforming triangulation \mathcal{T} . Denote by X_h the finite element space*

$$X_h := \{v_h \in \mathbb{C}^0(\bar{\Omega}); v_h|_T \in \mathbb{P}_k(T), T \in \mathcal{T}\}.$$

Then there exists a Lagrange interpolation operator \mathbb{I} , $\mathbb{I}u \in \mathbb{P}_k(T)$ and $\mathbb{I}u = u$ on the grid G_k of order k , $T \in \mathcal{T}$, such that \mathbb{I} leaves the space $\mathbb{P}_k(T)$ invariant, $\mathbb{I} \in L(\mathbb{H}^2(\Omega), X_h)$ and for $m, l \in \mathbb{N}_0$, where $0 \leq m \leq l + 1$ and $1 \leq l \leq k$, the following interpolation estimate holds

$$\begin{aligned} |u - \mathbb{I}u|_{H^m(\Omega)} &\leq C_1 \left(\sum_{T \in \mathcal{T}} (\sigma(T)^m h(T)^{l+1-m} |u|_{H^{l+1}(T)})^2 \right)^{\frac{1}{2}} \\ &\leq C_1 \sigma_0^m \underbrace{(h(\mathcal{T}))^{l+1-m}}_{\max_{T \in \mathcal{T}} h(T)} |u|_{H^{l+1}(\Omega)} \end{aligned}$$

Proof: Since $\mathbb{H}^2(\Omega) \hookrightarrow \mathbb{C}^0(\bar{\Omega})$, then the existence of an interpolation operator \mathbb{I} follows from the theorem 5.2. As for the estimate, we get

$$\begin{aligned} |u - \mathbb{I}u|_{H^m(\Omega)}^2 &= \sum_{T \in \mathcal{T}} |u - \mathbb{I}u|_{H^m(T)}^2 \\ &\leq C \sum_{T \in \mathcal{T}} (\sigma(T)^m h(T)^{l+1-m} |u|_{H^{l+1}(T)})^2 \\ &\leq C \max_{T \in \mathcal{T}} (\sigma(T)^m h(T)^{l+1-m})^2 |u|_{H^{l+1}(\Omega)}^2 \end{aligned}$$

The proof of the theorem is complete.

6.3 A priori error estimate

Putting Theorems 6.1 and 6.7 together, we derive the a priori error estimate:

Theorem 6.8. *Let $\Omega \subset \mathbb{R}^n$ a bounded polygonal domain, $X = H_0^1(\Omega)$ and $f \in X^*$. Let \mathcal{T} a proper, shape regular triangulation of Ω and $X_h = \{v_h \in \mathbb{C}^0(\bar{\Omega}), v_h|_T \in \mathbb{P}_k(T) \forall T \in \mathcal{T}\}$ the corresponding finite element space of piecewise polynomials of degree k .*

Let $u \in X$ the solution of Problem 6.1 and $u_h \in X_h$ the discrete solution of Problem 6.2. Then the following error estimate holds for all $l > \frac{n}{2} - 1$, $1 \leq l \leq k$, with $u \in \mathbb{H}^l(\Omega)$:

$$\|u - u_h\|_{H^1(\Omega)} \leq c h(\mathcal{T})^l |u|_{H^{l+1}(\Omega)}.$$

Now we want to consider again the more general elliptic problem from Section 5.3, where we were looking for solutions $u \in g + \mathring{X} \subset X = H^1(\Omega)$ and $u_h \in g_h + \mathring{X}_h \subset X$. How can we get error estimates in this case?

First of all, we reformulate the problem again.

Let $g \in X$ and $g_h \in X_h$ be continuations of the Dirichlet boundary values g and g_h . For the difference $u - g \in \mathring{X}$ holds

$$\langle L(u - g), v \rangle = \langle Lu, v \rangle - \langle Lg, v \rangle = \langle F, v \rangle - \langle Lg, v \rangle = \langle F - Lg, v \rangle$$

Defining the linear functionals $\tilde{F} \in X^*$ and $\tilde{F}_h \in X_h^*$ by

$$\tilde{F} := F - Lg, \quad \tilde{F}_h := F - Lg_h,$$

we arrive at continuous and discrete problems for $w := u - g$ and $w_h := u_h - g_h$.

Problem 6.3. *Find $w \in \mathring{X}$ such that*

$$\langle Lw, v \rangle = \langle \tilde{F}, v \rangle \text{ for all } v \in \mathring{X}.$$

Problem 6.4. *Find $w_h \in \mathring{X}_h$ such that*

$$\langle Lw_h, v_h \rangle = \langle \tilde{F}_h, v_h \rangle \text{ for all } v_h \in \mathring{X}_h.$$

Thus, the right hand sides of the continuous and the discrete problems are in general not the same. But we can prove the following generalization of Cea's Lemma:

Theorem 6.9. *Let X a Hilbert space and $X_h \subset X$, $\tilde{F} \in X^*$ and $\tilde{F}_h \in X_h^*$, $a(w, v) := \langle Lw, v \rangle$ a continuous and coercive bilinear form on X . Let $w \in X, w_h$ be the solutions to Problems 6.3 and 6.4. Then*

$$\|w - w_h\|_X \leq c \left(\inf_{v_h \in \mathring{X}_h} \|w - v_h\|_X + \|\tilde{F} - \tilde{F}_h\|_{X_h^*} \right).$$

By using $g_h = \mathbb{I}g$, we can estimate the error in right hand sides $\|\tilde{F} - \tilde{F}_h\|$ again by the interpolation estimates, and get the following error estimate.

Theorem 6.10. *Let w, w_h be the solutions to Problems 6.3 and 6.4, with $g_h = \mathbb{I}g$. Set $u := w + g$ and $u_h := w_h + g_h$, and let k, l as in Theorem 6.8 with $u, g \in \mathbb{H}^{l+1}(\Omega)$. Then*

$$\|u - u_h\|_{H^1(\Omega)} \leq c h(\mathcal{T})^l (|u|_{H^{l+1}(\Omega)} + |g|_{H^{l+1}(\Omega)}).$$

7 A posteriori error estimation for elliptic problems

In this chapter we give an introduction about adaptive finite element techniques for *elliptic* problems. Most of the principles are clear and easy to describe and understand in the context of elliptic problems. Afterwards we will apply the results to parabolic problems.

We consider the model problem: Find a solution u of

$$(7.1) \quad \begin{aligned} -\Delta u &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega \end{aligned}$$

where Ω is a bounded domain in \mathbb{R}^d with a polyhedral boundary and $f \in L_2(\Omega)$ is some given right hand side.

Let $\mathbb{H}^1(\Omega)$ be the Sobolev space of all functions with weak derivatives of first order and let $\mathbb{H}_0^1(\Omega)$ be the subspace of all those functions in $\mathbb{H}^1(\Omega)$ that vanish on the boundary of Ω . Then the weak formulation of (7.1) is stated as:

$$(7.2) \quad u \in \mathbb{H}_0^1(\Omega) : \quad \int_{\Omega} \nabla u \nabla \varphi = \int_{\Omega} f \varphi \quad \forall \varphi \in \mathbb{H}_0^1(\Omega).$$

Now, let $V_h \subset \mathbb{H}_0^1(\Omega)$ be a finite dimensional subspace. Then we have the discrete problem:

$$(7.3) \quad u_h \in V_h : \quad \int_{\Omega} \nabla u_h \nabla \varphi_h = \int_{\Omega} f \varphi_h \quad \forall \varphi_h \in V_h.$$

7.1 A posteriori error estimation in the energy norm

If V_h is for example the finite element space consisting of piecewise polynomials of degree $p \geq 1$ on a given triangulation \mathcal{T} (with zero boundary values) we have the following *a priori* estimate

$$(7.4) \quad |u - u_h|_{H^1(\Omega)} := \left(\int_{\Omega} |\nabla(u - u_h)|^2 \right)^{\frac{1}{2}} \leq c \left(\sum_{T \in \mathcal{T}} h_T^{2p} |u|_{H^{p+1}(\Omega)} \right)^{\frac{1}{2}}$$

where h_T is the diameter of a simplex T (see [21] e.g.).

The aim of *a posteriori* error estimation is to establish an estimation of the form

$$(7.5) \quad |u - u_h|_{H^1(\Omega)} \leq c \left(\sum_{T \in \mathcal{T}} \eta_T(u_h, f)^2 \right)^{\frac{1}{2}}$$

where the value of η_T does only depend on the discrete solution u_h on a simplex T and its adjacent neighbors and given data f on the simplex T . Thus, η_T is a computable value and we can control the adaptive procedure by these values (see Section 9).

The most important tool for such an a posteriori error estimation is an interpolation estimate for functions $v \in \mathbb{H}_0^1(\Omega)$. Since for $d \geq 2$ we do not have the embedding of $\mathbb{H}^1(\Omega)$ into $C^0(\bar{\Omega})$, we can not make use of the usual Lagrange interpolant. But we can use the Clément interpolant which avoids the pointwise evaluation of an H^1 function [22], see Section 6.2.1 for definition and interpolation estimates. On each element $T \in \mathcal{T}$ of the triangulation holds the local interpolation estimate

$$(7.6) \quad \|v - \mathbb{I}_h v\|_{L_2(T)} \leq ch_T |v|_{H^1(\omega_T)},$$

$$(7.7) \quad |\mathbb{I}_h v|_{H^1(T)} \leq c |v|_{H^1(\omega_T)},$$

where ω_T is the patch of all simplices $T' \in \mathcal{T}$ that have a non empty intersection with T .

Now we will derive the a posteriori estimate: Let $\mathbb{H}^{-1}(\Omega)$ be the dual space of $\mathbb{H}_0^1(\Omega)$, i.e. $\mathbb{H}^{-1}(\Omega) = (\mathbb{H}_0^1(\Omega))^*$. For $f \in L_2(\Omega)$ define $F \in \mathbb{H}^{-1}(\Omega)$ by

$$\langle F, \varphi \rangle_{\mathbb{H}^{-1}(\Omega) \times \mathbb{H}_0^1(\Omega)} := \int_{\Omega} f \varphi \quad \text{for all } \varphi \in \mathbb{H}_0^1(\Omega)$$

where $\langle \cdot, \cdot \rangle_{\mathbb{H}^{-1}(\Omega) \times \mathbb{H}_0^1(\Omega)}$ is the dual pairing on $\mathbb{H}^{-1}(\Omega) \times \mathbb{H}_0^1(\Omega)$.

We can look at $-\Delta$ as an operator

$$-\Delta : \mathbb{H}_0^1(\Omega) \rightarrow \mathbb{H}^{-1}(\Omega)$$

by defining $-\Delta v \in \mathbb{H}^{-1}(\Omega)$ for a function $v \in \mathbb{H}_0^1(\Omega)$ in the following way:

$$(7.8) \quad \langle -\Delta v, \varphi \rangle_{\mathbb{H}^{-1}(\Omega) \times \mathbb{H}_0^1(\Omega)} := \int_{\Omega} \nabla v \nabla \varphi \quad \text{for all } \varphi \in \mathbb{H}_0^1(\Omega).$$

It is clear that $-\Delta$ is a linear continuous operator. Moreover $-\Delta$ is invertible since (7.2) is uniquely solvable for a given right hand side $F \in \mathbb{H}^{-1}(\Omega)$ and it is an isometric isomorphism, i.e.

$$(7.9) \quad \|-\Delta v\|_{\mathbb{H}^{-1}(\Omega)} = |v|_{H^1(\Omega)}$$

because

$$\sup_{\varphi \in \mathbb{H}_0^1(\Omega) \setminus \{0\}} \frac{\langle -\Delta v, \varphi \rangle_{\mathbb{H}^{-1}(\Omega) \times \mathbb{H}_0^1(\Omega)}}{|\varphi|_{H^1(\Omega)}} = \sup_{\varphi \in \mathbb{H}_0^1(\Omega) \setminus \{0\}} \frac{\int_{\Omega} \nabla v \nabla \varphi}{|\varphi|_{H^1(\Omega)}}$$

$$\begin{cases} \leq |v|_{H^1(\Omega)} & \text{by Cauchy's inequality,} \\ \geq |v|_{H^1(\Omega)} & \text{taking } \varphi = v. \end{cases}$$

Remark: We can use such an abstract framework in more general situations also: Let V be an Hilbert space and $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$ a continuous V -coercive bilinear form, i.e.

$$a(v, \varphi) \leq c^* \|v\|_V \|\varphi\|_V \quad \text{and} \quad c_* \|v\|_V^2 \leq a(v, v) \quad \forall v, \varphi \in V,$$

Defining $A : V \rightarrow V^*$ by

$$\langle Av, \varphi \rangle_{V^* \times V} := a(v, \varphi) \quad \forall v, \varphi \in V,$$

we conclude

$$c_* \|v\|_V \leq \|Av\|_{V^*} \leq c^* \|v\|_V,$$

and the following analysis will also carry over to this situation.

Returning back to our model problem we rewrite (7.2) as:

$$u \in \mathbb{H}_0^1(\Omega) : \quad -\Delta u = F \quad \text{in } \mathbb{H}^{-1}(\Omega).$$

By this equation and by (7.9) we have for the error $e := u - u_h$

$$|e|_{H^1(\Omega)} = |u - u_h|_{H^1(\Omega)} = \|-\Delta(u - u_h)\|_{H^{-1}(\Omega)} = \|F + \Delta u_h\|_{H^{-1}(\Omega)}$$

Thus, we have an expression for the error in terms of u_h and data f . The problem is that we can not evaluate this expression because the norm on $\mathbb{H}^{-1}(\Omega)$ involves the evaluation of a supremum over all $\varphi \in \mathbb{H}_0^1(\Omega) \setminus \{0\}$. As a consequence we have to estimate this supremum.

For that we need the orthogonality of the error, i.e.

$$\begin{aligned} 0 &= \int_{\Omega} \nabla(u - u_h) \nabla \varphi_h = \langle -\Delta(u - u_h), \varphi_h \rangle_{H^{-1}(\Omega) \times H_0^1(\Omega)} \\ &= \langle F + \Delta u_h, \varphi_h \rangle_{H^{-1}(\Omega) \times H_0^1(\Omega)} \end{aligned}$$

for all $\varphi_h \in V_h$. Now, denote by $[\partial_\nu u_h]$ the jumps of the normal derivatives of the discrete solution u_h across a $(d-1)$ -simplex. We obtain by the orthogonality of the error, integration by parts, a scaled trace theorem, and the interpolation estimate (7.6)

$$\begin{aligned}
|e|_{H^1(\Omega)} &= \|F + \Delta u_h\|_{H^{-1}(\Omega)} \\
&= \sup_{\substack{\varphi \in H_0^1(\Omega) \\ |\varphi|_{H^1(\Omega)}=1}} \langle F + \Delta u_h, \varphi \rangle_{H^{-1}(\Omega) \times H_0^1(\Omega)} \\
&= \sup_{\substack{\varphi \in H_0^1(\Omega) \\ |\varphi|_{H^1(\Omega)}=1}} \langle F + \Delta u_h, \varphi - \mathbb{I}_h \varphi \rangle_{H^{-1}(\Omega) \times H_0^1(\Omega)} \\
&= \sup_{\substack{\varphi \in H_0^1(\Omega) \\ |\varphi|_{H^1(\Omega)}=1}} \sum_{T \in \mathcal{T}} \int_T f (\varphi - \mathbb{I}_h \varphi) - \int_T \nabla u_h \nabla (\varphi - \mathbb{I}_h \varphi) \\
&= \sup_{\substack{\varphi \in H_0^1(\Omega) \\ |\varphi|_{H^1(\Omega)}=1}} \sum_{T \in \mathcal{T}} \int_T (f + \Delta u_h) (\varphi - \mathbb{I}_h \varphi) - \frac{1}{2} \int_{\partial T \setminus \partial \Omega} [\partial_\nu u_h] (\varphi - \mathbb{I}_h \varphi) \\
&\leq c \sup_{\substack{\varphi \in H_0^1(\Omega) \\ |\varphi|_{H^1(\Omega)}=1}} \sum_{T \in \mathcal{T}} \left(h_T \|f + \Delta u_h\|_{L_2(T)} + \frac{1}{2} h_T^{\frac{1}{2}} \|[\partial_\nu u_h]\|_{L_2(\partial T \setminus \partial \Omega)} \right) |\varphi|_{H^1(M_T)} \\
&\leq c \left(\underbrace{\sum_{T \in \mathcal{T}} h_T^2 \|f + \Delta u_h\|_{L_2(T)}^2 + \frac{1}{2} h_T \|[\partial_\nu u_h]\|_{L_2(\partial T \setminus \partial \Omega)}^2}_{=: \eta_T(u_h, f)^2} \right)^{\frac{1}{2}}
\end{aligned}$$

where we used the fact that the overlap of different patches M_T is bounded by a constant. This establishes the a posteriori error estimate (7.5).

The above estimate makes sure that the error estimator $\eta := \left(\sum_{T \in \mathcal{T}} \eta_T(u_h, f)^2 \right)^{\frac{1}{2}}$ is *reliable*.

But we also have to answer the question whether the estimator is *efficient* also, i.e. can we estimate the estimator by the error itself. This is very important especially for higher order elements, because we only used the approximation property of the piecewise linear functions.

Let f_h be an approximation of the right hand side f belonging to some finite dimensional space (for example the piecewise L_2 projection on each element, or some other interpolant of the right hand side). Then we can prove

$$(7.10) \quad \eta_T(u_h, f_h) \leq c \left(|u - u_h|_{H^1(M(T))} + h_T \|f - f_h\|_{L_2(M(T))} \right)$$

where $M(T)$ now denotes the patch of all those simplices T' sharing a complete $(d-1)$ -simplex with T . The last term $h_T \|f - f_h\|_{L_2(M(T))}$ is of higher order if f is smooth. This term reflects that we first have to approximate given data sufficiently, i.e. $\|f - f_h\|_{L_2(M(T))}$ is small, and then we get an efficient error estimator which we can not expect for a poor approximation of given data. The proof of this estimate is very technical (one has to construct suitable cut-off functions to localize the element residual $f + \Delta u_h$ and the singular residual $[\partial_\nu u_h]$ and estimate them separately) and is omitted here (see [101] for example).

Remark: Usually, $\eta_T(u_h, f_h)$ is used as error estimator, since it is often not possible to compute the L_2 -norm of an arbitrary function exactly. By the triangle inequality it is clear that as well

$$\begin{aligned}\eta_T(u_h, f_h) &\leq \eta_T(u_h, f) + h_T \|f - f_h\|_{L_2(T)} & \text{as} \\ \eta_T(u_h, f) &\leq \eta_T(u_h, f_h) + h_T \|f - f_h\|_{L_2(T)}\end{aligned}$$

holds.

Since we usually can not compute the right hand side $\int_{\Omega} f \varphi_h$ of our discrete problem (7.3) exactly, the orthogonality of the error is disturbed. Applying an analysis which includes this defect will result in the a posteriori error estimation

$$|u - u_h|_{H^1(\Omega)} \leq c \left(\sum_{T \in \mathcal{T}} \eta_T(u_h, f_h)^2 \right)^{\frac{1}{2}} + c \|F - F_h\|_{V_h^*}$$

where we have replaced the right hand side of (7.3) by a computable value $\langle F_h, \varphi \rangle_{V_h^* \times V_h} := \int_{\Omega} f_h \varphi_h$.

The above analysis is not restricted to this simple model problem but can also be used for nonlinear problems (see [100]):

Let $F : \mathbb{H}_0^1(\Omega) \rightarrow \mathbb{H}^{-1}(\Omega)$ be an operator (maybe nonlinear) and let $u \in \mathbb{H}_0^1(\Omega)$ be a regular solution of

$$F(u) = 0 \quad \text{in } \mathbb{H}^{-1}(\Omega),$$

i.e. the Frechet-derivative of $DF(u)$ of F at u is invertible and bounded. Assume that DF and DF^{-1} are locally Lipschitz continuous. Now, let u_h be a discrete solution which is “near” u , i.e. $|u - u_h|_{H^1(\Omega)}$ is small enough. Then we get the following estimates:

$$c |u - u_h|_{H^1(\Omega)} \leq \|F(u_h)\|_{\mathbb{H}_0^1(\Omega)} \leq C |u - u_h|_{H^1(\Omega)}$$

where the constants c, C depend on the norms of $\|DF(u)\|$ and $\|(DF(u))^{-1}\|$ and the Lipschitz constants of DF and DF^{-1} . Again the error is represented in terms of given data and the discrete solution. Now using similar techniques to those used in the model problem will also establish efficient and reliable a posteriori error estimators for nonlinear problems.

8 Mesh refinement and coarsening

Finite element meshes may consist of geometric elements of various types:

- simplicial:** triangles or tetrahedra,
- quadrilateral:** rectangles, cubes, or general quadrilaterals,
- more general:** prisms, for example,
- mixed:** mixture of different types.

The choice of the mesh type for an application may depend on some special approximation properties or on the need for some special FE basis functions, which require a special local geometry. We will restrict ourselves here to the description of simplicial meshes, for several reasons:

- A simplex is one of the most simple geometric types.
- Complex domains may be approximated by a set of simplices quite easily.
- Simplicial meshes allow local refinement (see Figure 8.1) without the need of non-conforming meshes (hanging nodes), parametric elements, or mixture of element types (which is the case for quadrilateral meshes, for example, see Figure 8.2).
- Polynomials of a given degree are easily represented on a simplex using local (barycentric) coordinates. (On quadrilateral elements, the ‘standard’ type of ansatz spaces is a tensor product of one-dimensional polynomials.)

Refinement algorithms for non-simplicial meshes can be found in the literature.

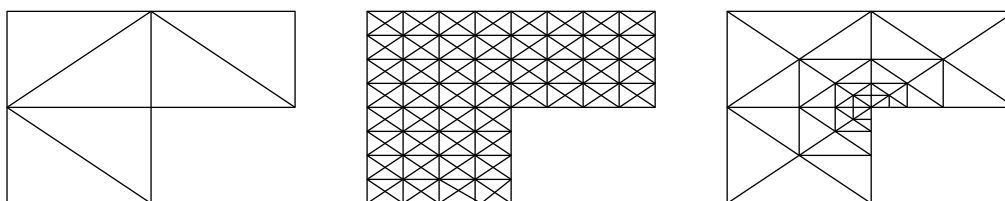


Figure 8.1: Global and local refinement of a triangular mesh.

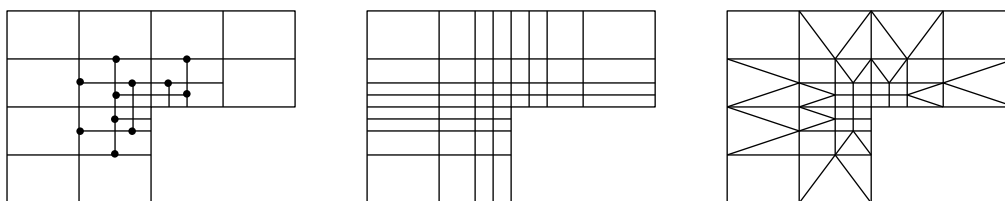


Figure 8.2: Local refinements of a rectangular mesh: with hanging nodes, conforming closure using bisected rectangles, and conforming closure using triangles. Using a conforming closure with rectangles, a local refinement has always global effects up to the boundary.

We will consider the following situation:

An initial (coarse) triangulation of the domain is given. We call it ‘macro triangulation’. It may be generated by hand or by some mesh generation algorithm. Some (or all) of the simplices are marked for refinement, depending on some error estimator or indicator. After several refinements, some other simplices may be marked for coarsening. Marking criteria and marking strategies are subject of Section 9.

8.1 Refinement algorithms for simplicial meshes

For simplicial elements, several refinement algorithms are widely used. One example is regular refinement (“red refinement”), which divides every triangle into four similar triangles, see Figure 8.3. The corresponding refinement algorithm in three dimensions cuts every tetrahedron into eight tetrahedra, and only a small number of similarity classes occur during successive refinements, see [10]. Unfortunately, hanging nodes arise during local regular refinement. To remove them and create a conforming mesh, in two dimensions some triangles have to be bisected (“green closure”). In three dimensions, several types of irregular refinement are needed for the green closure. This creates more similarity classes, even in two dimensions. Additionally, these bisected elements have to be removed before a further refinement of the mesh, in order to keep the triangulation regular.

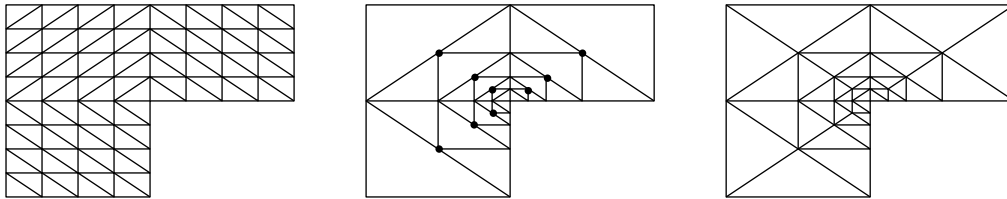


Figure 8.3: Global and local regular refinement of triangles and conforming closure by bisection.

Another possibility is to use bisection of simplices only. For every element (triangle or tetrahedron) one of its edges is marked as the refinement edge, and the element is refined into two elements by cutting this edge at its midpoint. There are several possibilities to choose such a refinement edge for a simplex, one example is to use the longest edge. Mitchell [68] compared different approaches. We will describe an algorithm where the choice of refinement edges on the macro triangulation prescribes the refinement edges for all simplices that are created during mesh refinement (the “newest vertex” bisection in Mitchell’s notation). This make sure that shape regularity of the triangulations is conserved.

The refinement by bisection can be implemented using recursive or non-recursive algorithms. For tetrahedra, the first description of such refinements was done in the non-recursive way by Bänsch [7]. It needs the intermediate handling of hanging nodes during the refinement process. Two recursive algorithms, which do not create such hanging nodes and are therefore easier to implement, are published by Kossaczky [59] and Maubach [67], which result in exactly the same tetrahedral meshes as the non-recursive algorithm.

Other refinement techniques for simplicial meshes, such as Delaunay techniques, are possible and described in the literature. We do not present details here.

In the following, we will describe the recursive refinement by bisection in detail, using the notation of Kossaczky. An implementation was done for example in [94].

The refinement algorithm is based on a recursive bisectioning of elements. For every element of the mesh, one of its edges is marked as its *refinement edge*. Elements are refined by bisecting this edge. To keep the mesh conforming, bisection of an edge is only allowed when this edge is the refinement edge for all elements which share this edge. Bisection of an edge and thus of all elements around the edge is the atomic refinement operation, and no other refinement operations are allowed. See Figures 8.4 and 8.5 for the two and three dimensional situations.

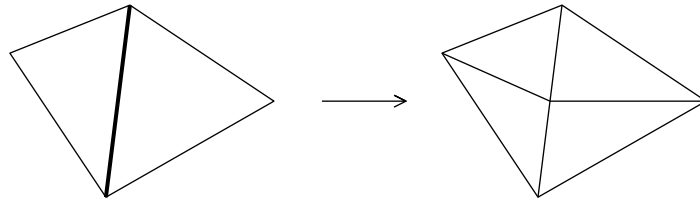


Figure 8.4: Atomic refinement operation in two dimensions. The common edge is the refinement edge for both triangles.

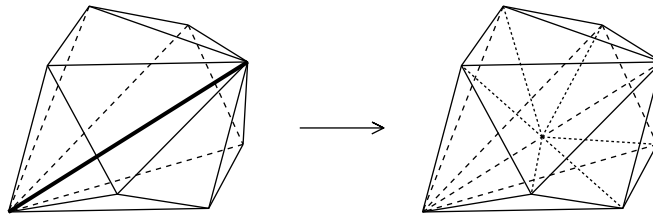


Figure 8.5: Atomic refinement operation in three dimensions. The common edge is the refinement edge for all tetrahedra around it.

If an element has to be refined, we first get all elements at this edge. In two dimensions this is just the neighbour opposite this edge or there is no other element at this edge in the case that the refinement edge belongs to the boundary. In three dimensions we have to loop around the edge and collect all neighbours at this edge. If for all collected neighbours this edge is the refinement edge also, we can refine the whole patch at same time by inserting one new vertex in the midpoint of the common refinement edge and bisecting every element of the patch. The resulting triangulation then is a conforming one.

If one of the collected neighbours has not the same refinement edge we first refine this neighbour recursively. Thus, we can formulate the refinement of an element in the following way

8.1 Algorithm. Recursive refinement of one simplex

```

subroutine recursive_refine(element)
{
  do
  {
    for all neighbours at refinement edge
      if neighbour has no compatible refinement edge
        recursive_refine(neighbour);
  } until all neighbours have a compatible refinement edge;

  bisect all elements at the refinement edge;
}

```

In two dimensions we used the so called newest vertex bisection and in three dimensions the algorithm described in [59]. For both variants it is proved, that for macro triangulation fulfilling certain criteria the recursion stops. Both algorithms are for special macro triangulations the recursive variants of the non recursive algorithms described in [7]. The beauty of the recursive approach is that we do not have to handle hanging nodes and not one to one adjacencies, since we can refine the whole refinement patch at same time.

In Figure 8.6 we show a two-dimensional situation where recursion is needed. For all triangles, the longest edge is the refinement edge. Let us assume that triangles A and B are marked for refinement. Triangle A can be refined at once, as its refinement edge is a boundary edge. For refinement of triangle B, we have to recursively refine triangles C and D. Again, triangle D can be directly refined, so recursion stops there. This is shown in the second part of the figure. Back in triangle C, this can now be refined together with its neighbour. After this, also triangle B can be refined together with its neighbour.

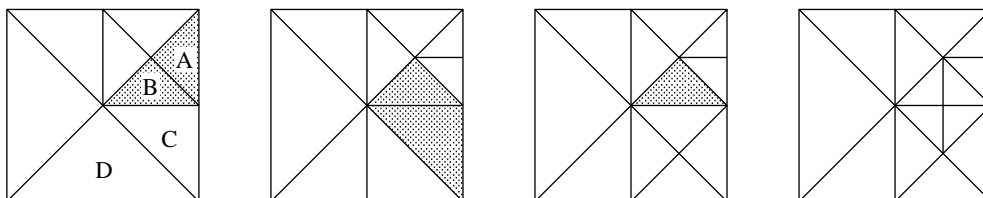


Figure 8.6: Recursive refinement in two dimensions. Triangles A and B are initially marked for refinement.

Now, the overall refinement algorithm can be formulated as follows:

8.2 Algorithm. Refinement of the mesh

```

subroutine refine_mesh()
{
  for all elements
    while element is marked for refinement
      recursive_refine(element);
}

```

We will use the convention, that all vertices of an element are given fixed *local indices*. Valid indices are 0, 1, and 2 for vertices of a triangle, and 0, 1, 2, and 3 for vertices of a tetrahedron. Now, the refinement edge for an element can be fixed to be the edge between the vertices with local indices 0 and 1.

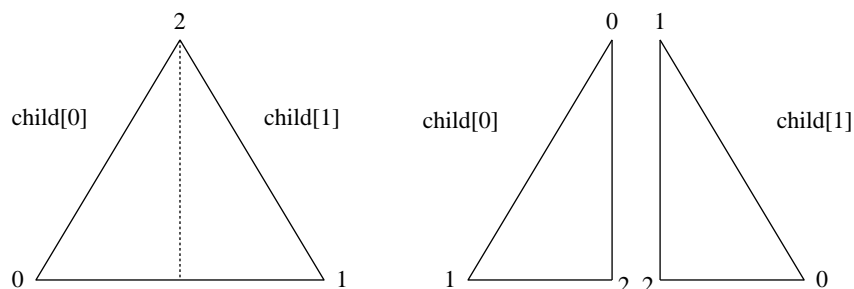


Figure 8.7: Numbering of nodes on parent and children triangles

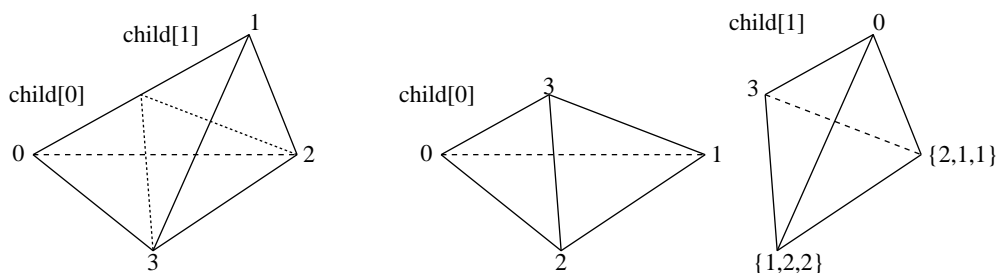


Figure 8.8: Numbering of nodes on parent and children tetrahedra

During refinement, the new vertex numbers for the newly created child simplices are prescribed by the refinement algorithm. This is done in such a way, that only a small number of similarity classes occur during successive refinement of one macro element. For both children elements, the index of the newly generated vertex at the midpoint of this edge has the highest local index (2 resp. 3 for triangles and tetrahedra). These numbers are shown in Figure 8.7 for 2d and in 8.8 for 3d. In 2d this numbering is the same for all refinement levels. In 3d, one has to make some special arrangements: the numbering of the second child's vertices does depend on the *generation* of the elements. There exist three different generations 0, 1, and 2, and the generation of a child element is always $((\text{parent's generation} + 1) \bmod 3)$. In Figure 8.8 we used the following convention: for the index set $\{1, 2, 2\}$ on `child[1]` of a tetrahedron of generation 0 we use the index 1 and for a tetrahedron of generation 1 and 2 the index 2. Figure 8.9 shows successive refinements of a generation 0 tetrahedron, producing tetrahedra of generations 1, 2, and 0 again.

Using the above refinement algorithm, the refinements of a mesh are totally determined by the local vertex numbering of the macro triangulation, plus a prescribed generation for every macro element in three dimensions.

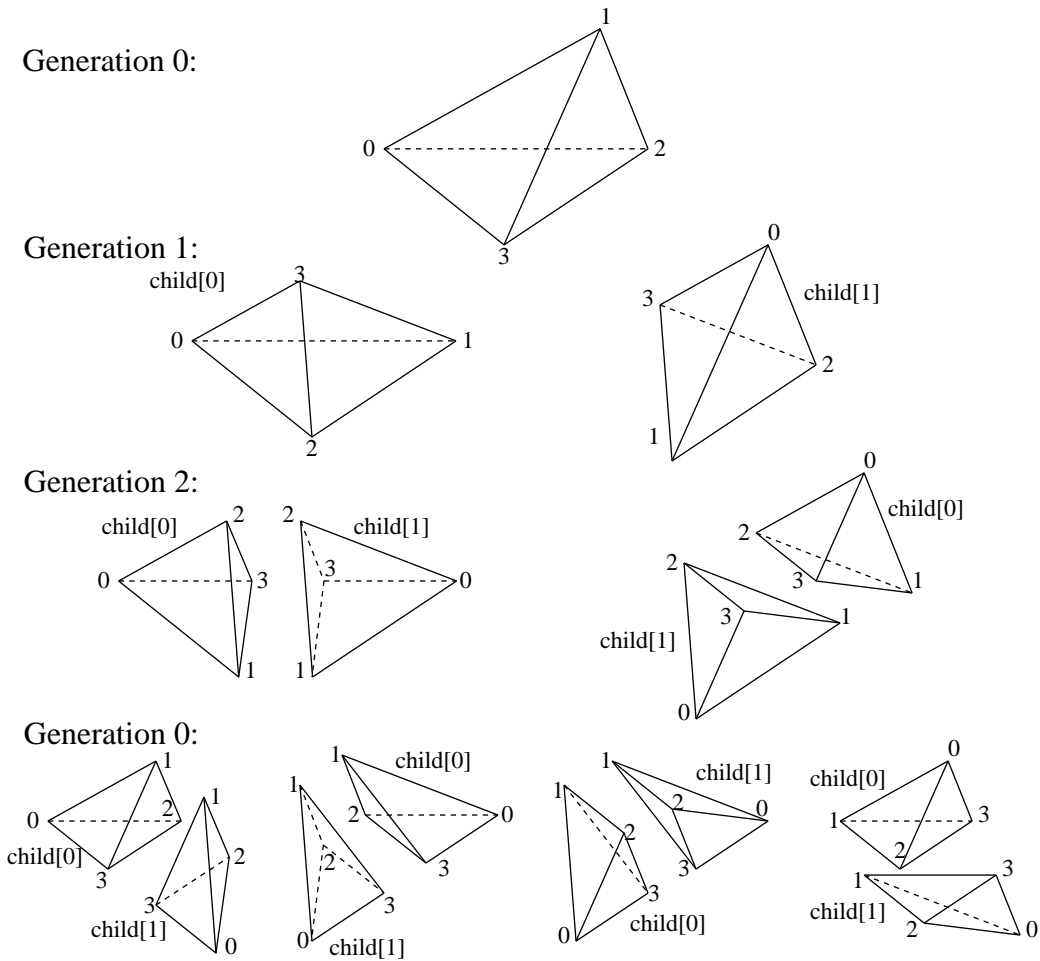


Figure 8.9: Successive refinements of a generation 0 tetrahedron

The numbering for tetrahedra was introduced by Kossaczky. In case of the “standard” triangulation of a (unit) square and cube into two triangles resp. six tetrahedra (see Figure 8.10), these numbering and the definition of the refinement edge during refinement of the elements guarantee that always the longest edge will be the refinement edge and will be bisected, see Figure 8.11. For the general case is proved:

8.3 Theorem. (Kossaczky [59], Mitchell [68])

1. The recursion stops if the macro triangulation fulfills certain criteria.
2. We obtain shape regularity for all elements at all levels.

In two dimensions, a triangulation where recursion does not stop is shown in Figure 8.12. The selected refinement edges of the triangles are shown by dashed lines. One can easily see, that there are no patches for the atomic refinement operation. This triangulation can only be refined if other choices of refinement edges are made, or by a non-recursive algorithm.

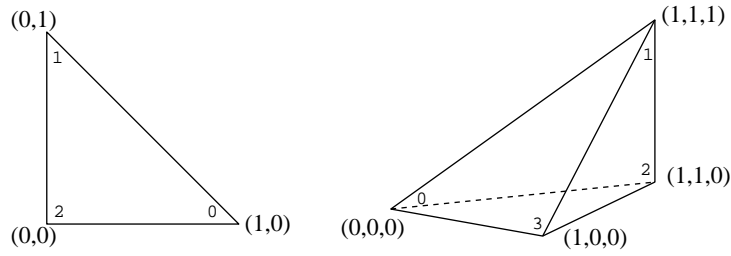


Figure 8.10: Standard elements in two and three dimensions

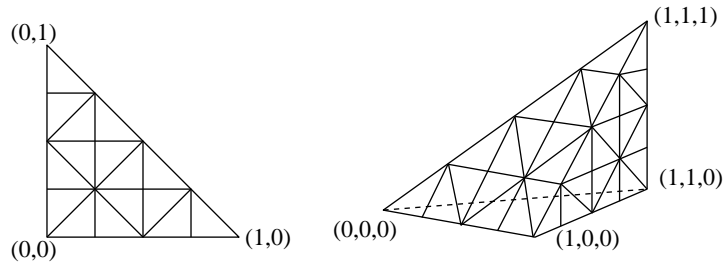


Figure 8.11: Refined standard elements in two and three dimensions

For using the refinement algorithm in a finite element package, we also need a numbering for edges, neighbours and faces. Edges and faces are needed for the implementation of higher order elements, for example, and neighbour information is used in the refinement algorithm itself and for error estimator calculation, for example.

In 2d the i -th edge/neighbour is the edge/neighbour opposite the i -th vertex; in 3d the i -th face/neighbour is the face/neighbour opposite the i -th vertex; edges in 3d are numbered in the following way:

- edge 0:** between vertex 0 and 1, **edge 3:** between vertex 1 and 1,
- edge 1:** between vertex 0 and 2, **edge 4:** between vertex 1 and 3,
- edge 2:** between vertex 0 and 3, **edge 5:** between vertex 2 and 3.

Figure 8.13 shows the numbering of the edges of child tetrahedra after refinement. The markers describe, which edge's degrees of freedom are changed during refinement, when higher order elements are used. For a more detailed description of handling higher order elements, see [94].

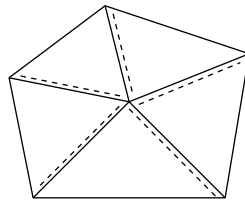


Figure 8.12: A macro triangulation where recursion does not stop

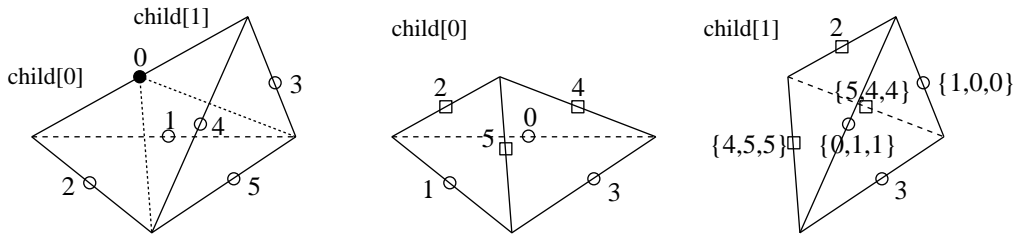


Figure 8.13: Edge numbers during refinement and degrees of freedom that are no longer needed \bullet , passed on from parent to child elements \circ , and newly generated \square

8.2 Prolongation of data during refinement

During refinement, finite element functions will have to be adjusted to the new mesh situation. Using hierarchically structured meshes, the finite element space of the coarse mesh is a subset of the space of the refined mesh (at least for typical polynomial ansatz spaces and refinement by bisection — there exist some finite elements where spaces are not nested, and the conforming closure needed by local regular refinements may lead to non-nested spaces, too). Thus, data can be represented identically on the refined mesh. During local refinement procedures, this *prolongation* of information from the old mesh to the new one is usually done directly together with the mesh changes.

After the geometrical part of the refinement is done on a patch around a refinement edge, we can prolongate data handled by the degrees of freedom from parents to child on the whole patch. We will describe the prolongation in detail for the case of piecewise linear finite elements; for higher order elements, everything is similar, but more degrees of freedom are involved.

For linear element, when degrees of freedom are located at vertices only, everything takes place on the bisected edge alone. Only one new vertex is created, the midpoint of the refinement edge. To determine the value of a function f_h at this new vertex, we can interpolate the function at this point. On the edge, f_h is a polynomial of degree 1, so the value at the midpoint is just the mean of the values at the edge endpoints:

$$f_h(\text{midpoint}) = \frac{1}{2}(f_h(\text{vertex } 0) + f_h(\text{vertex } 1)).$$

Using the nodal basis functions $\phi_i(v_j) = \delta_{i,j}$, then the coefficient f_n of the new basis function ϕ_n is just

$$f_n = \frac{1}{2}(f_0 + f_1).$$

8.3 Coarsening algorithms

The coarsening algorithm is more or less the inverse of the refinement algorithm. The basic idea is to collect all those elements that were created during the refinement at same time, i.e. the parents of these elements build a compatible refinement patch. If all the

elements are marked for coarsening, information is passed on the parents and the whole patch is coarsened at the same time.

If one of the elements is not marked for coarsening, we are not allowed to coarsen the patch. All element markers are reset. If one of the collected elements is not a leaf element but we are allowed to coarsen it more than one time, we first try to coarsen this element and then try to coarsen the newly collected patch.

This is the main difference between refinement and coarsening: Every element that is marked for refinement will be refined and this refinement may enforce a refinement of other elements that are not marked for refinement. An element that is marked for coarsening can only be coarsened if all elements of the coarsening patch may be coarsened together with this element. An element that is not marked for coarsening must not be coarsened, compare Section 9.3.

Thus, we can formulate the coarsening algorithm as follows:

```
8.4 Algorithm. Local coarsening around one edge
subroutine coarsen(element)
{
  get the parents of all elements at the coarsening edge
  for all parents
  {
    if the coarsening edge of the parent is not compatible
    {
      reset coarsening marks of all children of this patch;
      return false;
    }
  }
  for all parents
  {
    if the parent is refined more than once,
    and its children can be coarsened more than once
    return true;
  }
  coarsen all parents at the coarsening edge;
  return false;
}
```

The following routine coarsens as many elements as possible, even more than once if allowed:

8.5 Algorithm. Coarsening of the mesh

```

subroutine coarsen_mesh()
{
  do
  {
    do_coarsen_once_more = false;
    for all elements
      if element is marked for coarsening
        do_coarsen_once_more |= coarsen(element);
  } until do_coarsen_once_more is false
}

```

8.4 Restriction of data during coarsening

Also during coarsening, finite element functions will have to be adjusted to the new mesh situation. As now no longer the new finite element space is a superset of the old one, we lose some information. The marking strategies based on error estimators or indicators choose parts of the mesh to be coarsened, where the amount of lost information is not too big.

Nevertheless, finite element functions have to be restricted (transferred) from the fine to the coarse mesh. For linear finite elements, the easiest way to get around is just to ignore the value at the old vertex that will be removed, and interpolate the function in all remaining vertices.

In one special situation, information can be transferred identically from the old to the new mesh. If the values of a linear functional F applied to all basis functions are of interest, we can transform these values during coarsening, without making any error: If ϕ_0^{fine} , ϕ_1^{fine} , and ϕ_n^{fine} denote the basis functions corresponding to the endpoints and the midpoint of the edge inside the coarsened patch, then the new basis functions corresponding to the endpoints of the edge are

$$\phi_0^{\text{coarse}} = \phi_0^{\text{fine}} + \frac{1}{2}\phi_n^{\text{fine}}, \quad \phi_1^{\text{coarse}} = \phi_1^{\text{fine}} + \frac{1}{2}\phi_n^{\text{fine}}.$$

This can easily be seen by interpolation of the coarse basis functions. Now, if for some linear functional F the values $\langle F, \phi_0^{\text{fine}} \rangle$, $\langle F, \phi_1^{\text{fine}} \rangle$, and $\langle F, \phi_n^{\text{fine}} \rangle$ are available, the values of F applied to the new basis functions are

$$\langle F, \phi_0^{\text{coarse}} \rangle = \langle F, \phi_0^{\text{fine}} \rangle + \frac{1}{2}\langle F, \phi_n^{\text{fine}} \rangle, \quad \langle F, \phi_1^{\text{coarse}} \rangle = \langle F, \phi_1^{\text{fine}} \rangle + \frac{1}{2}\langle F, \phi_n^{\text{fine}} \rangle.$$

As one can easily see, the transformation matrix which transforms the old vector of functional values to the new one is just the transpose of the transformation matrix which was used for prolongation during refinement. This is the same for higher order elements.

One application of this procedure is time discretization, where scalar products with the solution u^{m-1} from the last time step appear on the right hand side of the discrete problem.

8.5 Storage methods for hierarchical meshes

There are basically two kinds of storing a finite element grid. One possibility is to store only the elements of the triangulation in a vector or a linked list. All information about elements is located at the elements. In this situation there is no direct information of a hierarchical structure for multigrid methods, e.g. Such information has to be generated and stored separately. During mesh refinement, new elements are added (at the end) to the vector or list of elements. During mesh coarsening, elements are removed. In case of an element vector, ‘holes’ may appear in the vector that contain no longer a valid element. One has to take care of them, or remove them by compressing the vector.

The other kind of storing the mesh is to keep the whole sequence of grids starting on the macro triangulation up to the actual one. Storing information about the whole hierarchical structure will need additional amount of computer memory, but on the other hand we can save computer memory by storing such information not explicitly on each element which can be produced by the hierarchical structure.

The simplicial grid is generated by refinement of a given macro triangulation. Refined parts of the grid can be derefined, but we can not coarsen elements of the macro triangulation. The refinement and coarsening routines construct a sequence of nested grids with a hierarchical structure. Every refined simplex is refined into two children. Elements that may be coarsened were created by refining the parent into these two elements and are now just coarsened back into this parent (compare Sections 8.1, 8.3).

Using this structure of the refinement/coarsening routines, every element of the macro triangulation is the root of a binary tree: every interior node of that tree has two pointers to the two children; the leaf elements are part of the actual triangulation, which is used to define the finite element space. The whole triangulation is a list (or vector) of given macro elements together with the associated binary trees.

Operations on elements can be performed by traversing the mesh, using standard tree traversing algorithms.

Some information is stored on the (leaf) elements explicitly, other information is located at the macro elements and is transferred to the leaf elements while traversing through the binary tree. All information that should be available for mesh elements is stored explicitly for elements of the macro triangulation. Thus, all information is present on the macro level and is transferred to the other tree elements by transforming requested data from one element to its children. These can be done by simple calculations using the hierarchic structure induced by the refinement algorithm.

An example of information which does not have to be stored for each element are the coordinates of the element’s vertices (in the case of non-parametric elements and polyhedral boundary). Going from parent to child only the coordinates of one vertex changes

and the new ones are simply the mean value of the coordinates of two vertices at the so called refinement edge of the parent. The other vertex coordinates stay the same.

Another example of such information is information about adjacent elements. Using adjacency information of the macro elements we can compute requested information for all elements of the mesh.

An implementation of the hierarchical mesh storage is done in [94].

9 Adaptive strategies for elliptic problems

The aim of adaptive methods is the generation of a mesh which is adapted to the problem such that a given criterion, like a tolerance for the estimated error between exact and discrete solution, is fulfilled by the finite element solution on this mesh. An optimal mesh should be as coarse as possible while meeting the criterion, in order to save computing time and memory requirements. For time dependent problems, such an adaptive method may include mesh changes in each time step and control of time step sizes.

The philosophy for the implementation should be the changing of meshes successively by local refinement or coarsening, based on error estimators or error indicators, which are computed a posteriori from the discrete solution and given data on the current mesh. In this section, we will present several strategies for the local refinement and coarsening of finite element meshes.

Let us assume that a triangulation \mathcal{T}_h of Ω , a finite element solution $u_h \in V_h$ to an elliptic problem, and an a posteriori error estimate

$$\|u - u_h\| \leq \eta(u_h) := \left(\sum_{T \in \mathcal{T}_h} \eta_T(u_h)^2 \right)^{1/2}$$

on this mesh are given. If ε is a given allowed tolerance for the error, and $\eta(u_h) > \varepsilon$, the problem arises

- where to refine the mesh in order to reduce the error,
- while at the same time the number of unknowns should not become too large.

A global refinement of the mesh would lead to the best reduction of the error, but the amount of new unknowns might be much larger than needed to reduce the error below the given tolerance. Using local refinement, we hope to do much better.

The design of an “optimal” mesh, where the number of unknowns is as small as possible to keep the error below the tolerance, is an open problem and will probably be much too costly. Especially in the case of linear problems, the design of an optimal mesh will be much more expensive than the solution of the original problem, since the mesh optimization is a highly nonlinear problem. Some heuristic arguments have to be used in the algorithm. The aim is to produce a result that is “not too far” from an optimal mesh, but with a relatively small amount of additional work to generate it.

9.1 Quasi-optimal meshes

Babuška and Rheinboldt [4] motivate that a mesh is almost optimal when the local errors are approximately equal for all elements.

For this purpose, assume that we have a discretization of $\Omega \subset \mathbb{R}^n$ with local mesh size $h(x)$. We assume now that h is a *smooth function*.

The size of the local mesh elements is $\approx h(x)^n$, and for the number of degrees of freedom holds

$$\#DOF = \mathcal{M}(h) = \int_{\Omega} \frac{\sigma(x)}{h(x)^n}$$

with $\sigma(x)$ bounded from above and below. The next assumption is that the error can be written as a functional of h in the way

$$\mathcal{E}(h) = \left(\int_{\Omega} (h(x)^\alpha E(x))^r \right)^{\frac{1}{r}}$$

where $r \geq 1$, $\alpha > 0$ and E is independent of h .

Problem 9.1. *Optimal mesh problem:*

Given a tolerance $\varepsilon > 0$, optimize h such that

$$\mathcal{M}(h) \text{ is minimal, while } \mathcal{E}(h) = \varepsilon.$$

This can be interpreted as a continuous optimization problem with restriction. The solution can be computed easily, and for that holds the condition, that the weighted pointwise error $h(x)^{n+\alpha r} E(x)^r / \sigma(x)$ must be constant over Ω . Integrating over mesh elements, this leads to the condition that the local element error

$$E_T^r = \int_T h(x)^{\alpha r} E(x)^r dx$$

must be equidistributed over all mesh elements.

By substituting the exact error by the local error indicators, this condition can be used to design an adaptive method which produces a quasi-optimal mesh.

9.1 Remark. We use here the notion of “quasi-optimality”, as the above derivation uses non-practical assumptions (for example that arbitrary smooth functions $h(x)$ can be used) and thus is only heuristic. Using only implementable variations of the mesh size, the whole procedure leads to a discrete optimization problem which is much harder to solve.

In the above consideration, optimality of the mesh is measured by counting the total number of unknowns, the dimension of the finite element space. Since the amount of computations necessary for most solution methods is a monotone function of this dimension, it makes sense to use this measure.

9.2 Mesh refinement strategies

Several adaptive strategies are proposed in the literature, that give criteria which mesh elements should be marked for refinement. All strategies are based on the idea of an equidistribution of the local error to all mesh elements. So, elements where the error estimate is large will be marked for refinement, while elements with a small estimated error are left unchanged.

The general outline of the adaptive algorithm is as follows. Starting from an initial triangulation \mathcal{T}_0 , we produce a sequence of triangulations \mathcal{T}_k , $k = 1, 2, \dots$, until the estimated error is below the given tolerance:

9.2 Algorithm. General adaptive refinement strategy

```

Start with  $\mathcal{T}_0$  and error tolerance  $\varepsilon$ 
 $k := 0$ 
do forever
  solve the discrete problem on  $\mathcal{T}_k$ 
  compute local error estimates  $\eta_T$ ,  $T \in \mathcal{T}_k$ 
  if  $\eta \leq \varepsilon$  then
    stop
  mark elements for refinement, according to a marking strategy
  refine mesh  $\mathcal{T}_k$ , producing  $\mathcal{T}_{k+1}$ 
   $k := k + 1$ 
enddo

```

Since a discrete problem has to be solved in every iteration of this algorithm, the number of iterations should be as small as possible. Thus, the marking strategy should select not too few mesh elements for refinement in each cycle. On the other hand, not much more elements should be selected than is needed to reduce the error below the given tolerance. In the sequel, we describe several marking strategies that are commonly used in adaptive finite element methods.

Maximum strategy: The simplest strategy is a maximum strategy. A threshold $\gamma \in (0, 1)$ is given, and all elements $T \in \mathcal{T}_k$ with

$$(9.1) \quad \eta_T > \gamma \max_{T' \in \mathcal{T}_k} \eta_{T'}$$

are marked for refinement. A small γ leads to more refinement and non-optimal meshes, while a large γ leads to more cycles until the error tolerance is reached, but produces a mesh with less unknowns. Typically, a threshold value $\gamma = 0.5$ is used [101, 103].

9.3 Algorithm. Maximum strategy

```

Start with parameter  $\gamma \in (0, 1)$ 
 $\eta_{\max} := \max(\eta_T, T \in \mathcal{T}_k)$ 
for all  $T$  in  $\mathcal{T}_k$  do
  if  $\eta_T > \gamma \eta_{\max}$  then mark  $T$  for refinement
enddo

```

Extrapolation strategy: Suppose that the local error estimates have an asymptotic behaviour

$$\eta_T = c h_T^\lambda \quad \text{as } h \rightarrow 0$$

for some $\lambda > 0$. If an element T with estimate η_T was generated by refining an element T^{old} in a previous mesh with corresponding estimate η_T^{old} , then the above behaviour suggests that the estimate at one of the children after refining T will be approximately

$$\eta_T^{\text{new}} = \frac{\eta_T^2}{\eta_T^{\text{old}}}.$$

Now, the idea is that no elements should be refined in the current iteration, where the estimated error is smaller than the largest local estimate that is expected after the next refinement. This leads to the following algorithm:

9.4 Algorithm. Extrapolation strategy [4]

```

cut := max( $\eta_T^{\text{new}}$ ,  $T \in \mathcal{T}_k$ )
for all  $T$  in  $\mathcal{T}_k$  do
  if  $\eta_T > \text{cut}$  then mark  $T$  for refinement
enddo

```

If η_T^{old} is unknown and thus η_T^{new} cannot be computed, some other marking strategy has to be used.

Equidistribution strategy: Let N_k be the number of mesh elements in \mathcal{T}_k . If we assume that the error is equidistributed over all elements, i. e. $\eta_T = \eta_{T'}$ for all $T, T' \in \mathcal{T}_k$, then

$$\eta = \left(\sum_{T \in \mathcal{T}_k} \eta_T^2 \right)^{1/2} = \sqrt{N_k} \eta_T \stackrel{!}{=} \varepsilon \quad \text{and} \quad \eta_T = \frac{\varepsilon}{\sqrt{N_k}}.$$

We can try to reach this equidistribution by refining all elements, where it is disturbed because the estimated error is larger than $\varepsilon/\sqrt{N_k}$. To make the procedure more robust, a parameter $\theta \in (0, 1)$, $\theta \approx 1$, is included in the method.

9.5 Algorithm. Equidistribution strategy [34]

```

Start with parameter  $\theta \in (0, 1)$ ,  $\theta \approx 1$ 
for all  $T$  in  $\mathcal{T}_k$  do
  if  $\eta_T > \theta\varepsilon/\sqrt{N_k}$  then mark  $T$  for refinement
enddo

```

If the error η is already near ε , then the choice $\theta = 1$ leads to the selection of only very few elements for refinement, which results in more iterations of the adaptive process. Thus, θ should be chosen smaller than 1, for example $\theta = 0.9$.

Guaranteed error reduction strategy: Usually, it is not clear whether the adaptive refinement strategy Algorithm 9.2 using a marking strategy (other than global refinement) will converge and stop, or how fast the convergence is. Dörfler [25] describes a strategy with a guaranteed relative error reduction for the Poisson equation.

We need the assumptions, that

- given data of the problem (like the right hand side) is sufficiently resolved by the current mesh \mathcal{T}_k ,
- all edges of marked mesh elements are at least bisected by the refinement procedure (using regular refinement or two/three iterated bisections of triangles/tetrahedra, for example).

The idea is to refine a subset of the triangulation that produces a considerable amount of the total error η . Given a parameter $\theta_* \in (0, 1)$, the procedure is:

$$\text{Mark a set } \mathcal{A} \subseteq \mathcal{T}_k \text{ such that } \sum_{T \in \mathcal{A}} \eta_T^2 \geq (1 - \theta_*)^2 \eta^2.$$

It follows from the assumptions that the error will be reduced by at least a factor $\kappa < 1$ depending of θ_* and data of the problem. Selection of the set \mathcal{A} can be done in the following way. The threshold γ is reduced in small steps of size $\nu \in (0, 1)$, $\nu \approx 0$, until the maximum strategy marks a set which is large enough. This inner iteration does not cost much time, as no computations are done in it.

9.6 Algorithm. Guaranteed error reduction strategy [25]

Start with given parameters $\theta_* \in (0, 1)$, $\nu \in (0, 1)$

$\eta_{\max} := \max(\eta_T, T \in \mathcal{T}_k)$

sum := 0

$\gamma := 1$

while sum < $(1 - \theta_*)^2 \eta^2$ do

$\gamma := \gamma - \nu$

 for all T in \mathcal{T}_k do

 if T is not marked

 if $\eta_T > \gamma \eta_{\max}$

 mark T for refinement

 sum := sum + η_T^2

 endif

 endif

 enddo

endwhile

Using the above algorithm, Dörfler [24] describes a robust adaptive strategy also for the *nonlinear* Poisson equation $-\Delta u = f(u)$. It is based on a posteriori error estimates and a posteriori saturation criteria for the approximation of the nonlinearity.

Other refinement strategies: Jarausch [51] describes a strategy which generates quasi-optimal meshes. It is based on an optimization procedure involving the increase of a cost function during refinement and the profit while minimizing an energy functional. For special applications, additional information may be generated by the error estimator and used by the adaptive strategy. This includes (anisotropic) directional refinement of

elements [58, 98], or the decision of local h - or p -enrichment of the finite element space [23].

9.3 Coarsening strategies

Up to now we presented only refinement strategies. For linear elliptic problems, no more is needed to generate a quasi-optimal mesh with nearly equidistributed local errors. Some of the refinement strategies described above can also be used to mark mesh elements for coarsening. Actually, elements will only be coarsened if all neighbour elements which are affected by the coarsening process are marked for coarsening, too. This makes sure that only elements where the error is small enough are coarsened, and motivates the coarsening algorithm in Section 8.3.

Equidistribution strategy: Equidistribution of the tolerated error ε leads to

$$\eta_T \approx \frac{\varepsilon}{\sqrt{N_k}} \quad \text{for all } T \in \mathcal{T}.$$

If the local error at an element is considerably smaller than this mean value, we may coarsen the element without producing an error that is too large. If we are able to estimate the error after coarsening, for example by assuming an asymptotic behavior like

$$\eta_T \approx c h_T^\lambda, \quad \lambda > 0,$$

we can calculate a threshold $\theta_c \in (0, \theta)$ such that the local error after coarsening is still below $\theta \varepsilon / \sqrt{N_k}$ if it was smaller than $\theta_c \varepsilon / \sqrt{N_k}$ before. If the error after coarsening gets larger than this value, the elements would directly be refined again in the next iteration.

9.7 Algorithm. Equidistribution refinement/coarsening strategy

Start with parameters $\theta \in (0, 1)$, $\theta \approx 1$, and $\theta_c \in (0, \theta)$
for all T in \mathcal{T}_k do
 if $\eta_T > \theta \varepsilon / \sqrt{N_k}$ then mark T for refinement
 if $\eta_T + \eta_{c,T} < \theta_c \varepsilon / \sqrt{N_k}$ then mark T for coarsening
enddo

When local h - and p -enrichment and coarsening of the finite element space is used, then the threshold θ_c depends on the local degree of finite elements. Thus, local thresholds $\theta_{c,T}$ have to be used.

Guaranteed error reduction strategy: Similar to the refinement in Algorithm 9.6, Dörfler [26] describes a marking strategy for coarsening. Again, the idea is to coarsen a subset of the triangulation such that the additional error after coarsening is not larger than a fixed amount of the given tolerance ε . Given a parameter $\theta_c \in (0, 1)$, the procedure is:

$$\text{Mark a set } \mathcal{B} \subseteq \mathcal{T}_k \text{ such that } \sum_{T \in \mathcal{B}} \eta_T^2 + \eta_{c,T}^2 \leq \theta_c^2 \varepsilon^2.$$

The selection of the set \mathcal{B} can be done similar to Algorithm 9.6. Under suitable assumptions, Dörfler proves that the adaptive algorithm with mesh refinement and coarsening leads to an error below the given tolerance [26].

Handling information loss during coarsening: Usually, some information is irreversibly destroyed during coarsening of parts of the mesh, compare Section 8.4. If the adaptive procedure iterates several times, it may occur that elements which were marked for coarsening in the beginning are not allowed to coarsen at the end. If the mesh was already coarsened, an error is produced which can not be reduced anymore.

One possibility to circumvent such problems is to delay the mesh coarsening until the final iteration of the adaptive procedure, allowing only refinements before. If the coarsening marking strategy is not too liberal (θ_c not too large), this should keep the error below the given bound.

Dörfler [26] proposes to keep all information until it is clear, after solving and by estimating the error on a (virtually) coarsened mesh, that the coarsening does not lead to an error which is too large.

10 Aspects of efficient implementation

10.1 Numerical integration (quadrature schemes)

We have seen above that for computing the system matrix and right hand side vector of the resulting linear system we need to calculate integrals being of the form

$$\sum_{T \in \mathcal{T}} \int_T \Phi(x) dx.$$

In the case when the domain has a curved boundary or for the general data of the problem, i.e. nonlinear problems or equations with variable coefficients, the integrals may be difficult to evaluate exactly. To evaluate such integrals, suitable quadrature formulas for numerical integration have to be used. Numerical integration in finite element method is done by looping over all grid elements and using a quadrature formula on each element T . The general quadrature formula for approximating the integrals takes the form

$$(10.1) \quad \int_T \Phi(x) dx \approx \sum_{i=1}^{n_q} \omega_i \Phi(y_i)$$

where ω_i , $i = 1, 2, \dots, n_q$ are certain weights, n_q is the number of quadrature nodes and the y_i are certain points (quadrature nodes) in the element T . The weights and nodes for any finite element T are derived from a simple quadrature formula defined over a *reference* element. Thus, we need to map an arbitrary triangle T onto a reference triangle \hat{T} and define the quadrature formula on the latter. For this purpose we use the notion of *barycentric coordinates* introduced in Section 5.4. We use the notations used there.

In the following we shall often drop the superscript T of λ^T and x^T . The mappings $\lambda(x) = \lambda^T(x)$ and $x(\lambda) = x^T(\lambda)$ are always defined with respect to the actual element $T \in \mathcal{T}$.

Definition 10.1 (Numerical quadrature). *A numerical quadrature \hat{Q} on \hat{T} is a set $\{(w_k, \lambda_k) \in \mathbb{R} \times \mathbb{R}^{d+1}; k = 0, \dots, n_Q - 1\}$ of weights w_k and quadrature points $\lambda_k \in \bar{\hat{T}}$ (i.e. given in barycentric coordinates) such that*

$$\int_{\hat{T}} f(\hat{x}) d\hat{x} \approx \hat{Q}(f) := \sum_{k=0}^{n_Q-1} w_k f(\hat{x}(\lambda_k)).$$

It is called exact of degree p for some $p \in \mathbb{N}$ if

$$\int_{\hat{T}} q(\hat{x}) d\hat{x} = \hat{Q}(q) \quad \text{for all } q \in \mathbb{P}_p(\hat{T}).$$

It is called stable if

$$w_k > 0 \quad \text{for all } k = 0, \dots, n_Q - 1.$$

Remark 10.1. A given numerical quadrature \hat{Q} on \hat{T} defines for each element S a numerical quadrature Q_S . Using the transformation rule we define Q_S on an element S which is parameterized by $F_S: \hat{S} \rightarrow S$ and a function $f: S \rightarrow \mathbb{R}$:

$$(10.2) \quad \int_S f(x) dx \approx Q_S(f) := \hat{Q}((f \circ F_S) | \det DF_S|) = \sum_{k=0}^{n_Q-1} w_k f(x(\lambda_k)) | \det DF_S(\hat{x}(\lambda_k))|.$$

For a simplex S this results in

$$(10.3) \quad Q_S(f) = d! |S| \sum_{k=0}^{n_Q-1} w_k f(x(\lambda_k)).$$

10.2 Efficient solvers for linear systems

We have seen that after the discretization of partial differential equations using finite elements, we obtain a system of algebraic equations. In general the coefficient matrix of such systems is large and sparse. By sparse matrix we mean that in each row of the matrix there are only a few number of entries that do not vanish.

The aim of the following chapter is to give a brief look at basic iterative methods for the solution of large linear systems.

10.2.1 Methods of Jacobi, Gauss-Seidel and Relaxation

Let suppose we want to solve a system of linear equations which in matrix form can be written as

$$(10.4) \quad Ax = b, \quad A \in \mathbb{C}^{n \times n}, \quad b \in \mathbb{C}^n$$

where A is the coefficient matrix, b is the given right-hand side vector and x is the vector to be calculated.

The methods we are going to describe are based on an idea to express the invertible matrix A by a splitting

$$(10.5) \quad A = B + (A - B)$$

where B is a “nice” matrix, i.e. it is an invertible matrix which can be relatively easy inverted (e.g. B can be chosen as a diagonal matrix). Thus, we have the following equivalences

$$(10.6) \quad \begin{aligned} Ax = b &\Leftrightarrow Bx = (B - A)x + B \Leftrightarrow x = \underbrace{B^{-1}(B - A)}_M x + \underbrace{B^{-1}b}_N \\ &\Leftrightarrow x = Mx + Nb \end{aligned}$$

If we take an arbitrary start vector x_0 , we can then consider the associated iterative scheme of (10.6)

$$(10.7) \quad x_{k+1} = Mx_k + Nb \quad k \geq 0$$

The iterative procedure (10.7) means that we pass from one iterate to the next and the next one must be, of course, a better approximation to the solution than the first iterate. In this process we hope to converge to the solution of the system (10.4).

More precisely, we start with a initial vector (iterate) $x^{(0)}$ and create the sequence of vectors $x^{(0)}, x^{(1)}, x^{(2)}, \dots$. If this sequence converges to the exact solution $x = A^{-1}b$, then we say that the iterative method *converges*. Thus, it is important to introduce some criteria for which the iterative method converges.

Definition 10.2. *If $\lambda_i, i = 1, 2, \dots, n$ are the eigenvalues of a matrix A , then we denote by $\rho(A)$ the spectral radius of the matrix A and define it as*

$$\rho(A) := \max_i |\lambda_i|, \quad i = 1, 2, \dots, n$$

Theorem 10.1. *The iterative method (10.7) converges if and only if*

$$\rho(M) < 1,$$

where $\rho(M)$ is the spectral radius of the matrix M .

Now we are free to choose the nonsingular matrix B . It is natural to choose such a “nice” B , so that the iterative equations (10.7) are easily solved. Each choice leads us to a different iterative method. Here we will discuss only a few basic iterative schemes. To start with we decompose the system matrix A in the form

$$(10.8) \quad A = D - L - R = \left[\begin{array}{c|c} \square & \\ \hline \square & \square \end{array} \right],$$

where L and R are strict lower and upper matrices and D is a diagonal matrix.

Jacobi Method: In the Jacobi method we assume that $a_{ii} \neq 0$ for $i = 1, 2, \dots, n$ and sets in (10.5) $B = D$. Then

$$A = D + (A - D) = \underbrace{D}_B - (L + R)$$

Therefore, from (10.6) we get

$$(10.9) \quad N = D^{-1} \quad M_J := M = B^{-1}(B - A) = D^{-1}(L + R)$$

With above notations we get the Jacobi Method

$$(10.10) \quad x_{k+1} = D^{-1}(L + R)x_k + D^{-1}b$$

It is convenient to write the equation (10.10) for components, i.e. for $i = 1, \dots, n$:

$$(10.11) \quad x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1, j \neq i}^n a_{ij} x_j^{(k)} \right).$$

Note that in order to calculate the new iterate x^{k+1} we need the values of previous iterates.

Theorem 10.2. (*Strong Row Sum Criterion*). *The Jacobi method converges for all matrices A with*

$$|a_{ii}| > \sum_{j \neq i} a_{ij} \text{ for } i = 1, 2, \dots, n$$

(*Strong Column Sum Criterion*). *The Jacobi method converges for all matrices A with*

$$|a_{jj}| > \sum_{i \neq j} a_{ij} \text{ for } j = 1, 2, \dots, n$$

If the above inequalities are satisfied, then A is called strictly diagonally dominant.

Gaus-Seidel Method: In Gauss-Seidel method one chooses $B = D - L$. Thus

$$A = \underbrace{(D - L)}_B + \underbrace{R}_{B-A}$$

Therefore, from (10.5)

$$M_{GS} := M = (D - L)^{-1}R, \quad N := N_{GS} = (D - L)^{-1}$$

The iteration scheme in matrix form looks

$$x_{k+1} = (D - L)^{-1}R x_k + (D - L)^{-1}b \quad \text{for } k = 1, 2, \dots$$

or which is the same

$$(10.12) \quad (D - L)x_{k+1} = R x_k + b \quad \text{für } k = 1, 2, \dots$$

Rewriting (10.12) for components $i = 1, \dots, n$, we obtain

$$\sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} + a_{ii} x_i^{(k+1)} = - \sum_{j=i+1}^n a_{ij} x_j^{(k)} + b_i$$

Note that in Gauss-Seidel method in order to calculate $x_i^{(k+1)}$ we use the already calculated $(k+1)$ -st iterates of previous components, namely for $k = 1, 2, \dots$ and $i = 1, 2, \dots, n$ we can get the equation for calculating $x_i^{(k+1)}$

$$(10.13) \quad x_i^{(k+1)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij} x_j^{(k)} \right]$$

Relaxation Method: In Jacobi and Gauss-Seidel methods the convergence is relatively slow, especially when the number of unknowns is large. To accelerate the convergence, we use the technique of the *relaxation method*. The idea is the following: we start with the initial iterative scheme and try to correct the iterate $x^{(k)}$ to get a better $x^{(k+1)}$:

$$x^{(k+1)} = B^{-1}(B - A)x^{(k)} + B^{-1}b = x^{(k)} + \underbrace{B^{-1}(b - Ax^{(k)})}_{\text{“correction”}}$$

Now we introduce a real parameter $\omega \neq 0$ and weight the correction vector with ω

$$x^{(k+1)} = x^{(k)} + \omega B^{-1}(B - Ax^{(k)})$$

After some manipulations we get

$$(10.14) \quad x^{(k+1)} = \underbrace{(I - \omega B^{-1}A)}_{M(\omega)} x^{(k)} + \underbrace{\omega B^{-1}b}_{N(\omega)}$$

This is the relaxation iterative method and ω is called the relaxation parameter.

As for the convergence, there is a sufficient condition for the convergence of the relaxation method.

Theorem 10.3. *If the system matrix A is Hermitian and positive definite, then the relaxation method converges if $0 < \omega < 2$.*

Example 10.1. (*Relaxation for Gauss-Seidel Method*)

We go back to Gauss-Seidel method again

$$x_i^{(k+1)} = x_i^{(k)} + \frac{1}{a_{ii}} \underbrace{\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)} \right]}_{\text{correction}}$$

Now we rewrite the equation weighted with a parameter ω

$$(10.15) \quad x_i^{(k+1)} = x_i^{(k)} + \frac{\omega}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)} \right]$$

Finally, the equation for $x^{(k+1)}$ in matrix writing takes the form

$$(10.16) \quad x^{(k+1)} = \underbrace{(D - \omega L)^{-1} [(1 - \omega)D + \omega R]}_{M_{GS}(\omega)} x^{(k)} + \underbrace{\omega(D - \omega L)^{-1} b}_{N_{GS}(\omega)}$$

That is what people call the Successive over-relaxation (SOR) method.

At the end of this section we would like to state a result which gives us a possibility to compare the Jacobi and Gauss-Seidel methods.

Theorem 10.4. *Let A be a tridiagonal matrix (or block tridiagonal). Then the spectral radii of the corresponding Jacobi and Gauss-Seidel matrices are related by*

$$\rho(M_{GS}) = \rho(M_J)^2.$$

We see from the theorem that both methods either converge or diverge at the same time. But what is important that when they converge, then the Gauss-Seidel converges more rapidly than the Jacobi method. This theorem is especially useful, because in the result of the discretization of partial differential equations very often we get matrices that are tridiagonal (or block tridiagonal).

10.2.2 Conjugate gradient method and other Krylov subspace iterations

The convergence rate of the iterative solvers described in the above section gets quite slow for finite element problems when the size of the system increases, i.e. when the local mesh element size decreases.

This is mainly due to the fact that high frequency parts of the error between the current iterate and the discrete solution are damped very well, while low frequency parts of the error are damped only very slowly.

For many problems, there are better iterative solvers which may give approximate solutions to a linear system much faster than the classical iterations. Examples of such methods are the Krylov subspace iteration methods, like the Conjugate Gradient (CG) method for symmetric positive definite matrices, or the GMRES method for arbitrary matrices. Details and more algorithms can be found for example in [92].

Both are based on successive minimization of residual norms $\|r_n\| = \|Ax_n - b\|$ on Krylov spaces $K_n = \text{span}\{r_0, Ar_0, A^2r_0, \dots, A^n r_0\}$. By using special properties of the matrix A or the iterative method, such minimizations can be done very efficiently.

The simplest example is the CG method for symmetric positive definite matrices, which reads as follows:

10.1 Algorithm. Conjugate Gradient algorithm for solution of $Ax = b$

Begin with $x_0 \in \mathbb{R}^N$ and set $d_0 = -g_0 = b - Ax_0$.

for $k = 0, 1, 2, \dots$ compute

$$\begin{aligned} \alpha_k &= \frac{g_k^T g_k}{d_k^T A d_k} \\ x_{k+1} &= x_k + \alpha_k d_k \\ g_{k+1} &= g_k + \alpha_k A d_k \\ \beta_k &= \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k} \\ d_{k+1} &= -g_{k+1} + \beta_k d_k \end{aligned}$$

while $g_k > \textit{tolerance}$

Note, that for each iteration, only one matrix-vector product Ad_k must be computed, plus some vector operations. So, especially for sparse finite element matrices, each iteration needs only $\mathcal{O}(N)$ arithmetic operations.

Preconditioners: All Krylov subspace iteration methods may get more efficient when a *preconditioning* of the linear system is used in a proper way. This means the transformation by a matrix C , such that the system $CAx = Cb$ is easier to solve than the original system $Ax = b$. The main idea is to reduce the condition of the iteration matrix, as this condition plays an important role for the convergence speed of the iteration. The best preconditioner would be the choice $C = A^{-1}$, but this is not known (and usually a full matrix, even when A is sparse). But A^{-1} can be approximated, for example by

- the inverse of the diagonal of A ,
- application of one iteration of a classical iterative solver,
- incomplete LU decompositions,
- one multilevel iteration (see below)
- and other ideas.

10.2.3 Multilevel preconditioners and solvers

Coming back to the fact, that the classical iterative solvers reduce only high frequency parts of the error efficiently, the idea was coming up to use few iterations of the iterative methods on a fine mesh as *smoothers*, while low frequency parts of the error can be reduced also on coarser meshes. Putting these ideas together, some of the most efficient solvers for systems of linear equations available nowadays were constructed.

Using the natural hierarchy of meshes and corresponding finite element spaces, which is generated by successive (local or global) refinement of a triangulation, multigrid solvers use the *smoothing* and *coarse grid correction* in a recursive way. Unfortunately, the multigrid methods work not for all classes of problems, sometimes one has to do very special smoothing operations. But, for example, symmetric positive definite problems with constant coefficients can be solved extremely efficient. It is possible to approximately solve a sparse $N \times N$ system of linear equations in $\mathcal{O}(N)$ arithmetic operations.

Details are not presented here, but the literature about multigrid solvers is very rich.

Multilevel iterations can also be used as preconditioners in Krylov subspace methods, see above.

11 Error estimates via dual techniques

All error estimates above gave estimates for the energy (\mathbb{H}^1) norm of the error. We will see, that the L^2 norm of the error (which is a weaker norm) can converge faster (with larger power of h) when the problem has enough regularity. This goes conforming with the better interpolation estimates in L^2 , compare Theorems 6.3 and 6.7.

Similar techniques to the one used in this section will later be used for error estimation in case of parabolic problems.

For simplicity, we consider here again just the Poisson problem.

11.1 Estimates for the L_2 norm of the error

Let $\Omega \subset \mathbb{R}^n$ a *convex* domain with polygonal boundary, and $f \in L^2(\Omega)$ given. Let $X = \mathbb{H}_0^1(\Omega)$ and $X_h \subset X$ a finite-dimensional finite element space with $\mathbb{P}_1(T) \subset X_h|_T$ for all $T \in \mathcal{T}$.

Let

$$\begin{aligned} u \in X : \quad (\nabla u, \nabla v)_{L^2(\Omega)} &= (f, v)_{L^2(\Omega)} \quad \forall v \in X, \\ u_h \in X_h : \quad (\nabla u_h, \nabla v_h)_{L^2(\Omega)} &= (f, v_h)_{L^2(\Omega)} \quad \forall v_h \in X_h. \end{aligned}$$

We want to estimate the L^2 norm of the error, $\|u - u_h\|_{L^2(\Omega)}$. For this purpose, we consider the

Problem 11.1. *Dual Problem:*

$$w \in X : \quad (\nabla v, \nabla w)_{L^2(\Omega)} = (u - u_h, v)_{L^2(\Omega)} \quad \forall v \in X.$$

Now let us assume that the solution w of the dual problem 11.1 is smooth, i.e.

$$w \in \mathbb{H}^2(\Omega) \cap \mathbb{H}_0^1(\Omega)$$

and there is an a priori estimate

$$(11.1) \quad \|w\|_{\mathbb{H}^2(\Omega)} \leq \|-\Delta w\|_{L^2(\Omega)} = \|u - u_h\|_{L^2(\Omega)}.$$

For a convex domain Ω , this follows from regularity theory of elliptic problems, see [46], e.g.

11.1 Remark. For general (non-symmetric) elliptic problems

$$u \in X : \quad a(u, v) = \langle f, v \rangle \quad \forall v \in X,$$

the corresponding dual problem is

$$w \in X : \quad a(v, w) = (u - u_h, v)_{L^2(\Omega)} \quad \forall v \in X.$$

Besides the convexity of the domain, the coefficients must be smooth enough in order to assure a dual solution $w \in \mathbb{H}^2(\Omega)$. Additionally, there will be constants (> 1) in the a priori estimate (11.1).

11.2 A priori error estimation in the L_2 norm

Let $w \in X$ be the solution of Problem 11.1. Then for test function $v := u - u_h$ holds

$$\begin{aligned} \|u - u_h\|_{L^2(\Omega)}^2 &= (u - u_h, u - u_h)_{L^2(\Omega)} \\ &= (\nabla(u - u_h), \nabla w)_{L^2(\Omega)} \\ &= (\nabla(u - u_h), \nabla(w - w_h))_{L^2(\Omega)} \quad \forall w_h \in X_h \\ &\leq \|\nabla(u - u_h)\|_{L^2(\Omega)} \|\nabla(w - w_h)\|_{L^2(\Omega)}. \end{aligned}$$

We can insert any discrete $w_h \in X_h$ because of the orthogonality of the error $(u - u_h)$ to the discrete space. Now, we choose an interpolation of w , $w_h = \mathbb{I}w \in X_h$, for example interpolation with piecewise linear finite element functions. Due to the regularity of w , we have the interpolation estimate

$$\|\nabla(w - w_h)\|_{L^2(\Omega)} \leq ch(\mathcal{T})|w|_{\mathbb{H}^2(\Omega)} \leq ch(\mathcal{T})\|u - u_h\|_{L^2(\Omega)}.$$

Putting the above estimates and the a priori error estimate together, we get

$$\begin{aligned} \|u - u_h\|_{L^2(\Omega)} &\leq ch(\mathcal{T})\|u - u_h\|_{L^2(\Omega)} \\ &\leq ch(\mathcal{T}) \left(\sum_{T \in \mathcal{T}} h_T^2 \|D^2 u\|_{L^2(T)^2} \right)^{\frac{1}{2}} \\ &\leq ch(\mathcal{T})^2 \|D^2 u\|_{L^2(\Omega)} \\ &\leq ch(\mathcal{T})^2 \|f\|_{L^2(\Omega)}. \end{aligned}$$

So, the estimate for the L^2 norm of the error behaves like $h(\mathcal{T})^2$, just like the interpolation estimate.

11.2 Remark. The above technique is named after Aubin and Nitsche, who were the first to introduce it (separately). For more details, see [72] or [21]).

When the solution of the dual problem w has less regularity, for example in case of non-convex corners in the domain's boundary, we get estimates with lower powers of $h(\mathcal{T})$. For example, when $\|w\|_{\mathbb{W}^{2,p}} \leq c\|u - u_h\|_{L^p}$ holds only for $1 < p \leq 2$, then the power of h will be like $2 + (\frac{1}{2} - \frac{1}{p})n$.

11.3 A posteriori error estimation in the L_2 norm

To establish an L_2 a posteriori estimate we use again the dual problem 11.1 and assume regularity of its solution w .

Defining for $g \in \mathbb{H}^{-2}(\Omega) := (\mathbb{H}^2(\Omega) \cap \mathbb{H}_0^1(\Omega))^*$

$$|g|_{H^{-2}(\Omega)} := \sup_{\substack{v \in H^2(\Omega) \cap H_0^1(\Omega) \\ \|v\|_{H^2(\Omega)} = 1}} \langle g, v \rangle_{H^{-2}(\Omega) \times (H^2(\Omega) \cap H_0^1(\Omega))},$$

using the fact that $w \in \mathbb{H}^2(\Omega)$, and setting $v = u - u_h$ in Problem 11.1, we conclude

$$\begin{aligned}
\|u - u_h\|_{L_2(\Omega)}^2 &= \int_{\Omega} \nabla(u - u_h) \nabla w \\
&= \left\langle -\Delta(u - u_h), \underbrace{w}_{\in H^2(\Omega)} \right\rangle_{H^{-1}(\Omega) \times H_0^1(\Omega)} \\
&= \left\langle F + \Delta u_h, w \right\rangle_{H^{-2}(\Omega) \times (H^2(\Omega) \cap H_0^1(\Omega))} \\
&\leq |F + \Delta u_h|_{H^{-2}(\Omega)} |w|_{H^2(\Omega)} \\
&\leq |F + \Delta u_h|_{H^{-2}(\Omega)} \|u - u_h\|_{L_2(\Omega)}.
\end{aligned}$$

On the other hand using the higher regularity of the test function v and integration by parts we have

$$\begin{aligned}
|F + \Delta u_h|_{H^{-2}(\Omega)} &= \sup_{\substack{v \in H^2(\Omega) \cap H_0^1(\Omega) \\ |v|_{H^2(\Omega)} = 1}} \left\langle F + \Delta u_h, v \right\rangle_{H^{-2}(\Omega) \times (H^2(\Omega) \cap H_0^1(\Omega))} \\
&= \sup_{\substack{v \in H^2(\Omega) \cap H_0^1(\Omega) \\ |v|_{H^2(\Omega)} = 1}} \int_{\Omega} \nabla(u - u_h) \nabla v \\
&= \sup_{\substack{v \in H^2(\Omega) \cap H_0^1(\Omega) \\ |v|_{H^2(\Omega)} = 1}} \int_{\Omega} (u - u_h) (-\Delta v) \\
&\leq \|u - u_h\|_{L_2(\Omega)}.
\end{aligned}$$

Combining these two estimates we achieve

$$(11.2) \quad \|u - u_h\|_{L_2(\Omega)} = |F + \Delta u_h|_{H^{-2}(\Omega)}.$$

In order to establish the L_2 estimate, we have to estimate now the term $|F + \Delta u_h|_{H^{-2}(\Omega)}$. This is done in the same manner as in the case of the a posteriori energy norm estimates. In contrast to that estimate we can use the fact that the test function v belongs to $\mathbb{H}^2(\Omega)$. Thus, for the interpolation of v we can make use of the usual Lagrange interpolant ($\mathbb{H}^2(\Omega)$ is embedded in $C^0(\bar{\Omega})$, $d = 2, 3!$) and we gain a higher power of h_T in front of the residuals since we can rely on second derivatives of v . As a result we have

$$(11.3) \quad \|u - u_h\|_{L_2(\Omega)} \leq c \left(\sum_{T \in \mathcal{T}} \tilde{\eta}_T(u_h, f)^2 \right)^{\frac{1}{2}}$$

where $\tilde{\eta}_T$ is defined to be

$$\tilde{\eta}_T(u_h, f)^2 := h_T^4 \|f + \Delta u_h\|_{L_2(T)}^2 + \frac{1}{2} h_T^3 \|[\partial_\nu u_h]\|_{L_2(\partial T \setminus \partial \Omega)}^2.$$

Again, using the finite dimensional approximation f_h of f we can prove the efficiency

$$(11.4) \quad \tilde{\eta}_T(u_h, f_h) \leq c \left(\|u - u_h\|_{L_2(M(T))} + h_T^2 \|f - f_h\|_{L_2(M(T))} \right)$$

where we also gain one additional power of h_T in front of the term $\|f - f_h\|_{L_2(M(T))}$. This analysis also carries over to nonlinear problems under suitable assumptions on the existence of the dual problem and the regularity of its solution. Under such assumptions we can prove

$$c \|u - u_h\|_{L_2(\Omega)} \leq \|F(u_h)\|_{H^2(\Omega) \cap H_0^1(\Omega)} \leq C \|u - u_h\|_{L_2(\Omega)}$$

where now c and C depend on the coercivity of the dual problem (which is associated to the norms of DF and DF^{-1}) and the regularity constant for the solution of the dual problem. This inequality now establishes L_2 error estimators for nonlinear problems using the same techniques as described above [8].

12 Parabolic problems - heat equation

In this section we introduce the weak formulation of a model parabolic problem (heat equation) derived in section 1. The heat equation is classically described by the following evolution equation plus initial and boundary conditions (for the sake of simplicity we set all material parameters equal to one)

$$(12.1) \quad u_t - \Delta u = f \text{ in } \Omega \times (0, T),$$

where $u = u(x, t)$ represents the temperature, t is the time variable, $T > 0$ is a fixed number, $\Omega \subset \mathbb{R}^d$ is as usual a bounded domain with a Lipschitz boundary $\partial\Omega$ and $f = f(x, t)$ describes the volume heat sources.

As for the boundary conditions, we impose Dirichlet boundary conditions on the whole boundary $\partial\Omega \times (0, T)$

$$(12.2) \quad u = 0 \text{ on } \partial\Omega \times (0, T),$$

The cases of other type of boundaries (Neumann, Robin) are treated analogously.

Finally, we assume that the temperature distribution at $t = 0$ is given

$$(12.3) \quad u(x, 0) = u_0(x) \quad x \in \Omega.$$

12.1 Weak solutions of heat equation

We start the formulation of weak solutions by the definition of Bochner integral.

Definition 12.1. *A function $u : [a, b] \rightarrow X$ is called Bochner measurable, if there exists a sequence of simple functions $\{u_k\}_{k \in \mathbb{N}}$ such that*

$$\lim_{k \rightarrow \infty} u_k(t) = u(t) \text{ for almost all } t \in [a, b].$$

Definition 12.2. *If additionally*

$$\lim_{k \rightarrow \infty} \int_a^b \|u(t) - u_k(t)\|_X dt = 0$$

then the function u is called Bochner-integrable and the integral

$$\int_a^b u(t) dt = \lim_{k \rightarrow \infty} \int_a^b u_k(t) dt$$

is said to be the Bochner-integral of u .

Let assume that $-\infty < a < b < \infty$, the function $u : [a, b] \rightarrow X$ is Bochner-measurable. Then we define the following functional spaces

$$\mathbb{L}^p(a, b; X) := \{u : [a, b] \rightarrow X \text{ Bochner-measurable; } \|u\|_{L^p(a,b;X)} < \infty\} \quad 1 \leq p < \infty$$

and

$$\mathbb{C}^0([a, b]; X) := \{u : [a, b] \rightarrow X \text{ continuous}\}$$

with corresponding norms

$$\|u\|_{L^p(a,b;X)} := \left(\int_a^b \|u(t)\|_X^p dt \right)^{\frac{1}{p}} \quad 1 \leq p < \infty.$$

$$\|u\|_{C^0([a,b];X)} := \max_{t \in [a,b]} \|u(t)\|_X$$

Theorem 12.1. *Let $-\infty < a < b < \infty$ and X is a Banach space. Then the spaces $(\mathbb{L}^p(a, b; X), \|\cdot\|_{L^p(a,b;X)})$ and $(\mathbb{C}^0([a, b]; X), \|\cdot\|_{C^0(a,b;X)})$ are Banach spaces.*

Further we define the weak derivative of a Bochner-integrable function.

Definition 12.3. *Let $(X, (\cdot, \cdot))$ be a separable Hilbert space, $0 < T < \infty$ and $u \in \mathbb{L}^1(0, T; X)$. A function $v \in \mathbb{L}^1(0, T; X)$ is called the weak derivative of u if*

$$\int_0^T u(t) \varphi'(t) dt = - \int_0^T v(t) \varphi(t) dt \quad \forall \varphi \in \mathbb{C}_0^\infty(0, T).$$

Then we write $u' = v$ (if such a v exists, of course).

We define the functional space

$$\mathbb{W}_2^1(0, T) := \{u \in \mathbb{L}^2(0, T; H_0^1(\Omega)), u' \in \mathbb{L}^2(0, T; H^{-1}(\Omega))\}$$

with a norm

$$\begin{aligned} \|u\|_{\mathbb{W}_2^1(0,T)} &:= \left(\int_0^T \|u(t)\|_{H_0^1(\Omega)}^2 dt + \int_0^T \|u'(t)\|_{H^{-1}(\Omega)}^2 dt \right)^{\frac{1}{2}} \\ &= \left(\int_0^T \int_\Omega |\nabla u|^2 + \int_0^T \left(\sup_{v \in H_0^1(\Omega)} \frac{\langle u'(t), v \rangle}{\|v\|_{H_0^1(\Omega)}} \right)^2 \right)^{\frac{1}{2}} \end{aligned}$$

Assume that we have a classical solution u of (12.1)+(12.2)+(12.3) and it is sufficiently smooth, more precisely $u \in \mathbb{C}^{2,1}(\overline{\Omega \times (0, T)})$. We now fix $t \in (0, T)$, multiply the equation (12.1) by a test function $v \in \mathbb{H}_0^1(\Omega)$ and do partial integration over Ω

$$\langle u'(t), v \rangle_{H^{-1} \times H_0^1} + \int_\Omega \nabla u \nabla v dx = \int_\Omega f v dx \quad \forall v \in \mathbb{H}_0^1(\Omega).$$

Similar to the elliptic case let us introduce a bilinear form

$$b(u(t), v; t) := \int_{\Omega} \nabla u \nabla v dx \quad \forall u, v \in \mathbb{H}_0^1(\Omega).$$

and a linear form

$$F(v) := \int_{\Omega} f v dx \quad \forall v \in \mathbb{H}_0^1(\Omega).$$

One can easily show that the bilinear form $b(\cdot, \cdot; t)$ is continuous on $\mathbb{H}_0^1(\Omega)$ and $u'(t) \in \mathbb{H}^{-1}(\Omega)$. Using these notations we arrive at

$$(12.4) \quad \langle u'(t), v \rangle + b(u(t), v; t) = F(v) \quad \forall v \in \mathbb{H}_0^1(\Omega).$$

Now the definition of the weak solution to the classical problem reads:

Definition 12.4. *A function $u \in \mathbb{W}_1^2(0, T)$ is called the weak solution of the classical problem (12.1)+(12.2)+(12.3), if u satisfies the equation (12.4) for almost every fixed $t \in (0, T)$ and the condition (12.3) for almost every $x \in \Omega$.*

We state the last result in this section (without proof though) which provides us with the existence and uniqueness of the weak solution.

Theorem 12.2. *For any right-hand side $f \in \mathbb{C}([0, T]; L^2(\Omega))$ and initial function $u_0 \in L^2(\Omega)$ there exists one and only one weak solution $u \in \mathbb{W}_1^2(0, T)$ of the classical heat equation and*

$$\int_0^T \|u(t)\|_{H_0^1(\Omega)}^2 dt + \int_0^T \|u'(t)\|_{H^{-1}(\Omega)}^2 dt \leq C(\Omega, d, \dots) \left(\|u_0\|_{L^2(\Omega)}^2 + \int_0^T \|f(t)\|_{L^2(\Omega)}^2 dt \right)$$

12.2 Discretization of heat equation

Let $X_h \subset \mathbb{H}_0^1(\Omega)$ be a finite dimensional subspace (it can be for example a finite element space) and $\{\varphi_1, \varphi_2, \dots, \varphi_N\}$ be the basis of X_h . Recall the weak formulation of the heat equation with vanishing Dirichlet boundary conditions: find a function $u \in \mathbb{W}_2^1(0, T)$ such that

$$(12.5) \quad \langle u'(t), v \rangle + (\nabla u(t), \nabla v)_{L^2(\Omega)} = F(v) \quad \forall v \in \mathbb{H}_0^1(\Omega) \text{ and a.e. } t \in (0, T).$$

and

$$u(x, 0) = u_0(x) \quad x \in \Omega$$

Space Discretization: The Galerkin approximation of the equation (12.4) consists of finding a discrete solution $u_h(x, t) \in X_h$ such that for almost every $t \in (0, T)$

$$(12.6) \quad \langle u_h'(t), v_h \rangle + (\nabla u_h(t), \nabla v_h)_{L^2(\Omega)} = F(v_h) \quad \forall v_h \in X_h,$$

and

$$(12.7) \quad u_h(x, 0) = u_{0,h} \in X_h.$$

This is called the *semi*-discrete Galerkin scheme. The reason to call it *semi*-discrete is that we have performed only a space discretization.

Since $u_h(t) \in X_h$, then we can present it with the help of the basis functions of X_h in the following way

$$(12.8) \quad u_h(t) = \sum_{i=1}^N \alpha_i(t) \varphi_i,$$

and the gradient of u_h is obtained to be

$$(12.9) \quad \nabla u_h(t) = \sum_{i=1}^N \alpha_i(t) \nabla \varphi_i,$$

Then the discrete equation (12.6) is equivalent to a system of ordinary differential equations for the unknowns $\alpha_1(t), \alpha_2(t), \dots, \alpha_N(t)$

$$(12.10) \quad \sum_{j=1}^N (\varphi_i, \varphi_j)_{L^2(\Omega)} \alpha_j'(t) + \sum_{j=1}^N (\nabla \varphi_i, \nabla \varphi_j)_{L^2(\Omega)} \alpha_j(t) = F(\varphi_i) \quad i = 1, 2, \dots, N.$$

Denoting by $\alpha(t) := (\alpha_1(t), \alpha_2(t), \dots, \alpha_N(t))^T$ the vector of unknowns, $M = (M_{ij}) := (\varphi_i, \varphi_j)_{L^2(\Omega)}$ the mass matrix, $S = (S_{ij}) := (\nabla \varphi_i, \nabla \varphi_j)_{L^2(\Omega)}$ the stiffness matrix and $F := (F(\varphi_1), F(\varphi_1), \dots, F(\varphi_N))^T$ the load vector, we can rewrite the equation (12.10) in the matrix form

$$(12.11) \quad M\alpha'(t) + S\alpha(t) = F(t).$$

It can be shown that there exists a unique vector $\alpha(t)$ which solves the system (12.11). Thus we arrive to the following result:

Theorem 12.3. *Let $T > 0$, $f \in \mathbb{L}^2(0, T; L^2(\Omega))$ and $u_0 = u_{0,h} \in X_h$. Then there exist unique functions $\alpha_i \in \mathbb{H}^1(0, T)$ for $1 \leq i \leq N$ such that the function $u_h(t) = \sum_{i=1}^N \alpha_i(t) \varphi_i$ solves the problem (12.6)+(12.7).*

Remark 12.1. *In the formulation of the theorem we require that the function u_0 belongs to the space X_h . If this condition fails, i.e. $u_0 \notin X_h$, then $u_{0,h}$ can be constructed as a projection of u_0 onto the space X_h . An example of such a projection is the interpolant $I_h u_0 \in X_h$ of u_0 .*

Time Discretization: Let Δt be the time step, α^n and F^n be the values of $\alpha(t)$ and $F(t)$ at the n -th time level, respectively. For discretization of the equation (12.11) in time it is common and natural to use the *explicit Euler* method where we replace the time derivative α' by the difference quotient

$$\alpha' = \frac{\alpha^{n+1} - \alpha^n}{\Delta t}$$

Substituting this in (12.11), we obtain

$$(12.12) \quad M \cdot \frac{\alpha^{n+1} - \alpha^n}{\Delta t} + S\alpha^n = F^n$$

The explicit Euler scheme is not a stable method, unless $\Delta t \leq Ch^2$. When this condition on time step fails then there will be instabilities. To avoid this problem, it is convenient to use the *implicit Euler* method which for the equation (12.11) can be written as

$$(12.13) \quad M \cdot \frac{\alpha^{n+1} - \alpha^n}{\Delta t} + S\alpha^{n+1} = F^{n+1}$$

or

$$(12.14) \quad (M + \Delta t S)\alpha^{n+1} = M\alpha^n + \Delta t F^{n+1},$$

with

$$(12.15) \quad \alpha^0 = \alpha(0) = (\alpha_1(0), \alpha_2(0), \dots, \alpha_n(0))^T, \quad n = 0, 1, 2, \dots$$

The implicit Euler method compared to the explicit one is unconditionally stable, i.e. no condition is forced on the time step Δt .

We now turn to *Crank-Nicholson* method. Here the semi-discrete equation (12.11) is discretized in a symmetric way around the point $(n + \frac{1}{2})\Delta t$. Here the equation for α^{n+1} can be written in matrix form as follows

$$(12.16) \quad (M + \frac{\Delta t}{2}S)\alpha^{n+1} = (M - \frac{\Delta t}{2}S)\alpha^n + \Delta t \frac{F^{n+1} + F^n}{2},$$

with

$$(12.17) \quad \alpha^0 = \alpha(0) = (\alpha_1(0), \alpha_2(0), \dots, \alpha_n(0))^T, \quad n = 0, 1, 2, \dots$$

This scheme will produce a second order accurate method in time.

A more general time discretization scheme is the so called θ -*scheme*. For $\theta \in [0, 1]$ and a variable time step $\Delta t_{n+1} = t_{n+1} - t_n$ we search for α^{n+1} such that

$$\frac{1}{\Delta t_{n+1}} M\alpha^{n+1} + \theta S\alpha^{n+1} = \frac{1}{\Delta t_{n+1}} M\alpha^n + (1 - \theta)S\alpha^n + F^{n+\theta\Delta t_{n+1}}$$

Note that for $\theta = 0$ we obtain the explicit Euler scheme, for $\theta = 1$ the implicit Euler scheme and for $\theta = \frac{1}{2}$ the Crank-Nicholson scheme. One can show that for $0.5 \leq \theta \leq 1$ the θ -scheme is unconditionally stable, while for $\theta < 0.5$ stability is only guaranteed if the time step is sufficiently small.

12.3 A priori error estimates

A priori estimates for the error between continuous and discrete solution can be derived by using the scheme described above,

- first proving an approximation result for the solution of the semi-discrete problem (where discretization is done only in space, not in time),
- and then using more or less the standard estimates for time discretization of systems of ordinary differential equations.

For example, for the backward Euler time discretization and piecewise linear finite element space over a triangulation \mathcal{T} in space, the error estimate reads like: Under the condition, that u is smooth enough, holds

$$\begin{aligned} \max_{0 \leq k \leq M} \|u(t_k) - u_{h,k}\|_{L^2(\Omega)}^2 &\leq c \left(\|u_o - u_{h,0}\|_{L^2(\Omega)}^2 + h(\mathcal{T})^4 \|u_o\|_{H^2(\Omega)}^2 \right. \\ &\quad \left. + h(\mathcal{T})^4 \int_0^{t_M} \left\| \frac{\partial}{\partial t} u \right\|_{H^2(\Omega)}^2 dt + (\Delta t)^2 \int_0^{t_M} \left\| \frac{\partial^2}{\partial t^2} u \right\|_{L^2(\Omega)}^2 dt \right) \end{aligned}$$

So, the error converges to zero quadratically in $h(\mathcal{T})$ and linearly in Δt . For the Crank-Nicholson scheme, you get also quadratic convergence in Δt , but under higher regularity assumptions on u .

We will not go into more detail here.

13 A posteriori error estimates for parabolic problems

13.1 Abstract error estimate for ordinary differential equations

Let \mathbb{H} be a Hilbert space and we are interested in the solution of the following parabolic equation

$$(13.1) \quad \begin{aligned} u'(t) + F(u(t)) &= f(t) \text{ in } \mathbb{H}^*, \quad t \in (0, T), \\ u(0) &= u_0, \end{aligned}$$

where $u_0 \in \mathbb{H}$ is the initial function, $f(t) \in \mathbb{H}^*$ and $F : \mathbb{H} \rightarrow \mathbb{H}^*$ is a smooth, Lipschitz continuous function, i.e. there is a constant l such that

$$\|F(u) - F(v)\|_{\mathbb{H}^*} \leq l \|u - v\|_{\mathbb{H}}.$$

After the time discretization of the equation (13.1) with a variable time step $\tau_n = t_n - t_{n-1} > 0$, $t_0 = 0$ and $t_N = T$ (implicit Euler Discretization), we obtain

$$(13.2) \quad \begin{aligned} u^n \in \mathbb{H} : \quad \frac{u^n - u^{n-1}}{\tau_n} + F(u^n) &= f_n := f(t_n), \quad n = 1, \dots, N, \\ u^0 &= u_0, \end{aligned}$$

Our goal is to obtain an equation for the error $e := u - U$, where $U := u^{n-1} + \frac{t-t_{n-1}}{\tau_n}(u^n - u^{n-1})$. In the interval (t_{n-1}, t_n) we have

$$U'(t) = \frac{u^n - u^{n-1}}{\tau_n}$$

Let \bar{U} denote the piecewise constant function (with respect to t) defined by

$$\bar{U}(t) := U^n \text{ for } t \in (t_{n-1}, t_n].$$

Then (13.2) is equivalent to

$$(13.3) \quad \begin{aligned} U' + F(\bar{U}) &= \bar{f} \text{ in } (0, T), \\ U^0 &= U_0. \end{aligned}$$

From (13.1) and (13.2) we get

$$\begin{aligned} (u - U)' + F(u) - F(\bar{U}) &= f - \bar{f}, \\ (u - U)' + F(u) - F(U) &= \underbrace{f - \bar{f} + F(\bar{U}) - F(U)}_{=-U'} \\ &= f - U' - F(U) := \underbrace{R(U)}_{Residual} \end{aligned}$$

For the initial condition we get

$$(u - U)(0) = 0.$$

Now we use the property of F being Lipschitz continuous, meaning that the Frechet derivative DF exists. Thus we can define a continuous linear operator $L : \mathbb{H} \rightarrow \mathbb{H}^*$

$$L := \int_0^1 DF(su + (1-s)U) ds.$$

Then

$$F(u) - F(U) = L(u - U),$$

and we arrive to the following equation for $e = u - U$

$$(13.4) \quad \begin{aligned} e' + Le &= R(U) \\ e(0) &= 0. \end{aligned}$$

In the case when F is linear then $L = F'$ independent on u . If F is non linear, then the operator L contains the exact solution u and therefore, we need a priori error estimates for L .

Let us rewrite the equation (13.4) in weak form:

$$\int_0^T \langle e', \varphi \rangle + \langle Le, \varphi \rangle dt = \int_0^T \langle R(U), \varphi \rangle dt \quad \text{for all } \varphi \in \mathbb{H}.$$

After integration by parts we obtain

$$(13.5) \quad \begin{aligned} \langle e(T), \varphi(T) \rangle - \langle e(0), \varphi(0) \rangle &= \int_0^T \langle e, \varphi' \rangle - \langle e, L^* \varphi \rangle dt + \int_0^T \langle R(U), \varphi \rangle dt, \\ \langle e(T), \varphi(T) \rangle &= \langle e(0), \varphi(0) \rangle + \int_0^T \langle e, \varphi' - L^* \varphi \rangle dt + \int_0^T \langle R(U), \varphi \rangle dt. \end{aligned}$$

Let φ be the solution of the *dual problem*

$$(13.6) \quad \begin{aligned} \varphi' - L^* \varphi &= 0 \text{ for almost all } t \in (0, T) \\ \varphi(T) &= e(T). \end{aligned}$$

With φ being the solution of (13.6) the equation (13.5) is equivalent to

$$\|e(T)\|^2 = \langle e(0), \varphi(0) \rangle + \int_0^T \langle R(U), \varphi \rangle dt$$

What we need is to find some stability estimation for the solution of the dual problem (13.6) which depends on the problem data and does not depend on the exact solution u . One example of such an estimation might be

$$\|\varphi(t)\|_H \leq C_s \|e(T)\|_H,$$

from which we could calculate

$$\|e(T)\|_H^2 \leq C_s \left(\|e(0)\|_{H^*} + \int_0^T \|R(U)\|_{H^*} dt \right) \|e(T)\|_H.$$

Finally, we would obtain a bound for $e(T)$

$$\|e(T)\|_H \leq C_s \left(\|e(0)\|_{H^*} + \int_0^T \|R(U)\|_{H^*} dt \right)$$

13.2 Weak formulation of the heat equation

Let again Ω be a bounded domain with polygonal boundary $\partial\Omega$. We set as before

$$\mathbb{W}_1^2(0, T) := \{v \in \mathbb{L}^2(0, T; H_0^1(\Omega)), v' \in \mathbb{L}^2(0, T; H^{-1}(\Omega))\}.$$

Suppose that the right-hand side $f \in \mathbb{L}^2(0, T; H^{-1}(\Omega))$ and the initial condition $u_0 \in \mathbb{H}_0^1(\Omega)$ are given and let $u \in \mathbb{W}_1^2(0, T)$ be the weak solution of the heat equation

$$(13.7) \quad \begin{aligned} u'(t) - \Delta u &= f \text{ in } \mathbb{H}^{-1}(\Omega), \quad t \in (0, T), \\ u(0) &= u_0, \end{aligned}$$

i.e. for $u \in \mathbb{W}_1^2(0, T)$ we consider the equation

$$(13.8) \quad \begin{aligned} \langle u', \varphi \rangle + \int_{\Omega} \nabla u \cdot \nabla \varphi - \int_{\Omega} f \varphi &= 0 \quad \forall \varphi \in \mathbb{H}_0^1(\Omega), \quad t \in (0, T), \\ u(0) &= u_0, \end{aligned}$$

13.3 Discretization

Let $0 = t_0 < t_1 < \dots < t_N = T$ define a subdivision of $(0, T)$, denote $I_n = (t_{n-1}, t_n)$ and $\tau_n = t_n - t_{n-1}$. For $n = 1, \dots, N$ let \mathcal{T}_n be a proper and shape regular triangulation of Ω and $X_n \subset \mathbb{H}_0^1(\Omega)$ a corresponding finite element space with piecewise polynomial functions of degree $p \geq 1$ on \mathcal{T}_n .

Let W_n denote the space of (constant in time) functions $w(t) = w \in X_n$, and $W_{h,\tau} := \{w \in L^2(0, t; \mathbb{H}_0^1(\Omega)) : w|_{I_n} \in W_n, n = 1, \dots, N\}$. For $w \in W_{h,\tau}$ define

$$w^+(x, t_n) := \lim_{s \searrow 0} w(x, t_n + s), \quad w^-(x, t_n) := \lim_{s \searrow 0} w(x, t_n - s),$$

and we define the *jump* of w at t_n by

$$[w]_n := w_n^+ - w_n^-, \quad \text{where} \quad w_n^+ := w^+(\cdot, t_n), \quad w_n^- := w^-(\cdot, t_n).$$

13.1 Remark. The whole procedure can be also done with higher order approximation on the time interval, for example piecewise linear in time. This would just change the definition of spaces W_n . Nevertheless, the discrete functions $w \in W_{h,\tau}$ will be in general not continuous at times t_n .

Now, the discrete problem reads:

Problem 13.1. *The discrete solution is $U \in W_{h,\tau}$ such that for all $w \in W_{h,\tau}$ holds*

$$(13.9) \quad \int_0^T (U_t, w) + (\nabla U, \nabla w) dt + \sum_{n=1}^N ([U]_{n-1}, w_{n-1}^+) = \int_0^T (f, w) dt,$$

where U_0^- is an approximation to u_0 in X_1 .

13.4 Error representation and dual problem

Let $u \in \mathbb{W}_1^2(0, T)$ be the solution of (13.8) and $U \in W_{h,\tau}$ the discrete solution of (13.9). Then it holds in $\mathbb{W}_1^2(0, T)^*$

$$\begin{aligned} u_t - \Delta u &= f, \\ U_t - \Delta U &= f - R, \end{aligned}$$

where the *residual* $R \in \mathbb{W}_1^2(0, T)^*$ is just defined by the second equation. So, for all $M = 1, \dots, N$ and $v \in \mathbb{W}_1^2(0, T)$ holds

$$\begin{aligned} \int_0^{t_M} (u_t, v) + (\nabla u, \nabla v) - (f, v) dt &= 0, \\ \int_0^{t_M} (U_t, v) + (\nabla U, \nabla v) - (f, v) dt &= -R(v). \end{aligned}$$

Subtracting the second from the first equality, we get

$$\int_0^{t_M} (u_t - U_t, v) + (\nabla(u - U), \nabla v) dt = R(v).$$

Integration by parts in time now gives

$$(13.10) \quad (u(t_M) - U(t_M), v(t_M)) = (u(0) - U(0), v(0)) + \int_0^{t_M} (u - U, v_t) - (\nabla(u - U), \nabla v) dt + R(v).$$

So, the corresponding *dual problem* like (13.6) is now (using $F(v) = -\Delta v$, $L = L^* = -\Delta$):

Problem 13.2. *Given final values χ at t_M , the discrete dual solution is $v \in W_1^2(0, t_M)$ such that for all $\varphi \in H_0^1(\Omega)$ and almost all $t \in (0, t_M)$ holds*

$$(13.11) \quad v(t_m) = \chi, \quad (v_t, \varphi) - (\nabla v, \nabla \varphi) dt = 0.$$

This is a *backward parabolic problem*, where *final values* are prescribed. Thus, it is easy to prove existence of a unique solution.

Now, using the dual solution for final value $\chi = u(t_M) - U(t_M)$ in the error representation (13.10), we get

$$\|u(t_M) - U(t_M)\|_{L^2(\Omega)}^2 \leq \|u(0) - U(0)\|_{L^2(\Omega)} \|v(0)\|_{L^2(\Omega)} + |R(v)|.$$

This gives an estimate of the error at time t_M in terms of the initial error and the dual solution.

Using now the definition of $R(v)$, we can show that for every $w \in W_{h,\tau}$ holds

$$(13.12) \quad \begin{aligned} & \|u(t_M) - U(t_M)\|_{L^2(\Omega)} \leq \\ & \leq \sup_{\chi \in \mathbb{L}^2(\Omega), \|\chi\|=1} \left| \sum_{n=1}^M \int_{I_n} (U_t, v_\chi - w) dt + (\nabla U, \nabla(v_\chi - w)) - (f, v_\chi - w) dt \right. \\ & \quad \left. + ([U]_{n-1}, v_\chi(t_{n-1}) - w_{n-1}^+) \right| + \|u(0) - U_0^-\|_{L^2(\Omega)}, \end{aligned}$$

where v_χ denotes the dual solution with final value χ .

In order to estimate the terms involving the dual solution, we need some a priori estimates for this:

Lemma 13.1. *Let $v \in \mathbb{W}_1^2(0, t_M)$ be the solution of (13.11) with final value $v(t_M) = \chi \in \mathbb{L}^2(\Omega)$. Then for almost all $t \in (0, t_M)$ we have that $v(t) \in \mathbb{H}^2(\Omega)$, $v'(t) \in \mathbb{L}^2(\Omega)$ and the following conditions hold:*

$$(i) \quad \|v(t)\|_{L^2(\Omega)} \leq \|\chi\|_{L^2(\Omega)},$$

$$(ii) \quad \int_t^{t_M} \|\nabla v(s)\|_{L^2(\Omega)}^2 ds \leq \frac{1}{2} \|\chi\|_{L^2(\Omega)}^2,$$

$$(iii) \quad \int_t^{t_M} (t_M - s) \|\Delta v(s)\|_{L^2(\Omega)}^2 ds \leq \frac{1}{4} \|\chi\|_{L^2(\Omega)}^2,$$

$$(iv) \quad \int_t^{t_M} (t_M - s) \|v_t\|_{L^2(\Omega)}^2 ds \leq \frac{1}{4} \|\chi\|_{L^2(\Omega)}^2,$$

$$(v) \quad \|\Delta v(t)\|_{L^2(\Omega)} \leq \frac{1}{\sqrt{2}(t_M - t)} \|\chi\|_{L^2(\Omega)},$$

$$(vi) \quad \int_t^{t_M} \|\Delta v(s)\|_{L^2(\Omega)} ds \leq \frac{1}{2} \left(\log \frac{t_M}{\tau_M} \right)^{\frac{1}{2}} \|\chi\|_{L^2(\Omega)},$$

$$\begin{aligned}
(vii) \quad & \int_t^{t_M} \|v_t(s)\|_{L^2(\Omega)} ds \leq \frac{1}{2} \left(\log \frac{t_M}{\tau_M} \right)^{\frac{1}{2}} \|\chi\|_{L^2(\Omega)}, \\
(viii) \quad & \sum_{n=1}^M \tau_n \|\Delta v(t_{n-1})\|_{L^2(\Omega)} \leq \left(\frac{1}{2} \left(\log \frac{t_M}{\tau_M} \right)^{\frac{1}{2}} + \frac{1}{\sqrt{2}} \right) \|\chi\|_{L^2(\Omega)}, \\
(ix) \quad & \sum_{n=1}^M \tau_n \|v_t(t_{n-1})\|_{L^2(\Omega)} \leq \left(\frac{1}{2} \left(\log \frac{t_M}{\tau_M} \right)^{\frac{1}{2}} + \frac{1}{\sqrt{2}} \right) \|\chi\|_{L^2(\Omega)}.
\end{aligned}$$

The proof is done by mainly using just the right choice of test functions φ in the weak formulation.

13.5 A posteriori error estimate

Now, we can choose in (13.12) a special $w \in W_{h,\tau}$: for $t \in I_n$, set $w := P_n v_\chi := P_{I_n} P_{X_n} v_\chi = P_{X_n} P_{I_n} v_\chi$, where P_{I_n} and P_{X_n} denote the L^2 projections onto the piecewise constant functions on I_n and on X_n . The following estimates hold:

$$\begin{aligned}
\|h^{-2}(v - P_{X_n} v)\| + \|h^{-1} \nabla(v - P_{X_n} v)\| &\leq c \|D^2 v\| \leq c \|\Delta v\| \quad \text{for } v \in \mathbb{H}^2(\Omega), \Omega \text{ convex}, \\
\|v - P_{I_n} v\|_{L^\infty(I_n)} &\leq c \|v\|_{L^\infty(I_n)} \quad \text{for } v \in L^\infty(I_n), \\
\|v - P_{I_n} v\|_{L^\infty(I_n)} &\leq c \int_{I_n} |v_t| dt \quad \text{for } v \in \mathbb{W}^{1,1}(I_n).
\end{aligned}$$

Using the a priori estimates from Lemma 13.1 and above approximation results, we have

13.2 Theorem. *Let $u \in \mathbb{W}_1^2(0, T)$ the solution of the heat equation (13.8) and $U \in W_{h,\tau}$ the discrete solution of (13.9). The the following a posteriori error estimate holds:*

$$\begin{aligned}
(13.13) \quad & \|u(t_N) - U_N^-\|_{L^2(\Omega)} \leq \|u(0) - U_0^-\|_{L^2(\Omega)} + \\
& c \left(\log \frac{t_N}{\tau_N} \right)^{\frac{1}{2}} \max_{1 \leq n \leq N} \left(\left(\sum_{T \in \mathcal{T}_n} h_T^4 \|U_t - \Delta U - f\|_{L^\infty(I_n, L^2(T))}^2 + h_T^3 \|[n \cdot \nabla U]\|_{L^\infty(I_n, L^2(\partial T \setminus \partial \Omega))}^2 \right)^{\frac{1}{2}} \right. \\
& \quad \left. + \|[U]_{n-1}\|_{L^2(\Omega)} + \left\| \frac{h^2}{\tau} [U]_{n-1} \right\|_{L^2(\Omega)}^* \right).
\end{aligned}$$

The final term $\|\cdot\|^*$ appears only when $X_{n-1} \not\subset X_n$, for example when the mesh was coarsened.

The proof of this theorem was first done by Eriksson and Johnson [34].

14 Adaptive methods for parabolic problems

In this chapter we will construct some adaptive methods for choosing the space and time steps in a finite element method for a linear time dependent problem.

In parabolic problems, the mesh is adapted to the solution in every time step using a posteriori error estimators or indicators. This may be accompanied by an adaptive control of time step sizes, see below.

Bänsch [6] lists several different adaptive procedures (in space) for time dependent problems:

- **Explicit strategy:** The current time step is solved once on the mesh from the previous time step, giving the solution u_h . Based on a posteriori estimates of u_h , the mesh is locally refined and coarsened. The problem is *not* solved again on the new mesh, and the solve–estimate–adapt process is *not* iterated.
This strategy is only usable when the solution is nearly stationary and does not change much in time, or when the time step size is very small. Usually, a given tolerance for the error can not be guaranteed with this strategy.
- **Semi–implicit strategy:** The current time step is solved once on the mesh from the previous time step, giving an intermediate solution \tilde{u}_h . Based on a posteriori estimates of \tilde{u}_h , the mesh is locally refined and coarsened. This produces the final mesh for the current time step, where the discrete solution u_h is computed. The solve–estimate–adapt process is *not* iterated.
This strategy works quite well, if the time steps are not too large, such that regions of refinement move too fast.
- **Implicit strategy A:** In every time step starting from the previous time step’s triangulation, a mesh is generated using local refinement and coarsening based on a posteriori estimates of a solution which is calculated on the current mesh. This solve–estimate–adapt process is iterated until the estimated error is below the given bound.
This guarantees that the estimated error is below the given bound. Together with an adaptive control of the time step size, this leads to global (in time) error bounds. If the time step size is not too large, the number of iterations of the solve–estimate–adapt process is usually very small.
- **Implicit strategy B:** In every time step starting from the macro triangulation, a mesh is generated using local refinements based on a posteriori estimates of a solution which is calculated on the current (maybe quite coarse) mesh; no mesh coarsening is needed. This solve–estimate–adapt process is iterated until the estimated error is below the given bound.
Like implicit strategy A, this guarantees error bounds. As the initial mesh for every time step is very coarse, the number of iterations of the solve–estimate–adapt

process becomes quite large, and thus the algorithm might become expensive. On the other hand, a solution on a coarse grid is fast and can be used as a good initial guess for finer grids, which is usually better than using the solution from the old time step.

Implicit strategy B can also be used with anisotropically refined triangular meshes, see [40]. As coarsening of anisotropic meshes and changes of the anisotropy direction are still open problems, this implies that the implicit strategy A can not be used in this context.

The following algorithm implements one time step of the implicit strategy A. The adaptive algorithm ensures that the mesh refinement/coarsening is done at least once in each time step, even if the error estimate is below the limit. Nevertheless, the error might be not equally distributed between all elements; for some simplices the local error estimates might be bigger than allowed.

14.1 Algorithm (Implicit strategy A).

```

Start with given parameters tol and time step size  $\tau$ ,
the solution  $u_n$  from the previous time step on grid  $\mathcal{T}_n$ 
 $\mathcal{T}_{n+1} := \mathcal{T}_n$ 
solve the discrete problem for  $u_{n+1}$  on  $\mathcal{T}_{n+1}$  using data  $u_n$ 
compute error estimates on  $\mathcal{T}_{n+1}$ 
do
  mark elements for refinement or coarsening
  if elements are marked then
    adapt mesh  $\mathcal{T}_{n+1}$  producing a modified  $\mathcal{T}_{n+1}$ 
    solve the discrete problem for  $u_{n+1}$  on  $\mathcal{T}_{n+1}$  using data  $u_n$ 
    compute error estimates on  $\mathcal{T}_{n+1}$ 
  end if
while  $\eta > tol$ 

```

14.1 Adaptive control of the time step size

A posteriori error estimates for parabolic problems usually consist of four different types of terms:

- terms estimating the initial error;
- terms estimating the error from discretization in space;
- terms estimating the error from mesh coarsening between time steps;
- terms estimating the error from discretization in time.

Thus, the total estimate can be split into parts

$$\eta_0, \eta_h, \eta_c, \text{ and } \eta_\tau$$

estimating these four different error parts.

Example: Recall the a posteriori error estimate of Eriksson and Johnson [34] for the heat equation, Theorem 13.2, which reads for piecewise linear (\mathbb{P}_1) finite elements and piecewise constant approximation in time:

$$\begin{aligned} \|u(t_N) - U_N\| \leq & \|u_0 - U_0\|_{L^2(\Omega)} + \\ & \max_{1 \leq n \leq N} \left(\sum_{T \in \mathcal{T}_n} C_1 h_T^4 \|f\|_{L^\infty(I_n, L^2(T))}^2 + C_2 h_T^3 \|[n \cdot \nabla U_n]\|_{L^\infty(I_n, L^2(\partial T \setminus \partial \Omega))}^2 \right)^{\frac{1}{2}} \\ & + C_3 \|U_n - U_{n-1}\| + C_4 \left\| h_n^2 \frac{[U_{n-1}]}{\tau_n} \right\|_{L^2(T)}^*, \end{aligned}$$

where U_n is the discrete solution on $I_n := (t_{n-1}, t_n)$, $\tau_n = t_n - t_{n-1}$ is the n^{th} time step size, $[\cdot]$ denotes jumps over edges or between time intervals, and $\|\cdot\|$ denotes the norm in $\mathbb{L}^2(\Omega)$. The last term $C_4 \|\dots\|^*$ is present only in case of mesh coarsening. The constants C_i depend on the time t_N and the size of the last time step: $C_i = C_i(\log(\frac{t_N}{\tau_N}))$.

This leads to the following error estimator parts:

$$\begin{aligned} \eta_0 &= \left(\sum_{T \in \mathcal{T}_0} \|u_0 - U_0\|_{L^2(T)}^2 \right)^{1/2}, \\ \eta_h &= \left(\sum_{T \in \mathcal{T}_n} C_1 h_T^4 \|f\|_{L^\infty(I_n, L^2(T))}^2 + C_2 h_T^3 \|[n \cdot \nabla U_n]\|_{L^\infty(I_n, L^2(\partial T \setminus \partial \Omega))}^2 \right)^{1/2}, \\ \eta_c &= \left(\sum_{T \in \mathcal{T}_n} C_4 \left\| h_T^2 \frac{[U_{n-1}]}{\tau_n} \right\|_{L^2(T)}^2 \right)^{1/2}, \\ \eta_\tau &= C_3 \|U_n - U_{n-1}\|_{L^2(\Omega)}. \end{aligned}$$

When a bound tol is given for the total error produced in each time step, the widely used strategy is to allow one fixed portion $\Gamma_h tol$ to be produced by the spatial discretization, and another portion $\Gamma_\tau tol$ of the error to be produced by the time discretization, with $\Gamma_h + \Gamma_\tau \leq 1.0$. The adaptive procedure now tries to adjust time step sizes and meshes such that in every time step

$$\eta_\tau \approx \Gamma_\tau tol \quad \text{and} \quad \eta_h^2 + \eta_c^2 \approx \Gamma_h^2 tol^2.$$

The adjustment of the time step size can be done via extrapolation techniques known from numerical methods for ordinary differential equations, or iteratively: The algorithm starts from the previous time step size τ_{old} or from an initial guess. A parameter $\delta_1 \in (0, 1)$ is used to reduce the step size until the estimate is below the given bound. If the error is smaller than the bound, the step size is enlarged by a factor $\delta_2 > 1$ (usually depending on the order of the time discretization). In this case, the actual time step is not recalculated, only the initial step size for the next time step is changed. Two additional parameters $\theta_1 \in (0, 1)$, $\theta_2 \in (0, \theta_1)$ are used to keep the algorithm robust, just like it is done in the

equidistribution strategy for the mesh adaption. The algorithm starts from the previous time step size τ_{old} or from an initial guess.

If $\delta_1 \approx 1$, consecutive time steps may vary only slightly, but the number of iterations for getting the new accepted time step may increase. Again, as each iteration includes the solution of a discrete problem, this value should be chosen not too large. For a first order time discretization scheme, a common choice is $\delta_1 \approx 0.5$, for example.

14.2 Algorithm (Time step size control).

```

Start with parameters  $\delta_1 \in (0, 1)$ ,  $\delta_2 > 1$ ,  $\theta_1 \in (0, 1)$ ,  $\theta_2 \in (0, \theta_1)$ 

 $\tau := \tau_{\text{old}}$ 
Solve time step problem and estimate the error
while  $\eta_\tau > \theta_1 \Gamma_\tau \text{tol}$  do
     $\tau := \delta_1 \tau$ 
    Solve time step problem and estimate the error
end while
if  $\eta_\tau \leq \theta_2 \Gamma_\tau \text{tol}$  then
     $\tau := \delta_2 \tau$ 
end if

```

The above algorithm controls only the time step size, but does not show the mesh adaption. There are several possibilities to combine both controls. An inclusion of the grid adaption in every iteration of Algorithm 14.2 can result in a large number of discrete problems to solve, especially if the time step size is reduced more than once. A better procedure is first to do the step size control with the old mesh, then adapt the mesh, and after this check the time error again. In combination with implicit strategy A, this procedure leads to the following algorithm for one single time step

14.3 Algorithm (Time and space adaptive algorithm).

```

Start with given parameter  $\text{tol}$ ,  $\delta_1 \in (0, 1)$ ,  $\delta_2 > 1$ ,  $\theta_1 \in (0, 1)$ ,  $\theta_2 \in (0, \theta_1)$ ,
the solution  $u_n$  from the previous time step on grid  $\mathcal{T}_n$  at time  $t_n$ 
with time step size  $\tau_n$ 

 $\mathcal{T}_{n+1} := \mathcal{T}_n$ 
 $\tau_{n+1} := \tau_n$ 
 $t_{n+1} := t_n + \tau_{n+1}$ 
solve the discrete problem for  $u_{n+1}$  on  $\mathcal{T}_{n+1}$  using data  $u_n$ 
compute error estimates on  $\mathcal{T}_{n+1}$ 

while  $\eta_\tau > \theta_1 \Gamma_\tau \text{tol}$ 
     $\tau_{n+1} := \delta_1 \tau_{n+1}$ 
     $t_{n+1} := t_n + \tau_{n+1}$ 
    solve the discrete problem for  $u_{n+1}$  on  $\mathcal{T}_{n+1}$  using data  $u_n$ 
    compute error estimates on  $\mathcal{T}_{n+1}$ 
end while

```

```

do
  mark elements for refinement or coarsening
  if elements are marked then
    adapt mesh  $\mathcal{T}_{n+1}$  producing a modified  $\mathcal{T}_{n+1}$ 
    solve the discrete problem for  $u_{n+1}$  on  $\mathcal{T}_{n+1}$  using data  $u_n$ 
    compute estimates on  $\mathcal{T}_{n+1}$ 
  end if
  while  $\eta_\tau > \theta_1 \Gamma_\tau tol$ 
     $\tau_{n+1} := \delta_1 \tau_{n+1}$ 
     $t_{n+1} := t_n + \tau_{n+1}$ 
    solve the discrete problem for  $u_{n+1}$  on  $\mathcal{T}_{n+1}$  using data  $u_n$ 
    compute error estimates on  $\mathcal{T}_{n+1}$ 
  end while
  while  $\eta_h > tol$ 
    if  $\eta_\tau \leq \theta_2 \Gamma_\tau tol$  then
       $\tau_{n+1} := \delta_2 \tau_{n+1}$ 
    end if
  end while
end do

```

The adaptive a posteriori approach can be extended to the adaptive choice of the order of the time discretization: Bornemann [11, 12, 13] describes an adaptive variable order time discretization method, combined with implicit strategy B using the extrapolation marking strategy for the mesh adaption.

15 The Stefan problem of phase transition

Let us recall the modelling of the heat equation from energy conservation law. The temporal change of energy density is equal to the spatial change of flux plus production,

$$\frac{\partial}{\partial t}e(x, t) = -\operatorname{div} q(x, t) + f(x, t).$$

In a homogeneous material, the energy density e was assumed to be proportional to the temperature θ , like $e(x, t) = \rho c \theta(x, t)$. But when the material undergoes a phase change from solid to liquid, then energy is consumed by the phase transition. Molecules need additional energy in order to leave the matrix of solid crystal, in order to flow around freely in the liquid state.

Experimentally it can be observed that when a solid material is heated to its melting temperature, and energy is added, then the temperature increases proportionally to the energy density like

$$\theta(x, t) = \frac{1}{\rho c_s} e(x, t). \quad (\theta < \theta_m).$$

When the melting temperature θ_m is reached (at energy e_m), additional energy is first consumed without raising the temperature, until the *latent heat of melting*, L , is added (and the material has changed its state from solid to liquid). Only then even more energy is added, the temperature increases again, like

$$\theta(x, t) = \theta_m + \frac{1}{\rho c_l} (e(x, t) - e_m - L) \quad (\theta > \theta_m).$$

Rescaling temperature and energy such that $\theta_m = 0$, $e_m = 0$, the relationship between temperature and energy can be described by

$$\theta(x, t) = \beta(e(x, t)) = \begin{cases} \frac{1}{\rho c_s} e(x, t) & \text{if } e(x, t) < 0, \\ 0 & \text{if } e(x, t) \in [0, L], \\ \frac{1}{\rho c_l} (e(x, t) - L) & \text{if } e(x, t) > L, \end{cases}$$

compare Figure 15.1.

Still, the heat flux can be modelled by Fourier's law $q = -\kappa \nabla \theta$ (with different heat conductivities in solid and liquid material).

So, the conservation law gives the differential equation

$$\frac{\partial}{\partial t}e(x, t) - \operatorname{div}(\kappa \nabla \beta(e(x, t))) = f(x, t).$$

As β vanishes for all $e \in [0, L]$, this equation is degenerate parabolic. In case that the solution is smooth and the set $\Gamma(t) := \{x : \theta(x, t) = 0\}$ is a smooth $n - 1$ -dimensional submanifold of the domain, the temperature $\theta = \beta(e)$ and the set $\Gamma(t)$ give a solution of the classical Stefan problem, compare Section 2.2.

For simplicity, we will set all constants to one in the following considerations.

15.1 Problem setting:

We consider the classical two phase Stefan problem, which describes the heat diffusion and phase change in a pure material. Let $\Omega \subset \mathbb{R}^d$ denote a bounded domain, u the enthalpy (or energy density) and θ the temperature, and $f(t, \cdot) \in \mathbb{H}^{-1}(\Omega)$ a given right hand side sufficiently smooth in time.

Problem 15.1. Two phase Stefan problem

Find $u \in \mathbb{L}_\infty(0, T; L_\infty(\Omega)) \cap \mathbb{W}^{1, \infty}(0, T; H^{-1}(\Omega))$ and $\theta \in \mathbb{L}_\infty(0, T; H_0^1(\Omega))$ such that

$$\frac{d}{dt}u - \Delta\theta = f \quad \text{in } \mathbb{H}^{-1}(\Omega),$$

with initial condition

$$u(\cdot, 0) = u_0$$

and

$$\theta = \beta(u),$$

where $\beta(s) = \min(s, 0) + \max(s - 1, 0)$, see Figure 15.1.

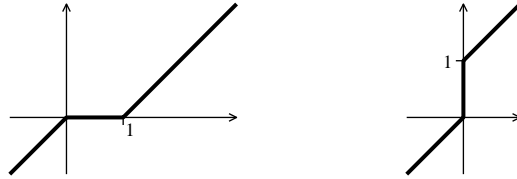


Figure 15.1: Graphs of β and β^{-1}

15.2 Discretization

We denote by τ_n the time step at the n -th step and set $t^n = \sum_{i=1}^n \tau_i$. Let \mathcal{M}^n be a uniformly regular partition of Ω into simplices, with mesh size density h_n , and \mathcal{B}^n be the collection of interior sides e of \mathcal{M}^n in Ω ; h_S (h_e) stands for the size of $S \in \mathcal{M}^n$ ($e \in \mathcal{B}^n$). Let $\mathbf{V}^n \subset \mathbb{H}_0^1(\Omega)$ be the piecewise linear finite element space over \mathcal{M}^n and I^n the Lagrange interpolation operator.

The discrete problem reads as follows. The discrete initial enthalpy $U^0 \in \mathbf{V}^0$ is an “interpolant” of u_0 . For time step (t_{n-1}, t_n) let $\tau_n = t_n - t_{n-1}$. Given $U^{n-1}, \Theta^{n-1} \in \mathbf{V}^{n-1}$, then \mathcal{M}^{n-1} and τ_{n-1} are modified as described below to get \mathcal{M}^n and τ_n and thereafter $U^n, \Theta^n \in \mathbf{V}^n$ computed according to $\Theta^n = I^n\beta(U^n)$ and

$$(15.1) \quad \frac{1}{\tau_n} \int_{\Omega} I^n((U^n - U^{n-1})\varphi) + \int_{\Omega} \nabla\Theta^n \cdot \nabla\varphi = \int_{\Omega} I^n(f(t^n)\varphi) \quad \forall \varphi \in \mathbf{V}^n.$$

Note that the nonlinear relation $\Theta^n = \beta(U^n)$ is enforced only in the vertices of the triangulation (otherwise not both U^n and Θ^n could be in \mathbb{P}_1 on each element of the triangulation).

Using mass lumping and enforcing $\Theta^n = I^n \beta(U^n)$ solely at the nodes introduces some consistency errors but amounts to having a monotone problem easy to implement and solved via an optimized nonlinear SOR [31, 74] or even with monotone multigrid methods [54, 55, 57].

15.3 Error control for Stefan problem

The presence of interfaces, and associated lack of regularity, is responsible for global numerical pollution effects for phase change problems. A cure consists of equidistributing discretization errors in adequate norms by means of highly graded meshes and varying time steps. Their construction relies on *a posteriori* error estimates, which are a fundamental component for the design of reliable and efficient adaptive algorithms for PDEs. These issues have been recently tackled in [79], [80], and are briefly discussed here.

We consider for simplicity the classical two-phase Stefan problem

$$(15.2) \quad \partial_t u - \Delta \beta(u) = f \quad \text{in } Q = \Omega \times (0, T),$$

where $\beta(s) = \min(s, 0) + \max(s - 1, 0)$; this corresponds to an ideal material with constant thermal properties and unit latent heat. A discrete solution U of (15.2) satisfies

$$(15.3) \quad \partial_t U - \Delta \beta(U) = f - \mathcal{R} \quad \text{in } Q,$$

where \mathcal{R} , a distribution with singular components and oscillatory behavior, is the so-called *parabolic residual*. Its size is to be determined in negative norms which entail averaging and thus quantify oscillations better.

In §15.4 we show how to represent the error $e_{\beta(u)} = \beta(u) - \beta(U)$ in terms of \mathcal{R} , and in §15.5 we state an error estimate of the form

$$\|e_{\beta(u)}\|_{L^2(Q)} \leq \mathcal{E}(U, f, T, \Omega; h, \tau),$$

with computable right hand side; hereafter h stands for the mesh size and τ for the time step. This formula is the basis of the adaptive algorithm of §15.6 and the simulations of §15.7.

15.4 Error Representation Formula

Upon subtracting (15.3) from (15.2) and integrating by parts against a smooth test function φ , the error $e_u = u - U$ satisfies the equation

$$(15.4) \quad \int_{\Omega} e_u(T) \varphi(T) = \int_{\Omega} e_u(0) \varphi(0) + \int_Q e_u (\partial_t \varphi + b \Delta \varphi) + \mathcal{R}(\varphi),$$

where $b(x, t) = \frac{\beta(u) - \beta(U)}{u - U}$ provided $u \neq U$ and $b(x, t) = 1$ otherwise, and

$$(15.5) \quad \mathcal{R}(\varphi) = \int_{\Omega} U(0) \varphi(0) - \int_{\Omega} U(T) \varphi(T) + \int_Q (f \varphi + U \partial_t \varphi + \beta(U) \Delta \varphi).$$

This motivates the study of the backward parabolic problem in *non-divergence* form [73], with vanishing diffusion coefficient $0 \leq b \leq 1$,

$$(15.6) \quad \partial_t \psi + (b + \delta) \Delta \psi = -b\chi \quad \text{in } Q, \quad \psi(T) = 0 \quad \text{in } \Omega,$$

where $\chi \in \mathbb{L}^2(Q)$ and $\delta \downarrow 0$. The theory of nonlinear strictly parabolic problems [60] yields existence of a unique solution ψ , which satisfies [79]

$$(15.7) \quad \|\nabla \psi(t)\|_{L^2(\Omega)}, \frac{1}{2} \|\partial_t \psi\|_{L^2(Q)}, \delta^{1/2} \|\Delta \psi\|_{L^2(Q)} \leq \|\chi\|_{L^2(Q)}.$$

Equation (15.6) does not exhibit a regularizing effect in that $\|\Delta \psi\|_{L^2(Q)}$ is never bounded uniformly in δ ; compare with [35]. Using (15.7), together with $be_u = e_{\beta(u)}$ and $|e_u| \leq 1 + |e_{\beta(u)}|$, (15.4) with $\varphi = \psi$ yields

$$(1 - \delta^{1/2}) \|e_{\beta(u)}\|_{L^2(Q)} \leq \|e_u(0)\|_{H^{-1}(\Omega)} + \sup_{\substack{\chi \in L^2(Q) \\ 0 < t < T}} \frac{|\mathcal{R}(\psi)|}{\|\nabla \psi(t)\|_{L^2(\Omega)}} + |Q|^{1/2} \delta^{1/2}.$$

Taking $\delta \downarrow 0$ we deduce a representation formula based on evaluating \mathcal{R} in the negative norm $L_t^1 H_x^{-1}$, and valid for *any* numerical method:

$$\|e_{\beta(u)}\|_{L^2(Q)} \leq \|e_u(0)\|_{H^{-1}(\Omega)} + \sup_{\substack{\chi \in L^2(Q) \\ 0 < t < T}} \frac{|\mathcal{R}(\psi)|}{\|\nabla \psi(t)\|_{L^2(\Omega)}}.$$

15.5 A Posteriori Error Estimators

If we set $U_t^n = (U^n - I^n U^{n-1})/\tau_n$ and $R^n = I^n f(t^n) - U_t^n$, integrate (16.9) by parts, and use Galerkin orthogonality (15.1), we easily arrive at

$$\begin{aligned} \mathcal{R}(\psi) &= \mathcal{C} + \sum_n \int_{t^{n-1}}^{t^n} \int_{\Omega} \left(R^n(\psi - \Psi) - \nabla \Theta^n \cdot \nabla(\psi - \Psi) \right) \\ &\quad + \sum_n \int_{\Omega} (U^{n-1} - I^n U^{n-1}) \psi^{n-1} + \sum_{n=1}^N \int_{t^{n-1}}^{t^n} \int_{\Omega} U_t^n (\psi - \psi^{n-1}), \end{aligned}$$

where Ψ is any discrete approximation of ψ and \mathcal{C} stands for consistency terms. Integrating $\nabla \Theta^n$ by parts element wise and using (15.7), we conclude

$$\mathcal{E}(U, f, T, \Omega; h, \tau) = C(\mathcal{E}_0 + \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 + \mathcal{E}_4) + \mathcal{E}_C.$$

Here C depends on the minimum angle of \mathcal{M}^n , the error indicators \mathcal{E}_i are

$$\begin{aligned} \mathcal{E}_0 &= \|u_0 - U^0\|_{H^{-1}(\Omega)} && \text{initial error,} \\ \mathcal{E}_1 &= \sum_n \tau_n \left(\sum_{e \in \mathcal{B}^n} h_e \|\llbracket \nabla \Theta^n \rrbracket_e \cdot \nu_e\|_{L^2(e)}^2 \right)^{1/2} && \text{jump residual,} \\ \mathcal{E}_2 &= \sum_n \tau_n \left(\sum_{S \in \mathcal{M}^n} h_S^2 \|R^n\|_{L^2(S)}^2 \right)^{1/2} && \text{interior residual,} \\ \mathcal{E}_3 &= \sum_n \|I^n U^{n-1} - U^{n-1}\|_{H^{-1}(\Omega)} && \text{coarsening,} \\ \mathcal{E}_4 &= \left(\sum_n \tau_n \|U^n - I^n U^{n-1}\|_{L^2(\Omega)}^2 \right)^{1/2} && \text{time residual,} \end{aligned}$$

and \mathcal{E}_c stands for consistency errors which we do not specify and could in principle be removed at the expense of complicating the implementation of (15.1); see [79]. All indicators \mathcal{E}_i can be evaluated in terms of the computed solution, initial datum, and source term, and are essential; they are also present for the heat equation [35]. The weights of $\mathcal{E}_1, \mathcal{E}_2$ correspond to \mathbb{H}_x^1 regularity of ψ , as opposed to \mathbb{H}_x^2 regularity for the heat equation, thereby reflecting the degenerate nature of (15.2). An alternative approach exploiting the additional, but nonuniform, regularity $\delta \|\Delta\psi\|_{L^2(Q)} \leq \|\chi\|_{L^2(Q)}$ is proposed in [79] and improves on this distinctive aspect of (15.2).

15.6 Adaptive Algorithm

The error estimators of §15.5 entail an L^1 or L^2 norm in time, which is impractical in that the entire history would be needed to control $\mathcal{E}(U, f, T, \Omega; h, \tau)$. We therefore resort to an L^∞ norm in time, and the ensuing equidistribution strategy of minimizing the spatial degrees of freedom for a uniform error distribution in time. For each $S \in \mathcal{M}^n$, the resulting *local* spatial error indicators are denoted by $E_0(S)$, for initial error, and

$$\begin{aligned} E_1(S) &= \frac{1}{2}T^2 h_S \|\llbracket \nabla \Theta^n \rrbracket_e \cdot \nu_e\|_{L^2(\partial S)}^2 && \text{local jump residual,} \\ E_2(S) &= T^2 h_S^2 \|R^n\|_{L^2(S)}^2 && \text{local interior residual,} \\ E_3(S) &= T^2 \tau_n^{-2} \|I^n U^{n-1} - U^{n-1}\|_{L^2(S)}^2 && \text{local coarsening.} \end{aligned}$$

Let ε be a given error tolerance and $M^n = \text{card } \mathcal{M}^n$. The objective is to select adaptively time steps and mesh densities in such a way that $E_i(S)$ have comparable size for all $S \in \mathcal{M}^n$ (equidistribution) and $\mathcal{E}(U, f, T, \Omega; h, \tau) \leq \varepsilon$. Elements S are either refined or coarsened via *bisection*. This algorithm creates *compatible* consecutive meshes, extends naturally from 2D to 3D, and is handy for combined refinement/coarsening operations [7].

Given refinement and coarsening parameters satisfying $\Gamma_0 + \Gamma_\tau + \Gamma_h \leq 1$, $\gamma_\tau < \Gamma_\tau$, $\gamma_h < \Gamma_h$, the initial mesh is obtained upon bisecting a coarse partition \mathcal{M}^0 until $E_0(S) \leq \Gamma_0^2 \varepsilon^2 / M^0$ for all $S \in \mathcal{M}^0$. Neglecting all terms \mathcal{E}_c for simplicity, the mesh size and time step are modified as follows. We first check whether

$$\frac{\gamma_h^2}{M^n} \varepsilon^2 < \max(E_1(S), E_2(S), E_3(S)) < \frac{\Gamma_h^2}{M^n} \varepsilon^2 \quad \forall S \in \mathcal{M}^n,$$

or not. If the rightmost constraint is violated, then S is bisected, whereas the leftmost constraint is used to flag S for coarsening. The latter will be allowed only if the local error incurred is sufficiently small; this is a critical computational issue. We next verify whether

$$\gamma_\tau \varepsilon < T^{1/2} \|U^n - I^n U^{n-1}\|_{L^2(\Omega)} < \Gamma_\tau \varepsilon,$$

or not. Failure of the rightmost inequality forces τ_n to diminish, whereas that of the leftmost constraint causes a corresponding increase of next time step τ_{n+1} but acceptance of the current time step.

15.7 Numerical Experiments

Our approach is able to detect the presence of interfaces, and refine accordingly, and is insensitive to topological changes such as merging, extinction, and mush or singularity formation. The interface velocity need not be computed explicitly for mesh design, which is a major improvement with respect to [74].

15.7.1 Example 1: Oscillating Circle

The interface is a circle of center $c(t)$ moving with velocity $V(t) = \dot{c}(t)$ and changing radius $R(t)$. The exact solution is given by

$$u(x, t) = \begin{cases} \alpha(r^2 - R^2) & \text{if } r \leq R \\ 1 + (2\alpha R - V \cdot \frac{x-c}{r} - \dot{R})(r - R) & \text{if } r > R, \end{cases}$$

where $r = |x - c|$ and $\alpha > 0$ is a constant, such that $u > 1$ when $r > R$. The numerical simulation was done in $\Omega = (-1, 1)^2$ for $t \in (0, 1)$ with

$$c(t) = (0.25, 0.4 \sin(10t)), \quad R(t) = 0.35 + 0.2 \sin(20t), \quad \alpha = 17.0.$$

Numerical parameters for the adaptive method were

$$\varepsilon = 175, \quad \Gamma_0 = 0.1, \quad \Gamma_\tau = 0.2, \quad \Gamma_h = 0.6, \quad \gamma_\tau = 0.155, \quad \gamma_h = 0.268.$$

Figures 15.2 and 15.3 show discrete isothermal lines and corresponding meshes. The innermost isothermal $\Theta = 0$ is the free boundary. Values (and derivatives) of the solution, along with normal interface velocities, exhibit a large variation in time, depending on R and V . This explains the various levels of mesh refinement and the variation in time of the element count.

15.7.2 Example 2: Oscillating Source

This is a phase change in a container $\Omega = (-1, 1)^2$, $T = 20.0$, with initial temperature $\Theta(x, 0) = 0.1 x_2$, prescribed temperature at three walls $\Theta(x, t) = 0.1 x_2$ for $x_2 > -1$, a fourth insulated wall $\partial_\nu \Theta(x, t) = 0$ for $x_2 = -1$, and two circular oscillating heat sources driving the evolution,

$$f(x, t) = \cos(0.2t) \max(0.0, 3.125 - 50|x - (-0.2, -0.5)|^2) \\ + \sin(0.2t) \max(0.0, 3.125 - 50|x - (-0.2, 0.5)|^2).$$

The exact solution is unknown. The upper pictures of Figure 15.4 show a topology change consisting of two liquid phases merging. The lower pictures show a mushy region produced by a cooling source term. Figure 15.5 depicts the corresponding meshes with highly refined regions; they capture the interface location along with the presence of strong heat sources.

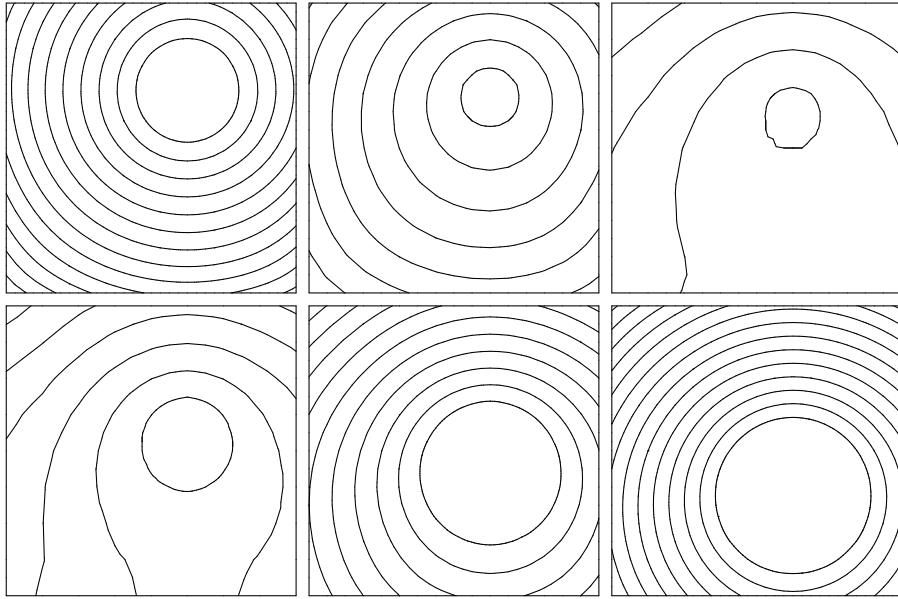


Figure 15.2: Example 1: Isothermal lines $\Theta = 0, 2, 4, 6, \dots$ at $t = 0.15, 0.20, 0.25, 0.30, 0.35, 0.40$

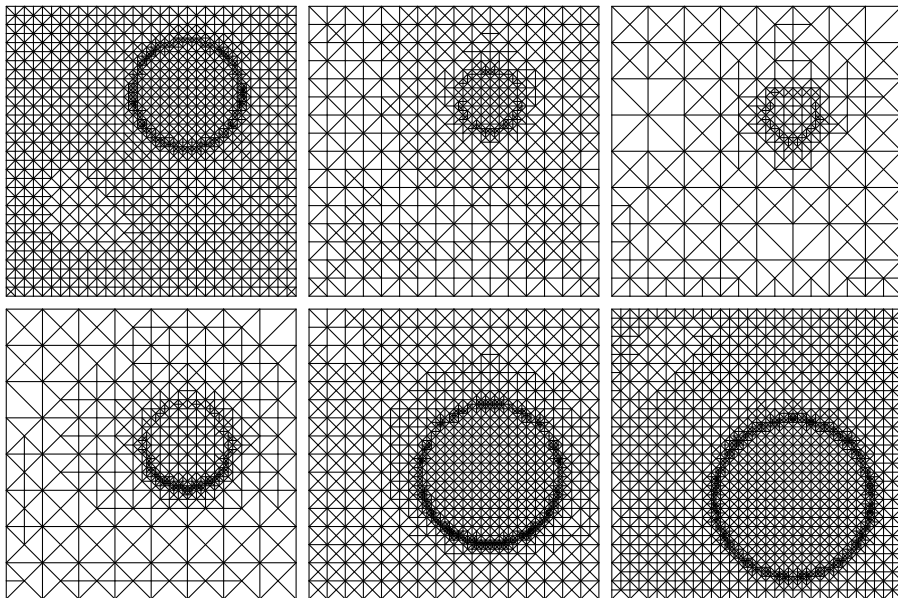


Figure 15.3: Example 1: Meshes at $t = 0.15, 0.20, 0.25, 0.30, 0.35, 0.40$; element count between 450 and 3600

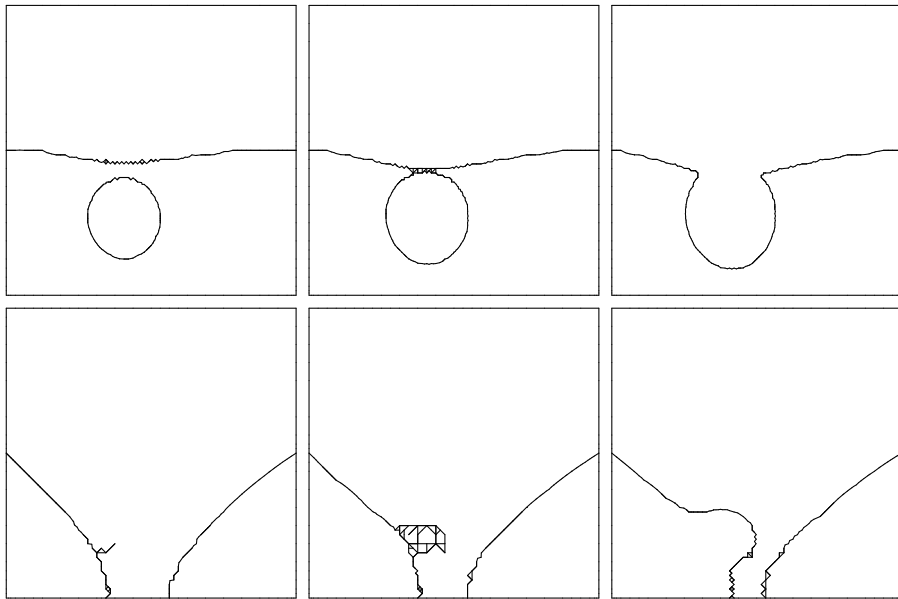


Figure 15.4: Example 2: Interfaces with topology changes and a mushy region at $t = 1.2, 1.6, 2.0, 8.6, 9.6, 10.6$

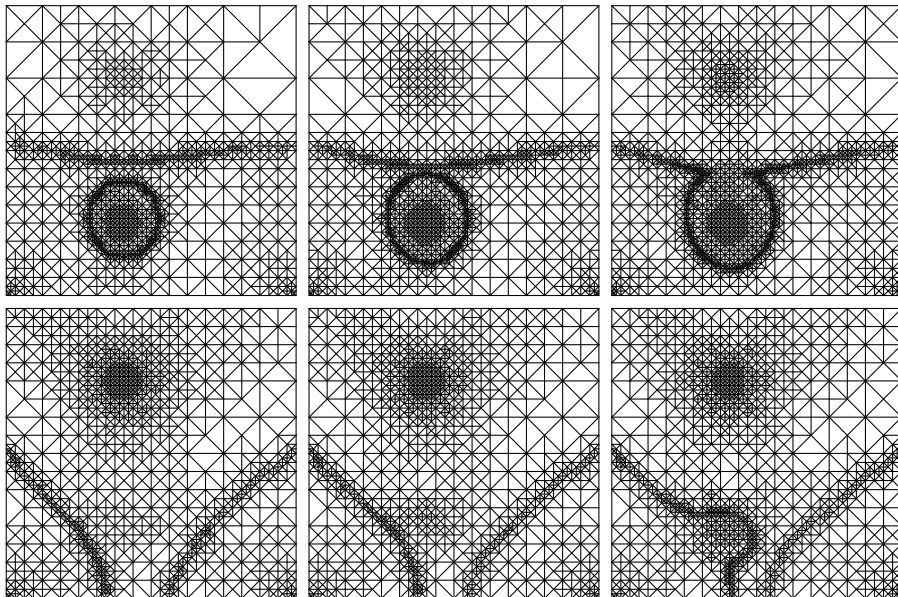


Figure 15.5: Example 2: Meshes at $t = 1.2, 1.6, 2.0, 8.6, 9.6, 10.6$

15.8 Nonlinear solver

As mentioned above, the system of nonlinear equations in each time step can be solved by a nonlinear Gauss-Seidel iteration.

Let $M = \text{diag}(m_i)$ denote the lumped mass matrix and $A = (a_{ij})$ the stiffness matrix in the n -th timestep. Then the nonlinear equation for unknown coefficient vectors U^n and Θ^n reads:

$$\frac{1}{\tau_n} M(U^n - U^{n-1}) + A\Theta^n = F^n, \quad \Theta_i^n = \beta(U_i^n), \quad i = 1, \dots, N.$$

The Gauss-Seidel equation for i -th coefficients U_i^n, Θ_i^n reads

$$\frac{m_i}{\tau_n} U_i^n + a_{ii} \Theta_i^n = \frac{m_i}{\tau_n} U_i^{n-1} + F_i^n - \sum_{j \neq i} a_{ij} \Theta_j^n =: \tilde{F}_i.$$

The additional condition $\Theta_i^n = \beta(U_i^n)$ leads to

$$\left(\frac{m_i}{\tau_n} Id + a_{ii} \beta \right) (U_i^n) = \tilde{F}_i.$$

As $(\frac{m_i}{\tau_n} Id + a_{ii} \beta)$ is a strictly monotone function, it is invertable and the solution can be computed by

$$U_i^n := \left(\frac{m_i}{\tau_n} Id + a_{ii} \beta \right)^{-1} (\tilde{F}_i), \quad \Theta_i^n := \beta(U_i^n).$$

Overrelaxation can be used, when not crossing the melting temperature [31, 74]. And monotone multigrid methods can be used as well [54, 55, 57].

16 The continuous casting problem

During continuous casting of steel ingots, hot molten steel is formed in a mold to a cylindrically shaped ingot, which is further cooled by spraying water onto it. In the interior of the ingot, a pool of liquid remains which must totally solidify before the ingot gets cut off.

Mathematically, the continuous casting problem leads to a convection-dominated nonlinearly degenerate diffusion problem. We will derive an a posteriori error estimate for a finite element discretization, and an adaptive method based on that. Most proofs are omitted here, they can be found in [18].

Remember that for the Stefan problem without convection, studied in the previous section, we were using L^2 norm estimates for both dual and primal problem. Here, the estimates will be based on L^∞ bounds for the dual problem, leading L^1 bounds for the error. This is the most natural norm when dealing with convection dominated problems. The application to the steel casting problem with physically realistic parameters is shown in Section 16.16.

Let the ingot occupy a cylindrical domain Ω with large aspect ratio. Let $0 < L < +\infty$ be the length of the ingot and $\Gamma \subset \mathbb{R}^d$ for $d = 1$ or 2 be its (polygonal) cross section. We show $\Omega = \Gamma \times (0, L)$ in Figure 16, and hereafter write $x = (y, z) \in \Omega$ with $y \in \Gamma$ and $0 < z < L$.

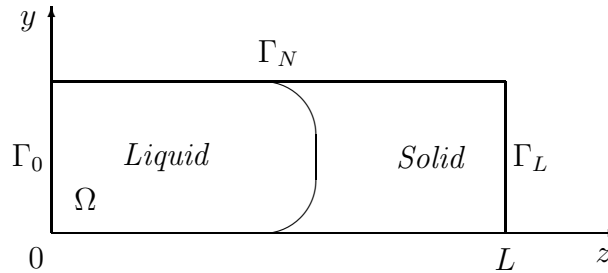


Figure : The domain Ω

We study the following *convection-dominated nonlinearly degenerate diffusion* problem

$$\begin{aligned}
 \partial_t u + v(t)\partial_z u - \Delta \theta &= 0 & \text{in } & Q_T, \\
 \theta &= \beta(u) & \text{in } & Q_T, \\
 \theta &= g_D & \text{on } & \Gamma_0 \times (0, T), \\
 \partial_\nu \theta + p(\theta - \theta_{\text{ext}}) &= 0 & \text{on } & \Gamma_N \times (0, T), \\
 u(x, 0) &= u_0(x) & \text{in } & \Omega,
 \end{aligned}$$

where

$$(16.1) \quad \Gamma_0 = \Gamma \times \{0\}, \quad \Gamma_L = \Gamma \times \{L\}, \quad \Gamma_N = \partial\Gamma \times (0, L), \quad Q_T = \Omega \times (0, T),$$

and $\theta + \theta_c$ is the absolute temperature, θ_c is the melting temperature, u is the enthalpy, $v(t) > 0$ is the extraction velocity of the ingot, ν is the unit outer normal to $\partial\Omega$, and

θ_{ext} is the external temperature. The mapping $\beta : \mathbb{R} \rightarrow \mathbb{R}$ is Lipschitz continuous and monotone increasing; since β is not strictly increasing, (16.1) is *degenerate* parabolic. The missing *outflow* boundary condition on Γ_L is unclear because the ingot moves at the casting speed and is cut shorter from time to time. It is thus evident that any standard boundary condition could only be an approximation. We impose either a Neumann

$$(16.2) \quad \partial_\nu \theta = g_N \leq 0 \quad \text{on} \quad \Gamma_L \times (0, T),$$

or a Dirichlet outflow condition

$$(16.3) \quad \theta = g_D < 0 \quad \text{on} \quad \Gamma_L \times (0, T).$$

Enforcing (16.2) with $g_N = 0$ is equivalent to assuming that the normal heat flux on Γ_L is entirely due to advection, which turns out to be an excellent approximation. Both boundary conditions lead to artificial boundary layers, with the second being more pronounced. In our simulations of §7 with real physical parameters, we take $g_N = 0$ and adjust g_D to minimize this effect. It is convenient to denote by Γ_D the Dirichlet part of $\partial\Omega$, that is Γ_0 for (16.2) and $\Gamma_0 \cup \Gamma_L$ for (16.3). The linear Robin condition (16.1) on that part of Γ_N in contact with air is just an approximation of the actual nonlinear Stefan-Boltzmann radiation law condition ($\sigma > 0$)

$$(16.4) \quad -\partial_\nu \theta = \sigma((\theta + \theta_c)_+^4 - \theta_{\text{ext}}^4) \quad \text{on} \quad \Gamma_N \times (0, T).$$

We see that linearizing (16.4) around a constant temperature leads to (16.1).

The importance of simulating and controlling the continuous casting process in the production of steel, copper, and other metals is recognized in industry. The extraction velocity $v(t)$ as well as the cooling conditions on the mold and water spray region are known to be decisive in determining material properties of the ingot. Avoiding excessive thermal stresses and material defects is an essential, and rather empirical, aspect of the continuous casting process.

If the extraction velocity $v(t)$ is assumed constant, and diffusion in the extraction direction z is ignored, then the resulting steady-state problem can be reformulated as a standard Stefan problem with a fictitious time $t = z/v$ [66], [88]. However, changes in the casting velocity v as well as in the cooling conditions are not only expected during a cycle of several hours of operation but are also desirable to handle late-arriving ladle, ladle or pouring problems, temporary malfunctions, etc. The casting machine must adjust to these demands and maintain production without degrading quality. The full non-stationary model (16.1)-(16.1) is thus more realistic than the steady state model in practical simulations and online control of continuous casting processes.

The system (16.1)-(16.1) is a special case of general Stefan problems with prescribed convection [91]. An outflow Dirichlet condition together with an inflow Neumann condition is assumed in [91] to guarantee uniqueness of weak solutions; our more realistic boundary data (16.1) and (16.2)-(16.3) violate this restriction. Under the additional assumption that the free boundary does not touch the inflow boundary Γ_0 , uniqueness of weak solutions to (16.1)-(16.1) and (16.3) is shown in [89].

A posteriori error estimates are computable quantities that measure the actual errors without knowledge of the limit solution. They are instrumental in devising algorithms for mesh and time-step modification which equidistribute the computational effort and so optimize the computations. Ever since the seminal paper [4] on elliptic problems, adaptivity has become a central theme in scientific and engineering computations. In particular, a posteriori error estimators have been derived in [34], [33] for linear and mildly nonlinear parabolic problems, and in [78], [17] for degenerate parabolic problems of Stefan type. Duality is the main tool in the analysis of [34],[33],[78], and so is in the present paper. We stress that the techniques of [77],[17] circumvent duality and thus apply to non-Lipschitz nonlinearities.

The purpose of this paper is twofold: we first introduce and analyze an *adaptive* method with error control, and second we apply it to steel casting, a concrete engineering application. We combine the method of characteristics for time discretization [27],[69],[87], with continuous piecewise linear finite elements for space discretization [21]. We derive a posteriori error estimators which provide the necessary information to modify the mesh and time step according to varying external conditions and corresponding motion of the solid-liquid interface. Our estimates exhibit a mild dependence on an upper bound V for the casting velocity $v(t)$, depending on the outflow conditions (16.2) and (16.3), which results from a novel and rather delicate analysis of a linearized dual problem - a convection-dominated degenerate parabolic with non-divergence form and a Dirichlet outflow condition. We stress that this mild as well as explicit dependence on V is a major improvement with respect to previous \mathbb{L}^2 -based a priori analyses of [16] for the continuous casting problem and of [27],[87] for parabolic PDE with large ratio advection/diffusion; they lead to an exponential dependence on V , unless $\Omega = \mathbb{R}^d$ or the characteristics do not intercept $\partial\Omega$ [69], which is not the case in Figure 16. We finally remark that convergence of a fully discrete finite element scheme for (16.4) is proved [20]; error estimates cannot in general be expected due to lack of compactness except on special cases [84].

The paper is organized as follows. In §2 we state the assumptions and set the problem. In §3 we discuss the fully discrete scheme, which combines the method of characteristics and finite elements. In §4 we introduce the concept of parabolic duality and prove several crucial stability estimates. In §5 we prove the a posteriori error estimates. In §6 we discuss an example with exact solution and document the method's performance. We conclude in §7 with applications to casting of steel with realistic physical parameters.

16.1 Setting

We start by stating the hypotheses concerning the data.

(H1) $\beta(s) = 0$ for $s \in [0, \lambda]$ and $0 < \beta_1 \leq \beta'(s) \leq \beta_2$ for a.e. $s \in \mathbb{R} \setminus [0, \lambda]$; $\lambda > 0$ is the latent heat.

(H2) $0 < v_0 V \leq v(t) \leq V$ for $t \in [0, T]$ and $|v'(t)| \leq v_1 V$ a.e. $t \in [0, T]$, with $v_0, v_1 > 0$ constants.

- (H3) $\theta_0 = \beta(u_0) \in W^{1,\infty}(\Omega)$, and the initial interface $F_0 := \{x \in \Omega : \theta_0(x) = 0\}$ is Lipschitz.
- (H4) $p \in \mathbb{H}^1(0, T; W^{1,\infty}(\Gamma_N)); p \geq 0$.
- (H5) $\theta_{\text{ext}} \in \mathbb{H}^1(0, T; C(\bar{\Gamma}_N))$.
- (H6) $g_D \in \mathbb{H}^1(0, T; C(\bar{\Gamma}_D)); g_D(x, 0) = \theta_0(x)$ on Γ_D .
- (H7) $g_N \in \mathbb{H}^1(0, T; C(\bar{\Gamma}_L))$.
- (H8) Uniqueness condition: $\exists \varepsilon_0, \rho_0 > 0$ such that $\theta \geq \rho_0$ a.e. in $\Gamma \times [0, \varepsilon_0] \times [0, T]$.
- (H9) Solidification condition: $\exists \varepsilon_1, \rho_1 > 0$ such that $\theta \leq -\rho_1$ a.e. in $\Gamma \times [L - \varepsilon_1, L] \times [0, T]$ and $\beta'(s) = \alpha > 0$ for $\beta(s) \leq -\rho_1$.
- (H10) $V \geq 1$.

We remark that (H8) is reasonable since it is satisfied for Stefan problems with $v = 0$ due to the continuity of θ and positivity of $\theta|_{\Gamma_0}$; heuristically the larger v , and so V , the larger the width ε_0 . The condition (H9) is an implicit assumption on data which corresponds to the ingot being solid in the vicinity of Γ_L , where it is to be cut, as well as having a constant conductivity β' . (H9) is only needed to handle (16.2). In addition, (H10) is not restrictive in that we are interested in the convection-dominated case. In view of (H4)-(H7) we may consider $p, \theta_{\text{ext}}, g_D, g_N$ extended to Ω in such a way that $\theta_{\text{ext}}, g_D, g_N \in \mathbb{H}^1(0, T; C(\bar{\Omega}))$, $p \in \mathbb{H}^1(0, T; W^{1,\infty}(\Omega))$.

Let $\mathbf{V}_0 = \{v \in \mathbb{H}^1(\Omega) : v = 0 \text{ on } \Gamma_D\}$ and \mathbf{V}^* the dual space of \mathbf{V}_0 . The weak formulation of (16.1)-(16.3) then reads as follows.

Continuous Problem. *Find u and θ such that*

$$\theta \in \mathbb{L}^2(0, T; H^1(\Omega)), \quad u \in \mathbb{L}^\infty(0, T; L^\infty(\Omega)) \cap \mathbb{H}^1(0, T; \mathbf{V}^*),$$

$$\theta(x, t) = \beta(u(x, t)) \quad \text{a.e. } (x, t) \in Q_T,$$

$$\theta(x, t) = g_D(x, t) \quad \text{a.e. } (x, t) \in \Gamma_D \times (0, T),$$

$$u(\cdot, 0) = u_0,$$

and for a.e. $t \in (0, T)$ and all $\phi \in \mathbf{V}_0$ the following equation holds

$$(16.5) \quad \langle \partial_t u, \phi \rangle + v(t) \langle \partial_z u, \phi \rangle + \langle \nabla \theta, \nabla \phi \rangle + \langle p \theta, \phi \rangle_{\Gamma_N} = \langle p \theta_{\text{ext}}, \phi \rangle_{\Gamma_N} + \langle g_N, \phi \rangle_{\Gamma_L}.$$

Hereafter, $\langle \cdot, \cdot \rangle$ stands for either the inner product in $\mathbb{L}^2(\Omega)$ or the duality pairing between \mathbf{V}^* and \mathbf{V}_0 , and $\langle \langle \cdot, \cdot \rangle \rangle_E$ denotes the inner product in $\mathbb{L}^2(E)$ with $E \subset \partial\Omega$; if $E = \partial\Omega$ we omit the subscript. We stress that the last term in (16.5) is absent when (16.3) is imposed. Existence and uniqueness of solutions (u, θ) to this problem satisfying $\theta \in C(\bar{Q}_T)$ are known [89].

16.2 Discretization

We now introduce the fully discrete problem, which combines continuous piecewise linear finite elements in space with characteristic finite differences in time. In fact, we use the method of characteristics to discretize the convection [27],[69],[87]. We denote by τ_n the n -th time step and set

$$t^n := \sum_{i=1}^n \tau_i, \quad \varphi^n(\cdot) := \varphi(\cdot, t^n)$$

for any function φ continuous in $(t^{n-1}, t^n]$. Let N be the total number of time steps, that is $t^N \geq T$. If e_z denotes the unit vector in \mathbb{R}^d in the z -direction, then $\frac{dx}{dt} = v(t)e_z$ defines the forward characteristics, and $U(t) = u(x(t), t)$ satisfies

$$(16.1) \quad \frac{dU}{dt} = \partial_t u + v \partial_z u.$$

The characteristic finite difference method is based on writing

$$\bar{x}^{n-1} = x - v^{n-1} \tau_n e_z, \quad \bar{u}^{n-1}(x) = u(\bar{x}^{n-1}, t^{n-1}),$$

for $n \geq 1$ and discretizing (16.1) by means of backward differences as follows:

$$\frac{dU^n}{dt} \approx \frac{U^n - U^{n-1}}{\tau_n} \Rightarrow \partial_t u^n + v^n \partial_z u^n \approx \frac{u^n - \bar{u}^{n-1}}{\tau_n}.$$

Therefore the discretization in time of (16.1)-(16.1) reads

$$(16.2) \quad \frac{u^n - \bar{u}^{n-1}}{\tau_n} - \Delta \beta(u^n) = 0 \quad \text{in } \Omega.$$

As $\bar{u}^{n-1}(x)$ is well defined only for $\bar{x}^{n-1} \in \bar{\Omega}$, one has to either restrict the time step size τ_n (at least locally) or extend u^{n-1} beyond the inflow boundary Γ_0 .

We denote by \mathcal{M}^n a uniformly regular partition of Ω into simplexes [21]. The mesh \mathcal{M}^n is obtained by refinement/coarsening of \mathcal{M}^{n-1} , and thus \mathcal{M}^n and \mathcal{M}^{n-1} are *compatible*. Given a triangle $S \in \mathcal{M}^n$, h_S stands for its diameter and ρ_S for its sphericity and they satisfy $h_S \leq 2\rho_S/\sin(\alpha_S/2)$, where α_S is the minimum angle of S ; h denotes the mesh density function $h|_S = h_S$ for all $S \in \mathcal{M}^n$. Uniform regularity of the family of triangulations is equivalent to $\alpha_S \geq \alpha^* > 0$, with α^* independent of n . We also denote by \mathcal{B}^n the collection of boundaries or sides e of \mathcal{M}^n in Ω ; h_e stands for the size of $e \in \mathcal{B}^n$. Let \mathbf{V}^n indicate the usual space of C^0 piecewise linear finite elements over \mathcal{M}^n and $\mathbf{V}_0^n = \mathbf{V}^n \cap \mathbf{V}_0$. Let $\{x_k^n\}_{k=1}^{K^n}$ denote the interior nodes of \mathcal{M}^n . Let $I^n : C(\bar{\Omega}) \rightarrow \mathbf{V}^n$ be the usual Lagrange interpolation operator, namely $(I^n \varphi)(x_k^n) = \varphi(x_k^n)$ for all $1 \leq k \leq K^n$. Finally, let the discrete inner products $\langle \cdot, \cdot \rangle^n$ and $\langle\langle \cdot, \cdot \rangle\rangle_E^n$ be the sum over $S \in \mathcal{M}^n$ of the element scalar products

$$\langle \varphi, \chi \rangle_S^n = \int_S I^n(\varphi \chi) dx, \quad \langle\langle \varphi, \chi \rangle\rangle_S^n = \int_{S \cap E} I^n(\varphi \chi) d\sigma,$$

for any piecewise uniformly continuous functions φ, χ . It is known that for all $\varphi, \chi \in \mathbf{V}^n$ [78]

$$\begin{aligned} \left| \int_S \varphi \chi dx - \int_S I^n(\varphi \chi) dx \right| &\leq \frac{1}{8} h_S^2 \|\nabla \varphi\|_{L^2(S)} \|\nabla \chi\|_{L^2(S)} \quad \forall S \in \mathcal{M}^n, \\ \left| \int_e \varphi \chi d\sigma - \int_e I^n(\varphi \chi) d\sigma \right| &\leq \frac{1}{8} h_S^2 \|\nabla \varphi\|_{L^2(e)} \|\nabla \chi\|_{L^2(e)} \quad \forall e \in \mathcal{B}^n. \end{aligned}$$

for any $S \in \mathcal{M}^n$ and $e \in \mathcal{B}^n$.

The discrete initial enthalpy $U^0 \in \mathbf{V}^0$ is defined at nodes x_k^0 of $\mathcal{M}^0 = \mathcal{M}^1$ to be

$$U^0(x_k^0) := u_0(x_k^0) \quad \forall x_k^0 \in \Omega \setminus F_0, \quad U^0(x_k^0) := 0 \quad \forall x_k^0 \in F_0.$$

Hence, U^0 is easy to evaluate in practice.

Discrete Problem. *Given $U^{n-1}, \Theta^{n-1} \in \mathbf{V}^{n-1}$, then \mathcal{M}^{n-1} and τ^{n-1} are modified as described below to get \mathcal{M}^n and τ_n and thereafter $U^n, \Theta^n \in \mathbf{V}^n$ computed according to the following prescription*

$$\Theta^n = I^n \beta(U^n), \quad \Theta^n - I^n g_D^n \in \mathbf{V}_0^n, \quad \bar{U}^{n-1} := U^{n-1}(\bar{x}^{n-1}),$$

$$(16.3) \quad \frac{1}{\tau_n} \langle U^n - I^n \bar{U}^{n-1}, \varphi \rangle^n + \langle \nabla \Theta^n, \nabla \varphi \rangle + \langle \langle p^n(\Theta^n - \theta_{\text{ext}}^n), \varphi \rangle \rangle_{\Gamma_N}^n = \langle \langle g_N^n, \varphi \rangle \rangle_{\Gamma_L}^n \quad \forall \varphi \in \mathbf{V}_0^n.$$

We stress that the right-hand side of (16.3) vanishes in case $\Gamma_D = \Gamma_0 \cup \Gamma_L$, and $p, \theta_{\text{ext}}, g_N$ need only be piecewise smooth. In view of the constitutive relation $\Theta^n = I^n \beta(U^n)$ being enforced only at the nodes, and the use of mass lumping, the discrete problem yields a monotone operator in \mathbb{R}^{K^n} which is easy to implement and solve via either nonlinear SOR [78] or monotone multigrid [57].

We conclude this section with some notation. Let the jump of $\nabla \Theta^n$ across $e \in \mathcal{B}^n$ be

$$(16.4) \quad \llbracket \nabla \Theta^n \rrbracket_e := (\nabla \Theta^n|_{S_1} - \nabla \Theta^n|_{S_2}) \cdot \nu_e.$$

Note that with the convention that the unit normal vector ν_e to e points from S_2 to S_1 , the jump $\llbracket \nabla \Theta^n \rrbracket_e$ is well defined. Let U and \hat{U} denote the piecewise constant and piecewise linear extensions of $\{U^n\}$, that is $U(\cdot, 0) = \hat{U}(\cdot, 0) = U^0(\cdot)$ and, for all $t^{n-1} < t \leq t^n$,

$$U(\cdot, t) := U^n(\cdot) \in \mathbf{V}^n, \quad \hat{U}(\cdot, t) := \frac{t^n - t}{\tau_n} U^{n-1}(\cdot) + \frac{t - t^{n-1}}{\tau_n} U^n(\cdot).$$

Finally, for any $\gamma > 0, k \geq 0$ and $\omega \subset \Omega$ we introduce the mesh dependent norms

$$\| \| h^\gamma \varphi \| \|_{H^k(\omega)} := \left(\sum_{e \subset \omega, e \in \mathcal{B}^n} h_e^{2\gamma} \|\varphi\|_{H^k(e)}^2 \right)^{1/2}, \quad \| h^\gamma \varphi \|_{H^k(\omega)} := \left(\sum_{S \subset \omega, S \in \mathcal{M}^n} h_S^{2\gamma} \|\varphi\|_{H^k(S)}^2 \right)^{1/2}.$$

16.3 Parabolic Duality

In this section we study a linear backward parabolic problem in non-divergence form, which can be viewed as the adjoint formal derivative of (16.1). For any $U \in BV(0, T; L^2(\Omega))$, we define

$$(16.1) \quad b(x, t) = \begin{cases} \frac{\beta(u) - \beta(U)}{u - U} & \text{if } u \neq U, \\ \beta_1 & \text{otherwise.} \end{cases}$$

It is clear from (H1) that $0 \leq b(x, t) \leq \beta_2$, for a.e. $(x, t) \in Q_T$. Let $b_\delta \in C^2(\bar{Q}_T)$ be a regularization of b satisfying

$$(16.2) \quad b_\delta \geq \delta > 0, \quad 0 \leq b_\delta - b \leq \delta \quad \text{a.e. in } Q_T,$$

where $0 < \delta \leq 1$ is a parameter to be chosen later. For arbitrary $t_* \in (0, T]$ and $\chi \in \mathbb{L}^\infty(Q_T)$, let ψ be the solution of the following linear backward parabolic problem

$$\begin{aligned} \mathcal{L}_\delta(\psi) &:= \partial_t \psi + v(t) \partial_z \psi + b_\delta \Delta \psi = -b^{1/2} \chi & \text{in } & \Omega \times (0, t_*), \\ \psi &= 0 & \text{on } & \Gamma_D \times (0, t_*), \\ \partial_\nu \psi + p\psi &= 0 & \text{on } & \Gamma_N \times (0, t_*), \\ \psi(x, t_*) &= 0 & \text{in } & \Omega, \end{aligned}$$

and

$$(16.3) \quad \partial_\nu \psi + \frac{v(t)}{b_\delta} \psi = 0 \quad \text{on } \Gamma_L \times (0, t_*)$$

provided (16.2) is enforced; we set $Q_* = \Omega \times (0, t_*)$. Existence of a unique solution $\psi \in W_q^{2,1}(Q_*)$ for any $q \geq 2$ of (16.3)-(16.3) follows from the theory of nonlinear strictly parabolic problems [60]. Note that we impose a Dirichlet outflow boundary condition on Γ_0 , which yields a boundary layer for ψ .

We now embark in the derivation of a priori estimates for the regularity of ψ . It turns out that such a technical endeavor depends on the boundary condition on Γ_L , which becomes inflow for (16.3). Consequently we distinguish the two cases on $\Gamma_L \times (0, t_*)$

$$\begin{aligned} \partial_\nu \psi + \frac{v(t)}{b_\delta} \psi &= 0 & \text{Robin inflow condition,} \\ \psi &= 0 & \text{Dirichlet inflow condition,} \end{aligned}$$

corresponding to (16.2) and (16.3): the former is more realistic but leads to worse stability bounds. We start with a simple, but essential, non-degeneracy property first proved in [16, Lemma 3.2].

Lemma 16.1. *Let $\xi, \rho \in \mathbb{R}$ satisfy $|\beta(\xi)| \geq \rho > 0$. Then we have*

$$(16.4) \quad |\xi - \eta| \leq \left(\frac{1}{\beta_1} + \frac{\lambda}{\rho} \right) |\beta(\xi) - \beta(\eta)|, \quad \forall \eta \in \mathbb{R}.$$

The following result is a trivial consequence of (16.1) and Lemma 16.1.

Corollary 16.1. *There exists $r > 0$ depending on ρ_0 of (H8) and ρ_1 of (H9) such that*

$$(16.5) \quad b(x, t) \geq r \quad \text{in } \Gamma \times ([0, \varepsilon_0] \cup [L - \varepsilon_1, L]) \times [0, T].$$

We observe that Corollary 16.1 only guarantees non-degeneracy of b but not its differentiability. If, in addition, $\beta(U) \leq -\rho_1$ on $\Gamma \times [L - \varepsilon_1, L] \times [0, T]$, which can be verified a posteriori, then (H9) leads to

$$(16.6) \quad b(x, t) = \alpha > 0 \quad \text{in } \Gamma \times [L - \varepsilon_1, L] \times [0, T].$$

This property will also be assumed for b_δ whenever it is valid for b .

16.4 Robin Inflow Condition

Throughout this section we assume that (16.4) is enforced. To motivate the estimates below consider the simplified PDE obtained from (16.3) by setting $b_\delta = 0$, $v(t) = V$ and $b\chi = 1$, namely,

$$(16.7) \quad \partial_t \Lambda + V \partial_z \Lambda = -1,$$

with terminal condition (16.3). If the inflow condition were $\partial_\nu \Lambda = 0$ then the method of characteristics would yield the solution $\Lambda(z, t) = t_* - t$ for the resulting transport problem. Such a Λ is an upper bound for the actual solution $\psi \geq 0$ of (16.7) satisfying $\partial_\nu \psi \leq 0$ on Γ_L . We then see that ψ is bounded uniformly in V , and expect a boundary layer of size $1/V$ due to the outflow Dirichlet condition on Γ_0 and the presence of non-vanishing diffusion (16.5) near Γ_0 ; so $|\partial_\nu \psi| \leq CV$ on Γ_0 . We now set $A = \|\chi\|_{L^\infty(Q_*)}$, and proceed to justify these heuristic arguments.

The proof of all following a priori L^∞ estimates are based on choosing an appropriate barrier function and using a comparison principle. To show the principle, the first (and simplest) proof is shown, all other ones are omitted.

Lemma 16.2. *The following a priori bound is valid*

$$\|\psi\|_{L^\infty(Q_*)} \leq \beta_2^{1/2} t_* \|\chi\|_{L^\infty(Q_*)}.$$

Proof. Consider the barrier function $\Lambda(t) = \beta_2^{1/2} A(t_* - t)$. In view of (H1), we easily get

$$\mathcal{L}_\delta(\Lambda \pm \psi) = -\beta_2^{1/2} A \mp b^{1/2} \chi \leq 0.$$

Since $\Lambda \pm \psi \geq 0$ on $\Gamma_0 \times (0, t_*)$ and $\Omega \times \{t_*\}$, along with

$$\partial_\nu(\Lambda \pm \psi) + q(\Lambda \pm \psi) = q\Lambda \geq 0,$$

where $q = \frac{v}{b_\delta}$ on $\Gamma_L \times (0, t_*)$ and $q = p$ on $\Gamma_N \times (0, t_*)$, the strong maximum principle yields the desired estimate

$$\Lambda \pm \psi \geq 0 \quad \text{in } Q_*. \quad \blacksquare$$

To obtain a bound for $\partial_z \psi$ on Γ_0 we modify a barrier technique in [89] to allow for variable velocity $v(t)$. We also explicitly trace the dependence on V and t_* .

Lemma 16.3. *There exists C independent of V and T such that the following a priori bound is valid for all $0 \leq t_* \leq T$*

$$(16.8) \quad |\partial_\nu \psi| \leq CVt_* \|\chi\|_{L^\infty(Q_*)} \quad \text{on } \Gamma_0 \times (0, t_*).$$

It turns out that we also need a bound on the tangential derivative $\partial_y \psi$ on Γ_L , which cannot be derived with a barrier technique. To this end we first prove a local gradient estimate in the vicinity of Γ_L , namely on the sets $\omega_1 := \Gamma \times (L - \varepsilon_1, L)$, $\omega_0 := \Gamma \times (L - \varepsilon_1/2, L)$. Let $\zeta \in C^\infty(\mathbb{R})$ be a cut-off function satisfying

$$0 \leq \zeta \leq 1, \quad \zeta(s) = 0 \quad \forall -\infty < s \leq L - \varepsilon_1, \quad \zeta(s) = 1 \quad \forall L - \frac{\varepsilon_1}{2} \leq s < \infty.$$

Lemma 16.4. *Let (16.6) hold for both b and b_δ . We then have the gradient estimate*

$$(16.9) \quad \int_0^{t_*} \int_{\omega_1} \zeta^2 |\nabla \psi|^2 + \int_0^{t_*} \int_{\Gamma_L} v \psi^2 \leq Ct_*^3 \|\chi\|_{L^\infty(Q_*)}^2.$$

Lemma 16.5. *Let (16.6) hold for both b and b_δ . Then there exists $C > 0$ independent of V and t_* such that the following a priori bounds are valid for all $0 \leq t_* \leq T$*

$$\max_{0 \leq t \leq t_*} \|\nabla \psi(\cdot, t)\|_{L^2(\Omega)}^2 + \int_t^{t_*} \int_\Omega b_\delta |\Delta \psi|^2 dx dt + \int_t^{t_*} \int_{\Gamma_L} v |\nabla \psi|^2 d\sigma dt \leq CV^3 t_*^3 \|\chi\|_{L^\infty(Q_*)}^2.$$

Corollary 16.2. *Let (16.6) hold for both b and b_δ . Then there exists $C > 0$ independent of V and t_* such that the following a priori bounds are valid for all $0 \leq t_* \leq T$*

$$\int_0^{t_*} \int_\Omega |\partial_t \psi + v(t) \partial_z \psi|^2 dx dt \leq CV^3 t_*^3 \|\chi\|_{L^\infty(Q_*)}^2.$$

Corollary 16.3. *Let (16.6) hold for both b and b_δ . Then there exists $C > 0$ independent of V and t_* such that the following a priori bounds are valid for all $0 \leq t_* \leq T$*

$$\int_0^{t_*} \delta \|D^2 \psi\|_{L^2(\Omega)}^2 dt \leq CV^3 t_*^3 \|\chi\|_{L^\infty(Q_*)}^2.$$

16.5 Dirichlet Inflow Condition

Throughout this section we assume that (16.3) is enforced. As in §16.4, we first examine the behavior of the simplified PDE (16.7) with inflow boundary condition $\psi = 0$ on Γ_L . If we allow $t_* = \infty$, then the method of characteristics gives the solution $\psi(z, t) = (L - z)/V$. Due to the effect of the terminal condition $\psi = 0$ at $t = t_* < \infty$, such a solution is larger than the actual one, and both exhibit an outflow boundary layer of size $1/V$ near Γ_0 . Since the size of the solution is also about $1/V$, we expect $\partial_z \psi$ to be bounded uniformly in V on Γ_0 . This heuristic reasoning is made rigorous below. We set again $A = \|\chi\|_{L^\infty(Q_*)}$.

Lemma 16.6. *The following a priori bound is valid for all $x \in \bar{\Omega}$ and $0 \leq t \leq t_* \leq T$*

$$|\psi(x, t)| \leq \frac{\beta_2^{1/2} L}{v_0 V} \|\chi\|_{L^\infty(Q_*)}.$$

A direct consequence of the barrier and comparison principle used for proving the last lemma and $\Lambda = 0$ on $\Gamma_L \times (0, t_*)$ is that

$$(16.10) \quad |\partial_\nu \psi| \leq \frac{\beta_2^{1/2}}{v_0 V} \|\chi\|_{L^\infty(Q_*)} \quad \text{on } \Gamma_L \times (0, t_*).$$

A similar bound holds on the outflow boundary Γ_0 .

Lemma 16.7. *There exists $C > 0$ independent of V and t_* such that for all $0 \leq t_* \leq T$*

$$(16.11) \quad |\partial_\nu \psi| \leq C \|\chi\|_{L^\infty(Q_*)} \quad \text{on } \Gamma_0 \times (0, t_*).$$

Lemma 16.8. *There exists $C > 0$ independent of V and t_* such that for all $0 \leq t_* \leq T$*

$$\max_{0 \leq t \leq t_*} \|\nabla \psi(\cdot, t)\|_{L^2(\Omega)}^2 + \int_0^{t_*} \int_\Omega b_\delta |\Delta \psi|^2 dx dt \leq CV t_* \|\chi\|_{L^\infty(Q_*)}^2.$$

Corollary 16.4. *There exists $C > 0$ independent of V and t_* such that for all $0 \leq t_* \leq T$*

$$\int_0^{t_*} \int_\Omega \left(|\partial_t \psi + v(t) \partial_z \psi|^2 + \delta |D^2 \psi|^2 \right) dx dt \leq CV t_* \|\chi\|_{L^\infty(Q_*)}^2.$$

16.6 Discontinuous p

We investigate the effect in 2d of a finite number of discontinuities of p along Γ_N ; this corresponds to abrupt changes in the cooling conditions as in the examples of §7. The estimates above remain all valid except for those in Corollaries 16.3 and 16.4, which involve second derivatives of ψ .

Using the intrinsic definition of fractional Sobolev spaces, together with the fact that p is piecewise $W^{1,\infty}$ over Γ_N , results in $\partial_\nu \psi = -p\psi \in \mathbb{H}^{1/2-\epsilon}$ in a vicinity of such discontinuities for $\epsilon > 0$. Elliptic regularity theory implies [64, p.188], [44]

$$(16.12) \quad \int_0^{t_*} \delta \|\psi\|_{H^{2-\epsilon}(\Omega)}^2 dt \leq C_\epsilon V^k t_*^k \|\chi\|_{L^\infty(Q_*)}^2 \quad \forall \epsilon > 0,$$

where $k = 3, 1$ for the Neumann and Dirichlet conditions, respectively. There is thus a slight loss of regularity with respect to the smooth case for both boundary conditions.

16.7 Error Representation Formula

We now derive an explicit representation formula for the error $\|\beta(u) - \beta(U)\|_{L^1(Q_*)}$ based on the linear backward parabolic problem (16.3)-(16.3). We only assume that $U(\cdot, t)$ is piecewise constant, so the derivation below applies to the solution U of (16.3).

We first multiply (16.3) by $-(u - U)$, and integrate in space and time from 0 to $t_* = t^m$. We examine the various contributions in turn. Since U is piecewise constant in time, we have

$$-\int_0^{t^m} \langle \partial_t \psi, u - U \rangle = \int_0^{t^m} \left(\langle \psi, \partial_t(u - \hat{U}) \rangle + \langle \partial_t \psi, U - \hat{U} \rangle \right) dt + \langle \psi^0, u_0 - U^0 \rangle.$$

Integrating by parts we get

$$-\int_0^{t^m} \langle v(t), \partial_z \psi(u - U) \rangle = \int_0^{t^m} v(t) \langle \psi, \partial_z(u - U) \rangle - \int_0^{t^m} v(t) \langle \psi, u - U \rangle_{\Gamma_L},$$

and using (16.1) we also obtain

$$\begin{aligned} -\int_0^{t^m} \langle b_\delta \Delta \psi, u - U \rangle &= \int_0^{t^m} \langle \nabla \psi, \nabla(\beta(u) - \beta(U)) \rangle \\ &\quad - \int_0^{t^m} \langle \partial_\nu \psi, \beta(u) - \beta(U) \rangle + \int_0^{t^m} \langle (b - b_\delta) \Delta \psi, u - U \rangle. \end{aligned}$$

Since

$$b^{1/2}|u - U| = |\beta(u) - \beta(U)|^{1/2}|u - U|^{1/2} \geq \beta_2^{-1/2}|\beta(u) - \beta(U)|,$$

collecting these estimates, and making use of (16.5), we easily end up with

$$(16.13) \quad \|\beta(u) - \beta(U)\|_{L^1(Q^m)} \leq \beta_2^{1/2} \sup_{\chi \in L^\infty(Q^m)} \frac{|\mathcal{R}(\psi)|}{\|\chi\|_{L^\infty(Q^m)}},$$

where \mathcal{R} , the *parabolic residual*, is the following distribution

$$\begin{aligned} \mathcal{R}(\psi) &= \langle u_0 - U^0, \psi^0 \rangle + \int_0^{t^m} \langle U - \hat{U}, \partial_t \psi \rangle dt + \int_0^{t^m} \langle u - U, (b - b_\delta) \Delta \psi \rangle dt \\ &\quad - \int_0^{t^m} \left(\langle \partial_t \hat{U} + v(t) \partial_z U, \psi \rangle + \langle \nabla \beta(U), \nabla \psi \rangle \right) dt \\ &\quad - \int_0^{t^m} \left(\langle p(\beta(U) - \theta_{\text{ext}}), \psi \rangle_{\Gamma_N} + \langle \partial_\nu \psi, g_D - \beta(U) \rangle_{\Gamma_0} - \langle g_N, \psi \rangle_{\Gamma_L} \right) dt. \end{aligned}$$

We conclude that an estimate of the error solely depending on discrete quantities and data may be obtained upon evaluating \mathcal{R} in suitable negative Sobolev norms. The latter are dictated by the a priori bounds of §§16.4 and 16.5. This program is carry out in §16.9 for the fully discrete solution U of (16.3).

16.8 A Posteriori Error Analysis

We first introduce the interior residual R^n and boundary residual B^n :

$$R^n := \frac{U^n - I^n \bar{U}^{n-1}}{\tau_n}, \quad B^n := \begin{cases} \partial_\nu \Theta^n + I^n p^n (\Theta^n - I^n \theta_{\text{ext}}^n) & \text{on } \Gamma_N, \\ \partial_\nu \Theta^n - I^n g_N^n & \text{on } \Gamma_L. \end{cases}$$

Theorem 16.1. (NEUMANN OUTFLOW) *Let (16.2) be enforced and $\Theta^n \leq -\rho_1$ in $\Gamma \times [L - \varepsilon_1, L]$ for any $n \geq 1$. Then there exists a constant $C > 0$ independent of V and t^m such that the following a posteriori error estimate holds for all $t^m \in [0, T]$,*

$$(16.1) \quad \int_0^{t^m} \|\beta(u) - \beta(U)\|_{L^1(\Omega)} dt \leq C(Vt^m)^{3/2} \left(\mathcal{E}_0 + \sum_{i=5}^{10} \mathcal{E}_i + \left(\Lambda_m \sum_{i=1}^4 \mathcal{E}_i \right)^{1/2} \right),$$

where

$$(16.2) \quad \Lambda_m = \left(\sum_{n=1}^m \tau_n (1 + \lambda|\Omega| + \|\Theta^n\|_{L^2(\Omega)}^2) \right)^{1/2}$$

and the error indicators \mathcal{E}_i are given by

$$\begin{aligned}
\mathcal{E}_0 &:= (V^3 t^m)^{-1/2} \|u_0 - U^0\|_{L^1(\Omega)} && \text{initial error,} \\
\mathcal{E}_1 &:= \left(\sum_{n=1}^m \tau_n \|h^{3/2} [\nabla \Theta^n]\|_{L^2(\Omega)}^2 \right)^{1/2} && \text{jump residual,} \\
\mathcal{E}_2 &:= \left(\sum_{n=1}^m \tau_n \|h^2 R^n\|_{L^2(\Omega)}^2 \right)^{1/2} && \text{interior residual,} \\
\mathcal{E}_3 &:= \left(\sum_{n=1}^m \tau_n \|h^{3/2} B^n\|_{L^2(\partial\Omega \setminus \Gamma_D)}^2 \right)^{1/2} && \text{boundary residual,} \\
\mathcal{E}_4 &:= \left(\sum_{n=1}^m \tau_n \|\beta(U^n) - I^n \beta(U^n)\|_{L^2(\Omega)}^2 \right)^{1/2} && \text{constitutive relation,} \\
\mathcal{E}_5 &:= \left(\sum_{n=1}^m \tau_n \|U^n - I^n U^{n-1}\|_{L^2(\Omega)}^2 \right. \\
&\quad \left. + \sum_{n=1}^m \frac{\tau_n}{V^2} \|U^n - I^n U^{n-1}\|_{L^2(\Gamma_L)}^2 \right)^{1/2} && \text{time residual,} \\
\mathcal{E}_6 &:= \left(\sum_{n=1}^m \tau_n \|U^{n-1} - I^n U^{n-1}\|_{L^2(\Omega)}^2 \right. \\
&\quad \left. + \sum_{n=1}^m \frac{\tau_n}{V^2} \|U^{n-1} - I^n U^{n-1}\|_{L^2(\Gamma_L)}^2 \right)^{1/2} && \text{coarsening,} \\
\mathcal{E}_7 &:= (V^3 t^m)^{-1/2} \sum_{n=1}^m \tau_n \|R^n - (\partial_t \hat{U} + v(t) \partial_z \hat{U})\|_{L^1(\Omega)} && \text{characteristic residual,} \\
\mathcal{E}_8 &:= \sum_{n=1}^m \tau_n \|h^2 \nabla R^n\|_{L^2(\Omega)} && \text{interior quadrature,} \\
\mathcal{E}_9 &:= \sum_{n=1}^m \tau_n \|h^{3/2} (\Theta^n - I^n \theta_{\text{ext}}^n)\|_{H^1(\Gamma_N)} \\
&\quad + \left(\sum_{n=1}^m \frac{\tau_n}{V} \|h^2 \partial_y (I^n g_N^n)\|_{L^2(\Gamma_L)}^2 \right)^{1/2} && \text{boundary quadrature,} \\
\mathcal{E}_{10} &:= (V^3 t^m)^{-1/2} \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \|\theta_{\text{ext}} - I^n \theta_{\text{ext}}^n\|_{L^1(\Gamma_N)} \\
&\quad + (V^3 t^m)^{-1/2} \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \|p - I^n p^n\|_{L^\infty(\Gamma_N)} \|\Theta^n - \theta_{\text{ext}}\|_{L^1(\Gamma_N)} \\
&\quad + (V^3 t^m)^{-1/2} \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \|g_N - I^n g_N^n\|_{L^1(\Gamma_L)} \\
&\quad + (V t^m)^{-1/2} \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \|g_D - I^n g_D^n\|_{L^1(\Gamma_0)} && \text{boundary discretization.}
\end{aligned}$$

Theorem 16.2. (DIRICHLET OUTFLOW) *Let (16.3) be satisfied. Then there exists a constant $C > 0$ independent of V and t^m such that the following a posteriori error*

estimate holds for all $t^m \in [0, T]$,

$$(16.3) \quad \int_0^{t^m} \|\beta(u) - \beta(U)\|_{L^1(\Omega)} dt \leq C(Vt^m)^{1/2} \left(\mathcal{E}_0 + \sum_{i=5}^{10} \mathcal{E}_i + \left(\Lambda_m \sum_{i=1}^4 \mathcal{E}_i \right)^{1/2} \right),$$

where Λ_m is defined in (16.2) and the error indicators \mathcal{E}_i are given by

$$\begin{aligned} \mathcal{E}_0 &:= (V^3 t^m)^{-1/2} \|u_0 - U^0\|_{L^1(\Omega)} && \text{initial error,} \\ \mathcal{E}_1 &:= \left(\sum_{n=1}^m \tau_n \|h^{3/2} \llbracket \nabla \Theta^n \rrbracket\|_{L^2(\Omega)}^2 \right)^{1/2} && \text{jump residual,} \\ \mathcal{E}_2 &:= \left(\sum_{n=1}^m \tau_n \|h^2 R^n\|_{L^2(\Omega)}^2 \right)^{1/2} && \text{interior residual,} \\ \mathcal{E}_3 &:= \left(\sum_{n=1}^m \tau_n \|h^{3/2} B^n\|_{L^2(\partial\Omega \setminus \Gamma_D)}^2 \right)^{1/2} && \text{boundary residual,} \\ \mathcal{E}_4 &:= \left(\sum_{n=1}^m \tau_n \|\beta(U^n) - I^n \beta(U^n)\|_{L^2(\Omega)}^2 \right)^{1/2} && \text{constitutive relation,} \\ \mathcal{E}_5 &:= \left(\sum_{n=1}^m \tau_n \|U^n - I^n U^{n-1}\|_{L^2(\Omega)}^2 \right)^{1/2} && \text{time residual,} \\ \mathcal{E}_6 &:= \left(\sum_{n=1}^m \tau_n \|U^{n-1} - I^n U^{n-1}\|_{L^2(\Omega)}^2 \right)^{1/2} && \text{coarsening,} \\ \mathcal{E}_7 &:= (V^3 t^m)^{-1/2} \sum_{n=1}^m \tau_n \|R^n - (\partial_t \hat{U} + v(t) \partial_z \hat{U})\|_{L^1(\Omega)} && \text{characteristic residual,} \\ \mathcal{E}_8 &:= \sum_{n=1}^m \tau_n \|h^2 \nabla R^n\|_{L^2(\Omega)} && \text{interior quadrature,} \\ \mathcal{E}_9 &:= \sum_{n=1}^m \tau_n \|h^{3/2} (\Theta^n - I^n \theta_{\text{ext}}^n)\|_{H^1(\Gamma_N)} && \text{boundary quadrature,} \\ \mathcal{E}_{10} &:= (V^3 t^m)^{-1/2} \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \|\theta_{\text{ext}} - I^n \theta_{\text{ext}}^n\|_{L^1(\Gamma_N)} \\ &\quad + (V^3 t^m)^{-1/2} \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \|p - I^n p^n\|_{L^\infty(\Gamma_N)} \|\Theta^n - \theta_{\text{ext}}\|_{L^1(\Gamma_N)} \\ &\quad + (V^3 t^m)^{-1/2} \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \|g_D - I^n g_D^n\|_{L^1(\Gamma_L)} \\ &\quad + (V t^m)^{-1/2} \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \|g_D - I^n g_D^n\|_{L^1(\Gamma_0)} && \text{boundary discretization.} \end{aligned}$$

Remark 5.1. We note that the quantity Λ_m in the estimates involves the \mathbb{L}^2 norm of the discrete temperature which is difficult to localize in practical computations. This bound, clearly achieved experimentally, can be proved by an a priori stability analysis which incorporates mesh changes.

Remark 5.2. If the meshes \mathcal{M}^n are of weakly acute type, or equivalently the stiffness matrix $(\nabla\phi_i, \nabla\phi_j)_{i,j}$ is an M-matrix, then the discrete maximum principle holds and guarantees the uniform boundedness of Θ^n ; thus $\Lambda_m \leq C\sqrt{t^m}$. If for all inter element sides e and corresponding pair of adjacent simplexes, the sum of angles opposite to e does not exceed π , then \mathcal{M}^n is weakly acute in 2D. Such a condition is not very restrictive in practice since it can be enforced with automatic mesh generators as long as the initial mesh exhibits this property.

16.9 Residuals

The error analysis hinges on the crucial estimate (16.13). To express the oscillatory character of \mathcal{R} in (16.14), we resort to Galerkin orthogonality. This replaces the evaluation of \mathcal{R} in negative Sobolev spaces by that on positive spaces plus weights depending on the mesh size h and the regularity of ψ . We first rewrite the discrete problem (16.3), for $t^{n-1} < t \leq t^n$, $\phi \in \mathbf{V}_0$, and $\varphi \in \mathbf{V}_0^n$, as follows:

$$\begin{aligned}
(16.4) \quad & \langle \partial_t \hat{U} + v(t) \partial_z U, \phi \rangle + \langle \nabla \Theta, \nabla \phi \rangle + \langle p(\Theta - \theta_{\text{ext}}), \phi \rangle_{\Gamma_N} - \langle g_N, \phi \rangle_{\Gamma_L} \\
& = \langle \partial_t \hat{U} + v(t) \partial_z \hat{U} - \tau_n^{-1}(U^n - I^n \bar{U}^{n-1}), \phi \rangle \\
& + v(t) \langle \partial_z(U - \hat{U}), \phi \rangle \\
& + \langle R^n, \phi - \varphi \rangle \\
& + \langle \nabla \Theta^n, \nabla(\phi - \varphi) \rangle + \langle I^n p^n(\Theta^n - I^n \theta_{\text{ext}}^n), \phi - \varphi \rangle_{\Gamma_N} - \langle I^n g_N^n, \phi - \varphi \rangle_{\Gamma_L} \\
& + \left(\langle R^n, \varphi \rangle - \langle R^n, \varphi \rangle^n \right) \\
& + \left(\langle I^n p^n(\Theta^n - I^n \theta_{\text{ext}}^n), \varphi \rangle_{\Gamma_N} - \langle I^n p^n(\Theta^n - I^n \theta_{\text{ext}}^n), \varphi \rangle_{\Gamma_N}^n \right) \\
& + \left(\langle I^n g_N^n, \varphi \rangle_{\Gamma_L}^n - \langle I^n g_N^n, \varphi \rangle_{\Gamma_L} \right) \\
& + \langle (p - I^n p^n)(\Theta^n - \theta_{\text{ext}}) + I^n p^n(I^n \theta_{\text{ext}}^n - \theta_{\text{ext}}), \phi \rangle_{\Gamma_N} + \langle I^n g_N^n - g_N, \phi \rangle_{\Gamma_L}.
\end{aligned}$$

This is the so-called Galerkin orthogonality, and reflects the essential property that the left-hand side is small in average. We next take $\phi = \psi$ and realize that to define φ we need to interpolate ψ under minimal regularity assumptions. We thus resort to the Clément interpolation operator $\Pi^n : L^2(\Omega) \rightarrow \mathbf{V}_0^n$, which satisfies the following local approximation properties [22], for $k = 1, 2$,

$$\begin{aligned}
& \|\psi - \Pi^n \psi\|_{L^2(S)} + h_S \|\nabla(\psi - \Pi^n \psi)\|_{L^2(S)} \leq C^* h_S^k \|\psi\|_{H^k(\tilde{S})}, \\
& \|\psi - \Pi^n \psi\|_{L^2(e)} \leq C^* h_e^{k-1/2} \|\psi\|_{H^k(\tilde{S})},
\end{aligned}$$

where \tilde{S} is the union of all elements surrounding $S \in \mathcal{M}^n$ or $e \in \mathcal{B}^n$. The constant C^* depends solely on the minimum angle of the mesh \mathcal{M}^n . An important by-product of shape regularity of \mathcal{M}^n is that the number of adjacent simplexes to a given one is bounded by a constant A independent of n , mesh-sizes and time-steps. Hence

$$\sum_{S \in \mathcal{M}^n} \|\xi\|_{L^2(\tilde{S})}^2 \leq A \|\xi\|_{L^2(\Omega)}^2 \quad \forall \xi \in L^2(\Omega).$$

This, in conjunction with (16.5) for $k = 1$, yields the \mathbb{H}^1 -stability bound

$$(16.5) \quad \|\nabla \Pi^n \psi(\cdot, t)\|_{L^2(\Omega)} \leq (1 + C^* A^{1/2}) \|\nabla \psi(\cdot, t)\|_{L^2(\Omega)}.$$

Consequently, we select φ in (16.4) to be

$$\varphi(\cdot, t) = \Pi^n \psi(\cdot, t) \quad \forall t^{n-1} < t \leq t^n.$$

Since $\beta(U^n) = \alpha U^n = I^n \beta(U^n)$ on Γ_L , we then obtain an explicit expression for the residual $\mathcal{R}(\psi) = \sum_{i=0}^{12} \mathcal{R}_i(\psi)$ of (16.14), where

(16.6)

$$\begin{aligned} \mathcal{R}_0(\psi) &= \langle u_0 - U_0, \psi^0 \rangle, \\ \mathcal{R}_1(\psi) &= \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \langle \nabla \Theta^n, \nabla(\Pi^n \psi - \psi) \rangle dt, \\ \mathcal{R}_2(\psi) &= \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \langle R^n, \Pi^n \psi - \psi \rangle dt, \\ \mathcal{R}_3(\psi) &= \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \langle \langle B^n - \partial_\nu \Theta^n, \Pi^n \psi - \psi \rangle \rangle dt, \\ \mathcal{R}_4(\psi) &= \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \left(\langle \nabla(I^n \beta(U^n) - \beta(U^n)), \nabla \psi \rangle - \langle \langle I^n \beta(U^n) - \beta(U^n) \rangle, \partial_\nu \psi \rangle \right) dt, \\ \mathcal{R}_5(\psi) &= \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \left(\langle U - \hat{U}, \partial_t \psi \rangle - v(t) \langle \partial_z(U - \hat{U}), \psi \rangle \right) dt, \\ \mathcal{R}_6(\psi) &= \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \langle \tau_n^{-1}(U^n - I^n \bar{U}^{n-1}) - \partial_t \hat{U} - v(t) \partial_z \hat{U}, \psi \rangle dt, \\ \mathcal{R}_7(\psi) &= \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \left(\langle R^n, \Pi^n \psi \rangle^n - \langle R^n, \Pi^n \psi \rangle \right) dt, \\ \mathcal{R}_8(\psi) &= \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \left(\langle \langle I^n p^n(\Theta^n - I^n \theta_{\text{ext}}^n), \Pi^n \psi \rangle \rangle_{\Gamma_N}^n - \langle \langle I^n p^n(\Theta^n - I^n \theta_{\text{ext}}^n), \Pi^n \psi \rangle \rangle_{\Gamma_N} \right) dt, \\ \mathcal{R}_9(\psi) &= \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \left(\langle \langle I^n g_N^n, \Pi^n \psi \rangle \rangle_{\Gamma_L} - \langle \langle I^n g_N^n, \Pi^n \psi \rangle \rangle_{\Gamma_L}^n \right) dt, \\ \mathcal{R}_{10}(\psi) &= \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \langle \langle (I^n p^n - p)(\Theta^n - \theta_{\text{ext}}) + I^n p^n(\theta_{\text{ext}} - I^n \theta_{\text{ext}}^n), \psi \rangle \rangle_{\Gamma_N} dt, \\ \mathcal{R}_{11}(\psi) &= \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \left(\langle \langle g_N - I^n g_N^n, \psi \rangle \rangle_{\Gamma_L} + \langle \langle I^n g_D^n - g_D, \partial_\nu \psi \rangle \rangle_{\Gamma_0} \right) dt, \\ \mathcal{R}_{12}(\psi) &= \int_0^{t^m} \langle u - U, (b - b_\delta) \Delta \psi \rangle dt. \end{aligned}$$

The rest of the argument consists of estimating each term $\mathcal{R}_i(\psi)$ separately. We rely on the regularity results of §16.4.

We decompose the integral $\langle \nabla \Theta^n, \nabla(\Pi^n \psi - \psi) \rangle$ over all elements $S \in \mathcal{M}^n$ and next integrate by parts to obtain the equivalent expression

$$(16.7) \quad \langle \nabla I^n \Theta^n, \nabla(\Pi^n \psi - \psi) \rangle = \sum_{e \in \mathcal{B}^n} \langle \llbracket \Theta^n \rrbracket_e, \psi - \Pi^n \psi \rangle_e + \langle \partial_\nu \Theta^n, \Pi^n \psi - \psi \rangle,$$

where $\langle \cdot, \cdot \rangle_e$ denotes the \mathbb{L}^2 -scalar product on $e \in \mathcal{B}^n$, and $\llbracket \Theta^n \rrbracket_e$ is defined in (16.4). In view of (16.5), we obtain

$$\sum_{n=1}^m \int_{t^{n-1}}^{t^n} \sum_{e \in \mathcal{B}^n} \langle \llbracket \Theta^n \rrbracket_e, \psi - \Pi^n \psi \rangle_e \leq C \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \| h^{3/2} \llbracket \Theta^n \rrbracket_e \|_{L^2(\Omega)} \| D^2 \psi \|_{L^2(\Omega)}.$$

Since the last term in (16.7) cancels out with a similar one in $\mathcal{R}_3(\psi)$, adding $\mathcal{R}_1(\psi)$ and $\mathcal{R}_3(\psi)$ and using (16.5) with $k = 2$ again, in conjunction with Corollary 16.3, we get

$$|\mathcal{R}_1(\psi) + \mathcal{R}_3(\psi)| \leq C(Vt^m)^{3/2} \delta^{-1/2} \| \chi \|_{L^\infty(Q^m)} (\mathcal{E}_1 + \mathcal{E}_3).$$

For $\mathcal{R}_2(\psi)$ we employ (16.5) with $k = 2$ and Corollary 16.3 to arrive at

$$|\mathcal{R}_2(\psi)| \leq C \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \| h^2 R^n \|_{L^2(\Omega)} \| D^2 \psi \|_{L^2(\Omega)} \leq C(Vt^m)^{3/2} \delta^{-1/2} \| \chi \|_{L^\infty(Q^m)} \mathcal{E}_2.$$

To estimate $\mathcal{R}_4(\psi)$, we integrate by parts and then use Lemma 16.5. We have

$$|\mathcal{R}_4(\psi)| \leq \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \| I^n \beta(U^n) - \beta(U^n) \|_{L^2(\Omega)} \| \Delta \psi \|_{L^2(\Omega)} \leq C(Vt^m)^{3/2} \delta^{-1/2} \| \chi \|_{L^\infty(Q^m)} \mathcal{E}_4.$$

These are all the terms involving $\delta^{-1/2}$. The remaining terms require lower regularity of ψ and are thus independent of δ , except for \mathcal{R}_{12} which is also of different character.

If $l(t)$ is the piecewise linear function $l(t) := \tau_n^{-1}(t^n - t)$, then $U - \hat{U} = l(t)(U^n - U^{n-1})$. Consequently, integration by parts and the fact that $\psi = 0$ on Γ_0 yield

$$-v(t) \langle \partial_z (U - \hat{U}), \psi \rangle = l(t) \langle U^n - U^{n-1}, v(t) \partial_z \psi \rangle - l(t) \langle \langle U^n - U^{n-1}, v(t) \psi \rangle \rangle_{\Gamma_L}.$$

Coupling the first term on the right-hand side with the remaining one in $\mathcal{R}_5(\psi)$, and writing $U^n - U^{n-1} = (U^n - I^n U^{n-1}) + (I^n U^{n-1} - U^{n-1})$, we obtain with the aid of Corollary 16.2

$$\begin{aligned} & \sum_{n=1}^m \int_{t^{n-1}}^{t^n} l(t) \langle U^n - U^{n-1}, \partial_t \psi + v(t) \partial_z \psi \rangle dt \\ & \leq C(\mathcal{E}_5 + \mathcal{E}_6) \left(\int_0^{t^m} \| \partial_t \psi + v(t) \partial_z \psi \|_{L^2(\Omega)}^2 \right)^{1/2} \leq C(Vt^m)^{3/2} \| \chi \|_{L^\infty(Q^m)} (\mathcal{E}_5 + \mathcal{E}_6). \end{aligned}$$

This is an essential step because neither $\partial_t \psi$ nor $\partial_z \psi$ are smooth alone, but rather their special combination above. In light of Lemma 16.4, the remaining boundary term in $\mathcal{R}_5(\psi)$ gives rise to

$$\begin{aligned} & - \sum_{n=1}^m \int_{t^{n-1}}^{t^n} l(t) \langle U^n - U^{n-1}, v(t)\psi \rangle_{\Gamma_L} \\ & \leq \left(\sum_{n=1}^m \frac{\tau_n V}{2} \|U^n - U^{n-1}\|_{L^2(\Gamma_L)}^2 \right)^{1/2} \left(\int_0^{t^m} \|v^{1/2}\psi\|_{L^2(\Gamma_L)}^2 \right)^{1/2} \\ & \leq C(Vt^m)^{3/2} \|\chi\|_{L^\infty(Q^m)} (\mathcal{E}_5 + \mathcal{E}_6). \end{aligned}$$

The term $\mathcal{R}_6(\psi)$ is easy to handle by Lemma 16.2, namely,

$$\begin{aligned} |\mathcal{R}_6(\psi)| & \leq \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \|\tau_n^{-1}(U^n - I^n \bar{U}^{n-1}) - \partial_t \hat{U} - v(t)\partial_z \hat{U}\|_{L^1(\Omega)} \|\psi\|_{L^\infty(\Omega)} \\ & \leq C(Vt^m)^{3/2} \|\chi\|_{L^\infty(Q^m)} \mathcal{E}_7. \end{aligned}$$

The next three terms $\mathcal{R}_7(\psi)$ to $\mathcal{R}_9(\psi)$ represent the effect of quadrature, and can be treated via (16.3) and (16.3). Hence, (16.5) and Lemma 16.5 imply

$$\begin{aligned} |\mathcal{R}_7(\psi)| & \leq C \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \|h^2 \nabla R^n\|_{L^2(\Omega)} \|\nabla \psi\|_{L^2(\Omega)} \leq C(Vt^m)^{3/2} \|\chi\|_{L^\infty(Q^m)} \mathcal{E}_8, \\ |\mathcal{R}_8(\psi)| & \leq C \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \|I^n p^n\|_{W^{1,\infty}(\Gamma_N)} \|h^{3/2}(\Theta^n - I^n \theta_{\text{ext}}^n)\|_{H^1(\Gamma_N)} \|\psi\|_{H^1(\Omega)} \\ & \leq C(Vt^m)^{3/2} \|\chi\|_{L^\infty(Q^m)} \mathcal{E}_9. \end{aligned}$$

Moreover, if we modify the boundary values of $\Pi^n \psi$ by using the \mathbb{L}^2 local projection over the sets $\text{supp}(\phi_k) \cap \partial\Omega$ instead of $\text{supp}(\phi_k)$, where $\{\phi_k\}_k$ is the canonical basis of \mathbf{V}^n , we achieve optimal approximability over $\partial\Omega$. If we now use Lemma 16.5, we obtain

$$|\mathcal{R}_9(\psi)| \leq CV^{-1/2} \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \|h^2 \partial_y(I^n g_N^n)\|_{L^2(\Gamma_L)} \|v^{1/2} \partial_y \psi\|_{L^2(\Gamma_L)} \leq C(Vt^m)^{3/2} \|\chi\|_{L^\infty(Q^m)} \mathcal{E}_9.$$

In addition, Lemma 16.2 yields

$$\begin{aligned} |\mathcal{R}_{10}(\psi)| & \leq \|\psi\|_{L^\infty(Q^m)} \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \left(\|I^n p^n - p\|_{L^\infty(\Gamma_N)} \|\Theta^n - \theta_{\text{ext}}\|_{L^1(\Gamma_N)} \right. \\ & \quad \left. + \|I^n p^n\|_{L^\infty(\Gamma_N)} \|\theta_{\text{ext}} - I^n \theta_{\text{ext}}^n\|_{L^1(\Gamma_N)} \right) dt \leq C(Vt^m)^{3/2} \|\chi\|_{L^\infty(Q^m)} \mathcal{E}_{10}, \end{aligned}$$

and

$$\begin{aligned} |\mathcal{R}_{11}(\psi)| & \leq \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \left(\|g_N - I^n g_N^n\|_{L^1(\Gamma_L)} \|\psi\|_{L^\infty(\Gamma_L)} + \|g_D - I^n g_D^n\|_{L^1(\Gamma_0)} \|\partial_\nu \psi\|_{L^\infty(\Gamma_0)} \right) dt \\ & \leq C(Vt^m)^{3/2} \|\chi\|_{L^\infty(Q^m)} \mathcal{E}_{10}. \end{aligned}$$

The last residual $\mathcal{R}_{12}(\psi)$ is of different nature from those above. We notice that (H1) and the a priori bound $\|\theta\|_{L^2(Q^m)} \leq C$ imply

$$\|u - U^n\|_{L^2(\Omega)}^2 \leq C \left(\lambda|\Omega| + \|\Theta^n - \theta\|_{L^2(\Omega)}^2 \right) \leq C \left(1 + \lambda|\Omega| + \|\Theta^n\|_{L^2(\Omega)}^2 \right) =: C\Xi_n^2,$$

whence Lemma 16.5 yields

$$|\mathcal{R}_{12}(\psi)| \leq C\delta^{1/2} \left(\sum_{n=1}^m \tau_n \Xi_n^2 \right)^{1/2} \left(\int_0^{t^m} \|b_\delta^{1/2} \Delta \psi\|_{L^2(\Omega)}^2 dt \right)^{1/2} \leq C\delta^{1/2} (Vt^m)^{3/2} \Lambda_m \|\chi\|_{L^\infty(Q^m)}.$$

16.10 Proof of Theorem 16.1

Collecting the above estimates for $\mathcal{R}_i(\psi)$, and inserting them back into (16.13), we obtain

$$\int_0^{t^m} \|\beta(u) - \beta(U)\|_{L^1(\Omega)} dt \leq C(Vt^m)^{3/2} \left(\mathcal{E}_0 + \sum_{i=5}^{10} \mathcal{E}_i + q(\delta) \right),$$

where

$$q(\delta) = \delta^{-1/2} \sum_{i=1}^4 \mathcal{E}_i + \delta^{1/2} \Lambda_m.$$

The asserted estimate follows from optimizing $q(\delta)$, namely from choosing $\delta = \Lambda_m^{-1} \sum_{i=1}^4 \mathcal{E}_i$.

16.11 Proof of Theorem 16.2

We first notice that the residual $\mathcal{R}(\psi)$ of (16.14) remains unaltered provided we remove $\langle\langle g_N, \psi \rangle\rangle_{\Gamma_L}$ and change Γ_0 by $\Gamma_D = \Gamma_0 \cap \Gamma_L$. The equality (16.4), expressing Galerkin orthogonality, is also valid provided all terms containing g_N are eliminated. We proceed as in §§16.9 and (16.1), but now using the regularity results of §16.5. The assertion follows immediately.

16.12 Discontinuous p

We examine now the case where p is piecewise Lipschitz as in §§16.6 and 16.16. In view of (16.12), the estimators in Theorems 16.1 and 16.2 which depend on second derivatives of ψ change as follows:

$$\begin{aligned} \mathcal{E}_1 &:= \left(\sum_{n=1}^m \tau_n \| \| h^{3/2-\epsilon} [\nabla \Theta^n] \| \|_{L^2(\Omega)}^2 \right)^{1/2}, \\ \mathcal{E}_2 &:= \left(\sum_{n=1}^m \tau_n \| h^{2-\epsilon} R^n \|_{L^2(\Omega)}^2 \right)^{1/2}, \\ \mathcal{E}_3 &:= \left(\sum_{n=1}^m \tau_n \| \| h^{3/2-\epsilon} B^n \| \|_{L^2(\partial\Omega \setminus \Gamma_D)}^2 \right)^{1/2}, \end{aligned}$$

for all $\epsilon > 0$. Moreover, the constants $C > 0$ in Theorems 16.1 and 16.2 depend also on ϵ , whereas the other estimators do not change.

16.13 Performance

In this section we explain how the estimators from §16.8 can be used for mesh and time-step modification, and document the performance of the resulting adaptive method.

16.14 Localization and adaption

For parabolic problems the aim of adaptivity is twofold: equidistribution of local errors in both space and time. We refer to [34],[33] for strictly parabolic problems and to [78],[81] for degenerate parabolic problems. On the basis of the a posteriori error estimates of §16.8, we can now design an adaptive method that meets these two goals and also keeps the error below a given tolerance.

The error estimators \mathcal{E}_i of both Theorems 16.1 and 16.2 can be split into contributions $E_i^n(S)$ for each element S and time t^n , and collected together to give rise to element indicators; see [78] for details. This way the error estimate is rewritten as

$$err := \int_0^{t^m} \|\beta(u) - \beta(U)\|_{L^1(\Omega)} dt \leq \sum_{S \in \mathcal{M}^0} \eta_S^0 + \max_{n=1, \dots, m} \left(\eta_\tau^n + \left(\sum_{S \in \mathcal{M}^n} (\eta_S^n)^2 \right)^{1/2} \right),$$

where η_τ^n includes all error indicators of time discretization (from $\mathcal{E}_5, \mathcal{E}_7, \mathcal{E}_{10}$) and η_S^n is the local indicator on element S of space discretization errors. We use them to equidistribute the space contributions by refinement/coarsening of the mesh \mathcal{M}^n and the time contributions by modifying the time step τ^n . Given a tolerance tol for the error err , the adaptive method adjusts time step sizes τ_n and adapts the meshes \mathcal{M}^n so as to achieve

$$(16.1) \quad \eta_S^0 \leq \frac{\Gamma_0 tol}{\#\mathcal{M}^0}, \quad \eta_\tau^n \leq \Gamma_\tau tol, \quad \eta_S^n \leq \frac{\Gamma_h tol}{\sqrt{\#\mathcal{M}^n}},$$

where $\Gamma_0 + \Gamma_\tau + \Gamma_h \leq 1$ are given parameters for the adaptive method. The mesh adaption in each time step is performed by local refinement and coarsening: all elements S violating (16.1) must be refined and those S with local indicators much smaller than the local tolerance may be coarsened. The time step may be enlarged in the latter case. The implementation uses local mesh refinement by bisectioning of elements; local mesh coarsening is the inverse operation of a previous local refinement. As meshes are nested, the interpolation of discrete functions such as U^{n-1} and $U^{\bar{n}-1}$ between consecutive meshes during local refinement or coarsening is a very simple operation. One new degree of freedom at the midpoint of the bisected edge is inserted during each local refinement, while one degree of freedom is deleted during a local coarsening. No other degrees of freedom are involved in such local operations.

16.15 Example: Traveling wave

An explicit solution for the nonlinearity $\beta(u) = \min(u, 0) + \max(u - 1, 0)$ is given by the traveling wave

$$\beta(u(x, y, t)) = \begin{cases} (1 - \exp(s)) & \text{if } s \leq 0 \text{ (liquid),} \\ 2(1 - \exp(s)) & \text{if } s > 0 \text{ (solid),} \end{cases} \quad \text{where } s = (\nu \cdot v - V)(\nu \cdot (x, y) - Vt),$$

with $\nu = (\cos(\alpha), \sin(\alpha))$ and parameters $v = (2, 0)$, $V = 0.4$, $\alpha = \pi/6$; V is the interface velocity in the normal direction ν . We solve the problem in the domain $\Omega = (0.0, 1.0) \times (0.0, 0.2)$ for time $t \in (1, 2)$ with Dirichlet boundary condition on $\partial\Omega$. To avoid any mesh effects, the interface normal ν is rotated from the horizontal direction by α . This way ν is never parallel to any mesh edge. As the domain in the applications of Section 16.16 has a very large aspect ratio, we explore here the use of elongated elements. We thus compare simulations with meshes of aspect ratios 1 and 5 originated from the macro triangulations of Figure 16.1, for the explicit traveling wave solution.

Figures 16.2 and 16.4 show adaptive meshes at time $t = 1.1$, generated with error tolerances $tol = 0.5$ and $tol = 0.25$, while Figure 16.3 depicts isothermal lines at the same time; the latter look the same for all simulations. Figure 16.5 displays the error $\|e_{\beta(u)}(t)\|_{L^1(\Omega)}$ and Figure 16.6 the mesh element counts for simulations with both aspect ratios. Finally, Figure 16.7 shows the total error for different given tolerances and mesh aspect ratios. Even though the triangle counts are larger for simulations with larger aspect ratio, the estimators and the adaptive method behave well. It is thus reasonable to use elongated elements in the following application. In any event, we do not employ specialized estimators or adaptive methods for anisotropic meshes such as [98]. The application of such methods to degenerate parabolic equations is still to be investigated.

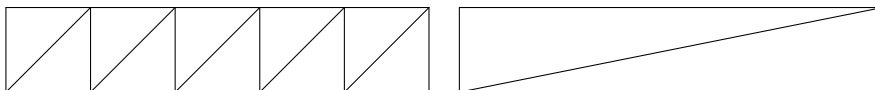


Figure 16.1: Example 16.15. Macro triangulations with aspect ratios 1 and 5.

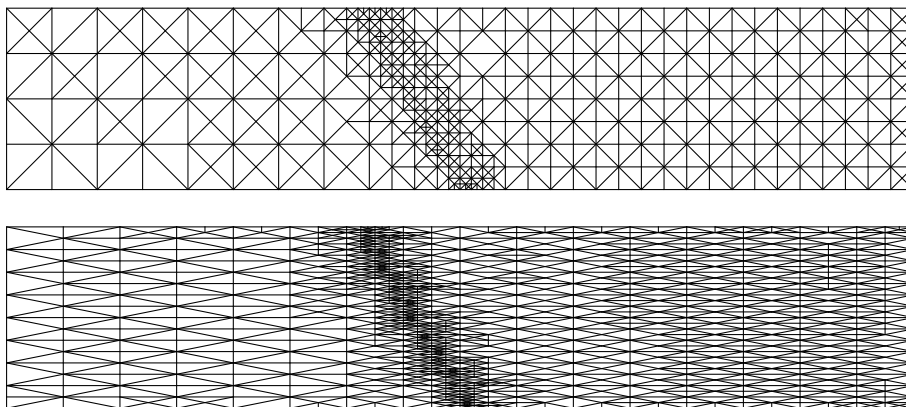


Figure 16.2: Example 16.15. Meshes with aspect ratios 1 and 5 for $tol = 0.5$ at $t = 1.1$.

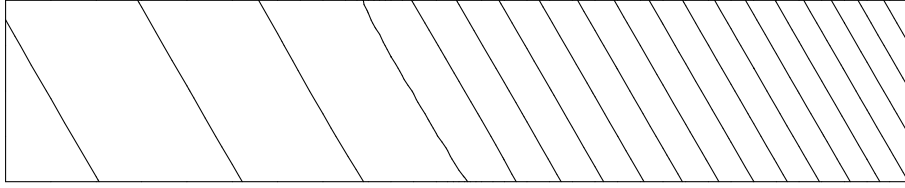


Figure 16.3: Example 16.15. Isothermal lines at $\beta(u) = k/8$, $k = -16 \dots 3$, at $t = 1.1$.

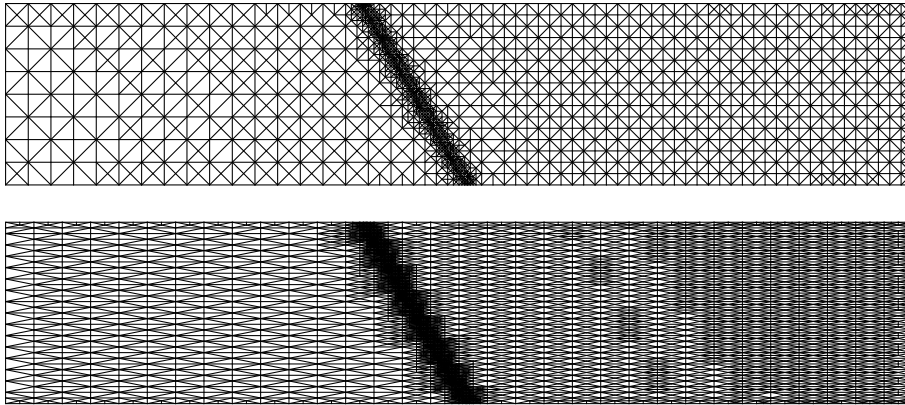


Figure 16.4: Example 16.15. Meshes with aspect ratios 1 and 5 for $tol = 0.25$ at $t = 1.1$.

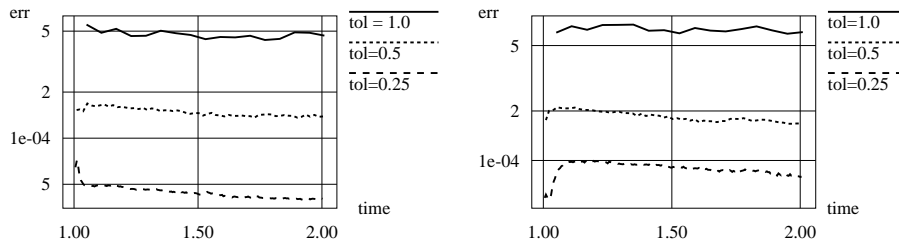


Figure 16.5: Example 16.15. $\|e_{\beta(u)}(t)\|_{L^1(\Omega)}$ for meshes with aspect ratios 1 (left) and 5 (right).

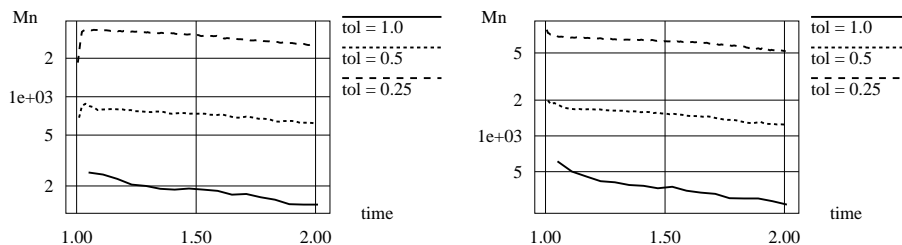


Figure 16.6: Example 16.15. Triangle counts for meshes with aspect ratios 1 (left) and 5 (right).

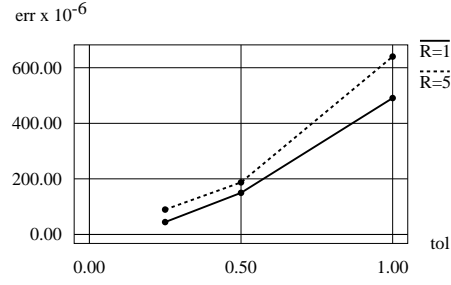


Figure 16.7: Example 16.15. $\|e_{\beta(u)}(t)\|_{L^1(L^1)}$ for meshes with aspect ratios $R = 1, 5$.

16.16 Applications to Casting of Steel

We study the casting of a slab in 2D. This problem was proposed in [61], and a similar problem with time dependent parameters was studied in [66]. In order to derive non-dimensional equations (16.1)-(16.3), we first rescale the physical equations with the material parameters.

16.17 Scaling

In this section, we mark all physical quantities by a tilde. The original equations with physical coefficients for temperature $\tilde{\theta}$ (in units $[\text{°K}]$) and energy density (or enthalpy) \tilde{u} (in units $[\text{kg}/\text{m}^2\text{s}^2]$) read

$$\begin{aligned}
 \partial_{\tilde{t}}\tilde{u} + \tilde{v}\partial_{\tilde{z}}\tilde{u} &= \tilde{\nabla} \cdot (\tilde{k}\tilde{\nabla}\tilde{\theta}) && \text{in } \tilde{\Omega} \times (0, \tilde{T}), \\
 \tilde{u} &= \tilde{\rho}(\tilde{c}\tilde{\theta} + \tilde{\chi}\tilde{\lambda}) && \text{in } \tilde{\Omega} \times (0, \tilde{T}), \\
 \tilde{\theta} &= \tilde{g}_D && \text{on } \tilde{\Gamma}_0 \times (0, \tilde{T}), \\
 \tilde{k}\partial_{\tilde{v}}\tilde{\theta} + \tilde{p}(\tilde{\theta} - \tilde{\theta}_{\text{ext}}) &= 0 && \text{on } \tilde{\Gamma}_N \times (0, \tilde{T}), \\
 \tilde{u}(\cdot, 0) &= \tilde{u}_0 && \text{in } \tilde{\Omega}, \\
 \tilde{k}\partial_{\tilde{v}}\tilde{\theta} &= \tilde{g}_N && \text{on } \tilde{\Gamma}_L \times (0, \tilde{T}) \\
 \text{or } \tilde{\theta} &= \tilde{g}_D && \text{on } \tilde{\Gamma}_L \times (0, \tilde{T}).
 \end{aligned}$$

The physical coefficients and their units are: casting speed $\tilde{v} [\text{m}/\text{s}]$, heat conductivity $\tilde{k} [\text{kg m}/\text{s}^3 \text{°K}]$, density $\tilde{\rho} [\text{kg}/\text{m}^3]$, specific heat $\tilde{c} [\text{m}^2/\text{s}^2 \text{°K}]$, latent heat $\tilde{\lambda} [\text{m}^2/\text{s}^2]$, melting temperature $\tilde{\theta}_m [\text{°K}]$, heat transfer coefficient $\tilde{p} [\text{kg}/\text{s}^3 \text{°K}]$, and external cooling temperature $\tilde{\theta}_{\text{ext}} [\text{°K}]$. Here $\tilde{\chi}$ stands for the characteristic function of the liquid phase. In the remainder of this section, subscripts s and l indicate the corresponding coefficients for the solid and liquid phase.

The simulations are done over a slab of length $\tilde{L} = 25 \text{ m}$ and height 0.21 m . We use material parameters for steel with 0.09% carbon content. Temperature-dependent data provided in [61] are approximated by piecewise constant data for the liquid and solid phase: $\tilde{k}_s = 30 \text{ kg m}/\text{s}^3 \text{°K}$, $\tilde{k}_l = 180 \text{ kg m}/\text{s}^3 \text{°K}$, $\tilde{c}_s = 660 \text{ m}^2/\text{s}^2 \text{°K}$, $\tilde{c}_l = 830 \text{ m}^2/\text{s}^2 \text{°K}$, $\tilde{\rho} = 7400 \text{ kg}/\text{m}^3$, $\tilde{\lambda} = 276\,000 \text{ m}^2/\text{s}^2$, and $\tilde{\theta}_m = 1733 \text{ °K}$.

The boundary condition on $\tilde{\Gamma}_N$ depends on the position along the slab; the model has a mold cooling zone ($1.15m$) and three water spray zones which include radiation. The (nonlinear) radiation condition (16.4) is linearized by using

$$(\tilde{\theta}^4 - \tilde{\theta}_{\text{ext}}^4) \approx (\tilde{\theta} - \tilde{\theta}_{\text{ext}})(\tilde{\theta}_m + \tilde{\theta}_{\text{ext}})(\tilde{\theta}_m^2 + \tilde{\theta}_{\text{ext}}^2).$$

The (linear) Robin conditions on $\tilde{\Gamma}_N$ in the mold and spray regions are then

$$-\tilde{k}\partial_{\tilde{v}}\tilde{\theta} = \begin{cases} \tilde{p}(\tilde{\theta} - \tilde{\theta}_{\text{mold}}) & \text{if } \tilde{z} < 1.15m, \\ \tilde{p}(\tilde{\theta} - \tilde{\theta}_{\text{H}_2\text{O}}) + \tilde{\sigma}\epsilon(\tilde{\theta} - \tilde{\theta}_{\text{rad}})(\tilde{\theta}_m + \tilde{\theta}_{\text{rad}})(\tilde{\theta}_m^2 + \tilde{\theta}_{\text{rad}}^2) & \text{if } \tilde{z} > 1.15m. \end{cases}$$

Casting and boundary parameters are given in Table 16.1. In this model, the quantities p and θ_{ext} exhibit discontinuities along Γ_N , which results in hypotheses (H4) and (H5) not being satisfied. But, as stated in §16.12, the estimators can be adjusted to the case of piecewise smooth boundary data. On the other hand, a refined model might include some mollifying effect of water spraying, which removes these discontinuities.

Quantity	Value	Unit	Description
\tilde{v}	0.0225	$\frac{m}{s}$	casting speed
\tilde{g}_D	1818	$^{\circ}K$	on $\tilde{\Gamma}_0$: inflow temperature
\tilde{g}_D	1250	$^{\circ}K$	on $\tilde{\Gamma}_L$: outflow temperature
\tilde{g}_N	0	$\frac{kg}{s^3}$	on $\tilde{\Gamma}_L$: outflow temperature flux
\tilde{p}	1500	$\frac{kg}{s^3 \text{ } ^{\circ}K}$	on $\tilde{\Gamma}_N$, $\tilde{z} \in (0, 1.15)m$: heat transfer in mold
$\tilde{\theta}_{\text{mold}}$	353	$^{\circ}K$	mold external temperature
\tilde{p}	700	$\frac{kg}{s^3 \text{ } ^{\circ}K}$	on $\tilde{\Gamma}_N$, $\tilde{z} \in (1.15, 4.4)m$: heat transfer in first spray region
\tilde{p}	350	$\frac{kg}{s^3 \text{ } ^{\circ}K}$	on $\tilde{\Gamma}_N$, $\tilde{z} \in (4.4, 14.6)m$: heat transfer in second spray region
\tilde{p}	50	$\frac{kg}{s^3 \text{ } ^{\circ}K}$	on $\tilde{\Gamma}_N$, $\tilde{z} \in (14.6, 25)m$: heat transfer in third spray region
$\tilde{\theta}_{\text{H}_2\text{O}}$	300	$^{\circ}K$	cooling water temperature
$\tilde{\sigma}$	5.67 E−8	$\frac{kg}{s^3 \text{ } ^{\circ}K^4}$	Stefan–Boltzmann constant
ϵ	0.8		emission factor
$\tilde{\theta}_{\text{rad}}$	370	$^{\circ}K$	$\tilde{z} \in (1.15, 14.6)m$: air temperature
$\tilde{\theta}_{\text{rad}}$	710	$^{\circ}K$	$\tilde{z} \in (14.6, 25)m$: air temperature in third spray region

Table 16.1: Casting and boundary parameters.

Using a length scale \bar{X} [m] and a time scale \bar{T} [s], the physical quantities can be transformed into dimensionless ones as follows:

$$x := \frac{\tilde{x}}{\bar{X}}, \quad t := \frac{\tilde{t}}{\bar{T}}, \quad u := \frac{\tilde{u}}{\tilde{\rho}\tilde{\lambda}} - \frac{\tilde{c}_s\tilde{\theta}_m}{\tilde{\lambda}}, \quad \theta := \begin{cases} \frac{\tilde{c}_s}{\tilde{\lambda}}(\tilde{\theta} - \tilde{\theta}_m) & \text{if } \tilde{\theta} \leq \tilde{\theta}_m, \\ \frac{\tilde{c}_l}{\tilde{\lambda}}(\tilde{\theta} - \tilde{\theta}_m) & \text{if } \tilde{\theta} \geq \tilde{\theta}_m, \end{cases}$$

$$v := \frac{\tilde{v}\bar{T}}{\bar{X}}, \quad k := \frac{\tilde{k}\bar{T}}{\tilde{\rho}\tilde{c}\bar{X}^2}, \quad p := \frac{\tilde{p}\bar{T}}{\tilde{\rho}\tilde{c}\bar{X}}, \quad g_D := \frac{\tilde{c}}{\tilde{\lambda}}(\tilde{g}_D - \tilde{\theta}_m), \quad g_N := \frac{\tilde{g}_N\bar{T}}{\tilde{\rho}\tilde{\lambda}\bar{X}}.$$

Using these new quantities, the dimensionless equation reads

$$u_t + v \partial_z u - \Delta \beta(u) = 0 \quad \text{in } \Omega \times (0, T), \quad \text{with } \beta(u) = \begin{cases} k_s u & \text{if } u < 0, \\ 0 & \text{if } u \in [0, 1], \\ k_l (u - 1) & \text{if } u > 1. \end{cases}$$

Dirichlet boundary conditions are transformed into

$$\beta(u) = k g_D \quad \text{on } \Gamma_0 \times (0, T),$$

and the scaled Robin and Neumann conditions are

$$\begin{aligned} k \partial_\nu \theta + p(\theta - \theta_{\text{ext}}) = 0 & \Leftrightarrow \partial_\nu \beta(u) + \frac{p}{k}(\beta(u) - k\theta_{\text{ext}}) = 0 \quad \text{on } \Gamma_N \times (0, T), \\ \partial_\nu \beta(u) = g_N & \quad \text{on } \Gamma_L \times (0, T). \end{aligned}$$

After scaling with $\bar{X} = 10 m$, $\bar{T} = 10^5 s \approx 28 h$, the non-dimensional domain is of size 0.021×2.5 and the slopes of β are $k_s = 0.006$, $k_l = 0.029$. A temperature range $\tilde{\theta} \in (1000, 1800) \text{ }^\circ\text{K}$ leads to scaled values $|\beta(u)| = O(10^{-2})$, while the scaled latent heat is $\lambda = 1$. The scaled convection speed is $v = 225$, so convection is *dominant*. The simulations run for $t \in (0, 0.1)$, which is equivalent to a final time $\tilde{T} = 10000s \approx 2\frac{3}{4}h$. Initial conditions are chosen piecewise linear in z direction, with a prescribed initial position of the interface at $z = L/10$. This is a convenient but totally unphysical initial condition: there is liquid in contact to water/air. The long-time behavior does not depend on the actual choice of initial conditions though.

Figure 16.8: Domain aspect ratio.

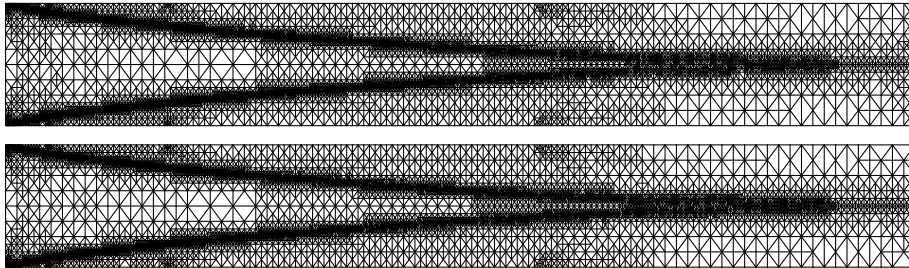


Figure 16.9: Adapted mesh for stationary solution, Dirichlet (top) and Neumann (bottom) outflow; vertical scale = 16.

The actual aspect ratio of the domain is depicted in Figure 16.8; so for visualization purposes, the height of all subsequent domains is scaled by a factor 16. The numerical simulations start from a macro triangulation of Ω into 20 triangles with aspect ratio ≈ 12 . Figures 16.9 and 16.10 compare adapted meshes and graphs of the temperature for Dirichlet and Neumann outflow conditions with error tolerances $tol = 2$ and $tol = 45$, respectively. It can be easily seen that the Dirichlet outflow condition generates a sharp boundary layer at Γ_L but no oscillations elsewhere; both solutions are indeed very similar away from Γ_L . To avoid this unphysical boundary layer, the following simulations were all done with a vanishing Neumann outflow condition. We conclude this paper with two simulations with time-dependent parameters.

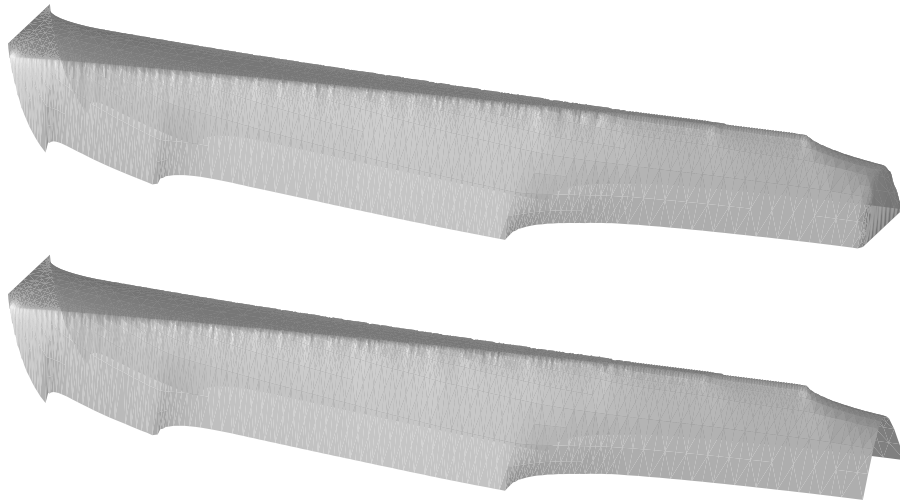


Figure 16.10: Graph of temperature for stationary solution with Dirichlet (top) and Neumann (bottom) outflow.

16.18 Example: Oscillating Velocity

First, we prescribe a variable casting speed

$$\tilde{v}(\tilde{t}) = 0.0175 + 0.005 * \sin(0.00175 \tilde{t}) [m/s],$$

which has a strong influence on the length of the liquid pool inside the slab. The largest velocity is chosen equal to the constant velocity in the problem with stationary casting speed; all other parameters are left unchanged. This guarantees that the liquid pool will not reach the outflow boundary. The variable velocity is shown in Figure 16.11, together with the number of elements in the adapted meshes $M(t^n) = \#(\mathcal{M}^n)$ and time step sizes. Due to a longer liquid pool (and interface), there are more mesh elements when the velocity is larger and a smaller time step size is needed. Figures 16.12 and 16.13 display adaptive meshes and temperature graphs for $t = 0.05$ and $t = 0.07$, corresponding to large and small velocity values. Some spurious oscillations can be seen in the temperature graphs near jumps of Robin boundary conditions. They are created by the method of characteristics which transport such cusps in the z direction. Therefore, an upper bound of 0.00025 ($= 25 s$) is imposed in this simulation.

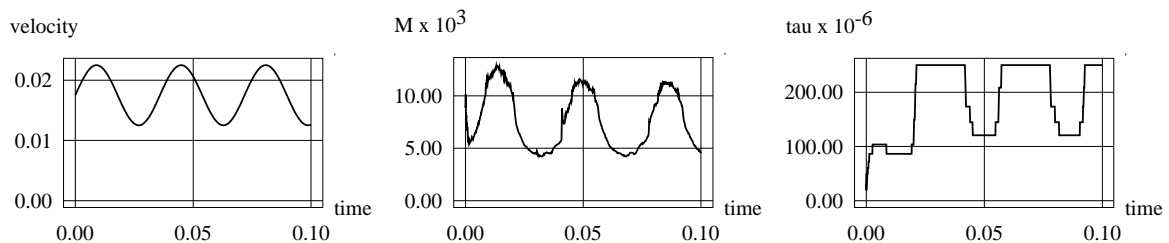


Figure 16.11: Variable casting speed. Velocity $\tilde{v}(t)$, element count, and time step sizes.

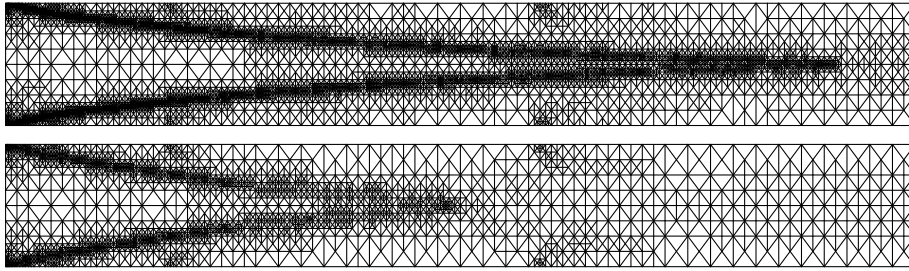


Figure 16.12: Variable casting speed. Adaptive meshes for $t = 0.05$ (top) and $t = 0.07$ (bottom).



Figure 16.13: Variable casting speed. Temperature graphs for $t = 0.05$ (top) and $t = 0.07$ (bottom).

16.19 Example: Oscillating Cooling

For constant casting velocity $\tilde{v} = 0.0225 \text{ m/s}$, we model a varying cooling water flow rate in the second spray region by a time dependent heat transfer coefficient:

$$\tilde{p}(\tilde{t}) = 550 + 200 \sin(0.00175 \tilde{t}) \text{ [kg/s}^3 \text{ }^\circ\text{K]} \quad \text{on } \Gamma_L, \tilde{z} \in (4.4, 14.6) \text{ m.}$$

Again, this has an influence on the length of the liquid pool inside the slab, which gets longer when the cooling coefficient is smaller, thereby representing a reduced water flow. Figure 16.14 shows the varying parameter $\tilde{p}(t)$ and the corresponding mesh element counts. Adaptive meshes and temperature graphs for $t = 0.05$ and $t = 0.07$ are displayed in Figures 16.15 and 16.16. As the liquid pool length does not depend so strongly on $\tilde{p}(t)$ as it did on $\tilde{v}(t)$, the mesh element count changes only slightly in this example; the larger changes for $t < 0.02$ are due to the given initial conditions. The time step size is not shown but equals the given upper bound 0.00025 for $t > 0.02$. The oscillations in Figure 16.16 near jumps of Robin boundary conditions along Γ_N uncover the undesirable condition $v\tau \gg h$ for the method of characteristics. We show the beneficial effect of reducing the time step in the bottom picture of Figure 16.16. This graph corresponds to

$t = 0.05$ for a simulation with smaller tolerance for the time error estimate, which leads to a time step size $\tau = 0.00006$ ($= 6$ s) for $t > 0.025$: $v\tau < h$ holds and the oscillations are removed.

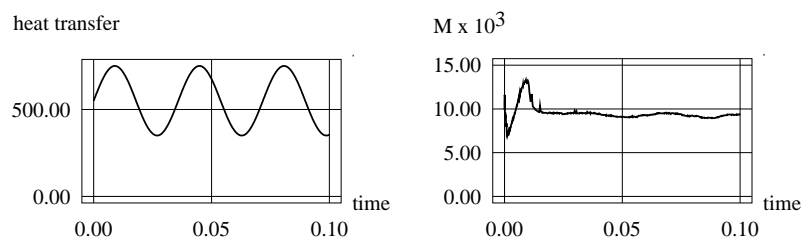


Figure 16.14: Variable cooling. Heat transfer coefficient $\tilde{p}(t)$ and element counts.

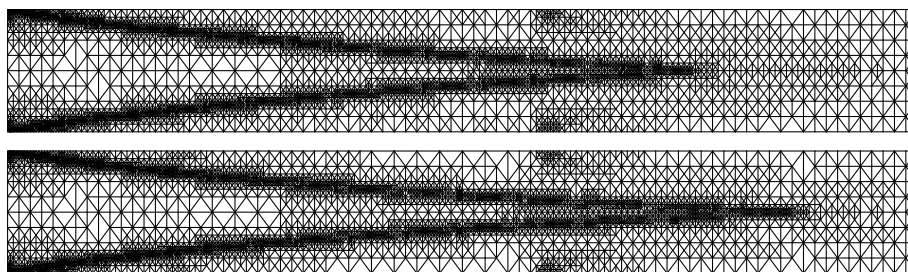


Figure 16.15: Variable cooling. Adaptive meshes for $t = 0.05$ (top) and $t = 0.07$ (bottom).

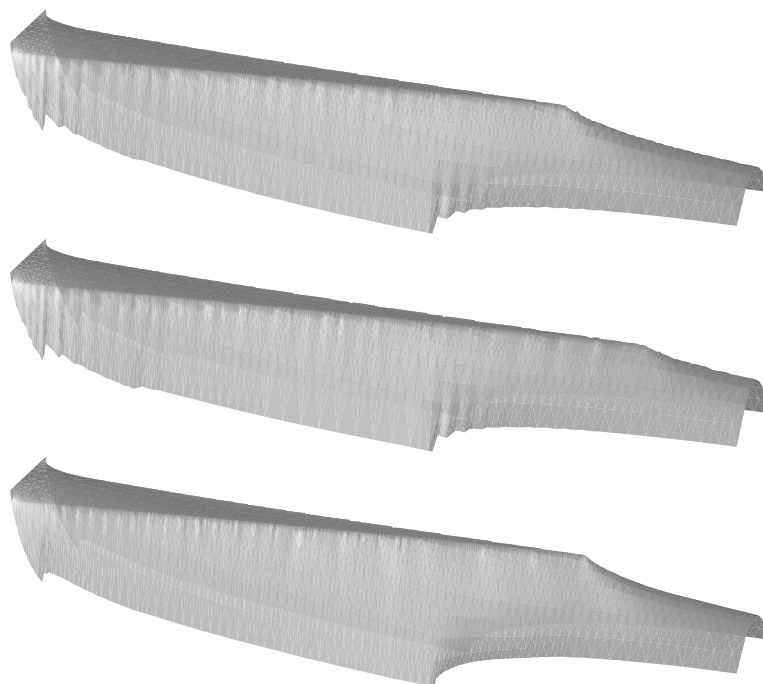


Figure 16.16: Variable cooling. Temperature graphs for $t = 0.05$ (top) and $t = 0.07$ (middle); Simulation with smaller time step size for $t = 0.05$ (bottom).

17 Mathematical modeling of thermal cutting

17.1 Introduction

There is a wide range of thermal cutting techniques available for the shaping of materials. One example is the plasma cutting. The cutting of the workpiece occurs as a result of melting/vaporizing the material by an extremely hot cylindrical plasma beam which burns and melts its way through the material, leaving a kerf in its wake.

The heat transfer from the plasma beam into the material accounts for most of the phenomena encountered subsequently: shrinkage, residual stresses, metallurgical changes (e.g. phase transition in the Heat Affected Zone, HAZ), mechanical deformations (e.g. the cut edge is not square as desired), chemical modifications, etc.

On the other hand the speed of moving plasma beam can cause a formation of high or low speed drosses, which is another problem as the removal of the dross is an additional operation which increases the cost of the cutting.

Investigations are needed for the prediction and control of the above mentioned phenomena during the plasma arc cutting process. To get the quantitative description of the process, one requires a mathematical model for it. Therefore a proper mathematical model has to be developed which must involve the different physical phenomena occurring in the cutting workpiece, i.e. heat conduction, convection and radiation effects, mechanical deformations, phase transition, etc.

While experiment reveals the particular features of every process, the developed model will permit the establishment of the general laws and thus will contribute to the fundamental knowledge of the process.

17.2 Problem description and physical modeling

Plasma cutting is desirable for many metal cutting applications. Carbon steels, aluminum and stainless steel are some examples of materials cut with thermal plasma. The heat source for plasma cutting is a high temperature and high velocity stream of partially ionized gas. The plasma stream appears as a result of a current which passes between a cathode and an anode workpiece. Due to the current, the plasma gas (mixture of nitrogen, hydrogen and argon) is heated to high temperature and the Lorenz forces propel it down towards the workpiece (anode) in the form of high velocity jet (see figure 17.2). Typical plasma temperatures are in the range of 10,000K to 30,000K.

Anyway, how does the thermal cutting work? The essential idea of cutting is to focus a lot of power onto a small area of surface of the material producing intense surface heating. First the material on the surface melts and then evaporates. As the vapor is puffed away

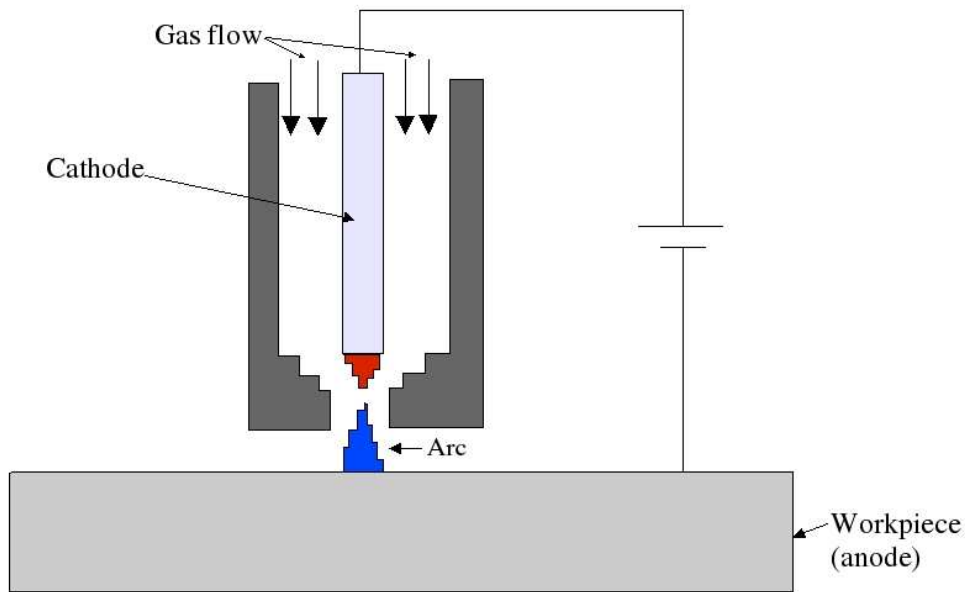


Figure 17.1: Thermal plasma system

or the molten metal is removed by the high speed gas flow, so a hole develops in the material (see figure 17.2).

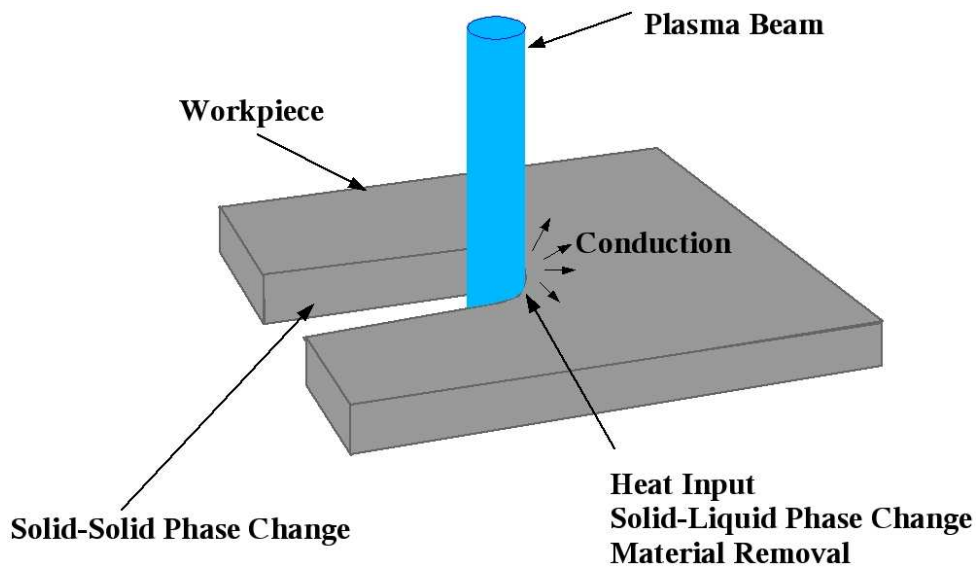


Figure 17.2: Real picture of thermal cutting

The problem which the industry faces at the moment is the deformations of the cut edges after the material is cut and cooled down. The deformations might be a result of shrinkage, residual stresses, metallurgical changes (e.g. phase transition), etc. Due to this deformations, the cut edges are not straight any more which makes a lot of difficulties

during the further applications of the metal. So, one needs to develop a mathematical model which must involve the different physical phenomena occurring in the workpiece, i.e. heat conduction, convection and radiation effects, phase transition, etc. The developed model will permit the establishment of the general laws and thus contribute to the fundamental knowledge of the cutting process. Results from the numerical simulations (verified by the experiments) of the model will provide a quantitative description of the process. The mathematical model should mainly include the

- temperature field analysis in the workpiece,
- effects of cutting on the geometry of the cut pieces,
- investigation of the properties of the material due to the solid-solid phase change during the cutting process.

17.3 Mathematical formulation of the problem

Let Ω be a given, open and bounded domain in \mathbb{R}^3 occupied by the workpiece, $0 < T < +\infty$. Denote: $\theta(x, t) :=$ the temperature of the workpiece, $I := (0, T)$, $\partial\Omega :=$ the boundary of Ω which is assumed to be piecewise smooth. The initial temperature distribution of the workpiece is given by $\theta_0(x)$, which is less than the melting temperature at all points. For every $t \in I$ the domain Ω is assumed to consist of three non-intersecting parts, namely $\Omega = \Omega_s(t) \cup \partial\Omega_s(t) \cup \Omega_c(t)$, where $\Omega_s(t)$ and $\Omega_c(t)$ are the domains occupied by the solid part of the workpiece and cut cavity at a time instant t , respectively, and $\partial\Omega_s(t)$ is the free interface at time t . Let $\partial\Omega_s(t)$ be also smooth. By ν we shall denote the unit outward normal vector of the domain $\Omega_s(t)$. Let j_{abs} be the heat flux density absorbed by the interface due to the plasma beam radiation. In addition to the terms defined above we will use the following notations: ρ is the density of the workpiece, c_s is the specific heat, k is the heat conductivity of the material, L_m is the latent heat of melting, h is the linear convective heat transfer coefficient from the surface of the workpiece, θ_m is the melting temperature, θ_a is the ambient temperature, $v \geq 0$ is the velocity of the free interface.

With the above mentioned notations and assuming no heat exchange between the workpiece and the exterior through $\partial\Omega_s(t)$, the classical mathematical formulation of the problem is the following:

find the function $\theta(x, t) \in \mathbb{C}_1^2(\Omega_s \times I) \cap \mathbb{C}(\overline{\Omega_s \times I})$, representing the temperature of the body, and the piecewise smooth surface $\partial\Omega_s(t)$ representing the free boundary of the solid domain $\Omega_s(t) = \{x; \theta < \theta_m\}$ such that the heat conduction equation is fulfilled

$$(17.1) \quad \rho c_s \frac{\partial \theta}{\partial t} = \nabla \cdot (k \nabla \theta) \quad x \in \Omega_s(t), \quad t \in I$$

with the following boundary conditions on $\partial\Omega_s(t)$:

$$\begin{aligned}
(17.2) \quad & \theta \leq \theta_m \\
& k\nabla\theta \cdot \nu + j_{abs} \cdot \nu \geq 0 \\
& (\theta - \theta_m)(k\nabla\theta \cdot \nu + j_{abs} \cdot \nu) = 0
\end{aligned}$$

called the **Signorini** boundary conditions, and

$$(17.3) \quad -k\nabla\theta \cdot \nu + \rho L_m v \cdot \nu = j_{abs} \cdot \nu$$

named the **Stefan** boundary condition,

and corresponding initial conditions

$$(17.4) \quad \theta(x, 0) = \theta_0(x) < \theta_m \quad x \in \Omega$$

$$(17.5) \quad s(x, 0) = s_0$$

Note that the heat flux density j_{abs} and v are equal to zero on the part of boundary where no heat input takes place.

Remark 17.1. *The Signorini boundary conditions (17.2) imply that at each time instant t there exist on $\partial\Omega_t$ two regions where on each region either $\theta = \theta_m$ or $j_{abs} \cdot \nu = -k\nabla\theta \cdot \nu$; on the regions where $j_{abs} \cdot \nu > -k\nabla\theta \cdot \nu$, i.e. the heat flux on the absorbing surface is greater than the heat conducted into the workpiece, the temperature is equal to the melting temperature (thus we have $\theta = \theta_m$ on this part of the boundary), on the other hand, on the regions where $\theta < \theta_m$, the entire heat flux absorbed by the interface is conducted within the material, thus yielding the condition $j_{abs} \cdot \nu > -k\nabla\theta \cdot \nu$.*

Remark 17.2. *The idea behind the Stefan boundary condition is rather simple; the total heat flux on the interface is decided into two parts: one part is conducted and the other part is used to melt the material.*

Note, that both Signorini and Stefan boundary conditions are non linear. Moreover, above mentioned regions are not prescribed, resulting in a “free boundary problem”.

Remark 17.3. *The above formulated problem could be referred to as a one phase Stefan problem, although there are some important differences between them. The one-phase Stefan problem represents a special case of the classical two-phase formulation, with the temperature constant in one of the phases, assuming the melting value. Here we have a different situation. First of all, we can not assume the value of the temperature in the cavity (where the melt is removed) equal to the melting temperature of the solid, because otherwise the cut edges will continue to melt and move forward, which does not correspond to the real situation of plasma cutting. Secondly, in our problem an additional heat source (plasma beam) is applied on the surface of the moving front which is not the case in classical one-phase Stefan problem.*

17.4 Variational inequalities and the weak formulation of the problem

The Signorini boundary conditions (17.2) allow us to rewrite the system (17.1)-(17.5) in the form of variational inequalities (VI). First we introduce some notations and a small fraction of the known theory of function spaces - just enough to establish the weak formulation of Stefan-Signorini system.

17.4.1 Notation and Functional Spaces

Since we are only interested in the modeling of solid region, for convenience we will write Ω_t instead of $\Omega_s(t)$. The time dependent domain Ω_t is clearly a bounded open subset of Ω and $\Omega = \cup_{t \in [0, T]} \Omega_t$.

Define by $\theta(t)$ the function $x \rightarrow \theta(x, t)$ and let us consider a bilinear form

$$(17.6) \quad \theta, \varphi \rightarrow b(\theta, \varphi)$$

$$(17.7) \quad b(\theta, \varphi) = \int_{\Omega_t} \nabla \theta \cdot \nabla \varphi dx$$

which is obviously a continuous, symmetric and coercive form on $\mathbb{H}^1(\Omega_t) \times \mathbb{H}^1(\Omega_t)$.

Let us define

$$(17.8) \quad \mathbb{V} = \left\{ \theta \mid \theta \in \mathbb{L}^2(0, T; H^1(\Omega_t)), \theta' = \frac{\partial \theta}{\partial t} \in \mathbb{L}^2(0, T; (H^1(\Omega_t))') \right\}$$

Provided with the scalar product

$$(17.9) \quad (\theta, \varphi)_{L^2(0, T; H^1(\Omega_t))} + (\theta', \varphi')_{L^2(0, T; (H^1(\Omega_t))')},$$

\mathbb{V} is a Hilbert space.

Remark 17.4. *The following embedding result of Dautrey and Lions ([63])*

$$(17.10) \quad \mathbb{L}^2(0, T; H^1(\Omega)) \cap \mathbb{H}^1(0, T; (H^1(\Omega))') \subset \mathbb{C}^0(0, T; L^2(\Omega))$$

implies that any function $\theta \in \mathbb{V}$ is almost everywhere equal to a continuous function from $[0, T] \rightarrow L^2(\Omega_t)$.

Thus, we can define a closed subspace \mathbb{V}_0 of \mathbb{V} in the following form

$$(17.11) \quad \mathbb{V}_0 = \{ \theta \mid \theta \in \mathbb{V}, \theta(0) = \theta_0, \text{ where } \theta_0 \text{ is a given function in } \mathbb{L}^2(\Omega) \}$$

In order to restate the initial problem in the form of variational inequality, we introduce the set $K \subset \mathbb{H}^1(\Omega_t)$ which is a closed and convex subset of $\mathbb{H}^1(\Omega_t)$

$$(17.12) \quad K = \{\varphi \mid \varphi \in \mathbb{H}^1(\Omega_t), \varphi \leq \theta_m \text{ on } \partial\Omega_t\}$$

For the functions $\theta \in \mathbb{V}$ we can define their trace on the lateral boundary $\Gamma := \partial\Omega_t \times I$ of $Q := \Omega_t \times I$. Recall that $\theta|_\Gamma \in \mathbb{L}^2(0, T; H^{\frac{1}{2}}(\partial\Omega_t))$. The latter allows us to define closed convex subsets \mathbb{B} and \mathbb{B}_0 in the following way

$$(17.13) \quad \mathbb{B} = \{\theta \mid \theta \in \mathbb{V}, \theta(t) \in K \text{ a.e. in } [0, T]\}$$

$$(17.14) \quad \mathbb{B}_0 = \{\theta \mid \theta \in \mathbb{V}_0, \theta(t) \in K \text{ a.e. in } [0, T]\}$$

17.4.2 A VI equivalent of Stefan-Signorini problem

We are interested in the solution of the following problem:

find $\theta \in \mathbb{B}_0$ such that

$$(17.15) \quad \int_0^T \int_{\Omega_t} \theta'(\varphi - \theta) dxdt + \int_0^T \int_{\Omega_t} \nabla\theta \cdot \nabla(\varphi - \theta) dxdt \geq \int_0^T \int_{\partial\Omega_t} g(\varphi - \theta) d\gamma dt$$

for all $\varphi \in \mathbb{B}$, $g \in \mathbb{H}^1(R^n) \cap \mathbb{L}^\infty(R^n)$.

It can be easily shown that the problems (17.1)-(17.2), (17.4)-(17.5) and (17.13), (17.14), (17.15) are equivalent. Another equivalent formulation of the problem can be written in the following way:

find $\theta(t) \in K$ for $t \in I$ such that

$$(17.16) \quad \int_{\Omega_t} \theta'(\varphi - \theta) dxdt + \int_{\Omega_t} \nabla\theta \cdot \nabla(\varphi - \theta) dx \geq \int_{\partial\Omega_t} g(\varphi - \theta) d\gamma$$

for all $\varphi \in K$, $\theta(0) = \theta_0$.

Remark 17.5. *The condition on θ' forcing it to belong to the space $\mathbb{L}^2((0, T; (H^1(\Omega_t))))$ seems to be too restrictive. It will be useful to consider a more general weak formulation. We observe that if θ is a solution of (17.13), (17.14), (17.15) and $\varphi \in \mathbb{B}$, then*

$$(17.17) \quad \begin{aligned} & \int_0^T \int_{\Omega_t} \varphi'(\varphi - \theta) dxdt + \int_0^T \int_{\Omega_t} \nabla\theta \cdot \nabla(\varphi - \theta) dxdt \\ & \geq \int_0^T (\varphi' - \theta')(\varphi - \theta) dt + \int_0^T \int_{\partial\Omega_t} g(\varphi - \theta) d\gamma dt \\ & \geq -\frac{1}{2} \|\varphi(0) - \theta_0\|^2 + \int_0^T \int_{\partial\Omega_t} g(\varphi - \theta) d\gamma dt \end{aligned}$$

If we now choose the function φ from the space \mathbb{B}_0 , then the term $-\frac{1}{2} \|\varphi(0) - \theta_0\|^2$ can be also avoided. Note that in the problem (17.17) we require only that $\theta \in \mathbb{L}^2(0, T; H^1(\Omega_t))$, nothing is said about θ' .

17.5 Level set formulation

Let assume at the moment that the temperature distribution θ in the workpiece is given. The problem of finding the second unknown of the Stefan-Signorini problem, namely the geometry of the cutting front, can be formulated as follows:

Problem 17.1. *Find a family of moving interfaces $\{\partial\Omega_t\}_{t \in (0, T)}$ such that*

$$(17.18) \quad -k\nabla\theta \cdot \nu + \rho L_m v \cdot \nu = j_{abs} \cdot \nu \text{ on } \partial\Omega_t$$

$$(17.19) \quad \partial\Omega_t|_{t=0} = \partial\Omega_0$$

The technique we are going to apply for the investigation of the condition (17.18) is the level-set theory introduced first by Sethian and Osher ([85]).

17.5.1 Stefan condition as level-set equation

The main idea of level-set method is to embed the evolving front $\partial\Omega_t$ into a surface, *the level-set surface*, which has the property that its zero-level set always yields the desired moving interface. Thus, in order to find the unknown interface at any time t , we only need to locate the set for which the level-set surface vanishes. What remains is to introduce an equation which will describe the motion of the level-set surface. For the derivation of this equation we follow the steps given in ([70]).

Let Γ_0 be the initial interface bounding some open domain. We wish to investigate and compute its motion under some velocity field which depends on several quantities. The idea of level-set approach is to introduce a scalar function (level-set function) $\phi(x, t)$, $x \in \mathbb{R}^2$ such that at any arbitrary time instant t the zero-level set of $\phi(x, t)$ is the desired curve Γ_t . To obtain the equation for the level-set function we consider some level-set $\phi(x, t) = C$. The trajectory $x(t)$ of a particle located on this level-set should satisfy the equation

$$(17.20) \quad \phi(x(t), t) = C.$$

After differentiating the equation (17.20) we get

$$(17.21) \quad \phi_t + x'(t) \nabla\phi = 0.$$

Finally, denoting by $w := x'(t)$ the particle velocity we arrive to the level-set equation

$$(17.22) \quad \phi_t + w \cdot \nabla\phi = 0.$$

In addition to the level-set equation (17.22) we need an initial condition for ϕ . Taking into account the fact that the level-set function $\phi(x, t)$ is positive for $x \in \Omega$, negative for $x \notin \overline{\Omega}$ and equal zero on the domain boundary Γ_t , it is natural to choose the signed

distance function as a good candidate for the initial value $\phi(x, t = 0)$. Thus, the initial condition for (17.22) takes the form

$$(17.23) \quad \phi(x, 0) = \begin{cases} d(x) & \text{for } x \in \Omega \\ 0 & \text{for } x \in \Gamma_0 \\ -d(x) & \text{for } x \notin \Omega \end{cases}$$

where $d(x)$ is the distance from the point x to the initial curve Γ_0 . An example of a distance function for a circle in the square is illustrated in figure 17.5.1.

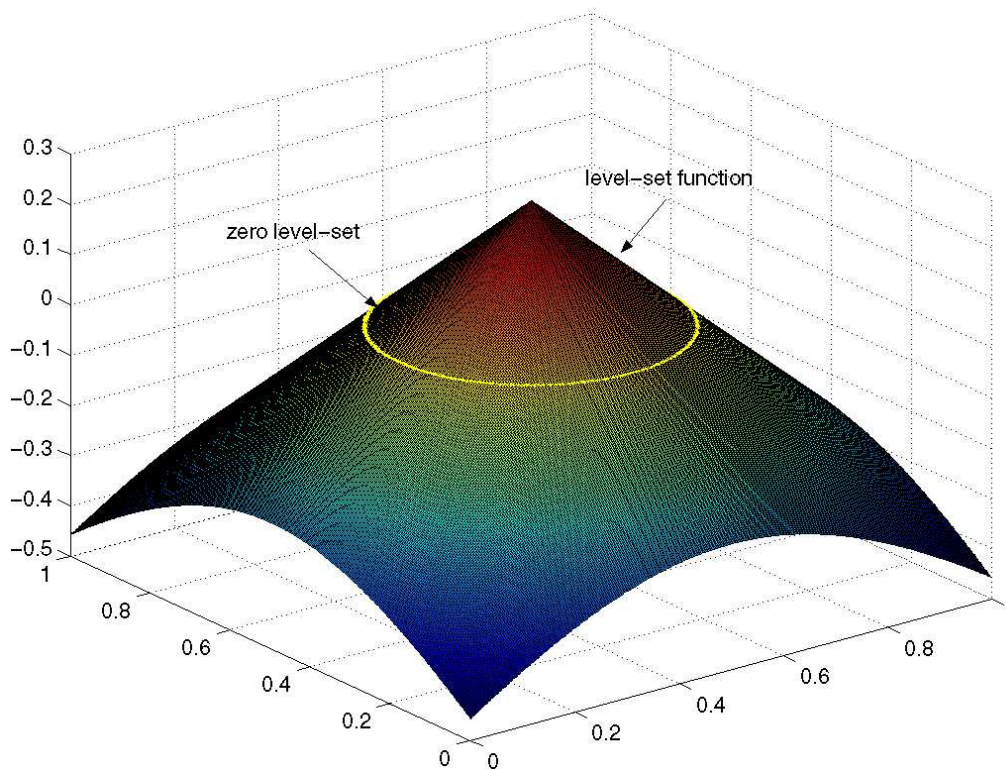


Figure 17.3: Distance Function for Circle

Coming back to Stefan condition (17.18), we suppose that the moving interface $\Gamma_t := \partial\Omega_t$, which bounds the domain Ω_t occupied by the workpiece, is built into a level-set function $\phi(x, t)$. Now we would like to redefine the terms with the help of the just introduced level-set function ϕ .

The workpiece: the domain $\Omega(t)$ is represented as

$$\Omega(t) = \{x; \phi(x, t) > 0\}.$$

The cutting interface : the advancing cut front Γ_t is given by

$$\Gamma_t = \{x; \phi(x, t) = 0\}.$$

The unit normal: the unit outward normal ν to Γ_t has the following level-set representation

$$\nu = \frac{\nabla\phi}{|\nabla\phi|}.$$

The velocity of interface: the velocity of the moving interface in the normal direction can be represented as

$$v \cdot \nu = -\frac{\frac{\partial\phi}{\partial t}}{|\nabla\phi|}.$$

Indeed, for two dimensional domain (analogy works for any dimension) the interface is given by

$$(17.24) \quad \phi(x, y, t) = 0.$$

By chain rule,

$$(17.25) \quad \frac{d\phi}{dt} = \frac{\partial\phi}{\partial x} \cdot \frac{dx}{dt} + \frac{\partial\phi}{\partial y} \cdot \frac{dy}{dt} + \frac{\partial\phi}{\partial t} = 0.$$

Using the definitions $\nabla\phi = \left(\frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}\right)$ and $v = \left(\frac{dx}{dt}, \frac{dy}{dt}\right)$ we arrive to

$$(17.26) \quad v \cdot \nabla\phi = -\frac{\partial\phi}{\partial t}$$

But

$$v \cdot \nabla\phi = v \cdot \frac{\nabla\phi}{|\nabla\phi|} \cdot |\nabla\phi| = v \cdot \nu \cdot |\nabla\phi|$$

Thus

$$v \cdot \nu = -\frac{\frac{\partial\phi}{\partial t}}{|\nabla\phi|}.$$

The terms in Stefan condition (17.18) are already defined by their level-set representation. All that remains is to substitute their expressions in (17.18).

$$\begin{aligned} -k\nabla\theta \cdot \frac{\nabla\phi}{|\nabla\phi|} - \rho L_m \frac{\frac{\partial\phi}{\partial t}}{|\nabla\phi|} &= j_{abs} \cdot \frac{\nabla\phi}{|\nabla\phi|} \\ \frac{\partial\phi}{\partial t} + \frac{1}{\rho L_m} (j_{abs} + k\nabla\theta) \cdot \nabla\phi &= 0. \end{aligned}$$

We can easily see that the last equation is simply the level-set equation with the velocity of convection equal to

$$w := \frac{1}{\rho L_m} (j_{abs} + k\nabla\theta).$$

Note, that w is the velocity on the advancing front and is equal to something arbitrary elsewhere.

Remark 17.6. *The velocity of propagating interface depends on the heat flux absorbed by the material at the cutting front, therefore, such parameters like the velocity with which the plasma beam advances in the cutting direction and the heat flux emitted by the beam are of central importance. On the other hand, since the molten material is removed immediately after it appears, the velocity of the front strongly depends on temperature distribution, more precisely, on the gradient of the temperature on the interface. You can see also the dependence of velocity on several material parameters which one would naturally expect.*

17.6 Weak formulation of Stefan-Signorini problem

Variational inequalities lead us to following weak formulation of the Stefan-Signorini problem (17.1)-(17.5):

find the pair $(\theta(x, t), \phi(x, t))$ representing the temperature and the moving interface, respectively, such that

1. $\theta \in \mathbb{B}_0$ and $\phi \in \mathbb{C}(Q)$ with $\phi(x, 0) = \phi_0(x)$ in Ω ,
2. $\nabla\theta$ belongs to the space $\mathbb{L}^2(0, T; H^1(\Omega_t))$,
3. $\theta \leq \theta_m$ on $\partial\Omega_t$ with $\theta \in \mathbb{L}^2(0, T; H^2(G_t))$ for any G_t with $G_t \subset\subset \Omega_t$,
4. θ satisfies the inequality

$$(17.27) \quad \int_0^T \int_{\Omega_t} \theta'(\varphi - \theta) dxdt + \int_0^T \int_{\Omega_t} \nabla\theta \cdot \nabla(\varphi - \theta) dxdt \geq \int_0^T \int_{\partial\Omega_t} g(\varphi - \theta) d\gamma dt$$

for all $\varphi \in \mathbb{B}$, $g \in \mathbb{H}^1(R^n) \cap \mathbb{L}^\infty(R^n)$,

5. ϕ is the solution of the equation

$$(17.28) \quad - \int_Q \phi v' dxdt + \int_Q \phi v \nabla \cdot \omega dxdt - \int_Q \phi \omega \cdot \nabla v dxdt = \int_\Omega \phi_0 v(x, 0) dx$$

for all $v \in \mathbb{C}^1(0, T; C_0^1(\Omega_t))$ with $v(x, T) = 0$.

where $\omega = q + \nabla\theta$ is the velocity of the zero level-set of ϕ .

Following the results in [65] and [15] one can show the existence and uniqueness of the weak solution of the above problem.

17.7 Heat Flux Density

A feature common to most plasma and laser cutting processes is that they occur as a result of removing the material by melting and/or vaporization as intense laser light or a high-temperature, partially ionized gas stream interacts with the material surface.

The kinetics of those cutting processes are combined mainly by i) the amount of heat generated by plasma arc or laser beam, and ii) the heat conducted through the workpiece.

Denote:

j_e - the heat flux density (the quantity of heat flowing across a unit area) emitted from the arbitrary point Q of the surface of the plasma jet,

j_{abs} - the heat flux density absorbed at the point P of the surface of the workpiece due to the emission from Q ,

r - the radius of the plasma beam (see figure 1).

There are several studies on the heat inputs from the different sources.

In relation to the measurable quantities (current voltage and power) Rosenthal [90] has made a study of the plasma arc and found that the energy delivered to the workpiece Q_w represents about 65% of the total energy Q_t supplied by the arc. Expressed in formula

$$(17.29) \quad Q_p = 0.65Q_t = 0.65 \cdot konstant \cdot VI_c$$

where V is the voltage drop in arc and I_c is the current intensity. Rosenthal discussed in his paper three types of moving heat sources: point source, line source and plane source, and for each type he gave the relation between the temperature distribution and heat Q_p delivered to the workpiece. For example, in the case of a point source the relation obtained is the following

$$(17.30) \quad \theta - \theta_0 = \frac{Q_p}{2\pi k} e^{-\lambda v \xi} \frac{e^{-\lambda v r}}{r}$$

where $\xi = x - vt$ and $\frac{1}{2\lambda} = \frac{k}{\rho c}$. Note, that this relation is valid only below $\theta = \theta_m$.

Arai *et all* [2] described two categories of heat flux density measurements: i) indirect, measurements made by calculating heat transfer rates, using fundamental theories together with measurements of temperature and thermo physical properties, and ii) direct measurements using heat flux density sensors placed in the thermal field.

In the model of Schulz *et all* [96] the heat flux density absorbed at the boundary is proportional to the laser beam intensity I via the absorption coefficient A_p

$$(17.31) \quad j_{abs} = -A_p I e_z \cdot \nu$$

where $e_z \cdot \nu$ is the angle of incidence of the laser beam.

The laser beam intensity I itself is characterized by the maximum intensity of the beam I_0 and the beam radius

$$(17.32) \quad I = -I_0(t) f\left(\frac{x - v_0 t}{r}\right)$$

where v_0 is the speed of feeding (the speed of the moving laser) and f is a distribution ($0 \leq f \leq 1$).

Bunting *et all* [3] developed a relationship between the power density incident on a material and the cut speed in terms of the thermal properties of the material. They used the technique of Rosenthal on moving heat sources and got the following relation

$$(17.33) \quad \frac{j_e}{h} = \frac{2k(\theta_m - \theta_0)}{r^2} \cdot \frac{1}{I(s)}$$

where h is the thickness of the material, $s = \frac{vr}{2\alpha}$ and $I(s)$ has been calculated by authors and could be expressed

$$(17.34) \quad I(s) = \int_0^1 r' dr' \int_0^{2\pi} \exp(-sr' \cos\phi) K_0 s (r'^2 - 2r' \sin\phi + 1)^{1/2} d\phi$$

where K_0 is the zeroth order, modified Bessel function of the second kind, α is the heat diffusivity and the equation is written in cylindrical coordinates (r', ϕ) with dimensionalized r' .

In studying the heat-affected zone during the laser cutting of stainless steel, Sheng *et all* [97] expressed the beam energy $E_b(x, y)$ as a function of spatial coordinates via the beam intensity $I(x, y)$ of Gaussian type

$$(17.35) \quad E_b(x, y) = \int I(u, y) \frac{du}{v} = \int \frac{A(u, y)P}{\pi r^2 v} \exp\left(-\frac{u^2 + y^2}{r^2}\right) du$$

where A is the absorptivity and P is the beam power.

In the following we describe a simple technique to calculate the heat flux density on the absorbing surface.

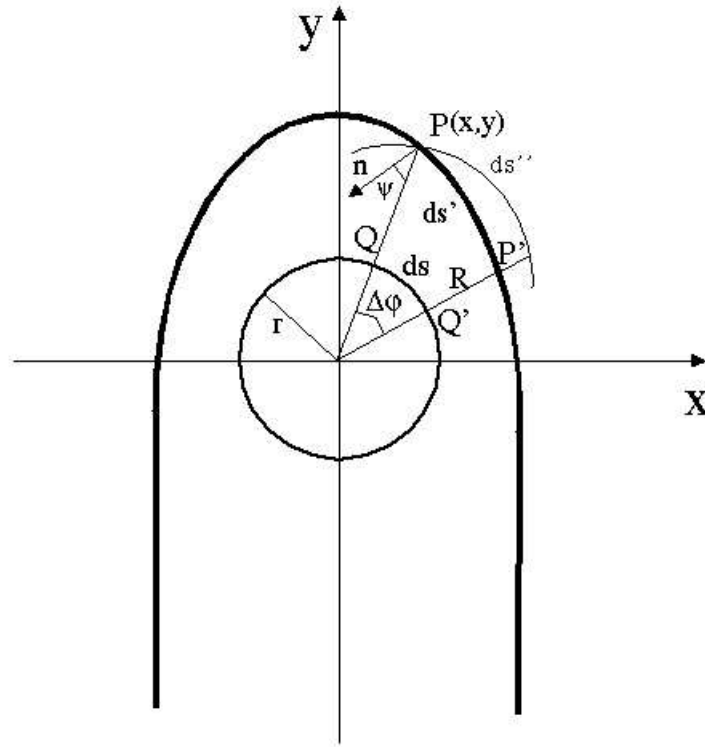
For the calculations it is convenient to discuss the topic not in terms of point source, but in terms of incremental surface elements. Therefore consider, as illustrated in the figure, a small emitting surface of area (length) ds , where the point Q is located. Further, let us assume that the cylindrical surface of the plasma beam is emitting heat in the radial direction, i.e. the heat flux density vector at any point of the beam surface has the direction of the normal to the plasma surface at that point.¹

Having in our disposal the heat flux density j_e at the point Q and the radius of the beam r , our aim is to calculate the heat flux density absorbed at the point P of the workpiece surface (see figure).

Let dq be the rate at which the energy leaves the incremental area ds . Then the average flux j_e^{av} leaving ds is defined as

$$(17.36) \quad j_e^{av} = \frac{dq}{ds}$$

¹This assumption is made only for the simplicity, the calculations can be done also for other flux density distributions



and the flux due to the point Q on the beam surface is defined to be

$$(17.37) \quad j_e = \lim_{ds \rightarrow 0} \frac{dq}{ds}$$

The flux emitted from ds is then completely absorbed (assume the material is a black body) by the surface of the material, more precisely, by the part of the surface which we denote by ds' . Then analogous to (17.36), the average heat flux density j_{abs}^{av} absorbed by ds' will be

$$(17.38) \quad j_{abs}^{av} = \frac{dq}{ds'}$$

and following (17.37) we obtain

$$(17.39) \quad j_{abs} = \lim_{ds' \rightarrow 0} \frac{dq}{ds'} = \lim_{ds' \rightarrow 0} \frac{j_e ds}{ds'}$$

For cylindrical heat source we have

$$(17.40) \quad ds = r \Delta\varphi$$

Now let ds'' be an element of the spherical surface which we obtain by projecting ds' normal to the direction PQ (the direction which the point P makes with the emitting point Q). In terms of the drawing in figure, the flux at the point P due to the energy

leaving the point Q may be determined in terms of energy falling on an element ds'' of the circular surface (center at origin, radius R) which passes through P . We then obtain for the surface element ds''

$$(17.41) \quad ds'' = ds' \cos\psi \quad \text{as} \quad \Delta\varphi \rightarrow 0$$

where ψ is the angle between the normal of the workpiece surface at point P and the line PQ .

Thus,

$$(17.42) \quad ds' = \frac{ds''}{\cos\psi}$$

If the interface is represented as a smooth graph, $y = f(x)$, then for $\cos\psi$, we obtain

$$(17.43) \quad \cos\psi = -\frac{1}{\sqrt{x^2 + f^2(x)}} \cdot \frac{1}{\sqrt{1 + (f'(x))^2}} \cdot \begin{pmatrix} x \\ f(x) \end{pmatrix} \cdot \begin{pmatrix} f'(x) \\ -1 \end{pmatrix}$$

where $\frac{1}{\sqrt{x^2 + f^2(x)}} \begin{pmatrix} x \\ f(x) \end{pmatrix}$ represents the unit normal vector to the surface ds'' at point $P(x, y)$ and $\frac{1}{\sqrt{1 + (f'(x))^2}} \begin{pmatrix} f'(x) \\ -1 \end{pmatrix}$ is the unit normal to the graph $y = f(x)$ at point $P(x, y)$. Inserting expressions for ds' from (17.42) and $\cos\psi$ from (17.43) with $ds'' = R\Delta\varphi$ into (17.39) and taking into account that $ds' \rightarrow 0$ is equivalent to $\Delta\varphi \rightarrow 0$, we obtain

$$(17.44) \quad \begin{aligned} j_{abs} &= -j_e \cdot \lim_{\Delta\varphi \rightarrow 0} \frac{r\Delta\varphi}{\frac{R\Delta\varphi}{(\sqrt{x^2 + f^2(x)})^{-1} \cdot (\sqrt{1 + (f'(x))^2})^{-1} \cdot (xf'(x) - f(x))}} \\ &= -j_e r \frac{xf'(x) - f(x)}{(x^2 + f^2(x)) \sqrt{1 + (f'(x))^2}} \end{aligned}$$

17.8 Solution algorithm

The following algorithm implements one time step only. One should repeat the steps in the algorithm for every time step using several adaptive procedures described in section 14.

17.1 Algorithm (Stefan-Signorini problem). *Step 1.* Start with initialization: take $\phi(x, 0) = \phi_0(x)$ to be the signed distance to the interface and $u(x, 0) = u_0(x)$ the initial temperature distribution,

Step 1. Given the old values θ^m and ϕ^m , in each time step t^{m+1} first compute the new temperature θ^{m+1} by numerically solving the inequality (17.27), where the interface $\partial\Omega_t$

is replaced with old discrete interface being the zero-level set of ϕ^m ,

Step 3. with the updated temperature θ^{m+1} solve the weak level-set equation (17.28) and get the new level-set function ϕ^{n+1} , the zero level set of θ^{m+1} is then the new discrete interface,

Step 4. Go to step 2.

Remark 17.7. *It is intuitively clear (also known from the physical experiments) that the temperature has big variations when we are close to the cutting front and varies very little far from the interface. Thus, for a good approximation of the solution we need a fine grid at least in the regions close to the cutting front. In the regions far away from the moving interface a relatively coarser grid would be satisfactory. Of course, at each time step we refine the mesh using the adaptive procedures and keep the time-space error below the given tolerance using a posteriori error estimates.*

The following figures illustrate the temperature distribution in the workpiece and adaptively refined mesh.

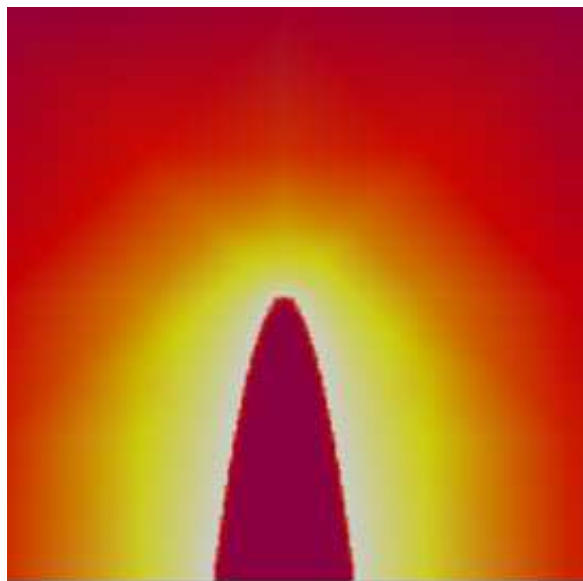


Figure 17.4: Temperature Distribution in Workpiece

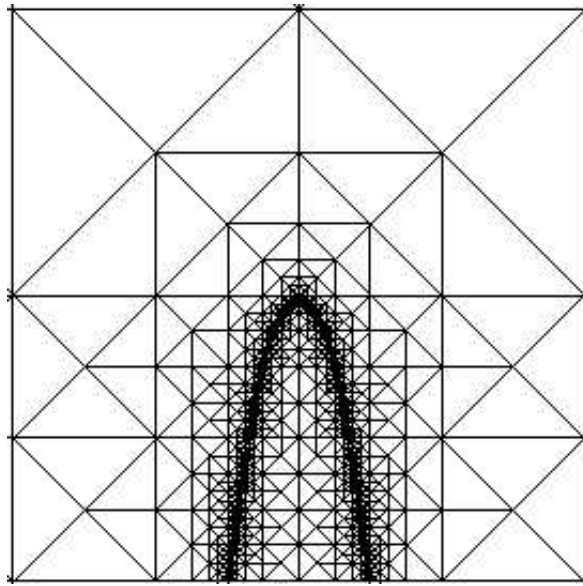


Figure 17.5: Adaptively Refined Mesh

References

- [1] R. ADAMS AND J. FOURNIER, *Sobolev spaces*, Academic press, 2003.
- [2] N. ARAI, A. MATSUNAMI AND S.W. CHURCHILL, *A Review of Measurements of Heat Flux Density Applicable to the Field of Combustion*, *Experimental Thermal and Fluid Science*, 12 (1996), pp 452–460
- [3] K.A. BUNTING, G. CORNFIELD, *Toward a General Theory of Cutting: A Relationship Between the Incident Power Density and the Cut Speed*, *Transactions of the ASME*, 116, (1975)
- [4] I. BABUŠKA AND W. RHEINBOLDT, *Error estimates for adaptive finite element computations*, *SIAM J. Numer. Anal.*, 15 (1978), pp. 736–754.
- [5] C. BANDLE, A. BRILLARD, G. DZIUK, AND A. SCHMIDT, *Course on mean curvature flow*. Lecture notes, Freiburg, 1994.
- [6] E. BÄNSCH, *Adaptive finite element techniques for the Navier–Stokes equations and other transient problems*, in *Adaptive Finite and Boundary Elements*, C. A. Brebbia and M. H. Aliabadi, eds., Computational Mechanics Publications and Elsevier, 193, pp. 47–76.
- [7] ———, *Local mesh refinement in 2 and 3 dimensions*, *IMPACT Comput. Sci. Engrg.*, 3 (1991), pp. 181–191.
- [8] E. BÄNSCH AND K. G. SIEBERT, *A posteriori error estimation for nonlinear problems by duality techniques*. Preprint 30, Universität Freiburg, 1995.
- [9] G. BELLETTINI AND M. PAOLINI, *Anisotropic motion by mean curvature in the context of Finsler geometry*. *Hokkaido Math. J.* 25 (1996), pp. 537–566.

- [10] J. BEY, *Tetrahedral grid refinement*. Report 18, SFB 382 Tübingen, 1995.
- [11] F. A. BORNEMAN, *An adaptive multilevel approach to parabolic equations I*, IMPACT Comput. Sci. Engrg., 2 (1990), pp. 279–317.
- [12] ———, *An adaptive multilevel approach to parabolic equations II*, IMPACT Comput. Sci. Engrg., 3 (1990), pp. 93–122.
- [13] ———, *An adaptive multilevel approach to parabolic equations III*, IMPACT Comput. Sci. Engrg., 4 (1992), pp. 1–45.
- [14] G. CAGINALP, *An analysis of a phase–field model of a free boundary*, Arch. Rat. Mech. Anal., 92 (1986), pp. 205–245.
- [15] B. AN TON, *A Stefan–Signorini problem with set-valued mappings in domains with intersecting fixed and free boundaries*, Bollettino U.M.I., 7 8-B (1994), pp. 231–249.
- [16] Z. CHEN AND L. JIANG, *Approximation of a two-phase continuous casting Stefan problem*, J. Partial Differential Equations 11 (1998), 59–72.
- [17] Z. CHEN, R.H. NOCHETTO AND A. SCHMIDT, *A posteriori error control and adaptivity for a phase relaxation model*, Math. Model. Numer. Anal. (M2AN) 34 (2000), pp. 775–797.
- [18] ———, *A characteristic Galerkin method with adaptive error control for the continuous casting problem*, Comp. Meth. Appl. Mech. Engrg. 189 (2000), 249–276.
- [19] X. CHEN AND F. REITICH, *Local existence and uniqueness of solutions of the Stefan problem with surface tension and kinetic undercooling*, J. Math. Anal. Appl., 164 (1992), pp. 350–362.
- [20] Z. CHEN, T.M. SHIH AND X. YUE, *Numerical methods for Stefan problems with prescribed convection and nonlinear flux*, IMA J. Numer. Anal. (1999).
- [21] P. G. CIARLET, *The finite element method for elliptic problems*, North-Holland, 1987.
- [22] P. CLÉMENT, *Approximation by finite element functions using local regularization*, R. A. I. R. O., 9 (1975), pp. 77–84.
- [23] L. DEMKOWICZ, J. T. ODEN, W. RACHOWICZ, AND O. HARDY, *Toward a universal h - p adaptive finite element strategy, Part 1 – Part 3*, Comp. Methods Appl. Mech. Engrg., 77 (1989), pp. 79–212.
- [24] W. DÖRFLER, *A robust adaptive strategy for the nonlinear Poisson equation*, Computing, 55 (1995), pp. 289–304.
- [25] ———, *A convergent adaptive algorithm for Poisson’s equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.
- [26] ———, *A time- and spaceadaptive algorithm for the linear time-dependent Schrödinger equation*, Numer. Math., 73 (1996), pp. 419–448.

- [27] J. DOUGLAS, JR. AND T.F. RUSSELL, *Numerical methods for convection-dominated diffusion problem based on combining the method of characteristic with finite element or finite difference procedures*, SIAM J. Numer. Anal. 19 (1982), pp. 871-885.
- [28] G. DZIUK, *Convergence of a semi discrete scheme for the curve shortening flow*, Math. Meth. Appl. Sc., 4 (1994), pp. 589-606.
- [29] ———, *Convergence of a semi-discrete scheme for the anisotropic curve shortening flow*.
- [30] I. EKELAND AND R. TEMAM, *Analyse convexe et problèmes variationnelles*, Dunod Gauthier-Villars, 1974.
- [31] C. M. ELLIOTT, *On the finite element approximation of an elliptic variational inequality arising from an implicit time discretization of the Stefan problem*, IMA J. Numer. Anal., 1 (1981), pp. 115-125.
- [32] C. M. ELLIOTT AND R. SCHÄTZLE, *The limit of the anisotropic double-obstacle Allen-Cahn equation*. Proc. Royal Soc. Edinburgh 126 A (1996), pp. 1217-1234.
- [33] K. ERIKSSON AND C. JOHNSON, *Adaptive finite element methods for parabolic problems IV: Nonlinear problems*, SIAM J. Numer. Anal. 32 (1995), pp. 1729-1749.
- [34] ———, *Adaptive finite element methods for parabolic problems I: A linear model problem*, SIAM J. Numer. Anal., 28 (1991), pp. 43-77.
- [35] ———, *Adaptive finite element methods for parabolic problems IV: nonlinear problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1750-1763
- [36] L. C. EVANS, H. M. SONER, AND P. E. SOUGANIDIS, *Phase transitions and generalized motion by mean curvature*, Comm. Pure and Appl. Math, 45 (1992), pp. 1097-1123.
- [37] L. C. EVANS AND J. SPRUCK, *Motion of level sets by mean curvature I*, J. Diff. Geom., 33 (1991), pp. 635-681.
- [38] M. FRIED, *A Level Set Based Finite Element Algorithm for the Simulation of Dendritic Growth*, Freiburg 2001, Preprint, to appear in Computing and Visualization in Science.
- [39] ———, *Berechnung des Krümmungsflusses von Niveauflächen*. Thesis, Freiburg, 1993.
- [40] J. FRÖHLICH, J. LANG, AND R. ROITZSCH, *Selfadaptive finite element computations with smooth time controller and anisotropic refinement*. Preprint SC 96-16, ZIB Berlin, 1996.
- [41] D. GILBARG AND N. S. TRUDINGER, *Elliptic partial differential equations of second order*, Springer, 1983.
- [42] M. E. GLICKSMAN, R. J. SCHAEFER, AND J. D. AYERS, *Dendritic growth — a test of theory*, Metal. Trans. A, 7A (1976), pp. 1747-1759.
- [43] R. GLOWINSKI, *Numerical Methods for Nonlinear Variational Problems*, Springer, 1984.

- [44] P. GRISVARD, *Elliptic Problems on Non-smooth Domains*, Pitman, Boston, 1985.
- [45] M. E. GURTIN, *Toward a nonequilibrium thermodynamics of two-phase materials*, Arch. Ration. Mech. Anal., 100 (1988), pp. 275–312.
- [46] W. HACKBUSCH, *Theorie und Numerik elliptischer Differentialgleichungen*, Teubner, 1986.
- [47] R. W. HOPPE, *A globally convergent multi-grid algorithm for moving boundary problems of two-phase Stefan type*, IMA J. Numer. Anal., 13 (1993), pp. 235–253.
- [48] R. W. HOPPE AND R. KORNHUBER, *Adaptive multilevel methods for obstacle problems*, SIAM J. Numer. Anal., 31 (1994), pp. 301–323.
- [49] G. HUISKEN, *Flow by mean curvature of convex surfaces into spheres*, J. Diff. Geom., 20 (1984), pp. 237–266.
- [50] J. E. HUTCHINSON, *Computing conformal maps and minimal surfaces*, Proc. C.M.A., Canberra, 26 (1991), pp. 140–161.
- [51] H. JARAUSCH, *On an adaptive grid refining technique for finite element approximations*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 1105–1120.
- [52] K. KALIK AND W. WENDLAND, *The approximation of closed manifolds by triangulated manifolds and the triangulation of closed manifolds*, Computing, 47 (1992), pp. 255–275.
- [53] J. KEVORKIAN, *Partial Differential Equations*, Chapman and Hall, London, 1990
- [54] R. KORNHUBER, *Monotone multigrid methods for elliptic variational inequalities I*, Numer. Math., 69 (1994), pp. 167–184.
- [55] —, *Monotone multigrid methods for elliptic variational inequalities II*, Numer. Math., 72 (1996), pp. 481–500.
- [56] —, *Adaptive monotone multigrid methods for some non-smooth optimization problems*. in R. Glowinski, J. Piaux, Z. Shi, O. Widlund (eds.), ‘Domain Decomposition Methods in Sciences and Engineering’, Wiley (1997), pp. 177–191.
- [57] —, *Monotone Multigrid Methods for Nonlinear Variational Problems*, Teubner, 1997.
- [58] R. KORNHUBER AND R. ROITZSCH, *On adaptive grid refinement in the presence of internal or boundary layers*, IMPACT Comput. Sci. Engrg., 2 (1990), pp. 40–72.
- [59] I. KOSSACZKÝ, *A recursive approach to local mesh refinement in two and three dimensions*, J. Comput. Appl. Math., 55 (1994), pp. 275–288.
- [60] O.A. LADYZENSKAJA, V. SOLONNIKOV AND N. URAL’CEVA, *Linear and Quasilinear Equations of Parabolic Type*, vol. TMM 23, AMS, Providence, 1968.
- [61] E.J. LAITINEN, *On the simulation and control of the continuous casting process*. Report 43, Math. Dept. Univ. Jyväskylä, 1989.

- [62] J. S. LANGER, *Instabilities and pattern formation in crystal growth*, Rev. Modern Phys., 52 (1980), pp. 1–28.
- [63] R. DAUTREY AND J.-L. LIONS, *Analyse Mathématique et Calcul Numérique pour les Sciences et les Techniques*, Masson, (1988)
- [64] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications I*, Springer, 1972.
- [65] J. L. LIONS AND G. STAMPACCHIA, *Variational inequalities*, Comm. on pure and appl. math., vol.XX (1972), pp. 493-519.
- [66] S. LOUHENKILPI, E. LAITINEN AND R. NIEMINEN, *Real-time simulation of heat transfer in continuous casting*, Metallurgical Trans. B, 24B (1993), pp. 685-693.
- [67] J. M. MAUBACH, *Local bisection refinement for n -simplicial grids generated by reflection*, SIAM J. Sci. Comput., 16 (1995), pp. 210–227.
- [68] W. MITCHELL, *A comparison of adaptive refinement techniques for elliptic problems*, ACM Trans. Math.Softw., 15 (1989), pp. 326–347.
- [69] K.W. MORTON AND E. SÜLI, *Evolution Galerkin methods and their supra-convergence*, Numer. Math. 64 (1995), pp. 1097-1122.
- [70] W. MULDER, S. OSHER, J.A. SETHIAN, *Computing Interface Motion in Compressible Gas Dynamics*, Journal of Computational Physics, 100(2) (1992), pp. 209-228
- [71] J. C. NEDELEC, *Curved finite element methods for the solution of integral singular equations on surfaces in \mathbf{R}^3* , Comput. Meth. Appl. Mech. Engrg., 8 (1976), pp. 61–80.
- [72] J. NITSCHKE, *Ein Kriterium für die Quasi-Optimalität des Ritzschen Verfahrens*, Numer. Math., 11 (1968), pp. 346–348.
- [73] R.H. NOCHETTO, *Error estimates for multidimensional singular parabolic problems*, Japan J. Indust. Appl. Math., 4 (1987), pp. 111–138
- [74] R. H. NOCHETTO, M. PAOLINI, AND C. VERDI, *An adaptive finite element method for two-phase stefan problems in two space dimensions. Part II: Implementation and numerical experiments*, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 1207–1244.
- [75] ———, *Double obstacle formulation with variable relaxation parameter for smooth geometric front evolutions: Asymptotic interface error estimates*, Asymptotic Anal, 10 (1995), pp. 173–198.
- [76] ———, *A dynamic mesh algorithm for curvature dependent evolving interfaces*, J. Comput. Phys., 123 (1996), pp. 296–310.
- [77] R.H. NOCHETTO, G. SAVARÉ AND C. VERDI, *Error control of nonlinear evolution equations*, C. R. Acad. Sci. Paris Sér. I Math 326 (1998), pp. 1437–1442.
- [78] R.H. NOCHETTO, A. SCHMIDT AND C. VERDI, *A posteriori error estimation and adaptivity for degenerate parabolic problems*, Math. Comp. 69 (2000) pp. 1-24.

- [79] ———, *Adaptive solution of parabolic free boundary problems with error control*. in I. Athanasopoulos, G. Makrakis, and J.F. Rodrigues (Eds.): *Free Boundary Problems: Theory and Applications*, Chapman and Hall / CRC Research Notes in Mathematics 409 (1999), 344-355.
- [80] ———, *Adapting meshes and time-steps for phase change problems*, *Rend. Mat. Acc. Lincei* s. 9, v. 8 (1997), 273-292
- [81] ———, *Adaptive solution of phase change problems over unstructured tetrahedral meshes*, in M. Bern, J.E. Flaherty, and M. Luskin (Eds.): *Grid Generation and Adaptive Algorithms*, IMA VMA 113 (1999), 163-181.
- [82] ———, *Error control for phase change problems*, in P. Argoul, F. Fremond, and Q.S. Nguyen (Eds.): *IUTAM Symposium on Variations of Domains and Free-Boundary Problems in Solid Mechanics*, *Solid Mechanics and its Applications* Vol. 66, Kluwer (1999), 53-61
- [83] R. H. NOCHETTO AND C. VERDI, *Convergence past singularities for a fully discrete approximation of curvature driven interfaces*. Quaderno 9/1994, Milano, 1994.
- [84] R.H. NOCHETTO, *Error estimates for the two-phase Stefan problems in several space variables II: Nonlinear flux conditions*, *Calcolo* 22 (1985), pp. 501-534.
- [85] S. OSHER AND J.A. SETHIAN, *Fronts Propogating with Curvature-Dependent Speed: Algorithms Based on Hamilton-Jacobi Formulations*, *Journal of Computational Physics*, 79 (1988), pp. 12-49
- [86] M. PAOLINI AND C. VERDI, *Asymptotic and numerical analyses of the mean curvature flow with a space-dependent relaxation parameter*, *Asymptotic Analysis*, 5 (1992), pp. 553–574.
- [87] O. PIRONNEAU, *On the transport-diffusion algorithm and its application to the Navier-Stokes equations*, *Numer. Math.* 38 (1982), pp. 309-332.
- [88] J.F. RODRIGUES, *Obstacle Problems in Mathematical Physics*, North-Holland, Amsterdam, 1987.
- [89] J.F. RODRIGUES AND F. YI, *On a two-phase continuous casting Stefan problem with nonlinear flux*, *Euro. J. Appl. Math.* 1 (1990), pp. 259-278.
- [90] D. ROSENTHAL, *Mathematical Theory of Heat Distribution During Welding and Cutting*, *Welding J.*, 20 (1941), pp. 220-234
- [91] J. RULLA, *Weak solutions to Stefan problems with prescribed convection*, *SIAM J. Math. Anal.* 18 (1987), pp. 1784-1800.
- [92] Y. SAAD, *Iterative Methods for Sparse Linear Systems*, PWS, 1996.
- [93] A. SCHMIDT, *Computation of threedimensional dendrites with finite elements*, *J. Comput. Phys.*, 125 (1996), pp. 293–312.
- [94] A. SCHMIDT AND K. G. SIEBERT, *Design of adaptive finite element software: The finite element toolbox ALBERTA*, Springer LNCSE Series 42, to appear.

- [95] ———, *Numerical Aspects of Parabolic Free Boundary Problems: Adaptive Finite Element Methods*, Lecture Notes, Summer School, Jyväskylä, 1996.
- [96] W. SCHULZ, V. KOSTRYKIN, ET ALL, *A Free Boundary Problem Related to Laser Beam Fusion Cutting:ODE Approximation*, Int. J. Heat Mass Transfer, vol.40, 12 (1997), pp. 2913-2928
- [97] P.S. SHENG, V.S. JOSHI, *Analysis of Heat-affected Zone Formation for Laser Cutting of Stainless Steel*, J. Material Processing Technology, 53 (1995), pp. 879-892
- [98] K. G. SIEBERT, *A posteriori error estimator for anisotropic refinement*, Numer. Math., 73 (1996), pp. 373–398.
- [99] J. STOER AND R. BULIRSCH, *Introduction to Numerical Analysis*, Texts in Applied Mathematics 12, Springer, 1993.
- [100] R. VERFÜRTH, *A posteriori error estimates for nonlinear problems: Finite element discretization of elliptic equations*, Math. Comp., 62 (1994), pp. 445–475.
- [101] ———, *A posteriori error estimation and adaptive mesh-refinement techniques*, J. Comp. Appl. Math., 50 (1994), pp. 67–83.
- [102] J. WLOKA, *Partial differential equations*, Cambridge University Press, 1987.
- [103] O. C. ZIENKIEWICZ, D. W. KELLY, J. GAGO, AND I. BABUŠKA, *Hierarchical finite element approaches, error estimates and adaptive refinement*, in The mathematics of finite elements and applications IV, J. Whiteman, ed., Academic Press, 1982, pp. 313–346.

Acknowledgments.

We gratefully acknowledge the support of this course by EU Tempus grant IMG-04-D1006.

The main part of these lecture notes is based on the references [94] and [95] as well as lectures given at the Universities of Freiburg and Bremen. Section 15 is based on [82] and Section 16 is based on [18].