Numerical methods for the solution of a three-dimensional anisotropic inverse heat conduction problem

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Diplom-Mathematiker
Marcus Soemers
aus Mönchengladbach

Berichter: Universitätsprofessor Dr. Arnold Reusken
Universitätsprofessor Dr.-Ing. Wolfgang Marquardt

Tag der mündlichen Prüfung: 04. September 2008

Für meine Großeltern
Martha und Wilhelm Bläsius

Danksagung

Die vorliegende Dissertation entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Geometrie und Praktische Mathematik (IGPM) der RWTH Aachen. Gefördert wurde die Arbeit von der deutschen Forschungsgemeinschaft DFG im Rahmen des Graduiertenkollegs GK 775 „Hierarchie und Symmetrie in mathematischen Modellen“.

Meinem Doktorvater, Herrn Prof. Dr. Arnold Reusken, der diese Arbeit ermöglichte und wissenschaftlich betreute, danke ich sehr herzlich. Er hat mein Interesse für die numerische Mathematik geweckt, meine Forschungsarbeit durch stets konstruktive Anregungen begleitet und durch seine kontinuierliche Diskussionsbereitschaft in allen Phasen unterstützt.

Ganz herzlich danke ich auch Herrn Prof. Dr.-Ing. Wolfgang Marquardt für die Übernahme des Korreferats und die Möglichkeit zur interdisziplinären Zusammenarbeit, nicht zuletzt im Rahmen des Sonderforschungsbereichs SFB 540. Die Arbeit an einem durch eine ingenieurtechnische Anwendung motivierten Forschungsthema war stets ein besonderer Ansporn für mich.

Ein besonderer Dank richtet sich an meine Kollegen und Freunde der legendären C6-Runde, an Adel Mhamdi, Sven Groß, Maka Karalashvili, Yi Heng und Faruk Al-Sibai für die einmalige Zusammenarbeit, die vielen spannenden Diskussionen im fachlichen Bereich und das alltägliche Leben betreffend, sowie die gemeinsamen Unternehmungen.

Allen Kollegen und Mitarbeitern des IGPM spreche ich meinen Dank für die äußerst angenehme Arbeitsatmosphäre, die mir jederzeit entgegengebrachte Hilfsbereitschaft und die gemeinsame Freizeit außerhalb des Instituts aus. Besonders danken möchte ich Jörg Grande für die langjährige Bürogemeinschaft und Volker Reichelt, der mir die HiWi-Verwaltung anvertraut hat.

Meine tiefe Dankbarkeit gilt meinen Großeltern Martha und Wilhelm Bläsius, ohne die das Entstehen der vorliegenden Arbeit nicht möglich gewesen wäre.

Aachen, im September 2008

Marcus Soemers

Contents

Chapter 1

Introduction

Most of the research presented in this thesis has been done in the graduate program "Hierarchy and symmetry in mathematical models"[1] funded by the German Research Foundation (DFG). The research topic is closely related to subjects studied in the collaborative research center (SFB) 540 "Model-based experimental analysis of kinetic phenomena in fluid multi-phase reactive systems"[2]. Fluid multi-phase systems are very common in process engineering, e.g. in reaction and separation processes. The methodology of the model-based experimental analysis, developed in SFB 540, combines high resolution measurement techniques, numerical simulation, and methods for model-supported planning and evaluation of experiments in a systematic way. Part of this methodology is the formulation and solution of inverse problems. The efficient solution of these typically very complex inverse problems allows the determination of interesting (for example, kinetic) quantities, which cannot be determined directly from the available experimental measurement data. One goal in SFB 540 is the successful modelling of selected (kinetic) phenomena on the basis of these estimated quantities leading to an improvement of existing and the development of new production processes. In the following we describe one particular physical process that is considered in this thesis and give an overview of the main research topics.

1.1 Problem motivation and main research topics

The modelling of kinetic phenomena in fluid multi-phase systems leads to problems of model structure and parameter identification, which belong to the class of inverse problems [Mar05]. Challenging inverse problems appear in the study of heat and mass transfer mechanisms in falling films, which are of special interest due to their technical relevance. These film systems are frequently used in the chemical industry and the production of food, e.g. in film evaporators used in the fruit juice and milk production. Many investigations have already been performed to analyze the heat and mass transfer in falling films [HR94, HME$^+$03, LMFA99, LM91]. It has been observed that in realistic wavy films both heat and mass transfer are significantly higher than predicted by simple one-dimensional models. In these studies semi-empirical correlations of the heat and mass transfer enhancement due to the waviness of the film are presented. However, the mechanisms of the transport processes and the flow characteristics of wavy films are still not well understood.

The research in this thesis is motivated by the question how the heat transfer in falling films is influenced by the wave characteristics. The approach used for answering this question is based on high resolution temperature measurements. These measurements were performed at the Chair of Heat and Mass Transfer (RWTH Aachen University) in a falling film apparatus specifically designed for this purpose (see Fig. 1.1).

[1] http://www.math.rwth-aachen.de/homes/Kolleg/
[2] http://www.sfb540.rwth-aachen.de/

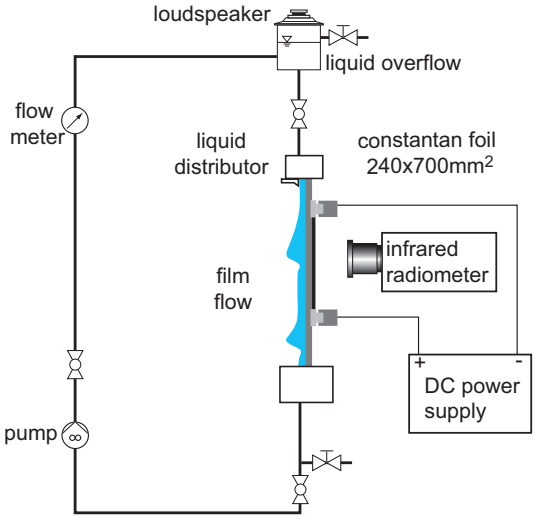


Figure 1.1: Schematic falling film experiment; adopted from [GSM+05]

In the falling film experiment a laminar wavy film travels along a thin foil that is heated electrically from the back side. The identification problem of estimating the heat flux on the film surface from the measurement data on the foil back side is coupled with the fluid dynamics of the falling film. Due to this coupling the identification problem is of huge complexity. Thus, a simplified but still very challenging and relevant problem is considered, in which the heat transfer is decoupled from the fluid dynamics. We investigate the unsteady heat transfer from the heating foil to the falling film (see Fig. 1.2 in which the z-coordinate denotes the flow direction of the laminar wavy falling film).

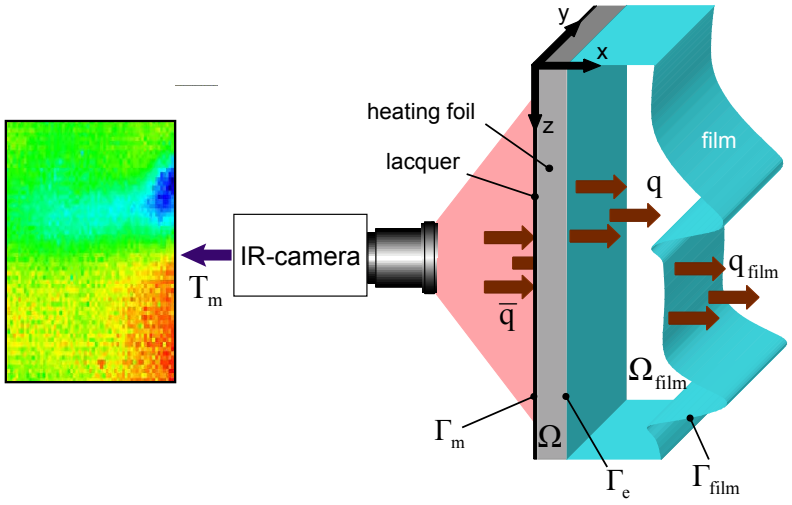


Figure 1.2: Measurement setup in the falling film experiment

More precisely, we estimate the space- and time-dependent heat flux q on the inaccessible film side Γ_e of the foil using infrared (IR) temperature measurements on the foil back side Γ_m, which are influenced by the transport phenomena in the falling film and the wavy film surface. This is a model-free estimation problem in the sense that no a-priori knowledge

of the heat transfer mechanism is used. In subsequent steps the estimated heat flux can be correlated with other characteristic quantities in the falling film such as the mean film temperature and the flow regime. This incremental approach of modelling, in which a model structure (e.g. involving the estimated heat flux) is refined gradually by specifying submodels, is investigated in SFB 540 and at the Chair of Process Systems Engineering (RWTH Aachen University). On the basis of this methodology an improved physical understanding of the related kinetic phenomena can be obtained and predictive mathematical models can be derived (see e.g. [KGM+08]). The results presented in this thesis are also meant to contribute to a better understanding of numerical solution methods for multi-dimensional inverse heat conduction problems. This is essential as a basis for future investigations of more complex inverse transport problems such as the heat transfer through the falling film, mass transfer from the film to the gas phase, or reaction processes inside the film.

There is a vast amount of studies devoted to inverse heat transfer problems. Most literature on numerical solution methods is restricted to one or two space dimensions. For three-dimensional problems only few publications are available (compare Chapter 2). In some of these publications *simulated* data with *few* pointwise measurement locations are used. Others are based on *realistic* experimental measurement data. However, the presented solution methods, e.g. using model reduction techniques to reduce the problem complexity, become computationally intractable when a fine space discretization is applied. In this thesis a *three*-dimensional *inverse* heat conduction problem (IHCP) for the reconstruction of a boundary heat flux with high resolution in space and time is solved numerically. As a key ingredient the realistic high resolution spatio-temporal measurements are used. For the solution of the IHCP efficient and robust numerical methods are needed. The combination of methods investigated in this thesis and outlined below is new.

The three-dimensional IHCP in falling films is formulated as an *optimal control* problem. The optimization is performed in cooperation with the Chair of Process Systems Engineering (RWTH Aachen University). The idea in this approach is to consider the unknown heat flux on the estimation boundary as a decision variable to minimize a specified defect functional on the measurement boundary of the heat conductor. In the defect functional the computed temperature for an approximation of the searched for boundary heat flux is needed, which is obtained from the solution of a direct heat conduction problem. We introduce the space-time domains $S_\omega := \Gamma_\omega \times (t_0, t_f]$, $\omega \in \{m, e\}$, where Γ_m and Γ_e denote the measurement and estimation boundary of the heat conductor, respectively. The variational formulation is as follows: find the optimal control $q \in L_2(S_e)$ such that

$$J(q) = \frac{1}{2}\|T(q)|_{S_m} - T_m\|^2_{L_2(S_m)} \to \min .$$

The norm in this objective functional is defined by

$$\|\cdot\|^2_{L_2(S_m)} := \int_{t_0}^{t_f} \int_{\Gamma_m} (\cdot)^2 d\mathbf{x}\, dt,$$

and $T(q)$ denotes the solution of the *direct* heat conduction problem

$$\frac{\partial T}{\partial t}(\mathbf{x}, t) = a\Delta T(\mathbf{x}, t), \quad (\mathbf{x}, t) \in \Omega \times (t_0, t_f], \tag{1.1}$$

$$T(\mathbf{x}, t_0) = T_0(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{1.2}$$

$$-\lambda \frac{\partial T}{\partial \mathbf{n}}(\mathbf{x}, t) = \bar{q}(\mathbf{x}, t), \quad (\mathbf{x}, t) \in S_m, \tag{1.3}$$

$$-\lambda \frac{\partial T}{\partial \mathbf{n}}(\mathbf{x}, t) = q(\mathbf{x}, t), \quad (\mathbf{x}, t) \in S_e, \tag{1.4}$$

with known initial and boundary data T_0 and $\bar{q}$, and strictly positive material constants a, λ. For an alternative formulation we use an abstract operator notation. Let $\mathcal{A}$ denote the operator, which maps a time- and space-dependent boundary heat flux $q \in L_2(S_e)$ on the

estimation boundary of the heat conductor to the direct problem solution $T(q)$ of (1.1)-(1.4) restricted to the measurement boundary, i.e.

$$\begin{aligned} \mathcal{A} : L_2(S_e) &\to L_2(S_m) \\ q &\mapsto T(q)|_{S_m}. \end{aligned} \tag{1.5}$$

With this operator we can reformulate the IHCP as follows: find the heat flux $q \in L_2(S_e)$ such that

$$J(q) = \frac{1}{2}\|\mathcal{A}q - T_m\|^2_{L_2(S_m)} \to \min, \tag{1.6}$$

where the temperature data $T_m \in L_2(S_m)$ are known. We will see that the operator defined by (1.5) is affine and bounded. However, for realistic (multi-dimensional) geometries and/or time- and space-dependent coefficients in the heat conduction equation (1.1) we are not able to give an *explicit* representation of this operator. For the solution of the optimal control problem we apply the conjugate gradient type method CGNE, in which the approximation quality depends on how many iterative steps are performed. Since the optimal control problem based on realistic measurement data is severely *ill-posed*, which means that a solution, if it exists, may be arbitrarily far from the exact solution, we use iterative *regularization*. Although a suitable problem discretization results in a kind of regularization, too, it is not sufficient for our problem. For the choice of the regularization parameter, i.e. the termination index of the CGNE iteration, we use the heuristic L-curve method which does not require any information about the noise level and turned out to be satisfactory for our problem class.

In the CGNE algorithm well-posed direct heat conduction problems of the form (1.1)-(1.4) have to be solved. However, the very different length scales of the heat conductor, i.e. the *thin* heating foil, result in an anisotropy effect. For the numerical treatment we use an implicit time integration method and a special space discretization based on (linear) *anisotropic* tetrahedral finite elements. In order to obtain a bound for the discretization error we use the Céa-lemma and error bounds for standard nodal Lagrangian interpolation. A *maximum* angle condition is formulated which is the essential condition to deduce suitable bounds for the interpolation error. Different approaches for the derivation of these error bounds are considered. We give a uniform presentation of results for anisotropic (triangular and) tetrahedral finite elements that can be found at different places in the literature. From these results we can conclude that anisotropic tetrahedral elements which satisfy a maximum angle condition are suitable for the space discretization of the direct heat conduction problems. Important phenomena concerning anisotropic finite elements are illustrated by means of systematic numerical experiments.

In each step of the implicit time integration method linear anisotropic reaction-diffusion problems have to be solved, which contain two important parameters: there is a parameter which describes the anisotropy of the domain and one that measures the size of the reaction term relative to that of the diffusion term. After discretization we have a third parameter, namely the mesh width. For the efficient solution of the corresponding discrete problems the convergence of a *multigrid* method with a special (symmetric line Gauss-Seidel) smoother is analyzed. In the analysis there is a need for other (better) interpolation operators than the standard Lagrangian interpolation operator. Thus, a modified Scott-Zhang interpolation operator is studied. Both, for the W- and the V-cycle method we derive contraction number bounds smaller than one uniform with respect to all three parameters. For the multigrid W-cycle the convergence is analyzed in the framework of the approximation and smoothing property. Most literature on convergence analyses of multigrid methods for anisotropic *pure* diffusion problems is based on a special (i.e. the standard Lagrangian) interpolation operator and (thus) is restricted to *two*-dimensional problems. In this thesis we treat the *three*-dimensional case and consider an additional *reaction* term.

The thesis is organized as follows. In order to point out some essential properties of IHCPs in more detail, in the next section we consider a *one*-dimensional inverse heat conduction

model problem. For this problem an explicit representation of the corresponding operator $\mathcal{A}$ in (1.5) is known. We finish the introduction with some notes on the CGNE approach used for solving the optimal control problem. A clear advantage of this iterative approach is that we only need *applications* of the operator $\mathcal{A}$ to given elements from the space $L_2(S_e)$, which requires the solution of well-posed direct heat conduction problems. The three-dimensional IHCP in falling films is formulated in Chapter 2. The experimental falling film setup and the optimization based solution algorithm are described. Different strategies for the choice of the CGNE stopping index (i.e. the regularization parameter) are considered. The basic theory of numerical solution methods for direct heat conduction problems using isotropic linear finite elements for the space discretization is discussed in Chapter 3. Standard techniques for the derivation of discretization error bounds are recalled. In Chapter 4 we investigate anisotropic finite elements. A maximum angle condition is introduced and more involved techniques that are needed for proving discretization error bounds for linear anisotropic (triangular and) tetrahedral finite elements are explained. A convergence analysis of a special multigrid method (using a robust line smoother) for the solution of the resulting discrete problems is presented in Chapter 5. The combination of these numerical methods is applied to realistic high resolution measurement data obtained from the falling film experiment in Chapter 6. A *multi-level* approach for the optimization procedure is considered, in which the entire optimization process is started on a relatively coarse level (space grid) in order to find a good initial approximation of the estimation quantity on a next finer level. Finally, some conclusions are given in Chapter 7.

1.2 An inverse heat conduction model problem

We want to determine an unknown heat source at the end of an insulated semi-infinite bar on the non-negative x-axis, given temperature measurements at position $x = 1$ of the bar. The mathematical formulation of this model problem is: find $f(t) := u(0,t)$ from

$$\begin{aligned} u_t &= u_{xx}, \quad x \in (0,\infty), \quad t \in (0,t_f], \\ u(x,0) &= 0, \quad x \geq 0, \\ u_x(0,t) &= 0, \quad t \in [0,t_f], \\ u(1,t) &= g(t), \quad t \in [0,t_f], \end{aligned}$$

where temperature data $g(t) \in L_2(0,t_f)$ up to the final time t_f are given. For convenience, the space domain is unbounded and we want to determine temperature values (Dirichlet data) instead of a boundary heat flux (Neumann data). If a solution $f(t) \in L_2(0,t_f)$ of the model problem exists, it can be determined from the first-kind Volterra integral equation

$$\int_0^t k(t,s) f(s)\, ds = g(t), \quad t \in [0,t_f], \tag{1.7}$$

where the kernel k is of *convolution* type, which means $k(t,s) = \kappa(t-s)$ for some continuous function κ. One can show that the kernel is given by

$$\kappa(t) = \frac{1}{2\sqrt{\pi}t^{3/2}} \exp\left(-\frac{1}{4t}\right), \quad t \in (0,t_f]. \tag{1.8}$$

In practice the measured signal $g(t)$ in equation (1.7) consists of the true signal and an unknown noise and the kernel function represents the so-called *instrument response function*. If we define the linear integral operator $\mathcal{K} : L_2(0,t_f) \to L_2(0,t_f)$ by

$$(\mathcal{K}f)(t) := \int_0^t \kappa(t-s) f(s)\, ds, \quad t \in [0,t_f], \tag{1.9}$$

then an equivalent formulation of the described model problem is: find the unknown heat source f from the integral equation

$$\mathcal{K}f = g. \tag{1.10}$$

Throughout this section, we assume that a solution of (1.10) exists for exact data g. Our goal is to show that the inverse problem is ill-posed due to *two* different effects. The searched for quantity does not depend continuously on the given data *and* it cannot be reconstructed at final time $t = t_f$ (compare [Eld83]). We emphasize that both effects are inherent to the problem and do not depend on the chosen solution (regularization) strategy.

Analysis of the continuous problem

The fact that small perturbations in the given data may lead to enormous deviations in the solution of the considered inverse heat conduction model problem can be seen as follows. The problem is to solve an integral equation of the first kind (see appendix). For this type of equation, the corresponding integral operator is *compact* in reasonable topologies and under weak conditions on the kernel [EHN96, Kir96]. It can be shown that equations of the form $\mathcal{K}f = g$, where $\mathcal{K}$ is a compact linear operator (with an infinite-dimensional range) are *always* ill-posed in the above described sense. This can be illustrated using a singular system for the operator $\mathcal{K}$.

In our case the integral operator is compact since the kernel in (1.8) is weakly singular [EHN96]. The kernel function $\kappa(t)$ vanishes exponentially for $t \downarrow 0$ and thus, for $f \in L_2(0, t_f)$, the integral in (1.9) exists and is finite. We introduce the *adjoint* operator $\mathcal{K}^* : L_2(0, t_f) \to L_2(0, t_f)$ of $\mathcal{K}$ w.r.t. the standard scalar product $(\cdot, \cdot)$ in $L_2(0, t_f)$

$$(\mathcal{K}^* g)(s) := \int_s^{t_f} \kappa(t - s) g(t)\, dt, \quad s \in [0, t_f]. \tag{1.11}$$

Let $\sigma_1 \geq \sigma_2 \geq \ldots > 0$ be the ordered sequence of positive singular values of $\mathcal{K}$ (i.e. the square roots of the non-zero eigenvalues of $\mathcal{K}^*\mathcal{K}$) counted with multiplicity. There exist orthonormal systems $\{v_n(t)\}_{n\in\mathbb{N}}$ and $\{u_n(t)\}_{n\in\mathbb{N}}$ with the properties:

$$\mathcal{K}v_n = \sigma_n u_n \quad \text{and} \quad \mathcal{K}^* u_n = \sigma_n v_n, \quad n \in \mathbb{N}.$$

The system $(\sigma_n; v_n, u_n)_{n\in\mathbb{N}}$ is called a singular system for $\mathcal{K}$. Now we can give the well-known *Picard criterion*: for the compact linear operator $\mathcal{K}$ with singular system denoted as before the problem (1.10) is solvable, i.e. $g \in \mathcal{R}(\mathcal{K})$, if and only if

$$g \in \overline{\mathcal{R}(\mathcal{K})} \quad \text{and} \quad \sum_{n=1}^{\infty} \tfrac{1}{\sigma_n^2} |(g, u_n)|^2 < \infty, \tag{1.12}$$

where $\mathcal{R}(\mathcal{K})$ denotes the range of the operator $\mathcal{K}$. If these conditions are fulfilled then the solution of (1.10) with minimum L_2-norm is given by

$$f = \sum_{n=1}^{\infty} \frac{1}{\sigma_n} (g, u_n) v_n. \tag{1.13}$$

Note that (1.12) requires a fast decay of the Fourier coefficients (g, u_n) relative to the singular values σ_n. Thus, the Picard criterion is quite restrictive since in practice the data g contain measurement errors. Error components that correspond to small singular values are amplified by the relatively large factors $1/\sigma_n$. When the range of $\mathcal{K}$ is finite-dimensional, i.e. $\dim(\mathcal{R}(\mathcal{K})) < \infty$, then there are only finitely many singular values. In this case the amplification factors are bounded even though they might be unacceptably large. However, the kernel in (1.8) is *non-degenerate* and thus for the corresponding integral operator in (1.9) we have $\dim(\mathcal{R}(\mathcal{K})) = \infty$ with $\lim_{n\to\infty} \sigma_n = 0$ (see [EHN96] for more details). For this reason data errors of a fixed size can be amplified *arbitrarily* much. The faster the singular values decay to zero, the more severe the ill-posedness becomes. This poor conditioning does not depend on special properties of the kernel function. The *direct* problem is to compute g from the application of $\mathcal{K}$ to a given element f. In our case this means *integration* of the given data, which is a *smoothing* process in the sense that highly oscillatory errors in f are damped and have only little effect on g. On the other hand, this smoothing property of $\mathcal{K}$

causes high frequency errors in g to create large oscillations in the solution f of the *inverse* problem.

We turn to the second effect that contributes to the ill-posedness of problem (1.10) and concerns the structure of the operator $\mathcal{K}$ in (1.9). This operator is *causal*, which means that for any $\bar{t} > 0$ the values of $g(t) = (\mathcal{K}f)(t)$ on the interval $[0, \bar{t}]$ are determined by the data $f(t)$ on the same interval. The kernel function is of convolution type and we have $\kappa(t) \in C^\infty((0, t_f))$ with

$$\lim_{t \downarrow 0} \kappa^{(\nu)}(t) = 0, \quad \nu = 0, 1, \ldots \tag{1.14}$$

For $n > 0$ we define

$$f_n(t) := \begin{cases} 0 & \text{if } t \le t_f - \dfrac{1}{n}, \\ n(t - (t_f - \dfrac{1}{n})) & \text{if } t > t_f - \dfrac{1}{n}. \end{cases} \tag{1.15}$$

Due to (1.14), for fixed $\varepsilon > 0$ (arbitrarily small) we can find n_0 such that $|\kappa(t)| \le \varepsilon$ for all $t \in [0, \frac{1}{n_0}]$. For $n \ge n_0$ we get

$$\begin{aligned} \|\mathcal{K}f_n\|^2_{L_2(0,t_f)} &= \int_0^{t_f} \left(\int_0^t \kappa(t-s) f_n(s)\, ds \right)^2 dt \\ &\le \varepsilon^2 \int_{t_f - \frac{1}{n}}^{t_f} \left(\int_{t_f - \frac{1}{n}}^t f_n(s)\, ds \right)^2 dt \\ &\le \frac{\varepsilon^2}{n} \left(\int_{t_f - \frac{1}{n}}^{t_f} f_n(s)\, ds \right)^2 \le (\frac{\varepsilon}{n})^2 \|f_n\|^2_{L_2(0,t_f)}. \end{aligned}$$

Thus, we can conclude

$$\|\mathcal{K}f_n\|_{L_2(0,t_f)} \le \frac{\varepsilon}{n} \|f_n\|_{L_2(0,t_f)} \le \frac{\varepsilon}{n\sqrt{n}} \to 0, \quad n \to \infty. \tag{1.16}$$

Now we take $c > 0$ (arbitrarily large), $\delta > 0$ (arbitrarily small) and n sufficiently large such that $\|\mathcal{K}f_n\|_{L_2(0,t_f)} \le \delta/c$. We consider (1.10) and define $g_\delta := g + c\mathcal{K}f_n$. Then we have $\|g_\delta - g\|_{L_2(0,t_f)} \le \delta$ while for the inverse problem solution, which is given by $f_\delta = f + cf_n$, we obtain

$$f_\delta(t_f) - f(t_f) = cf_n(t_f) = c.$$

In other words, arbitrarily small perturbations in the data g may cause arbitrarily large errors in the solution of the inverse problem at $t = t_f$, which means that f cannot be reconstructed at final time. However, note that $\|f_\delta - f\|_{L_2(0,t_f)} \le c/\sqrt{n} \to 0$, $n \to \infty$. For a discrete version of the integral equation (1.10) we illustrate the two above described effects, which are independent of each other, in the next paragraph.

A discrete version of the model problem

Let $N \ge 1$ and $h = t_f/N$. For the discretization of (1.10) we use the midpoint integration rule, which leads to the approximate solution

$$h \sum_{j=0}^{i-1} k(t_i, t_{j+1/2}) f_{j+1/2} = g(t_i), \quad i = 1, 2, \ldots, N, \tag{1.17}$$

where $t_j = jh$ and $t_{j+1/2} = t_j + h/2$. For a survey of regularization methods for Volterra equations of the first kind (including appropriate discretizations) we refer to [Lam00]. The linear system of equations (1.17) can be written in matrix-vector notation with a square lower triangular matrix $\mathbf{K}$ that has certain structure properties. The ill-posedness of (1.10) is reflected in the ill-conditioning of this matrix. For a quite similar inverse model problem

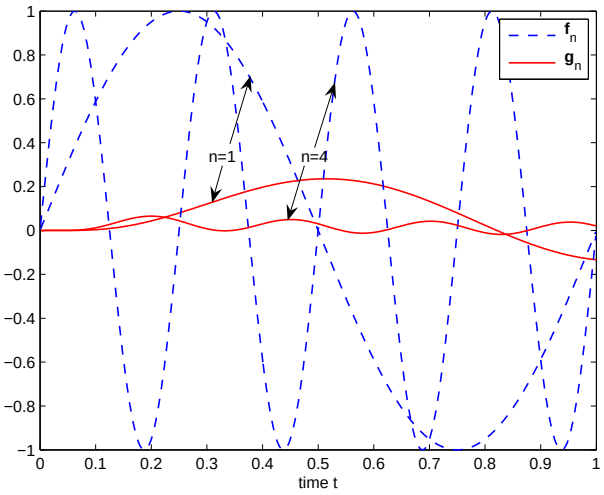


Figure 1.3: Smoothing property of the direct problem

in [Eld83] the structure and the pattern of the singular values of $\mathbf{K}$ are discussed. In the following numerical experiments we take $t_f = 1$ and $N = 500$.

First, we illustrate the smoothing property of the direct problem. For this reason we consider the vector $\mathbf{f}_n$ obtained from the discretization of

$$f_n(t) = \sin(2\pi n t), \quad n \in \mathbb{N}.$$

In Fig. 1.3 the corresponding direct problem solutions $\mathbf{g}_n = \mathbf{K}\mathbf{f}_n$ are shown for $n = 1, 4$. We observe that for increasing n the amplitude in the solution $\mathbf{g}_n$ becomes smaller, i.e. high frequencies are strongly damped. In order to demonstrate the second effect, we consider the vector $\mathbf{f}_n$ obtained from the discretization of (1.15). For $n \to \infty$ the values of f_n form approximations of a peak at final time.

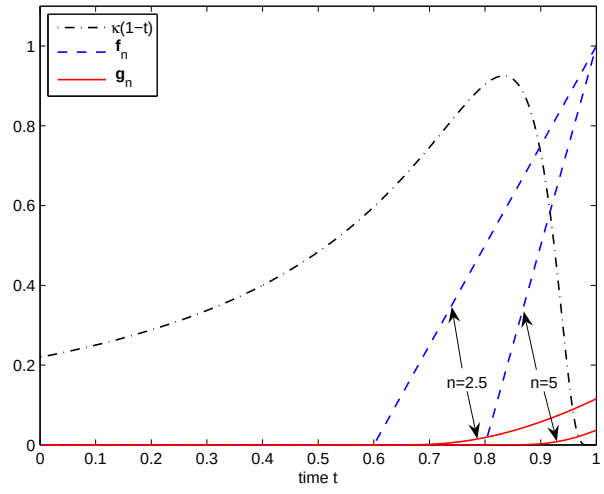


Figure 1.4: Direct problem solutions $\mathbf{g}_n$ for approximations $\mathbf{f}_n$ of a peak at final time

In Fig. 1.4 we can see the kernel function $\kappa(1-t)$ and the direct problem solutions $\mathbf{g}_n = \mathbf{K}\mathbf{f}_n$ for $n = 2.5$ and $n = 5$. The results indicate that the direct problem solutions converge to the zero-element in the L_2-sense (compare (1.16)).

1.3 The CGNE approach

For the "simple" one-dimensional inverse model problem of the previous section the corresponding integral operator is known explicitly. This model problem can be treated by using an appropriate discretization and standard regularization methods (dashed line in Fig. 1.5). However, direct methods are not appropriate and may even not be applicable in higher dimensions for the solution of the regularized discrete problem, if various computations for different a-priori chosen regularization parameters are needed. Thus, the discrete problems are often solved by an iterative technique. On the other hand, iterative methods have inherent regularization properties. In this thesis we apply the iterative CGNE method for solving the optimal control problem, in which the approximation quality depends on how many iterative steps are performed, and use a suitable stopping criterion (solid line in Fig. 1.5). A further clear advantage of this approach is that we do not need an explicit representation of the associated operator, which is not available for realistic (multi-dimensional) geometries and/or time- and space-dependent coefficients in the heat conduction equation. The CGNE algorithm merely requires the computation of well-posed direct heat conduction problems, i.e. different applications of the operator.

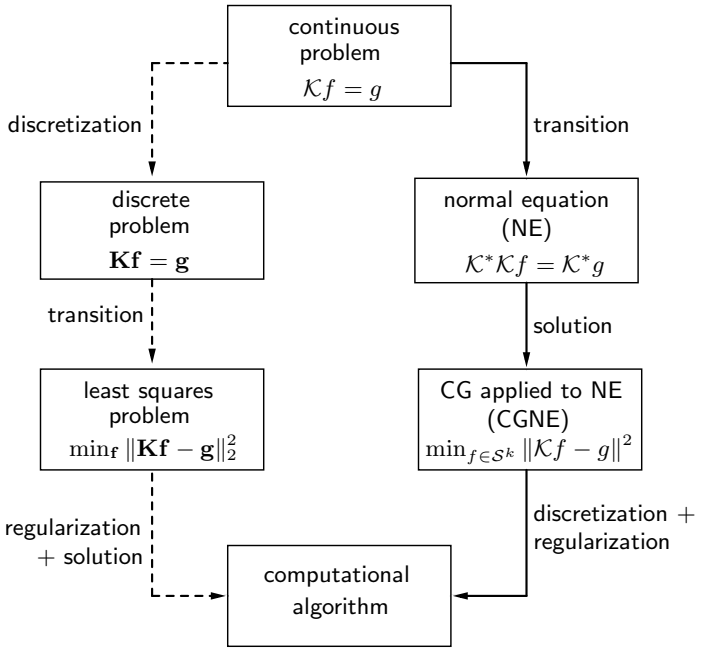


Figure 1.5: Proceeding in this thesis (solid line) and for a model problem (dashed line)

For the IHCP in falling films the linear operator in

$$\mathcal{K}f = g \tag{1.18}$$

is non-selfadjoint (compare Section 2.3.2). Throughout, we assume that a solution of (1.18) exists for exact data g. If we apply the adjoint operator $\mathcal{K}^*$ to this equation we get the normal equation (NE)

$$\mathcal{K}^*\mathcal{K}f = \mathcal{K}^*g \tag{1.19}$$

with the selfadjoint and positive semidefinite operator $\mathcal{K}^*\mathcal{K}$. Thus, the well-known conjugate gradient (CG) method can be applied to (1.19) resulting in the CGNE algorithm. In each iteration k of this algorithm the unique solution with minimum norm of the constrained minimization problem

$$\min_{f \in \mathcal{S}^k} \|\mathcal{K}f - g\|^2$$

is determined. Here, $\mathcal{S}^k$ denotes the corresponding Krylov subspace of dimension $k = 0, 1, \ldots$ (shifted by the initial guess). The computational algorithm used in this thesis is finally obtained from the discretization of the continuous CGNE approach.

Chapter 2

An anisotropic inverse heat conduction problem in 3D

In this chapter we formulate a three-dimensional inverse heat conduction problem (IHCP). The experimental falling film setup in which high resolution temperature measurements were performed is described. For the solution of the inverse problem an optimization method based on the numerical treatment of direct heat conduction problems in a thin foil is studied. The very different length scales in the geometry of this foil result in an anisotropy effect. Since the considered IHCP is severely ill-posed the inherent regularization properties of the applied iterative solution method are discussed.

There is a vast amount of studies devoted to inverse heat transfer problems (see e.g. [Ali94, BBH96] and the references therein). Most literature on numerical solution methods is restricted to one or two space dimensions (cf. e.g. [HR98]). For three-dimensional problems only few publications are available. In [Cha01, HW99, YC97] simulated data with few pointwise measurement locations are used. An efficient method for the solution of transient three-dimensional IHCPs based on experimental investigations of pool boiling can be found in [LMM05]. In that paper, which extends the studies of [BM97], a filter-based inversion method is presented. However, the achievable discretization resolution is limited by the used model reduction software. In contrast to that, in this thesis the reconstruction of boundary heat fluxes with high resolution in space and time based on high resolution spatio-temporal temperature measurements is investigated [GSM$^+$05]. Recently, the reconstruction of local boiling heat fluxes from pointwise experimental measurements applying the optimization based solution approach studied in this thesis, too, is considered in [HMG$^+$08].

2.1 Experimental setup

In order to analyze the influence of the wave characteristics of thin films on the heat transfer, measurements were performed in a falling film apparatus designed and operated at the Chair of Heat and Mass Transfer (RWTH Aachen University). In Fig. 2.1 a schematic representation of the falling film experiment and a realistic laminar wavy film surface are shown. The experimental setup consists of a fluid cycle with a loudspeaker that produces waves with a certain frequency. The laminar wavy falling film travels along a thin constantan foil that is heated electrically via a direct current (DC) power supply. On the back side of this foil high resolution temperature measurements, that are influenced by the transport phenomena in the falling film and the wavy film surface, are taken with an infrared camera CEDIP Jade II [ALR02, Als05, LAR01]. It has a HgCdTe focal-plane-array with a resolution of 320×240 *pixel* and is sensitive to radiation in the long wavelength range between 7.7 and 9.5 μm.

In Fig. 2.2 the measurement setup is illustrated. The z-coordinate denotes the flow direction

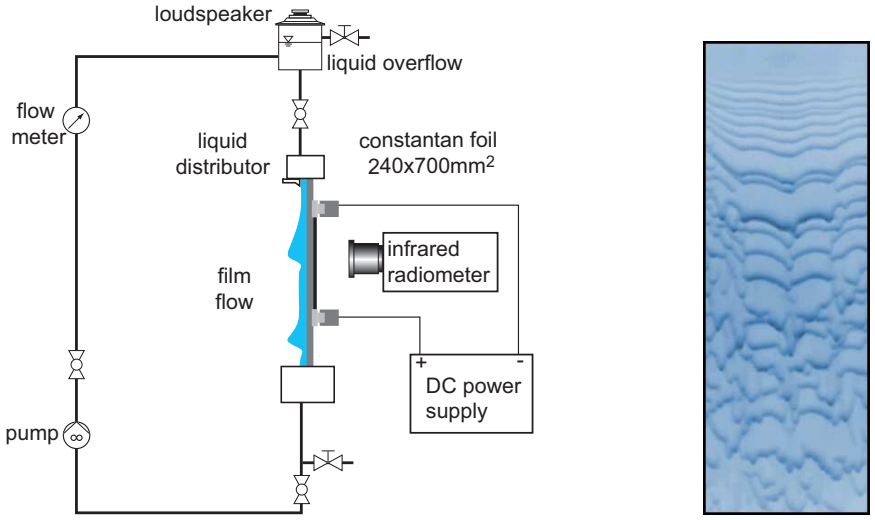


Figure 2.1: Schematic falling film experiment and surface of a laminar wavy film; adopted from [GSM+05]

of the laminar wavy falling film. The inverse heat conduction problem studied consists of determining the time- and space-dependent heat flux q on the inaccessible film side Γ_e of the foil from measurement data T_m, which are taken on the foil back side Γ_m. The remaining initial and boundary conditions are assumed to be known. There are two alternatives for modelling the electrical heating at the back side of the thin foil. We could consider a volumetric source inside the heat conductor or, as done in this thesis, we assume that we have a constant boundary heat flux $\bar{q}$ on Γ_m as shown in Fig. 2.2.

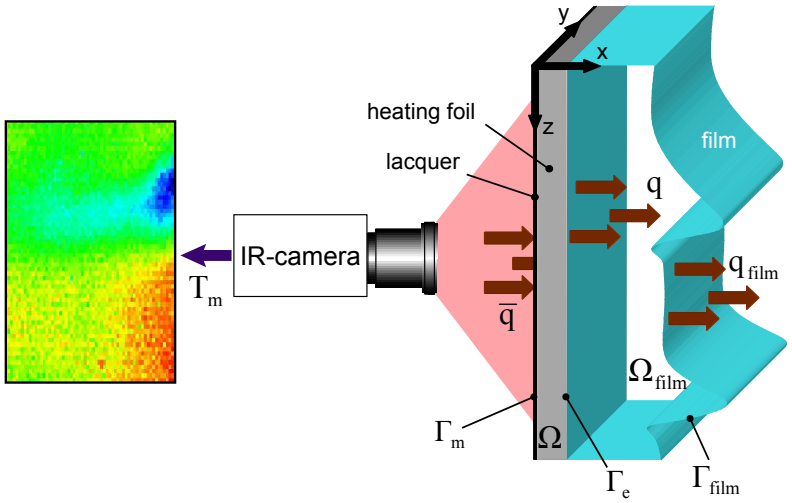


Figure 2.2: Measurement setup in the falling film experiment

In order to reach a small temperature gradient across the foil in x-direction it has a thickness of 25 μm while the measurement section on Γ_m has the dimension 19.5×39 mm^2 in the y, z-plane. In this way the measured temperature data already provide a good estimate of the temperature values on the inaccessible film side of the foil. Some measurements with

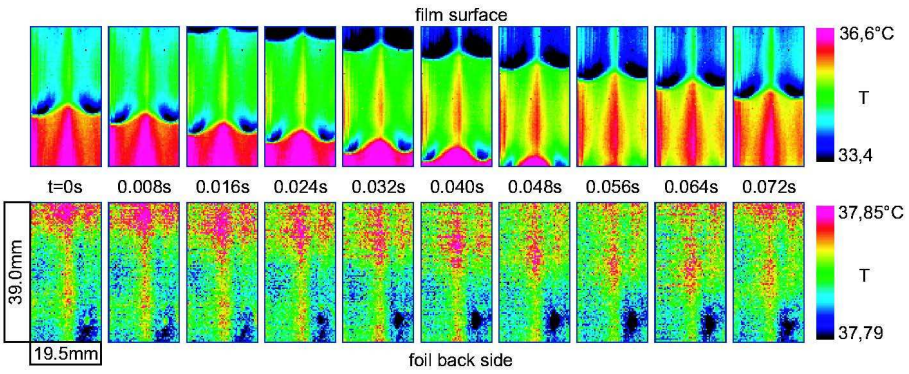


Figure 2.3: Temporal regimes of temperatures for a silicon oil; provided by F. Al-Sibai (compare [Als05])

infrared thermography are shown in Fig. 2.3. The sequence of pictures shows one wave of the falling film moving along the measurement section. There is an increasing surface temperature within the substrate region where the heat transfer is dominated by conduction. In the region of the wave front where the heat is transferred by convection, too, there is a strong cooling. The temperature fluctuations on the film surface are much higher than those on the back side of the foil, which are smaller than $0.06°C$. Thus, a clear wave structure cannot be observed at all on the foil back side, as Fig. 2.4 shows in more detail. Due to the very small temperature range in the falling film experiment the material properties of the heat conductor are assumed to be constant.

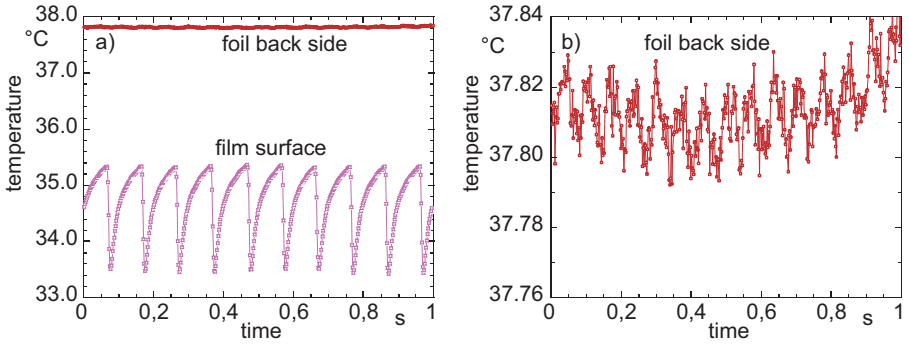


Figure 2.4: Temperature ranges on the film surface and the foil back side; provided by F. Al-Sibai (compare [Als05])

2.2 Mathematical problem formulation

Let $\Omega \subset \mathbb{R}^3$ be the parallelepiped (heat conductor) in Fig. 2.2. We consider the three-dimensional heat conduction equation

$$c\rho\frac{\partial T}{\partial t}(\mathbf{x},t) = \nabla\cdot(\lambda\nabla T)(\mathbf{x},t) + f(\mathbf{x},t), \quad (\mathbf{x},t)\in\Omega_{t_f} := \Omega\times(t_0,t_f], \tag{2.1}$$

where $T(\mathbf{x},t)$, $(\mathbf{x},t)\in\Omega_{t_f}$, denotes the temperature distribution of the heat conductor and $f\in L_2(\Omega_{t_f})$ models an inner heat source. The initial and final times are denoted by t_0 and t_f, respectively. In general the material properties which are given by the specific heat c,

the density ρ, and the heat conductivity λ are space- and time-dependent. However, in this thesis the heat conductor is assumed to be a homogeneous isotropic solid, which means that the material properties are positive constants.

The parabolic partial differential equation (2.1) together with an initial temperature distribution $T_0(\mathbf{x})$, $\mathbf{x} \in \Omega$, and Neumann boundary values $g(\mathbf{x},t)$, $(\mathbf{x},t) \in \Gamma_{t_f} := \partial\Omega \times (t_0, t_f]$, lead to the instationary initial boundary value problem

$$\begin{aligned} \frac{\partial T}{\partial t}(\mathbf{x},t) &= a\Delta T(\mathbf{x},t) + f(\mathbf{x},t), \quad (\mathbf{x},t) \in \Omega_{t_f}, && (2.2)\\ T(\mathbf{x},t_0) &= T_0(\mathbf{x}), \quad \mathbf{x} \in \Omega, && (2.3)\\ -\lambda\frac{\partial T}{\partial \mathbf{n}}(\mathbf{x},t) &= g(\mathbf{x},t), \quad (\mathbf{x},t) \in \Gamma_{t_f}, && (2.4) \end{aligned}$$

with $T_0 \in L_2(\Omega)$ and $g \in L_2(\Gamma_{t_f})$. In this representation the material properties enter the thermal diffusivity $0 < a := \lambda/c\rho < \infty$ and $\mathbf{n}$ denotes the outer normal on the boundary $\Gamma := \partial\Omega$.

Definition 2.1 (classical solution) *A solution T of problem (2.2)-(2.4) is called a classical solution if $T \in C^{2,1}(\Omega_{t_f}) \cap C^{1,0}(\overline{\Omega_{t_f}})$.*

The exponent in $C^{k,l}(\Omega_{t_f})$ denotes the kth and lth derivatives in space and time, respectively. Since we want to apply a *finite element* discretization (and a solution of problem (2.2)-(2.4) in the classical sense often does not exist, e.g. due to non-smooth initial or boundary data), we turn to the variational formulation of the introduced parabolic problem (compare [HR98, Hao98]).

For arbitrary $D \subset \Omega$ we use the notation $(\cdot,\cdot)_D$ and $\|\cdot\|_D$ for the standard scalar product and norm in $L_2(D)$. In an analogous manner we define the inner product and norm in $L_2(D_{t_*})$, where $D_{t_*} := D \times (t_0, t_*]$, $D \subset \Omega$, $t_* \in (t_0, t_f]$, is a time-dependent domain:

$$(u,v)_{D_{t_*}} := \int_{t_0}^{t_*}\int_D uv\, d\mathbf{x}\, dt, \quad \|u\|_{D_{t_*}} := (u,u)_{D_{t_*}}^{1/2}, \quad u,v \in L_2(D_{t_*}). \tag{2.5}$$

With this notation we introduce the following Hilbert spaces. Let $H^{1,0}(\Omega_{t_f})$ be the space of all functions $u \in L_2(\Omega_{t_f})$ that have generalized space derivatives in $L_2(\Omega_{t_f})$. The associated scalar product is defined by

$$(u,v)_{H^{1,0}(\Omega_{t_f})} := (u,v)_{\Omega_{t_f}} + (\nabla u, \nabla v)_{\Omega_{t_f}}, \quad u,v \in H^{1,0}(\Omega_{t_f}). \tag{2.6}$$

Let further $H^{1,1}(\Omega_{t_f})$ denote the space of all functions $u \in L_2(\Omega_{t_f})$ that have both, generalized space and time derivatives in $L_2(\Omega_{t_f})$. The associated scalar product is given by

$$(u,v)_{H^{1,1}(\Omega_{t_f})} := (u,v)_{\Omega_{t_f}} + (\nabla u, \nabla v)_{\Omega_{t_f}} + (\frac{\partial u}{\partial t}, \frac{\partial v}{\partial t})_{\Omega_{t_f}}, \quad u,v \in H^{1,1}(\Omega_{t_f}). \tag{2.7}$$

Note that all derivatives in (2.6) and (2.7) are understood in the weak sense (see appendix). With these definitions we are able to give the *weak formulation* of problem (2.2)-(2.4). The formal multiplication of the differential equation (2.2) with a test function $\eta \in H^{1,1}(\Omega_{t_f})$ leads to the identity

$$(\frac{\partial T}{\partial t}, \eta)_{\Omega_{t_*}} = (a\Delta T, \eta)_{\Omega_{t_*}} + (f,\eta)_{\Omega_{t_*}}, \quad t_* \in (t_0, t_f].$$

Integration by parts using Green's formula (see appendix) yields

$$(\frac{\partial T}{\partial t}, \eta)_{\Omega_{t_*}} + (a\nabla T, \nabla\eta)_{\Omega_{t_*}} = (a\frac{\partial T}{\partial \mathbf{n}}, \eta)_{\Gamma_{t_*}} + (f,\eta)_{\Omega_{t_*}}.$$

If we use the same technique for the time derivative and substitute the initial and boundary conditions in (2.3) and (2.4), we get the identity

$$\begin{aligned} -(T, \frac{\partial \eta}{\partial t})_{\Omega_{t_*}} + (a\nabla T, \nabla \eta)_{\Omega_{t_*}} + (T(\cdot, t_*), \eta(\cdot, t_*))_\Omega \\ = (T_0, \eta(\cdot, t_0))_\Omega - \frac{a}{\lambda}(g, \eta)_{\Gamma_{t_*}} + (f, \eta)_{\Omega_{t_*}}. \end{aligned} \tag{2.8}$$

A *weak solution* of problem (2.2)-(2.4) in the space $H^{1,0}(\Omega_{t_f})$ is a function $T \in H^{1,0}(\Omega_{t_f})$ that satisfies identity (2.8) for almost all $t_* \in (t_0, t_f]$ and all $\eta \in H^{1,1}(\Omega_{t_f})$. For the formulation of the inverse heat conduction problem we introduce the spaces

$$\begin{aligned} V(\Omega_{t_f}) &:= \left\{ u \in H^{1,0}(\Omega_{t_f}) \mid \text{ess sup}_{t_0 \le t \le t_f} \|u(\cdot, t)\|_\Omega + \|\nabla u\|_{\Omega_{t_f}} < \infty \right\}, \\ V^{1,0}(\Omega_{t_f}) &:= \left\{ u \in V(\Omega_{t_f}) \mid t \mapsto u(\cdot, t) \in L_2(\Omega) \text{ is continuous } \right\}. \end{aligned}$$

In [LSU68] it is shown that the space $V^{1,0}(\Omega_{t_f})$ is a Banach space with the norm

$$|u|_{V^{1,0}(\Omega_{t_f})} := \max_{t_0 \le t \le t_f} \|u(\cdot, t)\|_\Omega + \|\nabla u\|_{\Omega_{t_f}}.$$

Now we can formulate the following definition of a weak solution in the space $V^{1,0}(\Omega_{t_f})$.

Definition 2.2 (weak solution) *A function T is called a weak solution in $V^{1,0}(\Omega_{t_f})$ of problem (2.2)-(2.4) if it belongs to $V^{1,0}(\Omega_{t_f})$ and satisfies the identity*

$$-(T, \frac{\partial \eta}{\partial t})_{\Omega_{t_f}} + (a\nabla T, \nabla \eta)_{\Omega_{t_f}} = (T_0, \eta(\cdot, t_0))_\Omega - \frac{a}{\lambda}(g, \eta)_{\Gamma_{t_f}} + (f, \eta)_{\Omega_{t_f}},$$

for all $\eta \in H^{1,1}(\Omega_{t_f})$ with $\eta(\cdot, t_f) = 0$.

Remark 2.3 In [Hao98] it is shown that the formulation in Definition 2.2 is equivalent to the following one: a function T is called a weak solution in $V^{1,0}(\Omega_{t_f})$ of problem (2.2)-(2.4) if it belongs to $V^{1,0}(\Omega_{t_f})$ and satisfies the identity (2.8) for almost all $t_* \in (t_0, t_f]$ and all $\eta \in H^{1,1}(\Omega_{t_f})$.

The following theorem states that a weak solution of problem (2.2)-(2.4) in the space $V^{1,0}(\Omega_{t_f})$ exists and is unique. Moreover, a stability estimate is given.

Theorem 2.4 *Consider problem (2.2)-(2.4) with data $f \in L_2(\Omega_{t_f})$, $T_0 \in L_2(\Omega)$, and $g \in L_2(\Gamma_{t_f})$. Then there exists a unique weak solution $T \in V^{1,0}(\Omega_{t_f})$ and the inequality*

$$\sup_{t_0 \le t \le t_f} \|T(\cdot, t)\|_\Omega^2 + \|\nabla T\|_{\Omega_{t_f}}^2 + \|T\|_{\Gamma_{t_f}}^2 \le c\,(\|f\|_{\Omega_{t_f}}^2 + \|T_0\|_\Omega^2 + \|g\|_{\Gamma_{t_f}}^2)$$

holds with a constant c independent of T (and thus also of f, g and T_0).

A proof of this theorem in a more general setting (with space- and time-dependent coefficients and the general second order linear parabolic equation) can be found in [Hao98]. Existence, uniqueness, and stability results for parabolic equations are also given in [Eva98, LSU68].

Remark 2.5 We can formulate an analogon of Definition 2.2 for a weak solution of problem (2.2)-(2.4) in the space $H^{1,1}(\Omega_{t_f})$. However, stronger regularity assumptions w.r.t. the initial and boundary data are required ($T_0 \in H^1(\Omega)$ and $g \in H^{0,1}(\Gamma_{t_f})$) in order to show that a unique and stable weak solution of problem (2.2)-(2.4) in the space $H^{1,1}(\Omega_{t_f})$ exists [Hao98].

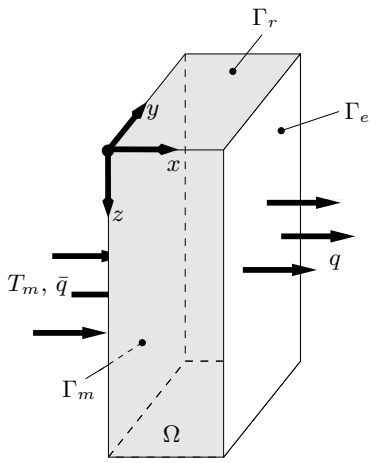


Figure 2.5: Heat conductor and boundary data; Γ_r consists of the (shaded) faces that are orthogonal to the y, z-plane

2.2.1 The inverse heat conduction problem

We now turn to the specific inverse heat conduction problem studied in this thesis. Let $\Omega \subset \mathbb{R}^3$ be the three-dimensional heat conductor shown in Fig. 2.5. This heat conductor is a parallelepiped whose boundary consists of pairwise disjoint parts $\Gamma := \Gamma_m \cup \Gamma_e \cup \Gamma_r$ which specify the measurement boundary Γ_m, the estimation boundary Γ_e, and the remaining boundary Γ_r orthogonal to the y, z-plane. Let the corresponding space-time domains be denoted by

$$S_\omega := \Gamma_\omega \times (t_0, t_f], \quad \omega \in \{m, e, r\}.$$

We assume the initial temperature distribution $T_0 \in L_2(\Omega)$ to be given. Further, we impose Dirichlet *and* Neumann boundary conditions on the measurement boundary Γ_m in time, i.e. we suppose the temperature $T_m \in L_2(S_m)$ and the boundary heat flux $\bar{q} \in L_2(S_m)$ to be known. Given these initial and boundary data, we want to find the temperature distribution $T(\mathbf{x}, t)$, $(\mathbf{x}, t) \in \Omega_{t_f}$, and in particular the boundary heat flux $q(\mathbf{x}, t) := -\lambda \frac{\partial T}{\partial \mathbf{n}}(\mathbf{x}, t)$, $(\mathbf{x}, t) \in S_e$, on the estimation boundary from

$$\begin{aligned}
\frac{\partial T}{\partial t}(\mathbf{x}, t) &= a\Delta T(\mathbf{x}, t), \quad (\mathbf{x}, t) \in \Omega_{t_f}, && (2.9)\\
T(\mathbf{x}, t_0) &= T_0(\mathbf{x}), \quad \mathbf{x} \in \Omega, && (2.10)\\
T(\mathbf{x}, t) &= T_m(\mathbf{x}, t), \quad (\mathbf{x}, t) \in S_m, && (2.11)\\
-\lambda \frac{\partial T}{\partial \mathbf{n}}(\mathbf{x}, t) &= \bar{q}(\mathbf{x}, t), \quad (\mathbf{x}, t) \in S_m, && (2.12)
\end{aligned}$$

where S_r is supposed to be adiabatic, i.e. $-\lambda \frac{\partial T}{\partial \mathbf{n}}(\mathbf{x}, t) = 0$, $(\mathbf{x}, t) \in S_r$. In the remainder of this chapter we always assume the boundaries not given explicitly in the problem formulation to be adiabatic. Note that we do not consider an inner heat source in equation (2.9). When the boundary heat flux $\bar{q}$ on S_m and the initial temperature distribution T_0 are given, then a (physically) imposed boundary heat flux q on S_e induces the temperature distribution of the heat conductor and thus in particular the temperature data T_m on S_m. In this respect, by solving problem (2.9)-(2.12), we want to determine the cause for an observed (measured) effect. For this reason, the latter problem is referred to as the *inverse problem*, which is known to be severely ill-posed (compare Section 1.2).

Remark 2.6 In the realistic falling film experiment a very thin heating foil is used in order to reach a small temperature gradient across the foil thickness (in x-direction). Thus, the

first temperature frame taken with the infrared camera and assumed to be constant in x-direction, should provide a good initial temperature distribution (see Section 6.3).

Definition 2.7 (solution of the inverse problem) *A function $q \in L_2(S_e)$ is called a solution of the inverse problem* (2.9)-(2.12) *if there exists a function $T \in V^{1,0}(\Omega_{t_f})$ that satisfies the identity*

$$-(T, \frac{\partial \eta}{\partial t})_{\Omega_{t_f}} + (a\nabla T, \nabla \eta)_{\Omega_{t_f}} = (T_0, \eta(\cdot, t_0))_\Omega - \frac{a}{\lambda}(\bar{q}, \eta)_{S_m} - \frac{a}{\lambda}(q, \eta)_{S_e},$$

for all $\eta \in H^{1,1}(\Omega_{t_f})$ with $\eta(\cdot, t_f) = 0$ and

$$T(\mathbf{x}, t) = T_m(\mathbf{x}, t), \quad (\mathbf{x}, t) \in S_m \quad \textit{(in the weak } L_2\textit{-sense).}$$

Since $T \in V^{1,0}(\Omega_{t_f})$ we have $T|_{S_m} \in L_2(S_m)$ and thus the definition of a solution of the inverse problem is meaningful (compare Lemma 3.2).

To prove existence of a solution of such ill-posed problems is a very challenging task. Most literature in the field of inverse heat conduction addresses problems in one space-dimension (see e.g. [HR98] and the references therein). In [Hao98] Háo studies a one-dimensional inverse problem with a general second order parabolic equation. The coefficients in the linear differential equation, which are time- and space-dependent, satisfy certain smoothness conditions in time. Roughly speaking, Háo succeeds in showing that a solution of the inverse problem exists if and only if the temperature data and the heat flux, imposed on one and the same side of the heat conductor, are also functions that are sufficiently smooth (w.r.t. the time variable) in a certain sense. However, the existence of a solution for the general multi-dimensional case with time- *and* space-dependent boundary data is not shown.

2.2.2 The associated direct and sensitivity problems

From the solution of the associated *direct problem*

$$\frac{\partial T}{\partial t}(\mathbf{x}, t) = a\Delta T(\mathbf{x}, t), \quad (\mathbf{x}, t) \in \Omega_{t_f}, \tag{2.13}$$

$$T(\mathbf{x}, t_0) = T_0(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{2.14}$$

$$-\lambda \frac{\partial T}{\partial \mathbf{n}}(\mathbf{x}, t) = \bar{q}(\mathbf{x}, t), \quad (\mathbf{x}, t) \in S_m, \tag{2.15}$$

$$-\lambda \frac{\partial T}{\partial \mathbf{n}}(\mathbf{x}, t) = q(\mathbf{x}, t), \quad (\mathbf{x}, t) \in S_e, \tag{2.16}$$

we obtain the temperature distribution $T(\mathbf{x}, t)$, $(\mathbf{x}, t) \in \Omega_{t_f}$, when the initial and boundary data $T_0 \in L_2(\Omega)$, $\bar{q} \in L_2(S_m)$, and $q \in L_2(S_e)$ are known. In this respect, by solving (2.13)-(2.16), we determine the effect of the imposed causes. Due to Theorem 2.4 there exists a unique weak solution $T \in V^{1,0}(\Omega_{t_f})$ of the direct heat conduction problem that satisfies the identity

$$-(T, \frac{\partial \eta}{\partial t})_{\Omega_{t_f}} + (a\nabla T, \nabla \eta)_{\Omega_{t_f}} = (T_0, \eta(\cdot, t_0))_\Omega - \frac{a}{\lambda}(\bar{q}, \eta)_{S_m} - \frac{a}{\lambda}(q, \eta)_{S_e},$$

for all $\eta \in H^{1,1}(\Omega_{t_f})$ with $\eta(\cdot, t_f) = 0$.

In this section we introduce a further heat conduction problem, which is used to describe the sensitivity of the associated direct problem solution w.r.t. a heat flux perturbation on the estimation boundary. Let $T(Q)$, $Q \in L_2(S_e)$, denote the solution of (2.13)-(2.16) with

$q \equiv Q$ in (2.16) and $S(Q)$, $Q \in L_2(S_e)$, the solution of the so-called *sensitivity problem*

$$\begin{aligned} \frac{\partial S}{\partial t}(\mathbf{x},t) &= a\Delta S(\mathbf{x},t), \quad (\mathbf{x},t) \in \Omega_{t_f}, && (2.17)\\ S(\mathbf{x},t_0) &= 0, \quad \mathbf{x} \in \Omega, && (2.18)\\ -\lambda\frac{\partial S}{\partial \mathbf{n}}(\mathbf{x},t) &= 0, \quad (\mathbf{x},t) \in S_m, && (2.19)\\ -\lambda\frac{\partial S}{\partial \mathbf{n}}(\mathbf{x},t) &= Q(\mathbf{x},t), \quad (\mathbf{x},t) \in S_e, && (2.20) \end{aligned}$$

with zero initial data and an adiabatic measurement boundary. We remark that $S(Q)$ depends linearly on Q. Let further $\delta q \in L_2(S_e)$ be a perturbation of the heat flux q in (2.16). The perturbed heat flux $\tilde{q} := q + \delta q$ results in a perturbed temperature distribution $T(\tilde{q})$, which can be written in the form

$$T(\tilde{q}) = T(q) + S(\delta q).$$

Hence, $S(\delta q)$ describes the effect of the heat flux perturbation δq in the associated direct heat conduction problem. The unique weak solution in $V^{1,0}(\Omega_{t_f})$ of the sensitivity problem fulfills the identity

$$-(S(Q), \frac{\partial \eta}{\partial t})_{\Omega_{t_f}} + (a\nabla S(Q), \nabla\eta)_{\Omega_{t_f}} = -\frac{a}{\lambda}(Q,\eta)_{S_e}, \tag{2.21}$$

for all $\eta \in H^{1,1}(\Omega_{t_f})$ with $\eta(\cdot, t_f) = 0$.

2.3 Solution strategy

In this thesis we apply a variational method for the solution of the IHCP in falling films (compare e.g. [EHN96, Hao98, NW99]). The idea in this approach is to consider the unknown heat flux on the estimation boundary as a decision variable to minimize a specified defect functional on the measurement boundary of the heat conductor. The optimal control problem is formulated in Section 2.3.1. There are different strategies to deal with this problem which is still ill-posed. Since an explicit representation of the integral operator used in the formulation of the IHCP is not available, we use an iterative technique. In that way only the applications of this operator are needed, which means that we have to compute well-posed direct heat conduction problems. Moreover, iterative methods have inherent regularization properties [Han95]. Thus, we can apply the conjugate gradient type method CGNE, in which the approximation quality depends on how many iterative steps are performed, directly to the original defect functional. In doing so, the choice of the regularization parameter is made during the iterative process. The algorithm is formulated and discussed in Section 2.3.2. Finally, in Section 2.3.3 we will see that the CGNE method (for linear ill-posed problems) in combination with an appropriate stopping criterion has even *order-optimal* regularization properties.

2.3.1 Variational formulation of the inverse problem

For the optimization-based formulation of the inverse heat conduction problem we introduce an abstract notation. We define the operator

$$\begin{aligned} \mathcal{A} : L_2(S_e) &\to L_2(S_m)\\ Q &\mapsto T(Q)|_{S_m}, \end{aligned}$$

mapping a time-dependent boundary heat flux on the estimation boundary of the heat conductor to the corresponding direct problem solution of (2.13)-(2.16) restricted to the

measurement boundary. This operator is an affine mapping, which can be written in the form

$$\mathcal{A} := \mathcal{A}_0 + \mathcal{A}_S, \tag{2.22}$$

with $\mathcal{A}_0 Q := T(0)|_{S_m}$, representing the temperature data on S_m for the case of an adiabatic boundary S_e. The linear part $\mathcal{A}_S$ in the splitting (2.22) is defined by

$$\begin{aligned} \mathcal{A}_S : L_2(S_e) &\to L_2(S_m) \\ Q &\mapsto S(Q)|_{S_m}, \end{aligned} \tag{2.23}$$

mapping $Q \in L_2(S_e)$ to the corresponding sensitivity problem solution of (2.17)-(2.20) restricted to the measurement boundary. With this notation we consider the following optimal control problem: determine the control $q \in L_2(S_e)$ such that

$$\begin{aligned} &J(q) := \frac{1}{2}\|\mathcal{A}q - T_m\|^2_{S_m} \to \min \\ &\text{with } \mathcal{A}q = T(q)|_{S_m} \text{ from (2.13)} - \text{(2.16)}. \end{aligned} \tag{2.24}$$

Here, the solution of (2.13)-(2.16) is understood in the weak sense in $V^{1,0}(\Omega_{t_f})$. For the definition of the norm $\|\cdot\|_{S_m}$ in (2.24) we refer to (2.5). Note that the defect functional can be written equivalently in the form $J(q) = \frac{1}{2}\|\mathcal{A}_S q - T_{m,s}\|^2_{S_m}$, with $T_{m,s} := T_m - \mathcal{A}_0 q$ denoting the shifted temperature data.

In [Hao98] Háo considers a multi-dimensional IHCP of the form (2.9)-(2.12), where the initial temperature distribution in addition to the boundary heat flux on S_e is unknown. If both quantities are considered as decision variables to minimize the defect functional, a simultaneous estimation can be performed. Háo proves the existence of a solution of the corresponding optimal control problem in a bounded and convex subspace. In this thesis we do not consider any restrictions and suppose that a solution of (2.24) in $L_2(S_e)$ exists. We remark that the variational problem has similar properties w.r.t. the ill-posedness as the original inverse problem (2.9)-(2.12) (compare [Hao98]).

2.3.2 The conjugate gradient type method CGNE

The CG method solves the least squares problem (2.24) by setting up an iteration sequence for the unknown optimal control $q \in L_2(S_e)$ until some stopping criteria are fulfilled. Let $q^k \in L_2(S_e)$ denote the approximate solution of (2.24) at iteration k. Then the next approximate solution $q^{k+1} \in L_2(S_e)$ is obtained from the previous one by the minimization of the defect functional along a certain search direction. For this search direction (the conjugate gradient) the gradient of the defect functional is needed. Further the optimal step length in this search direction has to be determined.

The CG algorithm which is illustrated in Fig. 2.6 consists of the following steps:

i. Set $k = 0$ and choose an initial value $q^0 \in L_2(S_e)$, e.g. $q^0 \equiv 0$.

ii. Compute $\mathcal{A}q^0 = T(q^0)|_{S_m}$ from the direct problem (2.13)-(2.16) and evaluate the defect functional $J(q^0)$.

iii. If the stopping criterion is satisfied then stop, otherwise continue.

iv. Calculate the new search direction $p^k = \nabla J(q^k) + \gamma^k p^{k-1}$. The gradient $\nabla J(q^k)$ is obtained from the solution of the *adjoint problem* (2.28)-(2.31). The *conjugate coefficient* γ^k is determined from the expression

$$\gamma^0 = 0; \quad \gamma^k = \frac{(\nabla J(q^k), \nabla J(q^k))_{S_e}}{(\nabla J(q^{k-1}), \nabla J(q^{k-1}))_{S_e}}, \; k \geq 1. \tag{2.25}$$

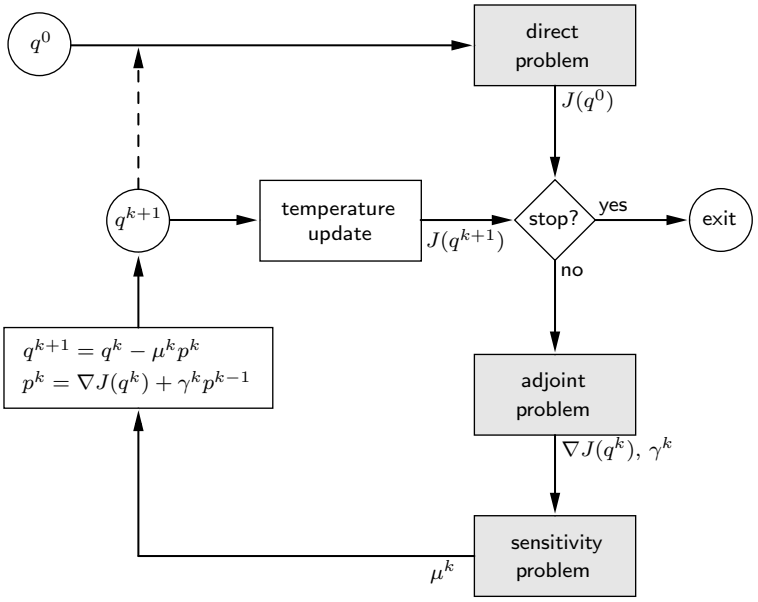


Figure 2.6: CGNE algorithm - flow chart

v. Compute the optimal step length in the search direction by solving the one-dimensional minimization problem

$$\mu^k = \min_{\mu \geq 0} J(q^k - \mu p^k), \tag{2.26}$$

leading to the expression

$$\mu^k = \frac{(\mathcal{A}q^k - T_m, \mathcal{A}_S p^k)_{S_m}}{(\mathcal{A}_S p^k, \mathcal{A}_S p^k)_{S_m}}, \tag{2.27}$$

where $\mathcal{A}_S p^k = S(p^k)|_{S_m}$ is obtained from solving the sensitivity problem (2.17)-(2.20).

vi. Update the approximation

$$q^{k+1} = q^k - \mu^k p^k.$$

vii. Since the considered inverse problem is affine, i.e. the operator $\mathcal{A}$ is an affine mapping in the control parameter, the new temperature distribution on the measurement boundary (i.e. the temperature update) can be determined via

$$\mathcal{A}q^{k+1} = \mathcal{A}q^k - \mu^k \mathcal{A}_S p^k,$$

using (2.22) without solving the direct problem again. Increase k by one and go back to iii.

For nonlinear problems, variants of this algorithm can be found in the literature (see e.g. [EHN96, NW99]).

Gradient determination via the adjoint problem

For the gradient determination of the defect functional in (2.24) we introduce a third heat conduction problem. Let $\psi(R)$, $R \in L_2(S_m)$, denote the solution of the so-called *adjoint*

problem

$$\begin{aligned} \frac{\partial \psi}{\partial t}(\mathbf{x},t) &= -a\Delta\psi(\mathbf{x},t), \quad (\mathbf{x},t)\in\Omega_{t_f}, && (2.28)\\ \psi(\mathbf{x},t_f) &= 0, \quad \mathbf{x}\in\Omega, && (2.29)\\ -\lambda\frac{\partial \psi}{\partial \mathbf{n}}(\mathbf{x},t) &= R(\mathbf{x},t), \quad (\mathbf{x},t)\in S_m, && (2.30)\\ -\lambda\frac{\partial \psi}{\partial \mathbf{n}}(\mathbf{x},t) &= 0, \quad (\mathbf{x},t)\in S_e, && (2.31) \end{aligned}$$

with zero temperature data at *final* time and an adiabatic estimation boundary. Then $\psi(R)$ satisfies the integral identity

$$-(\psi(R),\frac{\partial \xi}{\partial t})_{\Omega_{t_f}} - (a\nabla\psi(R),\nabla\xi)_{\Omega_{t_f}} = (\psi_0(R),\xi(\cdot,t_0))_\Omega + \frac{a}{\lambda}(R,\xi)_{S_m}, \tag{2.32}$$

for all $\xi\in H^{1,1}(\Omega_{t_f})$. In this representation $\psi_0(R)$ denotes the unknown solution at the initial time t_0. We emphasize the fact that the adjoint problem is well-posed due to the final time condition in (2.29) *and* the negative sign in front of the diffusion term in (2.28). By introducing a new time variable $t_b := t_f - t$, we get the heat equation forward in time with an initial temperature distribution. Thus, from Theorem 2.4 there exists a unique weak solution of the adjoint problem in $V^{1,0}(\Omega_{t_f})$.

We introduce the operator $\mathcal{A}_S^*$ mapping a boundary heat flux on the measurement boundary to the solution of the adjoint problem restricted to the estimation boundary, i.e.

$$\begin{aligned} \mathcal{A}_S^* : L_2(S_m) &\to L_2(S_e)\\ R &\mapsto \psi(R)|_{S_e}. \end{aligned} \tag{2.33}$$

The following lemma gives a relation between the sensitivity and the adjoint problem.

Lemma 2.8 *For any $Q\in L_2(S_e)$ and $R\in L_2(S_m)$ we have the identity*

$$(Q,\mathcal{A}_S^* R)_{S_e} = (\mathcal{A}_S Q, R)_{S_m}, \tag{2.34}$$

where the operators $\mathcal{A}_S$ and $\mathcal{A}_S^*$ are defined by (2.23) *and* (2.33)*, respectively.*

Proof: Let $Q\in L_2(S_e)$ and $R\in L_2(S_m)$. Then there exist sequences $\{Q^k\}_{k>0}$ in $H^{0,1}(S_e)$ and $\{R^k\}_{k>0}$ in $H^{0,1}(S_m)$ such that $Q^k\to Q$ in $L_2(S_e)$ and $R^k\to R$ in $L_2(S_m)$ for $k\to\infty$. Due to Theorem 2.4 there exist unique weak solutions $S(Q^k)$ of the sensitivity problem (2.17)-(2.20) and $\psi(R^k)$ of the adjoint problem (2.28)-(2.31) in the space $V^{1,0}(\Omega_{t_f})$. In [Hao98] it is shown that $S(Q^k)\to S(Q)$ and $\psi(R^k)\to\psi(R)$ in $V^{1,0}(\Omega_{t_f})$ for $k\to\infty$. From Remark 2.5 it follows that the solutions $S(Q^k)$ and $\psi(R^k)$ belong to the space $H^{1,1}(\Omega_{t_f})$. Thus, we can set $\eta=\psi(R^k)$ in (2.21) and $\xi=S(Q^k)$ in (2.32). Using integration by parts we get

$$\begin{aligned} \frac{a}{\lambda}(Q^k,\psi(R^k))_{S_e} &\overset{(2.21)}{=} (S(Q^k),\frac{\partial\psi(R^k)}{\partial t})_{\Omega_{t_f}} - (a\nabla S(Q^k),\nabla\psi(R^k))_{\Omega_{t_f}}\\ &= \int_\Omega [S(Q^k)\psi(R^k)]_{t_0}^{t_f}\,d\mathbf{x} - (\frac{\partial S(Q^k)}{\partial t},\psi(R^k))_{\Omega_{t_f}}\\ &\quad -(a\nabla S(Q^k),\nabla\psi(R^k))_{\Omega_{t_f}}\\ &\overset{(2.32)}{=} \frac{a}{\lambda}(R^k,S(Q^k))_{S_m}. \end{aligned}$$

The latter identity results from $\psi(R^k)=0$ at t_f (condition (2.29)) and thus in particular the integral term vanishes. Finally, letting $k\to\infty$ the lemma is proved. ◇

Due to Lemma 2.8 the operator $\mathcal{A}_S^*$ is the adjoint operator of $\mathcal{A}_S$, if the final time and boundary conditions of the corresponding adjoint problem are chosen as in (2.28)-(2.31).

Definition 2.9 (Fréchet derivative of a functional) *Let V be a Banach space with the norm $\|\cdot\|_V$ and V^* its dual space. Further, let the functional $J(v)$ be defined in a neighbourhood $\mathcal{U}(\hat{v},\gamma) := \{v \in V \mid \|v-\hat{v}\|_V < \gamma\}$ of a point $\hat{v} \in V$. Then J is said to be Fréchet differentiable at $\hat{v}$, if there exists an element $\nabla J(\hat{v}) \in V^*$ such that for any $\delta v \in V$, $\|\delta v\|_V < \gamma$, we have*

$$J(\hat{v}+\delta v) - J(\hat{v}) = (\nabla J(\hat{v}), \delta v)_{V^*} + o(\hat{v}, \delta v),$$

with

$$\lim_{\|\delta v\|_V \to 0} \frac{|o(\hat{v},\delta v)|}{\|\delta v\|_V} = 0.$$

The expression $(\nabla J(\hat{v}), \delta v)_{V^*}$ is called the differential of the functional J at the point $\hat{v}$, and $\nabla J(\hat{v})$ is called its gradient at $\hat{v}$.

Lemma 2.10 *Let $\mathcal{A}_S$ denote the linear operator defined by (2.23) and $T_{m,s} \in L_2(S_m)$. Then the functional $J(q) = \frac{1}{2}\|\mathcal{A}_S q - T_{m,s}\|^2_{S_m}$, $q \in L_2(S_e)$, is Fréchet differentiable at $\hat{q} \in L_2(S_e)$ and the gradient is given by*

$$\nabla J(\hat{q}) = \mathcal{A}_S^*(\mathcal{A}_S\hat{q} - T_{m,s}),$$

where $\mathcal{A}_S^*$ is defined by (2.33).

Proof: Let $\delta q \in L_2(S_e)$ be an increment of the control at $\hat{q} \in L_2(S_e)$. For the variation of the defect functional we get

$$\begin{aligned} J(\hat{q}+\delta q) - J(\hat{q}) &= (\mathcal{A}_S\hat{q} - T_{m,s}, \mathcal{A}_S\delta q)_{S_m} + \frac{1}{2}\|\mathcal{A}_S\delta q\|^2_{S_m} \\ &= (\mathcal{A}_S^*(\mathcal{A}_S\hat{q} - T_{m,s}), \delta q)_{S_e} + \frac{1}{2}\|\mathcal{A}_S\delta q\|^2_{S_m}. \end{aligned}$$

From Theorem 2.4 we get the following bound for the solution $S(\delta q)$ of the sensitivity problem

$$\sup_{t_0\le t\le t_f} \|S(\delta q)\|^2_\Omega + \|\nabla S(\delta q)\|^2_{\Omega_{t_f}} + \|S(\delta q)\|^2_{\Gamma_{t_f}} \le c\,\|\delta q\|^2_{S_e}. \tag{2.35}$$

Estimate (2.35) in particular yields $\|\mathcal{A}_S\delta q\|^2_{S_m}/\|\delta q\|_{S_e} \to 0$ for $\|\delta q\|_{S_e} \to 0$. Using the definition of the Fréchet derivative the lemma is proved. ◇

For the determination of the gradient in step iv. of the CGNE algorithm this means

$$\nabla J(q^k) = \mathcal{A}_S^*(\mathcal{A}_S q^k - T_{m,s}) = \mathcal{A}_S^*(\mathcal{A}q^k - T_m),$$

i.e. we have to solve the adjoint problem (2.28)-(2.31) setting $R := T(q^k)|_{S_m} - T_m$ in (2.30) and subsequently restrict the solution to the estimation boundary S_e.

The CGNE algorithm

For this CG method we get the following pseudocode:

Algorithm 2.11

choose q^0; $\gamma^0 := 0$;
$r^0 := \mathcal{A}_S^*(\mathcal{A}q^0 - T_m)$;
for $k = 0, 1, \ldots,$ unless $r^k \equiv 0$ do
  $p^k = r^k + \gamma^k p^{k-1}$;
  $\mu^k = \frac{(\mathcal{A}q^k - T_m, \mathcal{A}_S p^k)_{S_m}}{(\mathcal{A}_S p^k, \mathcal{A}_S p^k)_{S_m}}$;
  $q^{k+1} := q^k - \mu^k p^k$;
  if accurate then exit;
  $r^{k+1} = \mathcal{A}_S^*(\mathcal{A}q^{k+1} - T_m)$;
  $\gamma^{k+1} = \frac{(r^{k+1}, r^{k+1})_{S_e}}{(r^k, r^k)_{S_e}}$;
end for

CGNE algorithm in pseudocode notation

From this representation we can easily see that the solution algorithm is just the standard CG method applied to the normal equation (NE)

$$\mathcal{A}_S^*\mathcal{A}_S q = \mathcal{A}_S^* T_{m,s}, \tag{2.36}$$

with the selfadjoint positive semidefinite operator $\mathcal{A}_S^*\mathcal{A}_S$ and the shifted temperature data $T_{m,s} = T_m - \mathcal{A}_0 q$ (compare (2.22)).

The conjugate coefficient results from a certain optimality condition w.r.t. the new search direction. A detailed discussion of this issue is given in Section 3.3.2 for the discrete CG method. The continuous analogon of the conjugate coefficient in (3.52) for the normal equation directly leads to the expression in (2.25). The optimal step length in the search direction is determined from the minimization problem (2.26). For $\mu > 0$ we define

$$\begin{aligned} F(\mu) := J(q^k - \mu p^k) &= \frac{1}{2}(\mathcal{A}(q^k - \mu p^k) - T_m, \mathcal{A}(q^k - \mu p^k) - T_m)_{S_m} \\ &= \frac{1}{2}(\mathcal{A}q^k - \mu\mathcal{A}_S p^k - T_m, \mathcal{A}q^k - \mu\mathcal{A}_S p^k - T_m)_{S_m}. \end{aligned}$$

Setting $F'(\mu^k) = 0$ we get the identity

$$\mu^k(\mathcal{A}_S p^k, \mathcal{A}_S p^k)_{S_m} - (\mathcal{A}_S p^k, \mathcal{A}q^k - T_m)_{S_m} = 0,$$

which leads to the formula (2.27). Note, that $F''(\mu^k) = (\mathcal{A}_S p^k, \mathcal{A}_S p^k)_{S_m} > 0$ for $p^k \neq 0$.

Finally, we recall that the final time value of the initial guess q^0 is not modified by the algorithm due to the fact that the gradient of the defect functional $r^k = \mathcal{A}_S^*(\mathcal{A}q^k - T_m)$, $k \geq 0$, is always identically zero at t_f (see condition (2.29)). If r^κ becomes identically zero for some $\kappa \geq 0$ the iteration is terminated while we let $\kappa = \infty$ if no such finite κ exists. In the further context κ is called the *ultimate termination index* (compare [EHN96]).

2.3.3 Regularizing properties of CGNE

In this section we discuss the inherent regularization properties of the iterative CGNE method when an appropriate stopping rule is used in step iii. of the algorithm. We introduce some basic notions of regularization theory following the lines of [EHN96, Han95]. For a concise introduction to the CGNE method in an abstract Hilbert space setting we refer to [Kir96].

We consider the operator equation

$$\mathcal{K}f = g, \tag{2.37}$$

where $\mathcal{K}$ is a linear and bounded operator between the Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$. As we have already seen (in Section 1.2), in inverse problems the solution f typically does not depend continuously on the data g. In that case problem (2.37) is called *ill-posed*, a notion which was introduced by Hadamard [Had23] in 1923. Due to Hadamard a problem is said to be ill-posed, if one of the following postulates of *well-posedness* is violated: a solution exists for all admissible data, the solution is unique for all admissible data, and the solution depends continuously on the data. The existence of a solution often follows from physical considerations. However, when the operator $\mathcal{K}$ has a non-trivial nullspace then the solution is not uniquely defined. This is also the case for the studied IHCP in falling films. Those boundary heat fluxes which merely have a peak at final time belong to the nullspace of the corresponding operator defined by (2.23). Thus, one usually searches the solution of $\min_{f\in\mathcal{X}} \|\mathcal{K}f - g\|$ with minimum norm $\|f\|$. This uniquely determined element is called the *best-approximate solution* (see appendix). Let $\mathcal{K}^\dagger$ denote the generalized (Moore-Penrose) inverse of $\mathcal{K}$ and $\mathcal{D}(\mathcal{K}^\dagger)$ its domain. Then, for $g \in \mathcal{D}(\mathcal{K}^\dagger)$, the best-approximate solution of (2.37) is defined by

$$f^\dagger := \mathcal{K}^\dagger g. \tag{2.38}$$

With this notation the ill-posedness of problem (2.37) is equivalent to $\mathcal{K}^\dagger$ being unbounded. In that case, direct inversion is not adequate since the data g are usually not given exactly e.g. due to measurement or modelling errors. These considerations lead to the concept of regularization. The basic idea behind this concept is to approximate the original problem (2.37) by a "better posed" problem and to find a good compromise between the approximation and the propagated data errors. Let $\{g_\delta\}_{\delta>0}$ be a sequence of noisy data with noise level δ, i.e.

$$\|g_\delta - g\| \leq \delta, \quad \delta > 0. \tag{2.39}$$

A regularization method, involving a *regularization parameter* α, determines an approximation f_δ^α of $f^\dagger$ that depends continuously on g_δ and thus can be computed in a stable way with the property $f_\delta^\alpha \to f^\dagger$, $\delta \to 0$, for an appropriate choice of $\alpha = \alpha(\delta, g_\delta) > 0$. The replacement of the unbounded operator $\mathcal{K}^\dagger$ by a parameter-dependent family $\{\mathcal{R}^\alpha\}_{\alpha>0}$ of *continuous* operators leads to the approximation $f_\delta^\alpha := \mathcal{R}^\alpha g_\delta$. The precise definition of a regularization method is given in the appendix.

Besides the question how to construct appropriate regularization operators a main task is the practical choice of the regularization parameter for a given noisy data set. The optimal parameter $\alpha > 0$ which minimizes the term $\|\mathcal{R}^{\alpha(\delta,g_\delta)} g_\delta - \mathcal{K}^\dagger g\|$ cannot be determined since the exact solution $\mathcal{K}^\dagger g$ is unknown. Instead, one often considers the asymptotic convergence rate of $f_\delta^\alpha \to f^\dagger$, $\delta \to 0$. In [EHN96] it is shown that the convergence of a regularization method for solving (2.37) uniform in $g \in \mathcal{D}(\mathcal{K}^\dagger)$ is arbitrarily slow if $\mathcal{K}^\dagger$ is unbounded. Note, that $\mathcal{K}^\dagger$ being unbounded is equivalent to the range $\mathcal{R}(\mathcal{K})$ being non-closed (see appendix). Thus, convergence rates are given only on subsets of $\mathcal{D}(\mathcal{K}^\dagger)$, or rather under certain a-priori assumptions on the exact solution. For some $\mu > 0$ we define the *source set*

$$\mathcal{X}_\mu := \{f \in \mathcal{X} \mid f \in \mathcal{R}(|\mathcal{K}|^\mu) = \mathcal{R}((\mathcal{K}^*\mathcal{K})^{\mu/2})\}, \tag{2.40}$$

which enters the formulation of the following assumption (cf. [Han95]).

Assumption 2.12 *The solution $f^\dagger = \mathcal{K}^\dagger g$ with the unperturbed right-hand side $g \in \mathcal{R}(\mathcal{K})$ belongs to the set $\mathcal{X}_\mu$ in* (2.40) *for some $\mu > 0$, i.e. there exists $w \in \mathcal{X}$ with $f^\dagger = (\mathcal{K}^*\mathcal{K})^{\mu/2} w$; let $\omega := \|w\|$.*

We remark that in ill-posed problems the operator $\mathcal{K}$ is typically a smoothing operator and thus the assumption can be considered as a smoothness condition (compare Section 1.2). If $\mathcal{K}$ is compact this condition can be characterized in terms of the singular values (see appendix). In [EHN96] the following result is shown. If the generalized inverse $\mathcal{K}^\dagger$ is unbounded, i.e. $\mathcal{R}(\mathcal{K})$ is non-closed, then a regularization method cannot converge faster than $\mathcal{O}(\delta^{\mu/\mu+1})$ under the a-priori assumption $f^\dagger \in \mathcal{X}_\mu$ for some $\mu > 0$. This leads to the following definition of *order-optimal* methods.

Definition 2.13 *Let $\mathcal{R}(\mathcal{K})$ be non-closed, $\{\mathcal{R}^\alpha\}_{\alpha>0}$ be a family of regularization operators for $\mathcal{K}^\dagger$. For $\mu > 0$ and $g \in \mathcal{K}\mathcal{X}_\mu$, let α be a parameter choice rule for solving* (2.37)*. If $(\mathcal{R}^\alpha, \alpha)$ realizes the convergence rate*

$$\|\mathcal{R}^{\alpha(\delta,g_\delta)} g_\delta - \mathcal{K}^\dagger g\| = \mathcal{O}(\delta^{\mu/\mu+1}), \quad \delta \to 0,$$

it is called order-optimal (with respect to μ).

The CG method is known to be an efficient method for the solution of selfadjoint positive semidefinite and well-posed linear operator equations. We discuss the inherent regularization properties of the CGNE algorithm presented in Section 2.3.2 when applied to ill-posed problems. The following results can be found in [EHN96, Han95].

Theorem 2.14 *If the right-hand side g in* (2.37) *belongs to $\mathcal{D}(\mathcal{K}^\dagger)$ then the iterates f^k of the CGNE method converge to the exact solution $\mathcal{K}^\dagger g$ as $k \to \infty$.*

When problem (2.37) is ill-posed then $\mathcal{R}(\mathcal{K})$ is non-closed. Thus, even if $g \in \mathcal{D}(\mathcal{K}^\dagger)$, slightly perturbed data g_δ do no longer need to belong to $\mathcal{D}(\mathcal{K}^\dagger)$. In [Han95] it is shown that for $g_\delta \notin \mathcal{D}(\mathcal{K}^\dagger)$ the corresponding CGNE iterates f_δ^k diverge to infinity in norm as $k \to \infty$. However, if g_δ is an approximation of $g \in \mathcal{R}(\mathcal{K})$, then the data error propagation remains limited in the beginning of the iteration. This convergence behaviour is also called *semiconvergence*. In [Han95] the following result is proved.

Theorem 2.15 *Let Assumption 2.12 be satisfied, and let $\|g_\delta - g\| \le \delta$. Then there exists a constant c such that the iteration error of the CGNE method is bounded by*

$$\|f_\delta^k - \mathcal{K}^\dagger g\| \le c\,(|p_k'(0)|^{-\mu/2}\omega + |p_k'(0)|^{1/2}\delta), \quad 1 \le k \le \kappa, \tag{2.41}$$

with the ultimate termination index κ.

The so-called *residual polynomial* p_k of degree k in (2.41) is such that $|p_k'(0)| \ge k$ for every k. Thus, only for small values of k the second term in the right-hand side of (2.41) is of the order of δ and the iterates seem to converge while in the further iterations divergence sets in. If the iteration is terminated appropriately, and the approximation and propagated data errors are balanced in that way, even order-optimal approximations (in the sense of Definition 2.13) of the exact solution can be obtained. Below we discuss two different parameter choice rules for the determination of the CGNE stopping index. Both of them belong to the class of *a-posteriori* parameter choice rules, which means that $\alpha = \alpha(\delta, g_\delta)$ depends on the perturbed data g_δ (otherwise $\alpha = \alpha(\delta)$ is called *a-priori* parameter choice rule).

The discrepancy principle

The results discussed in this section go back to Nemirovskii (compare [EHN96, Han95] and the references therein). He suggested the so-called *discrepancy principle* of Morozov, originally used in Tikhonov regularization, for the determination of an appropriate CGNE stopping index. The idea behind this principle is to allow an error of the magnitude of the noise level δ in the data fit of the regularized solution f_δ^α in order to prevent an unacceptable data error amplification. In other words, it is not recommended to determine an approximate solution f_δ^α with $\|\mathcal{K} f_\delta^\alpha - g_\delta\| < \delta$, i.e. an approximation with a discrepancy that is smaller than the noise level. Thus, one should choose the largest possible parameter α such that the corresponding discrepancy is in the order of δ.

Stopping rule 2.16 (Discrepancy principle) *Assume $\|g_\delta - g\| \le \delta$. Fix $\tau > 1$, and terminate the iteration with index $\alpha = \alpha(\delta, g_\delta)$ when for the first time*

$$\|\mathcal{K} f_\delta^{\alpha(\delta, g_\delta)} - g_\delta\| \le \tau\delta. \tag{2.42}$$

In this principle an a-priori chosen parameter $\tau > 1$ is needed. For the so-called *Landweber iteration* the discrepancy principle with $\tau = 2$ determines the beginning of the transition from convergence to divergence [EHN96]. Since the iterates of the CGNE method do not depend linearly on the data g_δ (in contrast to Landweber) an analysis is much more involved. However, one can show that a unique and finite termination index is determined by (2.42) with $\alpha(\delta, g_\delta) \le \kappa$ even if κ is finite. For $g \in \mathcal{R}(\mathcal{K})$ (i.e. in the *attainable case*) the CGNE method in combination with Stopping rule 2.16 is order-optimal, as proved in [EHN96, Han95].

Theorem 2.17 *Let Assumption 2.12 be satisfied, and let $\|g_\delta - g\| \le \delta$. If the CGNE method (with right-hand side g_δ) is terminated after $\alpha(\delta, g_\delta)$ iterative steps according to Stopping rule 2.16, then*

$$\|f_\delta^{\alpha(\delta, g_\delta)} - \mathcal{K}^\dagger g\| = \mathcal{O}(\delta^{\mu/\mu+1}), \quad \delta \to 0.$$

Even if Assumption 2.12 does not hold, for exact data $g \in \mathcal{R}(\mathcal{K})$, the discrepancy principle leads to converging approximations $f_\delta^{\alpha(\delta, g_\delta)} \to \mathcal{K}^\dagger g$ as $\delta \to 0$. In [EHN96] also some remarks

about the *non-attainable case* with $g \in \mathcal{D}(\mathcal{K}^\dagger) \setminus \mathcal{R}(\mathcal{K})$ are given. We formulate the following consequence concerning so-called *semiiterative methods*, in which a basic step of the form $f_\delta^k = f_\delta^{k-1} + \mathcal{K}^*(g_\delta - \mathcal{K} f_\delta^{k-1})$, $k \in \mathbb{N}$, is performed followed by an averaging over all or some of the previous iterates. Since the CGNE algorithm minimizes the residual $\|\mathcal{K} f_\delta^k - g_\delta\|$ in the corresponding Krylov subspace shifted by the initial guess (see Section 3.3.2) it requires fewest iterations among all semiiterative methods, when the discrepancy principle is used for termination (compare [EHN96]).

In the numerical treatment of the IHCP besides the data error we also have discretization errors. Suppose that $\|\mathcal{K} - \mathcal{K}_\eta\| \leq \eta$, i.e. η characterizes the accuracy of the numerical schemes applied for the solution of the direct, the sensitivity, and the adjoint problems within the CGNE algorithm. For that case in [Hao98] (see also the references therein) the discrepancy principle with (2.42) replaced by

$$\|\mathcal{K}_\eta f_{\delta,\eta}^\alpha - g_\delta\| \leq \tau(\eta \|f_{\delta,\eta}^\alpha\| + \delta)$$

is used as the stopping rule. Although the problem discretization results in a kind of regularization it is usually not sufficient for practical purposes. In order to apply the discrepancy principle successfully a suitable bound δ for the noise level has to be known. If this information is not available or very inaccurate difficulties arise. An alternative approach, which does not require any knowledge about the noise level is the use of heuristic stopping rules.

The L-curve method

A brief description of the so-called *L-curve method* which is a heuristic and *error-free* parameter choice rule (i.e. it does not require information about the noise level) in the setting of continuous regularization methods can be found in [EHN96]. The idea of this method is to relate the norms of the computed approximations to the corresponding residual norms. This is done in a parametric plot of $\|f_\delta^\alpha\|$ versus $\|\mathcal{K} f_\delta^\alpha - g_\delta\|$ in a log-log scale for a large range of the regularization parameter. A schematic representation of this plot is shown in Fig. 2.7.

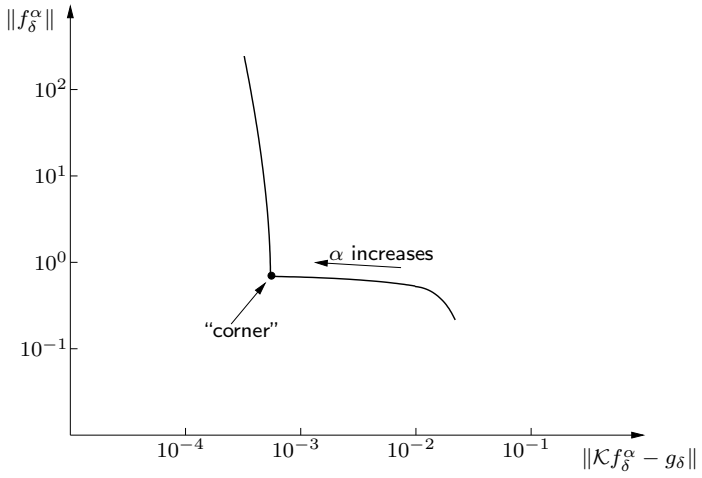


Figure 2.7: schematic L-curve (for CGNE)

Due to the semiconvergent behaviour of the CGNE method we can give the following heuristic explanation for the typical L-shape of the corresponding graph. Since the approximations seem to converge in the beginning of the iteration, i.e. for small values of the iteration index α, we expect the corresponding norms $\|f_\delta^\alpha\|$ to be close to $\|f^\dagger\|$ and thus vary only little

(see Theorem 2.15). On the other hand, if Assumption 2.12 is satisfied the bound

$$\|\mathcal{K}f_\delta^\alpha - g_\delta\| \leq \delta + c\,|p'_\alpha(0)|^{-(\mu+1)/2}\omega, \quad 1 \leq \alpha \leq \kappa,$$

for the residual norm is given in [Han95]. Thus, for small values of the index α the numbers $|p'_\alpha(0)|^{-(\mu+1)/2}$ essentially determine the rate of convergence and an observable decrease of the residual norm can be expected. The corresponding part of the graph, in which the solutions f_δ^α are dominated by approximation errors and said to be "oversmoothed", appears almost horizontal. In contrast to that, when α becomes larger, i.e. more and more steps of CGNE are applied with perturbed data $g_\delta \notin \mathcal{D}(\mathcal{K}^\dagger)$, then $\|f_\delta^\alpha\|$ blows up while $\|\mathcal{K}f_\delta^\alpha - g_\delta\|$ remains in the order of δ. The corresponding part of the graph, in which the solutions are dominated by perturbation errors, appears almost vertical. Altogether the above described parametric plot often has a shape that looks like the letter "L", which is the reason why it is called L-curve. For the choice of the regularization parameter we are interested in a good balance between the approximation and propagated data errors, i.e. in a parameter which corresponds to a point near the "corner" of the L-curve, which is defined to be the point of maximum curvature.

Stopping rule 2.18 (L-curve criterion) *Find the point on the L-curve with maximum curvature. The corresponding regularization parameter α provides a good balance between the approximation and perturbation errors.*

There is a vast amount of literature about the L-curve method and its pros and cons (see e.g. the references in [EHN96]). We only give a few comments. A clear advantage is that a bound on the noise level δ is not required. A study of the method when applied to discrete ill-posed problems is given by Hansen in [Han92]. He concludes that the discrepancy principle tends to oversmooth the real solution while the L-curve method may lead to slightly underregularized approximations. Further, in comparison to the generalized cross validation, the L-curve method is shown to be often more robust in the presence of correlated data errors. Due to the fact that we use iterative regularization only discrete points on the L-curve are known. In [HO93] a data fit to these discrete points using a cubic spline curve is recommended. Furthermore, an algorithm that generates the spline curve as well as an algorithm that tries to locate the corner of the L-curve efficiently is given in that paper. We emphasize that a distinct L-shape of the above described parametric plot is not guaranteed. The corner of the L-curve depends on the scale in which it is plotted, the chosen (semi-)norms, and it may even not appear at all [Reg96]. The numerical results of Chapter 6 show that the use of this parameter choice rule is satisfactory for our problem class.

Chapter 3

Numerical analysis and solution of the direct problems

We discuss standard numerical methods for the approximate solution of the continuous direct heat conduction problems. In each step of an implicit time integration method, when applied to these heat conduction problems, a linear stationary reaction-diffusion problem has to be solved. For this reason, in Section 3.1 a reaction-diffusion problem is studied. The variational formulation of this elliptic problem and its Galerkin discretization lead to the approximate solution u_h in the finite element space $V_h \subset V$. In this thesis we treat linear finite elements that are also called $\mathcal{P}_1$-elements [QV94]. For regular families of triangulations standard nodal interpolation and finite element discretization error bounds are presented. However, for the numerical treatment of the inverse heat conduction problem we have to take care of the very different length scales of the heat conductor that result in an anisotropy effect. For such anisotropic problems the given discretization error bounds do not hold in general (see Chapter 4). In Section 3.2 we give the weak formulation of the parabolic heat conduction problem. Using finite elements for the space discretization we get a semi-discrete approximation $u_h(\cdot, t) \in V_h$ for each point t in time. We get the fully discrete approximation scheme by discretizing the corresponding system of ordinary differential equations in time based on a one-step θ-method. Finally, in Section 3.3 we discuss different iterative solvers for the discrete problems.

3.1 The stationary boundary value problem

Let $\Omega \subset \mathbb{R}^n$, $n = 2, 3$, be an open and bounded parallelepiped. We introduce the finite element discretization of the linear stationary reaction-diffusion problem

$$\begin{aligned} -\Delta u + \mu u &= f \quad \text{in } \Omega, \\ -\lambda \frac{\partial u}{\partial \mathbf{n}} &= g \quad \text{on } \Gamma = \partial\Omega, \end{aligned} \tag{3.1}$$

with $f \in L_2(\Omega)$ and $g \in L_2(\Gamma)$. The constants μ and λ are assumed to be strictly positive. In order to derive the weak formulation, for the solution we consider the space $H^1(\Omega)$ which does not take the boundary conditions into account (see appendix). We remark that for $u \in H^1(\Omega)$ the weak derivative $D^\alpha u$, $|\alpha| = 1$, belongs to the space $L_2(\Omega)$ and due to this it is not possible to define $\frac{\partial u}{\partial \mathbf{n}}|_\Gamma$ unambiguously. If we multiply the differential equation in (3.1) with $v \in H^1(\Omega)$, integrate over Ω and use Green's formula, we obtain the variational formulation: find $u \in H^1(\Omega)$ such that

$$(\nabla u, \nabla v) + \mu(u, v) = (f, v) - \frac{1}{\lambda} \int_\Gamma g v \, d\Gamma, \quad \forall v \in H^1(\Omega).$$

In this representation $(\cdot,\cdot)$ denotes the standard scalar product in $L_2(\Omega)$. Defining the bilinear form $k(\cdot,\cdot)$ and the functional $l(\cdot)$ by the left- and right-hand sides

$$k(u,v) := (\nabla u, \nabla v) + \mu(u,v), \quad l(v) := (f,v) - \frac{1}{\lambda}\int_\Gamma gv\, d\Gamma, \tag{3.2}$$

we get the formulation: find $u \in H^1(\Omega)$ such that

$$k(u,v) = l(v), \quad \forall v \in H^1(\Omega). \tag{3.3}$$

In order to prove that the variational problem (3.3) has a unique solution we show that the bilinear form $k(\cdot,\cdot)$ is continuous and $H^1(\Omega)$-*elliptic*.

Definition 3.1 (V-elliptic bilinear form) *Let V be a Hilbert space with scalar product $(\cdot,\cdot)$ and corresponding norm $\|\cdot\|$. Let further $k : V \times V \to \mathbb{R}$ be a mapping with the properties:*

i. $k(u,\cdot)$ and $k(\cdot,u)$ are linear functionals on V for arbitrary $u \in V$.

ii. There exists a constant $M > 0$ such that

$$|k(u,v)| \le M\,\|u\|\|v\|, \quad \forall u,\, v \in V.$$

iii. There exists a constant $\gamma > 0$ such that

$$k(u,u) \ge \gamma\,\|u\|^2, \quad \forall u \in V.$$

A mapping $k(\cdot,\cdot)$ that fulfills assumptions i. and ii. is called continuous bilinear form on V. The property iii. is called V-ellipticity and the corresponding bilinear form coercive or V-elliptic.

Before we present the existence proof of a unique solution of problem (3.3) we give some elementary results (see e.g. [BS02, GR94]).

Lemma 3.2 (trace operator) *There exists a unique bounded linear operator*

$$\gamma : H^1(\Omega) \to L_2(\Gamma), \quad \|\gamma(\varphi)\|_{L_2(\Gamma)} \le c\,\|\varphi\|_{H^1(\Omega)},$$

with the property that for all $\varphi \in C^1(\bar\Omega)$ the equality $\gamma(\varphi) = \varphi|_\Gamma$ holds.

Lemma 3.3 (Lax-Milgram) *If the bilinear form $k(\cdot,\cdot)$ is bounded and coercive in the Hilbert space V, and $l(\cdot)$ is a bounded linear form in V, then there exists a unique $u \in V$ such that*

$$k(u,v) = l(v), \quad \forall v \in V,$$

is satisfied. Moreover, the energy estimate

$$\|u\|_V \le c\,\|l\|_{V^*}$$

holds, where $c = 1/\gamma$ with ellipticity constant γ.

Now we can prove the following theorem.

Theorem 3.4 *Consider the variational problem (3.3) with data $f \in L_2(\Omega)$ and $g \in L_2(\Gamma)$. Then this problem has a unique solution u and the inequality*

$$\|u\|_{H^1(\Omega)} \le c\,(\|f\|_{L_2(\Omega)} + \|g\|_{L_2(\Gamma)}) \tag{3.4}$$

holds with a constant c independent of f and g.

Proof: The linear functional $l(\cdot)$ is bounded:

$$\begin{aligned} |l(v)| &\leq \|f\|_{L_2(\Omega)}\|v\|_{L_2(\Omega)} + \frac{1}{\lambda}\|g\|_{L_2(\Gamma)}\|\gamma(v)\|_{L_2(\Gamma)} \\ &\leq c(\|f\|_{L_2(\Omega)} + \|g\|_{L_2(\Gamma)})\,\|v\|_{H^1(\Omega)}, \quad v \in H^1(\Omega). \end{aligned} \tag{3.5}$$

The latter inequalities follow from Cauchy-Schwarz and the continuity of the trace operator. The symmetric bilinear form $k(\cdot,\cdot)$ is obviously continuous:

$$\begin{aligned} |k(u,v)| &\leq |u|_{H^1(\Omega)}|v|_{H^1(\Omega)} + \mu\|u\|_{L_2(\Omega)}\|v\|_{L_2(\Omega)} \\ &\leq M\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)}, \quad u,\, v \in H^1(\Omega). \end{aligned} \tag{3.6}$$

And the bilinear form is coercive in $H^1(\Omega)$:

$$k(u,u) = |u|^2_{H^1(\Omega)} + \mu\|u\|^2_{L_2(\Omega)} \geq \gamma\|u\|^2_{H^1(\Omega)}, \quad \forall u \in H^1(\Omega), \tag{3.7}$$

with an ellipticity constant $\gamma > 0$. Thus, $k(\cdot,\cdot)$ is an inner product on $H^1(\Omega)$. Using (3.5), (3.6) and (3.7), from the Lax-Milgram lemma we obtain the existence of a unique solution such that (3.3) holds. The energy estimate finally yields inequality (3.4). ◇

Remark 3.5 One can show that the unique solution of problem (3.3) belongs to $H^2(\Omega)$ [Gri85] and that the variational problem is H^2-*regular*, i.e., for $f \in L_2(\Omega)$ the unique solution u of

$$k(u,v) = \int_\Omega f v \, d\mathbf{x}, \quad \forall v \in H^1(\Omega),$$

satisfies

$$\|u\|_{H^2(\Omega)} \leq c\,\|f\|_{L_2(\Omega)},$$

with a constant c independent of f. Such regularity results are discussed in Section 5.5 for *anisotropic* reaction-diffusion problems.

In the next section we introduce the space discretization of problem (3.3) based on finite elements.

3.1.1 Galerkin discretization and finite element method

For the space discretization of problem (3.3) we consider a finite dimensional subspace $V_h \subset V = H^1(\Omega)$ and the corresponding finite dimensional variational problem: find $u_h \in V_h$ such that

$$k(u_h, v_h) = l(v_h), \quad \forall v_h \in V_h. \tag{3.8}$$

This problem is called *Galerkin discretization* of (3.3). Since we assume $V_h \subset V$ the Galerkin discretization is said to be *conforming*. The following lemma gives a bound for the error between the continuous solution and the solution of problem (3.8), which is also referred to as the Galerkin approximation (compare [BS02, GR94]).

Lemma 3.6 (Céa-lemma) *Let $V_h \subset V$ be a finite dimensional subspace of the Hilbert space V and let the bilinear form $k(\cdot,\cdot)$ be continuous and V-elliptic. Then the variational problem*

$$k(u,v) = l(v), \quad \forall v \in V,$$

and its Galerkin discretization

$$k(u_h, v_h) = l(v_h), \quad \forall v_h \in V_h,$$

have unique solutions $u \in V$ and $u_h \in V_h$, respectively. Furthermore, the inequality

$$\|u - u_h\|_V \leq (1 + \frac{M}{\gamma}) \inf_{v_h \in V_h} \|u - v_h\|_V \tag{3.9}$$

holds.

On the one hand the Céa-lemma is a result on the existence and uniqueness of solutions for variational problems with continuous and V-elliptic bilinear forms. On the other hand inequality (3.9) allows us to give upper bounds for the discretization error $\|u - u_h\|_V$ in terms of the approximation error $\inf_{v_h \in V_h} \|u - v_h\|_V$ in the finite element space V_h. For linear finite elements approximation error bounds are discussed in the next section. In the remainder of this section we deduce the linear systems of equations resulting from a space discretization based on finite elements.

Let $\mathcal{T}_h$ denote a subdivision of Ω into a finite number of subsets e. Such a subdivision is called triangulation of Ω. In this thesis our main focus is devoted to triangular and tetrahedral triangulations. Since the domain Ω is assumed to be polygonal the triangulations $\mathcal{T}_h = \{e\}$ (where $\{e\}$ denotes a family of elements e) can be constructed such that

i. $\bar{\Omega} = \bigcup_{e \in \mathcal{T}_h} e$,

ii. int $e_1 \cap$ int $e_2 = \emptyset$ for all $e_1, e_2 \in \mathcal{T}_h$, $e_1 \neq e_2$,

iii. any edge (face) of any $e_1 \in \mathcal{T}_h$ is either subset of $\partial\Omega$ or an edge (face) of another $e_2 \in \mathcal{T}_h$.

If a triangulation fulfills these properties it is called *admissible*. For a family $\mathcal{F} = \{\mathcal{T}_h\}$ of admissible triangulations and $e \in \mathcal{T}_h$ we introduce the notations (compare Fig. 3.1)

$$\begin{aligned} h_e &:= \operatorname{diam}(e), \\ \rho_e &:= \sup\{\operatorname{diam}(S) \,|\, S \text{ is a ball contained in } e\}, \\ \sigma_e &:= \frac{h_e}{\rho_e} \in [1, \infty) \quad (\text{“aspect ratio”}). \end{aligned}$$

Here, the index parameter h of $\mathcal{T}_h$ is chosen such that $h = \max\{h_e \,|\, e \in \mathcal{T}_h\}$, i.e. h decreases if the triangulations get finer.

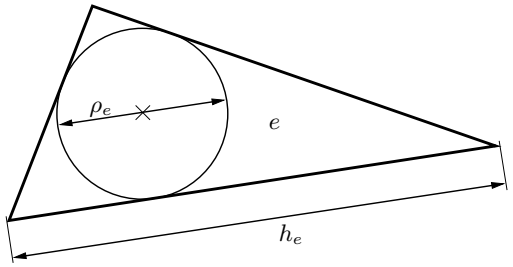


Figure 3.1: Notation for a single element e

Definition 3.7 (regular triangulations) *A family $\mathcal{F} = \{\mathcal{T}_h\}$ of admissible triangulations is called regular if*

i. h approaches zero: $\inf\{h \,|\, \mathcal{T}_h \in \mathcal{F}\} = 0$,

ii. the aspect ratio is bounded: there exists a constant $\kappa > 0$ such that $\sigma_e \leq \kappa$ for all $e \in \mathcal{T}_h$ and all $\mathcal{T}_h \in \mathcal{F}$.

A family $\mathcal{F} = \{\mathcal{T}_h\}$ of admissible triangulations is called quasi-uniform if there exists a constant $\tilde{\kappa} > 0$ such that

$$\frac{h}{\rho_e} \leq \tilde{\kappa}$$

for all $e \in \mathcal{T}_h$ and all $\mathcal{T}_h \in \mathcal{F}$.

Note, that the assumption of a bounded aspect ratio is quite restrictive. It prevents the triangles or tetrahedra of the triangulations to become degenerated, i.e. the angles of these elements are bounded from below and above independently of h. We define the general simplicial finite element spaces

$$\begin{aligned} X_h^0 &:= \{v \in L_2(\Omega) \,|\, v|_e \in \mathcal{P}_0 \text{ for all } e \in \mathcal{T}_h\}, \\ X_h^l &:= \{v \in C(\bar{\Omega}) \,|\, v|_e \in \mathcal{P}_l \text{ for all } e \in \mathcal{T}_h\}, \quad l \geq 1, \end{aligned} \tag{3.10}$$

where $\mathcal{P}_l$, $l \geq 0$, denotes the space of polynomials with degree less than or equal l. When the triangulations are regular then the elements of the associated spaces in (3.10) are referred to as *regular finite elements*. In the remainder of this thesis we restrict ourselves to $\mathcal{P}_1$-elements that are globally continuous and piecewise linear, i.e. X_h^1.

Let $\{\phi_i\}_{1\leq i\leq n}$ denote the standard nodal basis of X_h^1, i.e. if $\{N_i\}_{1\leq i\leq n}$ denote the vertices of the triangulation $\mathcal{T}_h$, the basis element ϕ_j is the pyramid (hat) function that takes the value one at the vertex N_j and vanishes at all other vertices. Then every $u_h \in X_h^1$ has the unique representation

$$u_h = \sum_{j=1}^{n} x_j \phi_j, \quad \text{with } \mathbf{x} := (x_1, \ldots, x_n)^T \in \mathbb{R}^n.$$

The reformulation of the Galerkin discretization (3.8) in the form: find $\mathbf{x} \in \mathbb{R}^n$ such that

$$\sum_{j=1}^{n} k(\phi_j, \phi_i) x_j = l(\phi_i), \quad \forall i = 1, \ldots, n,$$

then leads to a linear system of equations

$$\mathbf{A}\mathbf{x} = \mathbf{b},$$

where the matrix $\mathbf{A} = (A_{ij})_{i,j=1}^n$ with components $A_{ij} = k(\phi_j, \phi_i)$ is the so-called *stiffness matrix* and the right-hand side is given by $\mathbf{b} = (l(\phi_1), \ldots, l(\phi_n))^T$. Here n denotes the dimension of the finite element space $V_h = X_h^1$. For the Neumann boundary value problem (3.1) with the *natural boundary condition* $-\lambda \frac{\partial u}{\partial \mathbf{n}} = g$ on Γ, which is implicitly contained in the variational formulation, n equals the number of interior and boundary vertices of the given triangulation. The corresponding stiffness matrix is strictly symmetric positive definite. In order to see that this property holds, we write the term $\xi^T \mathbf{A} \xi$ for an arbitrary $\xi \in \mathbb{R}^n$ in the form

$$\sum_{j,k=1}^{n} A_{jk}\xi_j\xi_k = \|\nabla(\sum_{j=1}^{n} \xi_j\phi_j)\|_{L_2(\Omega)}^2 + \mu \|\sum_{j=1}^{n} \xi_j\phi_j\|_{L_2(\Omega)}^2 \geq 0. \tag{3.11}$$

Since we assume $\mu > 0$, equality only holds for $\xi \equiv 0 \in \mathbb{R}^n$. Note that for the pure diffusion problem with $\mu = 0$, equality also holds for $\xi \equiv c$ with an arbitrary constant $c \in \mathbb{R}^n$. In that case the stiffness matrix is symmetric positive semi-definite.

We briefly comment on the *Dirichlet* boundary value problem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma. \end{aligned}$$

The variational problem is: find $u \in H_0^1(\Omega)$ such that

$$(\nabla u, \nabla v) = (f, v), \quad \forall v \in H_0^1(\Omega).$$

Note, that for this problem we do not have degrees of freedom located on the boundary due to the *essential boundary condition* $u = 0$ on Γ that is imposed explicitly in the variational formulation. We give the following conclusion of the Friedrichs inequality [BS02, GR94].

Lemma 3.8 *There exists a constant $c > 0$ such that*

$$c \|v\|_{H^1(\Omega)} \leq \|\nabla v\|_{L_2(\Omega)}, \quad \forall v \in H_0^1(\Omega).$$

Let $\{\phi_i\}_{1 \leq i \leq \tilde{n}}$ be the standard nodal basis of $X_{h,0}^1 := X_h^1 \cap H_0^1(\Omega)$ (associated with the interior grid points only). From Lemma 3.8 it follows that $\xi_j = 0$, $j = 1, \ldots, \tilde{n}$, to obtain $\|\nabla(\sum_{j=1}^{\tilde{n}} \xi_j \phi_j)\|_{L_2(\Omega)}^2 = 0$ for the homogeneous Dirichlet boundary value problem. Thus, all eigenvalues of the corresponding stiffness matrix are strictly positive.

3.1.2 Basic error estimates

In this section we derive interpolation and finite element discretization error bounds based on the classical finite element theory that is restricted to regular finite elements. These results, which can be found at many places in the literature, e.g. in [Cia78, Tho97], serve as a preparation for the study of anisotropic finite elements in the next chapter.

Let $X_h^1 = V_h \subset V = H^1(\Omega)$. We already pointed out that inequality (3.9) in the Céa-lemma reduces the problem of estimating the discretization error $\|u - u_h\|_{H^1(\Omega)}$ to the problem of finding a bound for $\inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)}$, i.e. for the approximation error in the finite element space $V_h \subset V$. We introduce the standard nodal interpolation operator

$$I_h : C(\bar{\Omega}) \to X_h^1,$$

which is well-defined for $u \in H^2(\Omega)$ due to the embedding $H^2(\Omega) \hookrightarrow C(\bar{\Omega})$ (see appendix and Remark 3.5). The linear interpolant $I_h u$ has the same values as u at the vertices of the given triangulation and is linear on each $e \in \mathcal{T}_h$.

Definition 3.9 (linear interpolation) *Let $\mathcal{T}_h = \{e\}$ be an admissible triangulation of Ω. For $u \in C(\bar{\Omega})$ we define the function $I_h u$ by standard linear interpolation on each $e \in \mathcal{T}_h$, i.e.*

$$(I_h u)(\mathbf{x}) = \sum_{j=1}^{n} u(N_j)\phi_j(\mathbf{x}), \quad \mathbf{x} \in \Omega,$$

with $\{N_i\}_{1 \leq i \leq n}$ denoting the vertices of the triangulation $\mathcal{T}_h$ and $\{\phi_i\}_{1 \leq i \leq n}$ representing the corresponding standard nodal basis.

Since the linear interpolant of the exact solution $I_h u$ is an element of the space X_h^1 in combination with the Céa-lemma we get the inequality

$$\|u - u_h\|_{H^1(\Omega)} \leq c \|u - I_h u\|_{H^1(\Omega)}. \tag{3.12}$$

For the local interpolation operator I_e on a single element $e \in \mathcal{T}_h$ we have the identity

$$(I_h u)|_e = I_e(u|_e), \quad \forall e \in \mathcal{T}_h. \tag{3.13}$$

Thus, we can reformulate the term on the right-hand side of (3.12) in the form

$$\|u - I_h u\|_{H^1(\Omega)} = \left(\sum_{e \in \mathcal{T}_h} \|u - I_e u\|_{H^1(e)}^2 \right)^{\frac{1}{2}}$$

i.e., in terms of the local interpolation errors. A basic idea to obtain a bound for the local interpolation error on an arbitrary element e is to deduce an estimate on the so-called reference element $\hat{e}$ and subsequently convert it into an estimate on the original element. We illustrate this approach in Fig. 3.2 for the case $\Omega \subset \mathbb{R}^2$ and a decomposition of Ω into triangles. Let $F : \hat{e} \to e$ denote the affine transformation between the reference element $\hat{e}$ and an arbitrary element e. Note, that such a transformation exists for all $e \in \mathcal{T}_h$ when $\mathcal{T}_h$ is an admissible triangulation. Furthermore the reference element is independent of the discretization parameter h. The following lemmas can be found in [Cia78, GR94].

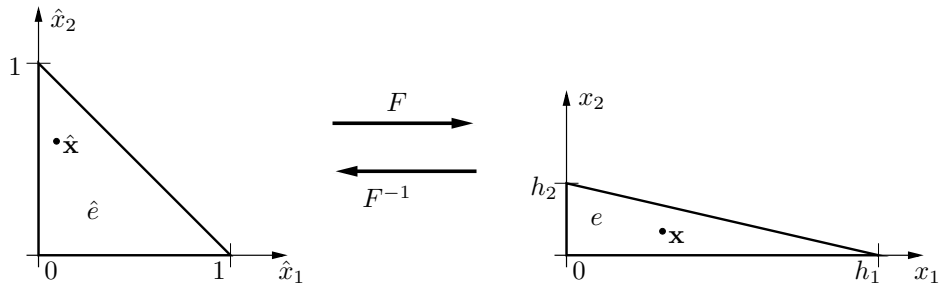


Figure 3.2: Transformation between $\hat{e}$ and e

Lemma 3.10 *Let e and $\hat{e}$ be two open domains in $\mathbb{R}^n$ that are affine-equivalent, i.e.*

$$F(\hat{e}) = e \text{ with } F(\hat{\mathbf{x}}) = \mathbf{B}\hat{\mathbf{x}} + \mathbf{c}, \quad \det(\mathbf{B}) \neq 0. \tag{3.14}$$

For $m \geq 0$, $p \in [1, \infty]$, and $v \in W_p^m(e)$ define $\hat{v} := v \circ F : \hat{e} \to \mathbb{R}$. Then $\hat{v} \in W_p^m(\hat{e})$ and there exists a constant $c = c(m, n)$ such that

$$\begin{aligned} |\hat{v}|_{W_p^m(\hat{e})} &\leq c\|\mathbf{B}\|_2^m |\det(\mathbf{B})|^{-\frac{1}{p}} |v|_{W_p^m(e)}, \quad \forall v \in W_p^m(e), \\ |v|_{W_p^m(e)} &\leq c\|\mathbf{B}^{-1}\|_2^m |\det(\mathbf{B})|^{\frac{1}{p}} |\hat{v}|_{W_p^m(\hat{e})}, \quad \forall \hat{v} \in W_p^m(\hat{e}). \end{aligned} \tag{3.15}$$

Here, the spaces $W_p^m(e)$ denote the usual Sobolev spaces of functions whose weak derivatives up to order m belong to $L_p(e)$ (see appendix). Thus, to obtain appropriate estimates for the transformations between e and $\hat{e}$ in particular the terms $\|\mathbf{B}\|_2$ and $\|\mathbf{B}^{-1}\|_2$ in (3.15) have to be bounded.

Lemma 3.11 *Let $e, \hat{e} \subset \mathbb{R}^n$ be two affine-equivalent bounded open domains and $F(\hat{\mathbf{x}}) = \mathbf{B}\hat{\mathbf{x}} + \mathbf{c}$ be an affine mapping such that $F(\hat{e}) = e$. Then the inequalities*

$$\|\mathbf{B}\|_2 \leq \frac{h_e}{\rho_{\hat{e}}}, \quad \|\mathbf{B}^{-1}\|_2 \leq \frac{h_{\hat{e}}}{\rho_e}$$

hold.

Due to this lemma we get $\|\mathbf{B}\|_2 \leq ch_e$ since $\rho_{\hat{e}}$ is a fixed quantity of the reference element that is independent of h_e. In contrast to that the term $\|\mathbf{B}^{-1}\|_2$ has to be treated more carefully. Note, that the terms in (3.15) containing the determinant of $\mathbf{B}$ cancel when we apply the transformation from the original element to the reference element and in the reverse direction (for more details we refer to the proof of Theorem 3.13).

We now present a local interpolation result that can be found in [BS02, Cia78] in a more general setting, e.g. including higher order finite element spaces. Note, that no regularity assumptions concerning the considered family of triangulations are made. Since it is a basic ingredient in the proof we also give the following lemma which was first proved by Deny and Lions [Cia78].

Lemma 3.12 *Let ω be a Lipschitz domain in $\mathbb{R}^n$ and $k \geq 0$, $p \in [1, \infty]$. Then there exists a constant $c = c(\omega)$ such that*

$$\inf_{q \in \mathcal{P}_k(\omega)} \|v + q\|_{W_p^{k+1}(\omega)} \leq c\,|v|_{W_p^{k+1}(\omega)}, \quad \forall v \in W_p^{k+1}(\omega). \tag{3.16}$$

We note that a Lipschitz domain is a domain whose boundary satisfies certain smoothness conditions. For the parallelepiped Ω and the unit n-simplex $\hat{e}$ with piecewise smooth boundaries these conditions are fulfilled.

Theorem 3.13 *Let $\mathcal{F} = \{\mathcal{T}_h\}$ be a family of triangulations of Ω consisting of n-simplices and X_h^1 the corresponding finite element space as defined by (3.10). Suppose $p \in [1, \infty]$ and either $n/p < 2$ when $p > 1$ or $n = 2$ when $p = 1$. Then for each $e \in \mathcal{T}_h$ there exists a constant c independent of e such that*

$$|u - I_e u|_{W_p^m(e)} \leq c \frac{h_e^2}{\rho_e^m} |u|_{W_p^2(e)}, \quad 0 \leq m \leq 2, \quad \forall u \in W_p^2(e). \tag{3.17}$$

Proof: Let $\hat{e}$ be the unit n-simplex and $F : \hat{e} \to e$ an affine transformation $F(\hat{e}) = \mathbf{B}\hat{e} + \mathbf{c}$ such that $F(\hat{e}) = e$. Note that $\|q\|_* := \sum_{x \in N(\hat{e})} |q(x)|$, where $N(\hat{e})$ denotes the set of vertices of $\hat{e}$, defines a norm on $\mathcal{P}_1$. Since all norms on $\mathcal{P}_1$ are equivalent there exists a constant c such that

$$\|q\|_{W_p^2(\hat{e})} \leq c \|q\|_*, \quad \forall q \in \mathcal{P}_1. \tag{3.18}$$

Since $n/p < 2$ when $p > 1$ or $n = 2$ when $p = 1$, the continuous embedding $W_p^2(\hat{e}) \hookrightarrow C(\bar{\hat{e}})$ (see appendix) yields the existence of a constant c such that

$$\|\hat{u}\|_{L_\infty(\hat{e})} \leq c \|\hat{u}\|_{W_p^2(\hat{e})}, \quad \forall \hat{u} \in W_p^2(\hat{e}). \tag{3.19}$$

Let $\hat{I}_{\hat{e}} : C(\hat{e}) \to \mathcal{P}_1(\hat{e})$ be the linear interpolation operator on $\hat{e}$. We then have

$$(I_e u) \circ F = \hat{I}_{\hat{e}}(u \circ F) = \hat{I}_{\hat{e}} \hat{u}$$

with $\hat{u} := u \circ F$. Define the linear operator $L := \mathrm{id} - \hat{I}_{\hat{e}} : W_p^2(\hat{e}) \to W_p^2(\hat{e})$. Using (3.18) and (3.19), the continuity of this operator follows from

$$\begin{aligned} \|L\hat{u}\|_{W_p^2(\hat{e})} &\leq \|\hat{u}\|_{W_p^2(\hat{e})} + \|\hat{I}_{\hat{e}}\hat{u}\|_{W_p^2(\hat{e})} \leq \|\hat{u}\|_{W_p^2(\hat{e})} + c \|\hat{I}_{\hat{e}}\hat{u}\|_* \\ &\leq \|\hat{u}\|_{W_p^2(\hat{e})} + c \sum_{x \in N(\hat{e})} |\hat{u}(x)| \leq \|\hat{u}\|_{W_p^2(\hat{e})} + c \|\hat{u}\|_{L_\infty(\hat{e})} \leq c \|\hat{u}\|_{W_p^2(\hat{e})}. \end{aligned}$$

Since $Lq = 0$ for all $q \in \mathcal{P}_1(\hat{e})$ we get

$$\hat{u} - \hat{I}_{\hat{e}}\hat{u} = L(\hat{u} + q), \quad \forall \hat{u} \in W_p^2(\hat{e}),\, q \in \mathcal{P}_1(\hat{e}).$$

From this identity, the continuity of the operator L and Lemma 3.12 we deduce that

$$\|\hat{u} - \hat{I}_{\hat{e}}\hat{u}\|_{W_p^2(\hat{e})} \leq c \inf_{q \in \mathcal{P}_1(\hat{e})} \|\hat{u} + q\|_{W_p^2(\hat{e})} \leq c |\hat{u}|_{W_p^2(\hat{e})}, \quad \forall \hat{u} \in W_p^2(\hat{e}). \tag{3.20}$$

For $u \in W_p^2(e)$ and $0 \leq m \leq 2$, using (3.20), we obtain

$$\begin{aligned} |u - I_e u|_{W_p^m(e)} &\overset{Lem.\ 3.10}{\leq} c\|\mathbf{B}^{-1}\|_2^m |\det(\mathbf{B})|^{\frac{1}{p}} |(u - I_e u) \circ F|_{W_p^m(\hat{e})} \\ &= c\|\mathbf{B}^{-1}\|_2^m |\det(\mathbf{B})|^{\frac{1}{p}} |\hat{u} - \hat{I}_{\hat{e}}\hat{u}|_{W_p^m(\hat{e})} \\ &\leq c\|\mathbf{B}^{-1}\|_2^m |\det(\mathbf{B})|^{\frac{1}{p}} \|\hat{u} - \hat{I}_{\hat{e}}\hat{u}\|_{W_p^2(\hat{e})} \\ &\leq \hat{c}(\hat{e})\|\mathbf{B}^{-1}\|_2^m |\det(\mathbf{B})|^{\frac{1}{p}} |\hat{u}|_{W_p^2(\hat{e})} \\ &\overset{Lem.\ 3.10}{\leq} \hat{c}\|\mathbf{B}^{-1}\|_2^m \|\mathbf{B}\|_2^2 |u|_{W_p^2(e)} \\ &\overset{Lem.\ 3.11}{\leq} \hat{c}\rho_e^{-m} h_e^2 |u|_{W_p^2(e)}. \end{aligned} \tag{3.21}$$

◇

We note that the constant c in (3.16) depends on the geometry of the associated domain and may "blow up" if this domain becomes degenerate. Thus, in the proof of Theorem 3.13 we use the transformation from the original to the reference element and apply Lemma 3.12. Subsequently, the obtained estimate is transformed back to the original element. In the

following example we consider the upper bound in Theorem 3.13 for an *anisotropic* element.

An example in $\mathbb{R}^2$:

We consider the special element e in Fig. 3.2 with $h_2 = o(h_1)$. The transformation F between e and $\hat{e}$ is given by

$$\mathbf{x} = F(\hat{\mathbf{x}}) = \mathbf{B}\hat{\mathbf{x}}, \quad \mathbf{B} = \begin{pmatrix} h_1 & 0 \\ 0 & h_2 \end{pmatrix}.$$

The example is chosen such that $\sigma_e \sim \frac{h_1}{h_2} \to \infty$ for $h_1 \to 0$, i.e. the element e has an unbounded aspect ratio. The condition number of the matrix $\mathbf{B}$ is

$$\mathrm{cond}_2(\mathbf{B}) = \|\mathbf{B}\|_2 \|\mathbf{B}^{-1}\|_2 = \frac{h_1}{h_2}$$

and thus it becomes worse for increasing degeneracy of the element e. Note, that already the factor

$$\|\mathbf{B}^{-1}\|_2^m |\det(\mathbf{B})|^{\frac{1}{p}} = h_2^{-m}(h_1 h_2)^{\frac{1}{p}}$$

in the upper bound (3.21) may become arbitrarily large for $h_1 \to 0$, e.g. if we set $p = 2$ and $m = 1$. Thus, if we want to use anisotropic elements that are characterized by an unbounded aspect ratio, we have to take special care of the bad asymptotics resulting from the transformation. For a detailed discussion of this issue we refer to Chapter 4.

In the remainder of this section we set $p = 2$ and assume that the family $\mathcal{F} = \{\mathcal{T}_h\}$ of triangulations is regular, i.e. $\mathrm{cond}_2(\mathbf{B}) \le c\frac{h_e}{\rho_e} \le \tilde{c}$ for some constants c and $\tilde{c}$.

Theorem 3.14 (interpolation error bound) *Let $\mathcal{F} = \{\mathcal{T}_h\}$ be a regular family of triangulations of Ω consisting of n-simplices and X_h^1 the corresponding finite element space as defined by* (3.10)*. Then there exists a constant c independent of h such that*

$$\|u - I_h u\|_{H^m(\Omega)} \le c h^{2-m} |u|_{H^2(\Omega)}, \quad m \in \{0, 1\}, \quad \forall u \in H^2(\Omega). \tag{3.22}$$

Proof: From Theorem 3.13 we get the *local* interpolation error bound

$$\|u - I_e u\|_{H^m(e)} \le c h_e^{2-m} |u|_{H^2(e)}, \quad 0 \le m \le 2, \quad \forall u \in H^2(e).$$

Using $h_e \le h$, $e \in \mathcal{T}_h$, this leads to the inequality

$$\begin{aligned} \left(\sum_{e \in \mathcal{T}_h} \|u - I_e u\|_{H^m(e)}^2 \right)^{\frac{1}{2}} &\le c h^{2-m} \left(\sum_{e \in \mathcal{T}_h} |u|_{H^2(e)}^2 \right)^{\frac{1}{2}} \\ &= c h^{2-m} |u|_{H^2(\Omega)}, \quad 0 \le m \le 2, \quad \forall u \in H^2(\Omega). \end{aligned} \tag{3.23}$$

For the *global* linear interpolation operator we have $I_h : C(\bar{\Omega}) \to X_h^1 \subset H^1(\Omega)$. Since $I_h u$ does not belong to $H^2(\Omega)$ for all $u \in H^2(\Omega)$ we have to restrict the local estimate to the choice $m \in \{0, 1\}$. With (3.13) and (3.23) we conclude

$$\begin{aligned} \|u - I_h u\|_{H^m(\Omega)} &= \left(\sum_{e \in \mathcal{T}_h} \|u - I_h u\|_{H^m(e)}^2 \right)^{\frac{1}{2}} \\ &\le c h^{2-m} |u|_{H^2(\Omega)}, \quad m \in \{0, 1\}, \quad \forall u \in H^2(\Omega). \end{aligned} \tag{3.24}$$

◇

Since $I_h u \in X_h^1$ for all $u \in H^2(\Omega)$, from (3.22) we directly obtain the following result for the approximation error in the space of linear finite elements.

Corollary 3.15 ($\mathcal{P}_1$-element approximation error bound) *For the approximation error of regular $\mathcal{P}_1$-elements we have*

$$\inf_{v_h \in X_h^1} \|u - v_h\|_{H^m(\Omega)} \leq ch^{2-m} |u|_{H^2(\Omega)}, \quad m \in \{0, 1\}, \quad \forall u \in H^2(\Omega). \tag{3.25}$$

The following elementary results concerning finite element discretization error bounds can be found in [BS02, Cia78].

Theorem 3.16 *Let u be the solution of the continuous problem* (3.3) *and u_h the solution of the discrete problem* (3.8) *with $V_h = X_h^1$. If $u \in H^2(\Omega)$ then the error estimate*

$$\|u - u_h\|_{H^1(\Omega)} \leq ch\, |u|_{H^2(\Omega)}$$

holds, with a constant c independent of h.

Proof: From the Céa-lemma 3.6 it follows that the continuous and the discrete problems have unique solutions and that

$$\|u - u_h\|_{H^1(\Omega)} \leq c \inf_{v_h \in X_h^1} \|u - v_h\|_{H^1(\Omega)}$$

holds. Using (3.25) the proof is complete. ◇

In order to estimate the L_2-error we consider a duality argument applied in the Aubin-Nitsche lemma (see e.g. [BS02, Cia78]). For $k(\cdot, \cdot)$ as in (3.2) we define the problem: given any element $g \in L_2(\Omega)$ find $\tilde{u} \in H^1(\Omega)$ such that

$$k(v, \tilde{u}) = \int_\Omega gv \, d\mathbf{x}, \quad \forall v \in H^1(\Omega). \tag{3.26}$$

The problem (3.26) is called the *dual problem* of (3.3). Note, that the arguments in the bilinear form $k(\cdot, \cdot)$ are interchanged. The dual problem is said to be *H^2-regular* if there exists a constant c independent of g such that

$$\|\tilde{u}\|_{H^2(\Omega)} \leq c\, \|g\|_{L_2(\Omega)}. \tag{3.27}$$

Theorem 3.17 *Let u be the solution of* (3.3) *with $u \in H^2(\Omega)$ and u_h the solution of* (3.8) *with $V_h = X_h^1$. Then the dual problem* (3.26) *is H^2-regular and there exists a constant c independent of h such that*

$$\|u - u_h\|_{L_2(\Omega)} \leq ch^2\, |u|_{H^2(\Omega)}.$$

Proof: Since the bilinear form $k(\cdot, \cdot)$ of the reaction-diffusion problem is symmetric, the original and the dual problem only differ in the corresponding right-hand sides. Thus, problem (3.26) has a unique solution for every $g \in L_2(\Omega)$ and it is H^2-regular (compare Remark 3.5). Define $e_h := u - u_h$ and let $\tilde{u} \in H^2(\Omega)$ be the solution of the dual problem

$$k(v, \tilde{u}) = \int_\Omega e_h v \, d\mathbf{x}, \quad \forall v \in H^1(\Omega).$$

Using the orthogonality $k(e_h, v_h) = 0$, $\forall v_h \in X_h^1$, we get

$$\begin{aligned}
\|e_h\|^2_{L_2(\Omega)} &= \int_\Omega e_h^2 \, d\mathbf{x} \\
&= k(e_h, \tilde{u}) = k(e_h, \tilde{u} - I_h\tilde{u}) \\
&\leq c\, \|e_h\|_{H^1(\Omega)} \|\tilde{u} - I_h\tilde{u}\|_{H^1(\Omega)} \qquad (3.28) \\
&\overset{(3.22)}{\leq} ch\, \|e_h\|_{H^1(\Omega)} |\tilde{u}|_{H^2(\Omega)} \\
&\overset{(3.27)}{\leq} ch\, \|e_h\|_{H^1(\Omega)} \|e_h\|_{L_2(\Omega)}, \qquad (3.29)
\end{aligned}$$

which gives the desired statement using Theorem 3.16. ◇

3.2 The instationary boundary value problem

We now consider the time-dependent direct heat conduction problem from Section 2.2 which is given by

$$\begin{aligned} \frac{\partial u}{\partial t} - a\Delta u &= f \quad \text{in } \Omega_{t_f} = \Omega \times (t_0, t_f], \\ u(\cdot, t_0) &= u_0 \quad \text{in } \Omega, \\ -\lambda \frac{\partial u}{\partial \mathbf{n}} &= g \quad \text{on } \Gamma_{t_f} = \partial\Omega \times (t_0, t_f], \end{aligned} \tag{3.30}$$

with $f \in L_2(\Omega_{t_f})$, $u_0 \in L_2(\Omega)$ and $g \in L_2(\Gamma_{t_f})$. For the constant thermal diffusivity we assume $0 < a < \infty$. We deduce the fully discrete approximation scheme in two steps.

3.2.1 The space discretization

First we apply the Galerkin discretization to obtain a semi-discrete solution $u_h(\cdot, t)$ for each point t in time. We define

$$k(u, v) := (a\nabla u, \nabla v), \quad l(v) := (f, v) - \frac{a}{\lambda} \int_\Gamma g v \, d\Gamma.$$

The weak formulation of problem (3.30) is: for $t \in (t_0, t_f]$ find $u(t) \in H^1(\Omega)$ such that

$$\begin{aligned} (\frac{\partial u}{\partial t}, v) + k(u, v) = l(v), \quad \forall v \in H^1(\Omega), \\ u(t_0) = u_0. \end{aligned} \tag{3.31}$$

In order to gain the spatially semi-discrete problem we formulate the approximate solution: for $t \in (t_0, t_f]$ find $u_h(t) \in V_h \subset H^1(\Omega)$ such that

$$\begin{aligned} (\frac{\partial u_h}{\partial t}, v_h) + k(u_h, v_h) = l(v_h), \quad \forall v_h \in V_h, \\ u_h(t_0) = u_h^0, \end{aligned} \tag{3.32}$$

where u_h^0 is some approximation of u_0 in V_h. With $\{\phi_i\}_{1 \le i \le n}$ denoting a basis of V_h we can write

$$u_h(t) = \sum_{j=1}^{n} x_j(t)\phi_j, \quad \text{with } \mathbf{x}(t) := (x_1(t), \ldots, x_n(t))^T \in \mathbb{R}^n.$$

Note that the coefficient vector $\mathbf{x}(t)$ is time-dependent. Substitution of this representation in the Galerkin discretization (3.32) finally leads to the problem: for $t \in (t_0, t_f]$ find $u_h(t) \in V_h \subset H^1(\Omega)$ such that

$$\sum_{j=1}^{n} (\phi_j, \phi_i) \frac{\partial x_j}{\partial t}(t) + \sum_{j=1}^{n} k(\phi_j, \phi_i) x_j(t) = l(\phi_i), \quad \forall i = 1, \ldots, n.$$

Using a matrix-vector notation this is equivalent to solving the system of coupled ordinary differential equations

$$\mathbf{M}\mathbf{x}'(t) + \mathbf{A}\mathbf{x}(t) = \mathbf{b}(t), \quad t \in (t_0, t_f], \quad \mathbf{x}(t_0) = \boldsymbol{\alpha}, \tag{3.33}$$

where the matrix $\mathbf{M} = (M_{ij})_{i,j=1}^n$ with components $M_{ij} = (\phi_j, \phi_i)$ is the so-called *mass matrix* and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ contains the coefficients of the initial approximation u_h^0 with respect to the chosen basis of V_h. The matrix $\mathbf{A}$ denotes the symmetric positive semi-definite stiffness matrix (compare Section 3.1.1). Note that the mass matrix $\mathbf{M}$ is symmetric and strictly positive definite. In order to derive the fully discrete approximation scheme we treat

the time discretization subsequently.

We remark that the entries of the stiffness matrix represent the couplings between the grid points (unknowns) and their neighbouring points. In case of a regular triangulation the number of neighbouring grid points is bounded by a constant which is independent of the triangulation. Thus, the number of nonzero entries in the corresponding stiffness matrix is proportional to the number of unknowns and the matrix is said to be *sparse*. This sparsity property also holds for the mass matrix.

3.2.2 The time discretization

For the time discretization we apply a so-called *one-step θ-scheme* (see e.g. [MM94]). Let the time step size be denoted by τ and let u_h^n be the approximation of $u_h(t)$ at time $t = t_n = n\tau$, $n \geq 0$. In the same manner we define l^n to be the approximation of the time-dependent right-hand side functional $l(t)$ at time $t = t_n$. Given u_h^n we determine the approximate solution $u_h^{n+1} \approx u_h(t_{n+1})$ replacing the time derivative in (3.32) by a backward difference quotient introducing the parameter $0 \leq \theta \leq 1$ to control the implicitness of the scheme. This leads to the variational equation

$$\left(\frac{u_h^{n+1} - u_h^n}{\tau}, v_h\right) + \theta k(u_h^{n+1}, v_h) + (1-\theta)k(u_h^n, v_h) = \theta l^{n+1}(v_h) + (1-\theta)l^n(v_h),$$

and – after a separation of the terms containing u_h^{n+1} and u_h^n – we obtain

$$\begin{aligned}(u_h^{n+1}, v_h) + \theta\tau k(u_h^{n+1}, v_h) \;&= (u_h^n, v_h) + (1-\theta)\tau(l^n(v_h) - k(u_h^n, v_h)) \\ &\quad + \theta\tau l^{n+1}(v_h), \quad \forall v_h \in V_h, \quad n \geq 0.\end{aligned} \tag{3.34}$$

Using the notation introduced for the semi-discrete approximation equation (3.34) may be written as

$$\begin{aligned}[\mathbf{M} + \theta\tau\mathbf{A}]\mathbf{x}^{n+1} \;&= [\mathbf{M} - (1-\theta)\tau\mathbf{A}]\mathbf{x}^n \\ &\quad + \theta\tau\mathbf{b}^{n+1} + (1-\theta)\tau\mathbf{b}^n, \quad n \geq 0, \quad \mathbf{x}^0 = \mathbf{x}(t_0).\end{aligned} \tag{3.35}$$

In particular, $\theta = 0$ yields the explicit Euler scheme, $\theta = 0.5$ the Crank-Nicholson scheme, and $\theta = 1$ leads to the implicit Euler scheme. Because of the strong stiffness of the system in (3.33) only implicit schemes ($\theta \neq 0$) should be used to solve the instationary heat conduction problems. The implicit Euler scheme is only first order accurate in time, but is strongly A-stable. The Crank-Nicholson scheme is of second order, is A-stable, but does not have the strong A-stability property, which may lead to stability problems in certain situations. In the remainder of this thesis we consider $\theta = 0.5$ and $\theta = 1$. For the numerical analysis we study the implicit Euler scheme and give the following result.

Theorem 3.18 *Let u_h^n be the approximation of $u_h(t)$ at time $t = t_n$ with $\theta = 1$ in (3.34) and u the solution of the continuous problem (3.31). If $\|u_0 - u_h^0\|_{L_2(\Omega)} \leq ch^2 \|u_0\|_{H^2(\Omega)}$ then the inequality*

$$\|u(t_n) - u_h^n\|_{L_2(\Omega)} \leq ch^2 \left(\|u_0\|_{H^2(\Omega)} + \int_{t_0}^{t_n} \|\frac{\partial u}{\partial t}\|_{H^2(\Omega)}\, ds\right) + \tau \int_{t_0}^{t_n} \|\frac{\partial^2 u}{\partial t^2}\|_{L_2(\Omega)}\, ds \tag{3.36}$$

holds for $n \geq 0$.

For problem (3.30) with homogeneous Dirichlet boundary conditions a proof of this theorem can be found in [GR94, LT03, Tho97]. In [Tho97] there is a remark that an analysis of Neumann boundary value problems follows the same lines.

Proof: Define $X_{h,0}^1 := X_h^1 \cap H_0^1(\Omega)$ and let $X_{h,0}^1 = V_h \subset V = H_0^1(\Omega)$. We introduce the

elliptic projection R_h from $H_0^1(\Omega)$ onto $X_{h,0}^1$ as the orthogonal projection w.r.t. the inner product $k(\cdot,\cdot)$. For $w \in H_0^1(\Omega)$ it is defined by

$$k(R_h w, v_h) = k(w, v_h), \quad \forall v_h \in X_{h,0}^1. \tag{3.37}$$

In other words, if w is the exact solution of the corresponding elliptic problem, then $R_h w$ is the finite element approximation of w in the space $X_{h,0}^1$. Thus, for regular $\mathcal{P}_1$-elements, using the Céa-lemma 3.6 and the analogon of (3.25) in $X_{h,0}^1$, we can conclude

$$\|v - R_h v\|_{H^m(\Omega)} \leq c h^{2-m} \|v\|_{H^2(\Omega)}, \quad m \in \{0,1\}, \quad \forall v \in H^2(\Omega) \cap H_0^1(\Omega). \tag{3.38}$$

In order to derive the desired error bound we use the representation

$$u(t_n) - u_h^n = (u(t_n) - R_h u(t_n)) + (R_h u(t_n) - u_h^n), \quad n \geq 0,$$

and give separate bounds for the right-hand side terms. We get

$$\begin{aligned} \|u(t_n) - R_h u(t_n)\|_{L_2(\Omega)} &\leq c h^2 \|u(t_n)\|_{H^2(\Omega)} \\ &\leq c h^2 \left(\|u_0\|_{H^2(\Omega)} + \int_{t_0}^{t_n} \|\frac{\partial u}{\partial t}(s)\|_{H^2(\Omega)}\, ds\right). \end{aligned} \tag{3.39}$$

For the estimation of $\rho^n := R_h u(t_n) - u_h^n$ we consider the identity

$$\left(\frac{\rho^{n+1} - \rho^n}{\tau}, v_h\right) + k(\rho^{n+1}, v_h) = (w^n, v_h), \quad \forall v_h \in X_{h,0}^1, \quad n \geq 0, \tag{3.40}$$

where w^n is defined by

$$\begin{aligned} w^n &:= \frac{R_h u(t_{n+1}) - R_h u(t_n)}{\tau} - \frac{\partial u}{\partial t}(t_{n+1}) \\ &= \left(\frac{R_h u(t_{n+1}) - R_h u(t_n)}{\tau} - \frac{u(t_{n+1}) - u(t_n)}{\tau}\right) + \left(\frac{u(t_{n+1}) - u(t_n)}{\tau} - \frac{\partial u}{\partial t}(t_{n+1})\right) \\ &=: w_1^n + w_2^n. \end{aligned}$$

The identity (3.40) can easily be verified by the substitution of ρ^n, using (3.37), (3.31) and (3.34) with $\theta = 1$. If we set $v_h = \rho^{n+1}$ in (3.40), Cauchy-Schwarz leads to the inequality

$$\|\rho^{n+1}\|_{L_2(\Omega)} \leq \|\rho^n\|_{L_2(\Omega)} + \tau \|w^n\|_{L_2(\Omega)},$$

and, by repeated application,

$$\|\rho^{n+1}\|_{L_2(\Omega)} \leq \|\rho^0\|_{L_2(\Omega)} + \tau \sum_{l=1}^{n} \|w^l\|_{L_2(\Omega)}, \quad n \geq 0.$$

Note that

$$\|\rho^0\|_{L_2(\Omega)} \leq \|R_h u(t_0) - u_0\|_{L_2(\Omega)} + \|u_0 - u_h^0\|_{L_2(\Omega)} \leq c h^2 \|u_0\|_{H^2(\Omega)},$$

using (3.39) and the assumption of the theorem with respect to the initial value approximation. It remains to give adequate bounds for $\|w^l\|_{L_2(\Omega)}$. We note that the operator R_h commutes with the time differentiation [Tho97] and conclude

$$\tau \|w_1^l\|_{L_2(\Omega)} = \|\int_{t_l}^{t_{l+1}} (R_h \frac{\partial u}{\partial t}(s) - \frac{\partial u}{\partial t}(s))\, ds\|_{L_2(\Omega)} \overset{(3.38)}{\leq} c h^2 \int_{t_l}^{t_{l+1}} \|\frac{\partial u}{\partial t}(s)\|_{H^2(\Omega)}\, ds.$$

Using integration by parts, we get the identity

$$\int_{t_l}^{t_{l+1}} (t_l - s) \frac{\partial^2 u}{\partial t^2}(s)\, ds = u(t_{l+1}) - u(t_l) - \tau \frac{\partial u}{\partial t}(t_{l+1}) = \tau w_2^l.$$

Thus, we have

$$\tau\|w_2^l\|_{L_2(\Omega)} = \|\int_{t_l}^{t_{l+1}} (t_l - s)\frac{\partial^2 u}{\partial t^2}(s)\,ds\|_{L_2(\Omega)} \leq \tau \int_{t_l}^{t_{l+1}} \|\frac{\partial^2 u}{\partial t^2}(s)\|_{L_2(\Omega)}\,ds.$$

Finally, summation over l completes our considerations. ◇

From this theorem we can conclude that for a sufficiently smooth solution u of problem (3.31) the discretization error using piecewise linear finite elements and the implicit Euler scheme is of second order in space and first order in time, i.e. the error is of order $\mathcal{O}(h^2+\tau)$. Analogously, one obtains an $\mathcal{O}(h^2+\tau^2)$ convergence behaviour for the Crank-Nicholson scheme. Note, that the approximation error bound (3.25) is crucial for the derivation of the first term on the right-hand side of (3.36) containing the discretization parameter h. For the gradient approximation using the implicit Euler scheme one can prove

$$|u(t_n) - u_h^n|_{H^1(\Omega)} = \mathcal{O}(h+\tau), \quad n \geq 0,$$

using (3.38) for the estimation of $|u(t_n) - R_h u(t_n)|_{H^1(\Omega)}$ (see [Tho97] for more details).

Remark 3.19 When the time integration scheme is applied first, we obtain a linear elliptic reaction-diffusion problem of the form (3.2)-(3.3) with $\mu := 1/\tau\theta a > 0$ in each time step. We restrict our investigations (of degenerated finite elements) to this linear elliptic problem class. Note that the parameter μ measures the size of the reaction term relative to that of the diffusion term. The subsequent Galerkin discretization of the corresponding reaction-diffusion problem results in the variational formulation (3.34), too.

3.3 Solution methods for the discrete problems

For an introduction of iterative solution methods we consider the *isotropic* reaction-diffusion problem (3.2)-(3.3) with $\mu = \mathcal{O}(1)$. The Galerkin discretization of this problem with regular (linear) finite elements results in a linear system of equations

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{3.41}$$

with a regular coefficient matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and the right-hand side $\mathbf{b} \in \mathbb{R}^n$. Note that the diagonal entries of $\mathbf{A}$ are nonzero since this matrix is symmetric positive definite. Let the unique solution of (3.41) be denoted by $\mathbf{x}^*$.

3.3.1 Basic iterative solution methods

In this section we introduce some of the classical iterative solution methods for the approximate solution of large (sparse) systems of equations (see e.g. [Reu05, Saa03]). These methods are rather inefficient and thus not very suitable when applied to (3.41). However, the point relaxation methods that we introduce will be needed in Section 3.3.3 as smoothers within multigrid methods.

Using a further regular matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ we can rewrite (3.41) in the form

$$\mathbf{M}\mathbf{x} = (\mathbf{M} - \mathbf{A})\mathbf{x} + \mathbf{b}.$$

The latter representation leads to the frequently used basic model

$$\mathbf{M}\mathbf{x}^{k+1} = (\mathbf{M} - \mathbf{A})\mathbf{x}^k + \mathbf{b}, \quad k \geq 0, \quad \mathbf{x}^0 \in \mathbb{R}^n \text{ starting vector}, \tag{3.42}$$

for the iterative solution of (3.41). Here the matrix $\mathbf{M}$ is assumed to be such that the linear systems $\mathbf{M}\mathbf{y} = \mathbf{c}$ can be solved with relatively low arithmetic costs for an arbitrarily chosen right-hand side $\mathbf{c} \in \mathbb{R}^n$. For the error $\mathbf{e}^k := \mathbf{x}^k - \mathbf{x}^*$, using (3.42), we get

$$\mathbf{e}^{k+1} = (\mathbf{I} - \mathbf{M}^{-1}\mathbf{A})\mathbf{e}^k$$

and thus we have a linear iterative method with iteration matrix

$$\mathbf{C} = \mathbf{I} - \mathbf{M}^{-1}\mathbf{A}.$$

Let $\sigma(\mathbf{A})$ denote the *spectrum* of $\mathbf{A}$, i.e. the set of all of its eigenvalues, and $\rho(\mathbf{A})$ the *spectral radius*, which is defined by

$$\rho(\mathbf{A}) := \max\{|\lambda| \mid \lambda \in \sigma(\mathbf{A})\}.$$

We give an elementary lemma that can be found e.g. in [Saa03].

Lemma 3.20 *A linear iterative method is convergent for an arbitrary starting vector* $\mathbf{x}^0 \in \mathbb{R}^n$ *if and only if for the corresponding iteration matrix we have*

$$\rho(\mathbf{C}) < 1.$$

In the following we introduce some basic point iteration schemes which are based on the additive splitting of the coefficient matrix

$$\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U} \tag{3.43}$$

into the strictly lower triangular part $-\mathbf{L}$, the nonsingular diagonal matrix $\mathbf{D} := \operatorname{diag}(\mathbf{A})$, and the strictly upper triangular part $-\mathbf{U}$.

The Jacobi method

The Jacobi iteration is defined by setting

$$\mathbf{M}_{\text{Jac}} := \mathbf{D}$$

in (3.42) and thus this method can be written in the form

$$\begin{cases} \mathbf{x}^0 \text{ given starting vector,} \\ \mathbf{D}\mathbf{x}^{k+1} = (\mathbf{L} + \mathbf{U})\mathbf{x}^k + \mathbf{b}, \quad k \geq 0. \end{cases}$$

Note that for the evaluation of the new approximation x_i^{k+1} only values x_j^k, $j \neq i$, of the previous approximation are used. The corresponding iteration matrix is given by

$$\mathbf{C}_{\text{Jac}} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}.$$

A variant of the Jacobi method, in which a real damping parameter $\omega \neq 0$ is used, is given by

$$\mathbf{M}_{\omega\,\text{Jac}} := \frac{1}{\omega}\mathbf{D}, \quad \omega \in (0, 1].$$

For $\omega = 1$ we obtain the original Jacobi method while for $0 < \omega < 1$ we get the *damped Jacobi method*.

The Gauss-Seidel method

The Gauss-Seidel method is given by setting

$$\mathbf{M}_{\text{GS}} := \mathbf{D} - \mathbf{L}$$

in (3.42) leading to the representation

$$\begin{cases} \mathbf{x}^0 \text{ given starting vector,} \\ (\mathbf{D} - \mathbf{L})\mathbf{x}^{k+1} = \mathbf{U}\mathbf{x}^k + \mathbf{b}, \quad k \geq 0. \end{cases} \tag{3.44}$$

Thus, in contrast to the Jacobi method, for the evaluation of x_i^{k+1} the new approximation x_j^{k+1}, $j < i$, is already used. The iteration matrix of the Gauss-Seidel relaxation is given by

$$\mathbf{C}_{\text{GS}} = \mathbf{I} - (\mathbf{D} - \mathbf{L})^{-1}\mathbf{A}.$$

In many cases the convergence rate of the Gauss-Seidel iteration is at least as high as for the Jacobi relaxation. Thus, in general the Gauss-Seidel method is preferred. It is important to note that the results of the Gauss-Seidel iteration depend on the ordering of the unknowns while this is not the case for the Jacobi relaxation. Again there exists a damped version of the Gauss-Seidel iteration which is called *successive overrelaxation method* (SOR) and is described by

$$\mathbf{M}_{\mathrm{SOR}} := \frac{1}{\omega}\mathbf{D} - \mathbf{L}, \quad \omega > 0.$$

The symmetric Gauss-Seidel method

One symmetric Gauss-Seidel iteration consists of two Gauss-Seidel steps. In the first step an original Gauss-Seidel iteration as in (3.44) is applied. In the second step again one Gauss-Seidel iteration is applied but with the reversed ordering of the unknowns. Thus, the symmetric Gauss-Seidel method is described by

$$\begin{cases} \mathbf{x}^0 \text{ given starting vector,} \\ x_i^{k+\frac{1}{2}} = \left(-\sum_{j<i} a_{ij} x_j^{k+\frac{1}{2}} - \sum_{j>i} a_{ij} x_j^k + b_i\right)/a_{ii}, \qquad i = 1, 2, \ldots, n, \\ x_i^{k+1} = \left(-\sum_{j<i} a_{ij} x_j^{k+1} - \sum_{j>i} a_{ij} x_j^{k+\frac{1}{2}} + b_i\right)/a_{ii}, \quad i = n, n-1, \ldots, 1, \quad k \geq 0. \end{cases}$$

This relaxation method results if we set

$$\mathbf{M}_{\mathrm{SGS}} := (\mathbf{D} - \mathbf{L})\mathbf{D}^{-1}(\mathbf{D} - \mathbf{U})$$

in (3.42). If we apply the above described procedure to the SOR scheme we obtain the so-called *symmetric successive overrelaxation method* (SSOR) which corresponds to the matrix splitting

$$\mathbf{M}_{\mathrm{SSOR}} := \frac{1}{\omega(2-\omega)}(\mathbf{D} - \omega\mathbf{L})\mathbf{D}^{-1}(\mathbf{D} - \omega\mathbf{U}), \quad \omega > 0.$$

Note that for symmetric positive definite $\mathbf{A}$ the matrix $\mathbf{M}_{\mathrm{SSOR}}$ has this property, too, and thus it can be used as a preconditioner for the CG method (see Section 3.3.2).

Block iterative methods

If the matrix $\mathbf{A}$ has a block structure – with square diagonal blocks which are assumed to be nonsingular – and the splitting in (3.43) is based on block-matrices we obtain the so-called *block-variants* of the above introduced *point relaxation schemes*. More precisely, we use the block-diagonal $\mathbf{D}$, the lower block-triangular matrix $\mathbf{L}$, and the upper block-triangular matrix $\mathbf{U}$. In this way whole sets of unknowns (associated with the blocks) are updated simultaneously within the iteration while for the point relaxation schemes this is done separately for each unknown.

3.3.2 Krylov subspace methods

In this section we give a brief introduction to Krylov subspace methods with focus on the conjugate gradient (CG) method and its preconditioned version (PCG). For a more detailed analysis of these methods using the so-called polynomial based approach we refer to [Saa03, vdV03].

The conjugate gradient method

Following the lines of [AB84] we introduce the CG method as a method which minimizes the quadratic functional

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x} + \mathbf{c} \tag{3.45}$$

over $\mathbf{x} \in \mathbb{R}^n$. For this functional we can conclude

$$DF(\mathbf{x}) = \nabla F(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b} \quad \text{and} \quad D^2F(\mathbf{x}) = \mathbf{A},$$

with a symmetric positive definite matrix $\mathbf{A}$ and thus the minimization of the quadratic functional F in (3.45) is equivalent to solving the linear system $\mathbf{Ax} = \mathbf{b}$. The functional minimization is accomplished as already described in Section 2.3.2 for the solution of the IHCP. For a given initial approximation the next approximate solution is given by

$$\begin{cases} \mathbf{x}^0 \text{ given starting vector,} \\ \mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_{\text{opt}}(\mathbf{x}^k, \mathbf{p}^k)\mathbf{p}^k, \quad k \geq 0, \end{cases} \tag{3.46}$$

where $\mathbf{p}^k \neq 0$ is the search direction at $\mathbf{x}^k$ and $\alpha_{\text{opt}}(\mathbf{x}^k, \mathbf{p}^k)$ is the optimal steplength in this direction. We use the notation

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^T\mathbf{x}, \quad \langle \mathbf{x}, \mathbf{y} \rangle_A = \mathbf{y}^T\mathbf{Ax}, \quad \|\mathbf{x}\|_A = \langle \mathbf{x}, \mathbf{x} \rangle_A^{\frac{1}{2}}.$$

Since $-\infty < \alpha_{\text{opt}} < \infty$ minimizes the function $F(\mathbf{x}^k + \alpha_{\text{opt}}\mathbf{p}^k)$ it is determined by the expression

$$\alpha_{\text{opt}}(\mathbf{x}^k, \mathbf{p}^k) := -\frac{\langle \mathbf{p}^k, \mathbf{r}^k \rangle}{\langle \mathbf{p}^k, \mathbf{Ap}^k \rangle},$$

with $\mathbf{r}^k := \mathbf{Ax}^k - \mathbf{b} = \nabla F(\mathbf{x}^k)$. For the residuals we have

$$\mathbf{r}^{k+1} = \mathbf{r}^k + \alpha_{\text{opt}}(\mathbf{x}^k, \mathbf{p}^k)\mathbf{Ap}^k, \quad k \geq 0. \tag{3.47}$$

In order to explain the idea behind the construction of the search directions we define the notion:

$$\mathbf{x} \text{ is } \textit{optimal} \text{ w.r.t. the direction } \mathbf{p} \quad \Leftrightarrow \quad \langle \mathbf{p}, \mathbf{Ax} - \mathbf{b} \rangle = 0. \tag{3.48}$$

Note that the term on the right-hand side of equivalence (3.48) means that $\alpha_{\text{opt}}(\mathbf{x}, \mathbf{p}) = 0$. In the conjugate gradient method the search directions are chosen such that $\mathbf{x}^k$ is optimal w.r.t. the subspace

$$V_k := \text{span}\{\mathbf{p}^0, \ldots, \mathbf{p}^{k-1}\}$$

and not – as it is the case for the steepest descent method – only w.r.t the subspace $\text{span}\{\mathbf{p}^{k-1}\}$. For the construction of these directions we remark that at $\mathbf{x}^0$ the steepest descent direction of the functional F is determined by setting $\mathbf{p}^0 = \mathbf{r}^0 = \nabla F(\mathbf{x}^0)$. Using (3.47) we can conclude that $\mathbf{x}^1$ from (3.46) is optimal w.r.t. $\text{span}\{\mathbf{p}^0\}$. Now we assume that search directions $\mathbf{p}^0, \ldots, \mathbf{p}^{k-1}$ are given such that $\mathbf{x}^k$ is optimal w.r.t. V_k and $\mathbf{r}^k \neq 0$. A unique new search direction $\mathbf{p}^k$ is given by the *A–orthogonal* (or *conjugate*) projection of the residual $\mathbf{r}^k$ on $V_k^{\perp_A}$, i.e.

$$\mathbf{p}^k \in V_k^{\perp_A} \text{ such that } \|\mathbf{p}^k - \mathbf{r}^k\|_A = \min_{\mathbf{p} \in V_k^{\perp_A}} \|\mathbf{p} - \mathbf{r}^k\|_A. \tag{3.49}$$

This choice of the new search direction leads to the formula

$$\mathbf{p}^k = \mathbf{r}^k - \sum_{j=0}^{k-1} \frac{\langle \mathbf{p}^j, \mathbf{r}^k \rangle_A}{\langle \mathbf{p}^j, \mathbf{p}^j \rangle_A} \mathbf{p}^j = \mathbf{r}^k - \sum_{j=0}^{k-1} \frac{\langle \mathbf{p}^j, \mathbf{Ar}^k \rangle}{\langle \mathbf{p}^j, \mathbf{Ap}^j \rangle} \mathbf{p}^j. \tag{3.50}$$

One can easily show that the new iterand $\mathbf{x}^{k+1}$ is optimal w.r.t. V_{k+1} for $\mathbf{p}^k$ defined by (3.49). Before we give the resulting CG algorithm we consider some properties of the method (compare [Reu05]).

Lemma 3.21 *Let $\mathbf{x}^0 \in \mathbb{R}^n$ be given and $m < n$ be such that for $k = 0, 1, \ldots, m$ we have $\mathbf{r}^k \neq 0$ and $\mathbf{x}^{k+1}, \mathbf{p}^k$, as in* (3.46), (3.50). *Then the following holds for all $k = 1, \ldots, m+1$:*

$$\dim(V_k) = k, \tag{3.51a}$$
$$\mathbf{x}^k \in \mathbf{x}^0 + V_k, \tag{3.51b}$$
$$F(\mathbf{x}^k) = \min\{F(\mathbf{x}) \mid \mathbf{x} \in \mathbf{x}^0 + V_k\}, \tag{3.51c}$$
$$V_k = \text{span}\{\mathbf{r}^0, \ldots, \mathbf{r}^{k-1}\} = \text{span}\{\mathbf{r}^0, \mathbf{Ar}^0, \ldots, \mathbf{A}^{k-1}\mathbf{r}^0\}, \tag{3.51d}$$
$$\langle \mathbf{p}^j, \mathbf{r}^k \rangle = 0, \quad \textit{for all } j = 0, 1, \ldots, k-1, \tag{3.51e}$$
$$\langle \mathbf{r}^j, \mathbf{r}^k \rangle = 0, \quad \textit{for all } j = 0, 1, \ldots, k-1, \tag{3.51f}$$
$$\mathbf{p}^k \in \text{span}\{\mathbf{r}^k, \mathbf{p}^{k-1}\} \quad (k \leq m). \tag{3.51g}$$

The subspace

$$V_k = \text{span}\{\mathbf{r}^0, \mathbf{A}\mathbf{r}^0, \ldots, \mathbf{A}^{k-1}\mathbf{r}^0\} =: \mathcal{K}^k(\mathbf{A}; \mathbf{r}^0)$$

is called the *Krylov subspace* of dimension k corresponding to $\mathbf{r}^0$. Due to statement (3.51g) of the previous lemma only the last summand in (3.50) is nonzero, i.e.

$$\mathbf{p}^k = \mathbf{r}^k - \frac{\langle \mathbf{p}^{k-1}, \mathbf{A}\mathbf{r}^k \rangle}{\langle \mathbf{p}^{k-1}, \mathbf{A}\mathbf{p}^{k-1} \rangle} \mathbf{p}^{k-1}.$$

Using the orthogonality properties (3.51e) and (3.51f) we can derive the equivalent representations

$$\alpha_{\text{opt}}(\mathbf{x}^k, \mathbf{p}^k) = -\frac{\langle \mathbf{r}^k, \mathbf{r}^k \rangle}{\langle \mathbf{p}^k, \mathbf{A}\mathbf{p}^k \rangle}, \qquad \mathbf{p}^k = \mathbf{r}^k + \frac{\langle \mathbf{r}^k, \mathbf{r}^k \rangle}{\langle \mathbf{r}^{k-1}, \mathbf{r}^{k-1} \rangle} \mathbf{p}^{k-1}, \tag{3.52}$$

for the optimal steplength and the search direction and thus we finally obtain the following CG algorithm.

Algorithm 3.22

choose $\mathbf{x}^0$;
$\mathbf{p}^0 := \mathbf{r}^0 := \mathbf{A}\mathbf{x}^0 - \mathbf{b}$;
for $k = 0$ to k_{max} do
  $\alpha_{\text{opt}}(\mathbf{x}^k, \mathbf{p}^k) = -\frac{\langle \mathbf{p}^k, \mathbf{r}^k \rangle}{\langle \mathbf{p}^k, \mathbf{A}\mathbf{p}^k \rangle}$;
  $\mathbf{x}^{k+1} := \mathbf{x}^k + \alpha_{\text{opt}}(\mathbf{x}^k, \mathbf{p}^k)\mathbf{p}^k$;
  if accurate then exit;
  $\mathbf{r}^{k+1} = \mathbf{r}^k + \alpha_{\text{opt}}(\mathbf{x}^k, \mathbf{p}^k)\mathbf{A}\mathbf{p}^k$;
  $\mathbf{p}^{k+1} = \mathbf{r}^{k+1} + \frac{\langle \mathbf{r}^{k+1}, \mathbf{r}^{k+1} \rangle}{\langle \mathbf{r}^k, \mathbf{r}^k \rangle} \mathbf{p}^k$;
end for

CG algorithm in matrix-vector notation

We note that in the above presented CG algorithm the recursive choice of the search directions is done so that at each step the approximate solution has smallest error $\|\mathbf{x}^k - \mathbf{x}^*\|_A$ if we minimize over $\mathbf{x}^k \in \mathbf{x}^0 + \mathcal{K}^k(\mathbf{A}; \mathbf{r}^0)$. When the algorithm is applied to the normal equation then the error norm

$$\|\mathbf{x}^k - \mathbf{x}^*\|_{A^T A} = \|\mathbf{A}\mathbf{x}^k - \mathbf{b}\|_2 =: J(\mathbf{x}^k)$$

is minimized in each step. In the continuous setting with $\mathcal{A}$ (or $\mathcal{A}_S$) denoting a linear operator we obtain the CGNE algorithm from Section 2.3.2. In contrast to the basic iterative methods presented in Section 3.3.1 the CG method determines the exact solution in at most n iterations due to (3.51a), (3.51c), and it is nonlinear.

Before we discuss a main result from the convergence analysis of the CG method we briefly give an idea of the polynomial based introduction of CG (cf. [vdV03]). The iteration scheme (3.42) can be written in the form

$$\begin{aligned} \mathbf{x}^{k+1} &= \mathbf{x}^k - \mathbf{M}^{-1}(\mathbf{A}\mathbf{x}^k - \mathbf{b}) \\ &= \mathbf{x}^k - (\tilde{\mathbf{A}}\mathbf{x}^k - \tilde{\mathbf{b}}) =: \mathbf{x}^k - \tilde{\mathbf{r}}^k, \end{aligned} \tag{3.53}$$

with $\tilde{\mathbf{A}} := \mathbf{M}^{-1}\mathbf{A}$ and $\tilde{\mathbf{b}} := \mathbf{M}^{-1}\mathbf{b}$. From (3.53) we get

$$\tilde{\mathbf{r}}^{k+1} = (\mathbf{I} - \tilde{\mathbf{A}})\tilde{\mathbf{r}}^k = (\mathbf{I} - \tilde{\mathbf{A}})^{k+1}\tilde{\mathbf{r}}^0 =: p_{k+1}(\tilde{\mathbf{A}})\tilde{\mathbf{r}}^0,$$

where p_{k+1} is a polynomial of degree $k+1$ with $p_{k+1}(0) = 1$. For the error we have

$$\mathbf{x}^{k+1} - \mathbf{x}^* = p_{k+1}(\tilde{\mathbf{A}})(\mathbf{x}^0 - \mathbf{x}^*).$$

Thus, the error reduction depends on how well the polynomial p_{k+1} damps the initial error components. Note that for the basic iterative methods of Section 3.3.1 this polynomial is the $(k+1)$-th power of the iteration matrix. In the CG method (applied to the original problem, i.e. $\mathbf{M} = \mathbf{I}$,) the approximate solution $\mathbf{x}^{k+1}$ is constructed such that the A–norm error is minimal over all polynomials of degree $k+1$, with $p_{k+1}(0) = 1$. The next lemma can be found in [AB84].

Lemma 3.23 *Define* $\mathcal{P}_k^* := \{p \in \mathcal{P}_k \mid p(0) = 1\}$. *Let further* $\mathbf{x}^k, k \geq 0$, *denote the iterands of the CG method and* $\mathbf{e}^k = \mathbf{x}^k - \mathbf{x}^*$ *the error. Then we have*

$$\begin{aligned} \|\mathbf{e}^k\|_A &= \min_{p_k \in \mathcal{P}_k^*} \|p_k(\mathbf{A})\mathbf{e}^0\|_A \\ &\leq \min_{p_k \in \mathcal{P}_k^*} \max_{\lambda \in \sigma(\mathbf{A})} |p_k(\lambda)| \|\mathbf{e}^0\|_A \qquad (3.54) \\ &\leq 2\left(\frac{\sqrt{\mathrm{cond}_2(\mathbf{A})} - 1}{\sqrt{\mathrm{cond}_2(\mathbf{A})} + 1}\right)^k \|\mathbf{e}^0\|_A. \qquad (3.55) \end{aligned}$$

The upper bound (3.55) is obtained from the selection of an appropriate polynomial in (3.54) (the suitably scaled Chebyshev polynomial of degree k). In [AB84] it is shown that this bound can be improved for certain distributions (clusters) of the eigenvalues. Thus, one is interested (w.r.t. preconditioning) in having a small spectral condition number of the system matrix, or, in having eigenvalue distributions with well-separated (clusters of) eigenvalues.

The preconditioned conjugate gradient method

In the preconditioned conjugate gradient algorithm (PCG) we need a symmetric positive definite preconditioner $\mathbf{W} \approx \mathbf{A}$ which should be such that a system with this matrix can be solved with low computational costs. If we use a basic iterative method for preconditioning we can take $\mathbf{W} := \mathbf{M}_{\mathrm{SSOR}}$ with an appropriate damping parameter. The solution of $\mathbf{M}_{\mathrm{SSOR}}\mathbf{x} = \mathbf{y}$ results from one step of the SSOR iteration with $\mathbf{x}^0 = 0$ and $\mathbf{b} = \mathbf{y}$. One can show that the higher the rate of convergence of the iterative method (3.42), the better the quality of $\mathbf{M}$ as a preconditioner of $\mathbf{A}$.

Algorithm 3.24

choose $\mathbf{x}^0$;
$\mathbf{p}^0 := \mathbf{r}^0 := \mathbf{A}\mathbf{x}^0 - \mathbf{b}$;
for $k = 0$ to k_{max} do
  solve $\mathbf{z}^k$ from $\mathbf{W}\mathbf{z}^k = \mathbf{r}^k$;
  $\alpha_{\mathrm{opt}}(\mathbf{x}^k, \mathbf{p}^k) = -\frac{\langle \mathbf{z}^k, \mathbf{r}^k \rangle}{\langle \mathbf{p}^k, \mathbf{A}\mathbf{p}^k \rangle}$;
  $\mathbf{x}^{k+1} := \mathbf{x}^k + \alpha_{\mathrm{opt}}(\mathbf{x}^k, \mathbf{p}^k)\mathbf{p}^k$;
  if accurate then exit;
  $\mathbf{r}^{k+1} = \mathbf{r}^k + \alpha_{\mathrm{opt}}(\mathbf{x}^k, \mathbf{p}^k)\mathbf{A}\mathbf{p}^k$;
  $\mathbf{p}^{k+1} = \mathbf{z}^{k+1} + \frac{\langle \mathbf{z}^{k+1}, \mathbf{r}^{k+1} \rangle}{\langle \mathbf{z}^k, \mathbf{r}^k \rangle}\mathbf{p}^k$;
end for

PCG algorithm in matrix-vector notation

3.3.3 Multigrid methods

In this section we give a brief introduction and some fundamental results concerning multigrid methods for solving scalar elliptic boundary value problems. The main ideas of multigrid methods are described. The approach of analyzing the convergence of a multigrid method based on the approximation and smoothing property (introduced by Hackbusch [Hac85]) is discussed. Finally, the method of nested iteration is explained. For a convergence analysis of a multigrid algorithm for solving an *anisotropic* reaction-diffusion problem we refer to Chapter 5.

Introduction

We introduce the *two-grid* and *multigrid* algorithms and discuss the related main ideas. One of the characteristic properties of these methods is the use of discretizations of the given continuous boundary value problem on two or several grids with different mesh sizes. We consider the finite element discretization of the isotropic reaction-diffusion model problem. The continuous variational formulation of this problem is: find $u \in H^1(\Omega)$ such that

$$k(u,v) = l(v) \quad \text{for all} \quad v \in H^1(\Omega), \tag{3.56}$$

where the bilinear form and the right-hand side functional are defined by (3.2). For the discretization we use standard conforming finite elements on a *nested* family of grids. Let this family of nested triangulations be denoted by $\{\mathcal{T}_k\}_{k\geq 0}$. With $\mathcal{T}_k$ we associate the mesh size parameter h_k. For the corresponding hierarchy of finite element spaces $U_k := X^l_{h_k}$, $l \geq 0$, (compare (3.10)) we have

$$U_k \subset U_{k+1} \qquad \text{for all } k.$$

More details about the Galerkin discretization can be found in Section 3.1.1. The standard nodal basis $\{\phi_i\}_{1\leq i\leq n_k}$ in these finite element spaces (of dimension n_k) induces the isomorphism

$$P_k : X_k := \mathbb{R}^{n_k} \to U_k, \quad P_k\mathbf{x} = \sum_{i=1}^{n_k} x_i\phi_i.$$

The discrete problem on level k is: find $u_k \in U_k$ such that

$$k(u_k, v_k) = l(v_k) \quad \text{for all} \quad v_k \in U_k,$$

which can be represented as a linear system of equations

$$\mathbf{A}_k\mathbf{x}_k = \mathbf{b}_k. \tag{3.57}$$

Remark 3.25 When the hierarchy of nested triangulations is *uniform*, we can give an explicit representation of an orthogonal eigenvector basis of the stiffness matrix $\mathbf{A}_{FD,k}$ obtained by a finite difference discretization of problem (3.56). In Section 5.3 we treat the anisotropic case. Let $\Omega \subset \mathbb{R}^3$ be the unit cube, $N_k := h_k^{-1} - 1$ and ν, η, $\xi \in \mathbb{N}_0$. For the isotropic pure diffusion problem (with $\mu = 0$) the discrete *Fourier modes* on $\mathcal{T}_k$, which are eigenvectors of $\mathbf{A}_{FD,k}$, are given by

$$\mathbf{e}^{\nu\eta\xi}(x,y,z) = \cos(\nu\pi x)\cos(\eta\pi y)\cos(\xi\pi z), \quad (x,y,z) \in \mathcal{T}_k, \quad 0 \leq \nu,\, \eta,\, \xi \leq N_k + 1.$$

These vectors do not change if we consider the case with reaction, too (since in that case $\mu > 0$ times the identity is added to the stiffness matrix of the pure diffusion problem). We define the frequency $f := \max\{\nu, \eta, \xi\}$. If for this frequency we have $f < \frac{1}{2}N_k$ then it is called *low* frequency and for $f \geq \frac{1}{2}N_k$ it is called *high* frequency.

Without going into more detail here we remark that the damped Jacobi method is not adequate when applied as a solver for (3.57). However, if we use an appropriate damping parameter then the error is *smoothed* by this basic iterative method, i.e. the high frequency

modes are strongly damped by the iteration matrix $\mathbf{C}_{\omega\,\mathrm{Jac}}$. Before we discuss the two-grid algorithm we introduce some additional notation.

For the transfer between two different grids we introduce the canonical *prolongations* and *restrictions*

$$\mathbf{p}_k \;:\; X_{k-1} \to X_k, \quad \mathbf{p}_k = P_k^{-1} P_{k-1}, \tag{3.58}$$
$$\mathbf{r}_k \;:\; X_k \to X_{k-1}, \quad \mathbf{r}_k = \mathbf{p}_k^T. \tag{3.59}$$

The prolongation is based on the identity operator between the corresponding nested finite element spaces while the restriction is the one and only which satisfies the Galerkin condition

$$\mathbf{r}_k \mathbf{A}_k \mathbf{p}_k = \mathbf{A}_{k-1}. \tag{3.60}$$

For smoothing we use a linear iterative method of the form

$$\mathbf{x}^{new} = \mathbf{x}^{old} - \mathbf{M}_k^{-1}(\mathbf{A}_k \mathbf{x}^{old} - \mathbf{b}_k) =: \mathcal{S}_k(\mathbf{x}^{old}, \mathbf{b}_k).$$

The basic idea of the *two-grid* algorithm is the following. If we have an approximation of the exact solution on a fine grid and the corresponding error is sufficiently smooth, then this error can be approximated on a coarser grid. Solving the so-called *coarse grid* problem for the error we get a new approximation on the fine grid. Let $\mathbf{x}_k^*$ denote the exact solution of (3.57) and $\bar{\mathbf{x}}_k$ the result of ν_1 (pre-)smoothing iterations applied to an initial vector $\mathbf{x}_k^0$. For the error $\mathbf{e}_k := \bar{\mathbf{x}}_k - \mathbf{x}_k^*$ we get the so-called *defect*

$$\mathbf{A}_k \mathbf{e}_k = \mathbf{A}_k \bar{\mathbf{x}}_k - \mathbf{b}_k =: \mathbf{d}_k. \tag{3.61}$$

If $\mathbf{e}_k$ is sufficiently smooth we can make the approximation $\mathbf{e}_k \approx \mathbf{p}_k \tilde{\mathbf{e}}_{k-1}$ with an appropriate vector $\tilde{\mathbf{e}}_{k-1}$. Using (3.60) and (3.61) for the latter vector we get the linear system

$$\mathbf{A}_{k-1} \tilde{\mathbf{e}}_{k-1} = \mathbf{r}_k \mathbf{d}_k. \tag{3.62}$$

Finally, we take $\mathbf{x}_k := \bar{\mathbf{x}}_k - \mathbf{p}_k \tilde{\mathbf{e}}_{k-1}$ for the new iterand and thus we obtain the two-grid algorithm. If ν_2 smoothing iterations are applied after the coarse grid correction this is called post-smoothing.

Algorithm 3.26

procedure $\mathrm{TGM}_k(\mathbf{x}_k, \mathbf{b}_k)$
  if $k = 0$ then $\mathbf{x}_0 := \mathbf{A}_0^{-1}\mathbf{b}_0$ else
  begin
    $\mathbf{x}_k := \mathcal{S}_k^{\nu_1}(\mathbf{x}_k, \mathbf{b}_k)$; (pre-smoothing)
    $\mathbf{d}_{k-1} := \mathbf{r}_k(\mathbf{A}_k \mathbf{x}_k - \mathbf{b}_k)$; (defect restriction)
    $\tilde{\mathbf{e}}_{k-1} := \mathbf{A}_{k-1}^{-1} \mathbf{d}_{k-1}$; (coarse grid solution)
    $\mathbf{x}_k := \mathbf{x}_k - \mathbf{p}_k \tilde{\mathbf{e}}_{k-1}$; (correction)
    $\mathbf{x}_k := \mathcal{S}_k^{\nu_2}(\mathbf{x}_k, \mathbf{b}_k)$; (post-smoothing)
    $\mathrm{TGM}_k := \mathbf{x}_k$;
  end;

The two-grid algorithm in pseudocode notation

We observe that the coarse grid system (3.62) is of the same form as the original system (3.57). For this reason the above described two-grid algorithm can be used recursively for the approximate solution of (3.62). This idea leads to the *multigrid* algorithm. We only consider the choices $\tau = 1$ (V-cycle) and $\tau = 2$ (W-cycle) in the recursive call for the coarse grid solution.

Algorithm 3.27

procedure $\mathrm{MGM}_k(\mathbf{x}_k, \mathbf{b}_k)$
  if $k = 0$ then $\mathbf{x}_0 := \mathbf{A}_0^{-1}\mathbf{b}_0$ else
  begin
    $\mathbf{x}_k := \mathcal{S}_k^{\nu_1}(\mathbf{x}_k, \mathbf{b}_k)$;
    $\mathbf{d}_{k-1} := \mathbf{r}_k(\mathbf{A}_k\mathbf{x}_k - \mathbf{b}_k)$;
    $\tilde{\mathbf{e}}_{k-1}^0 := \mathbf{0}$; for $i = 1$ to τ do $\tilde{\mathbf{e}}_{k-1}^i := \mathrm{MGM}_{k-1}(\tilde{\mathbf{e}}_{k-1}^{i-1}, \mathbf{d}_{k-1})$;
    $\mathbf{x}_k := \mathbf{x}_k - \mathbf{p}_k\tilde{\mathbf{e}}_{k-1}$;
    $\mathbf{x}_k := \mathcal{S}_k^{\nu_2}(\mathbf{x}_k, \mathbf{b}_k)$;
    $\mathrm{MGM}_k := \mathbf{x}_k$;
  end;

The multigrid algorithm in pseudocode notation

Convergence analysis

We analyze the convergence of the multigrid W-cycle in the framework of the approximation and smoothing property (compare [Hac85, Reu05]). From this analysis one can see that the two-grid convergence directly implies the convergence of the multigrid W-cycle. However, for the multigrid V-cycle a more involved analysis is needed.

It can easily be shown that the two-grid method is a linear iterative method with iteration matrix

$$\mathbf{C}_{TG,k}(\nu_2, \nu_1) = \mathbf{S}_k^{\nu_2}(\mathbf{I} - \mathbf{p}_k\mathbf{A}_{k-1}^{-1}\mathbf{r}_k\mathbf{A}_k)\mathbf{S}_k^{\nu_1}, \tag{3.63}$$

where $\mathbf{S}_k = \mathbf{I} - \mathbf{M}_k^{-1}\mathbf{A}_k$ denotes the iteration matrix of the smoother. The multigrid algorithm can be formulated (see [Hac94]) with an iteration matrix that satisfies the recursion

$$\begin{aligned}
&\mathbf{C}_{MG,0}(\nu_2, \nu_1) = 0,\\
&\mathbf{C}_{MG,k}(\nu_2, \nu_1) = \mathbf{S}_k^{\nu_2}(\mathbf{I} - \mathbf{p}_k(\mathbf{I} - \mathbf{C}_{MG,k-1}^{\tau})\mathbf{A}_{k-1}^{-1}\mathbf{r}_k\mathbf{A}_k)\mathbf{S}_k^{\nu_1}, \qquad (3.64)\\
&\qquad\qquad = \mathbf{C}_{TG,k} + \mathbf{S}_k^{\nu_2}\mathbf{p}_k\mathbf{C}_{MG,k-1}^{\tau}\mathbf{A}_{k-1}^{-1}\mathbf{r}_k\mathbf{A}_k\mathbf{S}_k^{\nu_1}, \quad k = 1, 2, \ldots
\end{aligned}$$

If we consider (3.63) without post-smoothing and $\nu_1 = \nu$ pre-smoothing steps we get the splitting

$$\begin{aligned}
\|\mathbf{C}_{TG,k}(0, \nu)\|_2 &= \|(\mathbf{I} - \mathbf{p}_k\mathbf{A}_{k-1}^{-1}\mathbf{r}_k\mathbf{A}_k)\mathbf{S}_k^{\nu}\|_2\\
&\leq \|\mathbf{A}_k^{-1} - \mathbf{p}_k\mathbf{A}_{k-1}^{-1}\mathbf{r}_k\|_2\|\mathbf{A}_k\mathbf{S}_k^{\nu}\|_2 \qquad (3.65)
\end{aligned}$$

of the two-grid iteration matrix. For the multigrid convergence we are interested in bounds for the two terms in (3.65) of the form

$$\begin{aligned}
\|\mathbf{A}_k^{-1} - \mathbf{p}_k\mathbf{A}_{k-1}^{-1}\mathbf{r}_k\|_2 &\leq C_A\|\mathbf{A}_k\|_2^{-1}, \qquad (3.66)\\
\|\mathbf{A}_k\mathbf{S}_k^{\nu}\|_2 &\leq g(\nu)\|\mathbf{A}_k\|_2, \qquad (3.67)
\end{aligned}$$

where $g(\nu)$ is a monotonically decreasing function with $\lim_{\nu\to\infty} g(\nu) = 0$. The result in (3.66) is called *approximation property* and the one in (3.67) *smoothing property*.

Remark 3.28 For the further analysis in terms of the approximation and smoothing property some assumptions and technical results have to be formulated (see e.g. [Hac85]). Since these issues are discussed extensively in Chapter 5 for an *anisotropic* reaction-diffusion problem, we only give some comments. The family $\{\mathcal{T}_k\}_{k\geq 0}$ of triangulations is assumed to be quasi-uniform. Furthermore, uniform bounds (w.r.t. the level k) of the isomorphism P_k and its inverse, and a result concerning the scaling of the stiffness matrix are needed. In order to prove that the approximation property (3.66) holds, the H^2-regularity of the continuous

problem (3.56) and an approximation error bound as in Theorem 3.17 are required. In [Hac85] the derivation of the smoothing property for a few basic *point* iterative schemes and symmetric positive definite matrix $\mathbf{A}_k$ can be found. However, we will see that these point relaxation methods are not appropriate smoothers within the multigrid algorithm for solving the anisotropic problem. For that case, in Section 5.6 we prove the smoothing property (3.67) for a symmetric *block* Gauss-Seidel iteration.

We now turn to a convergence result for the multigrid method with $\tau \geq 2$ (see [Hac85] and the references therein). Note that for the proof of this result in addition to (3.66) and (3.67) we need a stability bound for the iteration matrix of the smoother.

Theorem 3.29 *Consider the multigrid method with iteration matrix defined by* (3.64) *and parameters* $\nu_2 = 0, \nu_1 = \nu > 0, \tau \geq 2$. *Assume that there are constants* C_A, C_S *and a monotonically decreasing function* $g(\nu)$ *with* $\lim_{\nu\to\infty} g(\nu) = 0$ *such that for all* k *we have*

$$\begin{aligned} \|\mathbf{A}_k^{-1} - \mathbf{p}_k\mathbf{A}_{k-1}^{-1}\mathbf{r}_k\|_2 &\leq C_A\|\mathbf{A}_k\|_2^{-1}, \\ \|\mathbf{A}_k\mathbf{S}_k^{\nu}\|_2 &\leq g(\nu)\|\mathbf{A}_k\|_2, \quad \nu \geq 1, \\ \|\mathbf{S}_k^{\nu}\|_2 &\leq C_S, \quad \nu \geq 1. \end{aligned}$$

Then for any fixed $\psi \in (0,1)$ *there exists* $\nu_0 > 0$ *such that for all* $\nu \geq \nu_0$

$$\|\mathbf{C}_{MG,k}\|_2 \leq \psi, \quad k = 0, 1, \ldots$$

holds.

Proof: For the two-grid iteration matrix we have

$$\|\mathbf{C}_{TG,k}\|_2 \leq \|\mathbf{A}_k^{-1} - \mathbf{p}_k\mathbf{A}_{k-1}^{-1}\mathbf{r}_k\|_2\|\mathbf{A}_k\mathbf{S}_k^{\nu}\|_2 \leq C_A g(\nu),$$

i.e. if sufficiently many smoothing iterations are applied the two-grid method converges. Standard arguments yield that the isomorphism P_k and its inverse are uniformly bounded and thus for $\mathbf{p}_k = P_k^{-1}P_{k-1}$ we can conclude

$$c_1\|\mathbf{x}\|_2 \leq \|\mathbf{p}_k\mathbf{x}\|_2 \leq c_2\|\mathbf{x}\|_2 \quad \text{for all } \mathbf{x} \in \mathbb{R}^{n_{k-1}}, \tag{3.68}$$

with constants $c_1 > 0$ and c_2 independent of k. We define $\xi_k := \|\mathbf{C}_{MG,k}\|_2$. Due to (3.64) we have $\xi_0 = 0$ and for $k \geq 1$, using (3.68), we get

$$\begin{aligned} \xi_k &\leq C_A g(\nu) + \|\mathbf{p}_k\|_2\xi_{k-1}^{\tau}\|\mathbf{A}_{k-1}^{-1}\mathbf{r}_k\mathbf{A}_k\mathbf{S}_k^{\nu}\|_2 \\ &\leq C_A g(\nu) + c_2c_1^{-1}\xi_{k-1}^{\tau}\|\mathbf{p}_k\mathbf{A}_{k-1}^{-1}\mathbf{r}_k\mathbf{A}_k\mathbf{S}_k^{\nu}\|_2 \\ &\leq C_A g(\nu) + c_2c_1^{-1}\xi_{k-1}^{\tau}\left(\|(\mathbf{I} - \mathbf{p}_k\mathbf{A}_{k-1}^{-1}\mathbf{r}_k\mathbf{A}_k)\mathbf{S}_k^{\nu}\|_2 + \|\mathbf{S}_k^{\nu}\|_2\right) \\ &\leq C_A g(\nu) + c_2c_1^{-1}\xi_{k-1}^{\tau}(C_A g(\nu) + C_S) \leq C_A g(\nu) + c^*\xi_{k-1}^{\tau}, \end{aligned}$$

with $c^* := c_2c_1^{-1}(C_A g(1) + C_S)$. We consider the sequence

$$\begin{cases} x_0 := 0, \\ x_i := C_A g(\nu) + c^* x_{i-1}^{\tau}. \end{cases} \tag{3.69}$$

For $\tau = 1$ this sequence is obviously unbounded. However, for $\tau \geq 2$ the sequence is bounded by any $\psi \in (0,1)$ if $g(\nu)$ is sufficiently small. To see this we consider Fig. 3.3 and use a fixed point argument. For fixed $\psi \in (0,1)$ we consider the function $\varphi(x) := C_A g(\nu) + c^* x^{\tau}$, $\tau \geq 2$. If $g(\nu)$ is sufficiently small, then there obviously exists a point $x^* = \varphi(x^*) \leq \psi$, which is the limit of sequence (3.69). ◇

Using Theorem 3.29 we get a bound for the spectral radius of the iteration matrix $\mathbf{C}_{MG,k}(0,\nu)$. Now we consider the iteration matrix of the multigrid method with ν_1 pre- and ν_2 post-smoothing iterations. Setting $\nu := \nu_1 + \nu_2$ we conclude

$$\rho(\mathbf{C}_{MG,k}(\nu_2,\nu_1)) = \rho(\mathbf{C}_{MG,k}(0,\nu)) \leq \|\mathbf{C}_{MG,k}(0,\nu)\|_2.$$

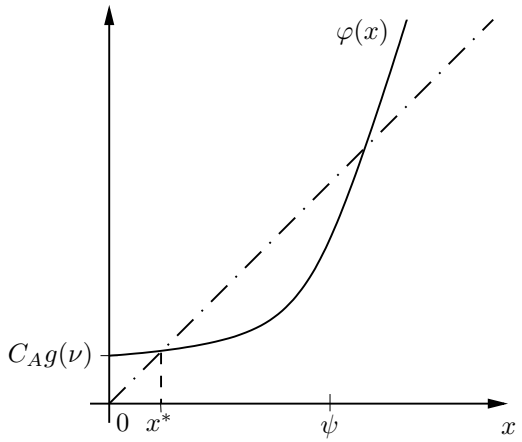


Figure 3.3: Schematic fixed point argument

Thus, for $\tau \geq 2$, Theorem 3.29 also yields a bound for the spectral radius of the iteration matrix $\mathbf{C}_{MG,k}(\nu_2, \nu_1)$.

We are interested in a convergence result for the multigrid V-cycle, too. However, the convergence analysis for $\tau = 1$ in (3.64) is more involved. We consider the case with a symmetric positive definite system matrix $\mathbf{A}_k$ and $\nu_1 = \nu_2 \geq 1$ pre- and post-smoothing iterations. In [Hac85, Hac94] the contraction number of a multigrid method, including the V-cycle, is shown to be uniformly bounded w.r.t. the $\|\cdot\|_A$-norm by a constant smaller than one independent of the level k. Here, we only give the result for the two-grid method. These considerations carry over to the multigrid algorithm.

We define the energy scalar product and corresponding norm by

$$\langle \mathbf{x}, \mathbf{y} \rangle_A := \langle A_k \mathbf{x}, \mathbf{y} \rangle, \quad \|\mathbf{x}\|_A := \langle \mathbf{x}, \mathbf{x} \rangle_A^{\frac{1}{2}}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^{n_k}.$$

In the analysis only smoothers with an iteration matrix $\mathbf{S}_k = \mathbf{I} - \mathbf{M}_k^{-1}\mathbf{A}_k$ and symmetric positive definite $\mathbf{M}_k$ are considered. Let $\mathbf{B}$, $\mathbf{C} \in \mathbb{R}^{n_k \times n_k}$ be matrices. If these matrices are symmetric (w.r.t. $\langle \cdot, \cdot \rangle$) we use the notation $\mathbf{B} \leq \mathbf{C}$ iff $\langle \mathbf{B}\mathbf{x}, \mathbf{x} \rangle \leq \langle \mathbf{C}\mathbf{x}, \mathbf{x} \rangle$ for all $\mathbf{x} \in \mathbb{R}^{n_k}$. If these matrices are symmetric w.r.t. $\langle \cdot, \cdot \rangle_A$ we write $\mathbf{B} \leq_A \mathbf{C}$ iff $\langle \mathbf{B}\mathbf{x}, \mathbf{x} \rangle_A \leq \langle \mathbf{C}\mathbf{x}, \mathbf{x} \rangle_A$ for all $\mathbf{x} \in \mathbb{R}^{n_k}$. One can easily show that for the damped Jacobi method (with appropriate chosen damping parameter) and the symmetric Gauss-Seidel relaxation we have

$$\mathbf{A}_k \leq \mathbf{M}_k, \qquad \text{for all } k, \tag{3.70a}$$

$$\exists\, C_M : \|\mathbf{M}_k\|_2 \leq C_M \|\mathbf{A}_k\|_2, \qquad \text{for all } k. \tag{3.70b}$$

Furthermore, for these smoothers the standard approximation property (3.66) and (3.70b) yield the *modified* approximation property:

$$\exists\, \tilde{C}_A : \|\mathbf{M}_k^{\frac{1}{2}}(\mathbf{A}_k^{-1} - \mathbf{p}_k \mathbf{A}_{k-1}^{-1} \mathbf{r}_k)\mathbf{M}_k^{\frac{1}{2}}\|_2 \leq \tilde{C}_A \quad \text{for all } k = 1, 2, \ldots \tag{3.71}$$

Due to symmetry arguments only the two-grid method with $\nu_1 = \nu_2 = \frac{1}{2}\nu$ and $\nu > 0$ even in (3.63) is discussed.

Theorem 3.30 *Assume that* (3.70a) *and* (3.71) *hold. Then we have*

$$\|\mathbf{C}_{TG,k}(\nu)\|_A \leq \max_{y \in [0,1]} y(1 - \tilde{C}_A^{-1} y)^\nu = \begin{cases} (1 - \tilde{C}_A^{-1})^\nu & \text{if } \nu \leq \tilde{C}_A - 1, \\ \dfrac{\tilde{C}_A}{\nu + 1} \left(\dfrac{\nu}{\nu + 1} \right)^\nu & \text{if } \nu \geq \tilde{C}_A - 1. \end{cases}$$

Sketch of the proof: We define $\mathbf{Q}_k := \mathbf{I} - \mathbf{p}_k \mathbf{A}_{k-1}^{-1} \mathbf{r}_k \mathbf{A}_k$. For this matrix we have $\mathbf{Q}_k^2 = \mathbf{Q}_k$ and $(\mathbf{A}_k \mathbf{Q}_k)^T = \mathbf{A}_k \mathbf{Q}_k$. Note that the latter identity holds iff $\mathbf{Q}_k$ is symmetric w.r.t. the energy scalar product. It follows that $\mathbf{Q}_k$ is an orthogonal projection w.r.t. $\langle \cdot, \cdot \rangle_A$. The matrix $\mathbf{M}_k^{-1} \mathbf{A}_k$ is symmetric w.r.t. $\langle \cdot, \cdot \rangle_A$, too, and from (3.70a) one can conclude that

$$0 \leq_A \mathbf{M}_k^{-1} \mathbf{A}_k \leq_A \mathbf{I} \tag{3.72}$$

holds. Further, one can show that property (3.71) is equivalent to

$$0 \leq_A \mathbf{Q}_k \leq_A \tilde{C}_A \mathbf{M}_k^{-1} \mathbf{A}_k. \tag{3.73}$$

From (3.72), (3.73) and the fact that the A-orthogonal projection $\mathbf{Q}_k$ is not identically zero, we get $\tilde{C}_A \geq 1$. For the iteration matrix of the two-grid method $\mathbf{C}_{TG,k}(\nu) = \mathbf{S}_k^{\frac{1}{2}\nu} \mathbf{Q}_k \mathbf{S}_k^{\frac{1}{2}\nu}$ we obtain

$$\begin{aligned} 0 \leq_A \mathbf{C}_{TG,k}(\nu) &\leq_A \mathbf{S}_k^{\frac{1}{2}\nu} \tilde{C}_A \mathbf{M}_k^{-1} \mathbf{A}_k \mathbf{S}_k^{\frac{1}{2}\nu} \\ &= (\mathbf{I} - \mathbf{M}_k^{-1} \mathbf{A}_k)^{\frac{1}{2}\nu} \tilde{C}_A \mathbf{M}_k^{-1} \mathbf{A}_k (\mathbf{I} - \mathbf{M}_k^{-1} \mathbf{A}_k)^{\frac{1}{2}\nu}. \end{aligned}$$

Thus, we finally get

$$\|\mathbf{C}_{TG,k}(\nu)\|_A \leq \max_{x \in [0,1]} \tilde{C}_A \, x(1-x)^\nu = \max_{y \in [0, \tilde{C}_A^{-1}]} y(1 - \tilde{C}_A^{-1} y)^\nu,$$

and the proof is finished with elementary calculus. ◇

Nested iteration

The method of nested iteration uses coarse grid problems to produce start approximations for problems on finer levels. This idea is not restricted to multigrid methods but we mention it here since it fits into the multigrid framework in which discretizations of a continuous problem on different levels are used. In order to obtain a start approximation of the problem $\mathbf{A}_k \mathbf{x}_k = \mathbf{b}_k$ on level k we can apply some steps of an iterative solution method to a coarse grid problem $\mathbf{A}_j \mathbf{x}_j = \mathbf{b}_j$ with $j < k$. Let $\mathbf{x}_j^i$ denote the result of i steps on level j. For the start approximation on level k we then use $\mathbf{x}_k^0 := \tilde{\mathbf{p}}_k \mathbf{x}_j^i$, where $\tilde{\mathbf{p}}_k$ denotes an appropriate prolongation. In Fig. 3.4 a schematic representation of this method is given.

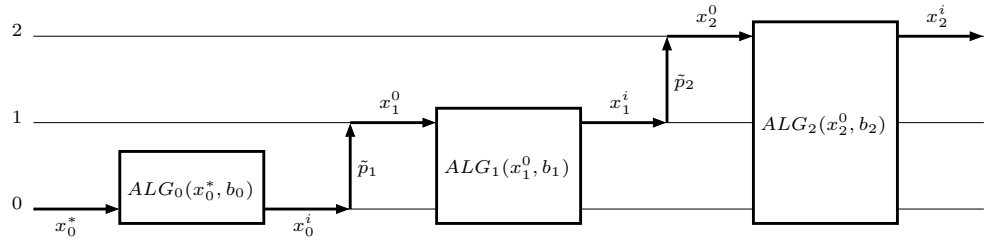


Figure 3.4: Nested iteration; ALG_k denotes an iterative solution algorithm on level k

We will make use of the nested iteration method for the solution of the inverse heat conduction problem. In order to get a good start estimation for the heat flux on a given discretization level we consider the whole inverse problem on a coarser level. A comparison w.r.t. efficiency of the nested optimization approach and a one-level procedure is given in Chapter 6.

Chapter 4

Anisotropic finite elements

In this thesis a three-dimensional IHCP in falling films is solved numerically based on realistic high resolution temperature measurements (compare Section 2.1). In the falling film experiment a *thin* heating foil is used in order to reach a small temperature gradient across the foil thickness. The measurement data on the foil back side are influenced by the transport phenomena in the falling film and the wavy film surface. However, the very different length scales of the thin heat conductor result in an *anisotropy* effect. Thus, in this chapter we investigate anisotropic finite elements and treat the question how these elements affect the quality of the space discretization. In order to obtain a bound for the discretization error we use the Céa-lemma and error bounds for standard nodal Lagrangian interpolation. We give a uniform presentation of results for (triangular and) tetrahedral finite elements that can be found at different places in the literature and illustrate important phenomena by means of systematic numerical experiments. A *maximum* angle condition is formulated which is the essential condition to get suitable bounds for the interpolation error. Different approaches for the derivation of these error bounds are considered.

4.1 Introduction and historical review

In the previous chapter we introduced basic error estimates for isotropic finite elements with a uniformly bounded aspect ratio. Now we study anisotropic finite elements e that are characterized by the property

$$\lim_{h_e \to 0} \sigma_e = \lim_{h_e \to 0} \frac{h_e}{\rho_e} = \infty. \tag{4.1}$$

Again the focus is on linear finite elements. For anisotropic $\mathcal{P}_1$-elements approximation error bounds of the form

$$\inf_{v_h \in X_h^1} \|u - v_h\|_{H^m(\Omega)} \leq c h^{2-m} |u|_{H^2(\Omega)}, \quad m \in \{0, 1\}, \quad \forall u \in H^2(\Omega), \tag{4.2}$$

do not hold in general. Note that such bounds are used to prove standard finite element discretization error bounds (compare Theorem 3.16 and Theorem 3.17).

Remark 4.1 In the next chapter we consider a domain with a given fixed tetrahedral triangulation. When the thickness of this domain becomes very small due to a one-directional distortion, then the tetrahedra volumes decrease (with increasing degeneracy) while the tetrahedra diameters do not change. This phenomenon can be compared to situations where in a domain with equal length scales in all directions the grid is adapted to solutions that have an anisotropic behaviour. The grid adaption results in a rapid refinement in the direction(s) of rapid change and a relatively slow refinement in those direction(s) where the solution is smooth. Since one always is restricted to a finite grid resolution and the heat conductor is

very thin but has a fixed geometry the corresponding finite elements are not anisotropic in the asymptotical sense of (4.1). However, $\sigma_e \to \infty$ for increasing degeneracy, and thus the constant in (4.2) may be arbitrarily large when highly distorted finite elements are used.

In the following we give a brief historical review of some results obtained in the field of anisotropic finite elements without claiming completeness (see also [Ape99]). Starting point of our review is Zlámal's work [Zla68] of 1968 concerning the approximate solution of a linear second order elliptic boundary value problem in two dimensions using the finite element method. For a family of triangular elements the so-called Zlámal condition, which states that the minimum angle of every triangle is uniformly bounded from below by a positive constant, is introduced.

Minimum angle condition (in 2D):

Let there exist a constant $\bar{\alpha}$ such that

$$\alpha_e \geq \bar{\alpha} > 0 \quad \text{for all } e \in \mathcal{T}_h, \tag{4.3}$$

where α_e denotes the minimum angle of the element e.

This condition is equivalent to the condition of a uniformly bounded aspect ratio. Assuming that (4.3) holds for a given triangulation $\mathcal{T}_h$ and using piecewise quadratic finite elements, Zlámal gives the bound

$$\|u - u_h\|_{H^1(\Omega)} \leq c \frac{1}{\sin \bar{\alpha}} h^2$$

for the discretization error. Here, u_h denotes the finite element approximation and the constant c is independent of the parameter h, which denotes the largest side of all the triangles in the given triangulation, but depends on u. The exact solution u of the problem is supposed to be sufficiently smooth. However, if $\bar{\alpha} > 0$ gets small, the factor $1/\sin(\bar{\alpha})$ in the upper bound becomes large. The minimum angle condition (4.3) was considered to be essential for quite some time. But in 1976 Babuška and Aziz elaborated upon a result from interpolation theory that was investigated by Synge [Syn57] in 1957 for the case of piecewise linear interpolation. This result can also be found in [Kri92] and assumes that the maximum angle of every triangle in a given triangulation is uniformly bounded from above by a constant that is smaller than π.

Maximum angle condition (in 2D):

Let there exist a constant $\bar{\gamma}$ such that

$$\gamma_e \leq \bar{\gamma} < \pi \quad \text{for all } e \in \mathcal{T}_h, \tag{4.4}$$

where γ_e denotes the maximum angle of the element e.

We remark that Synge's condition (4.4) is weaker than Zlámal's condition (4.3): in a triangle with *one* arbitrarily small angle, the other angles can always be chosen such that there is no angle "close" to π. Note that if a triangle has *two* small angles, the third angle is necessarily a large one. For a family of triangular elements that satisfy condition (4.4) in [Syn57] Synge proves the interpolation error estimate

$$\|u - I_h u\|_{W_p^1(\Omega)} \leq c(\bar{\gamma}) h \, |u|_{W_p^2(\Omega)}, \quad \forall u \in W_p^2(\Omega), \tag{4.5}$$

for the parameter $p = \infty$. Here, $c(\bar{\gamma}) > 0$ is independent of h but depends on $\bar{\gamma}$ and I_h denotes the standard linear interpolation operator. In [BA76] Babuška and Aziz give an extension of (4.5) for $p = 2$ and consider the obtained interpolation result in the context of the finite element method. For piecewise linear finite elements that fulfill the maximum angle condition (4.4) they succeed in showing the discretization error bound

$$\|u - u_h\|_{H^1(\Omega)} \leq c \, \Gamma(\bar{\gamma}) h \, \|u\|_{H^2(\Omega)}, \quad \forall u \in H^2(\Omega). \tag{4.6}$$

In this result the constant c is independent of h and $\bar{\gamma}$ while the increasing function $\Gamma(\bar{\gamma})$ depends on the maximum angle. We emphasize that (4.5) gives an interpolation error bound while the result in (4.6) concerns the finite element discretization error. The latter result in fact allows anisotropic finite elements that violate the minimum angle condition (4.3) without loss of the discretization quality (in the sense that the H^1-error is of first order).

Some numerical experiments to point out *two*-dimensional anisotropic effects are presented in Section 4.2. A generalization of (4.4) to tetrahedral elements and a special error bound for standard Lagrangian interpolation, which was given by Křížek [Kri92] in 1992, is studied in Section 4.3. A more general approach to anisotropic estimates due to the work of Apel [Ape99] is discussed in Section 4.4. It is shown that for *three*-dimensional problems and the nodal Lagrangian interpolation operator a result as in (4.5) with $p = 2$ does *not* hold in the anisotropic case. However, for this operator and any $p \in (2, \infty]$ an estimate of the form

$$\|u - I_h u\|_{W_p^1(\Omega)} \leq ch\,|u|_{W_p^2(\Omega)}, \quad \forall u \in W_p^2(\Omega),$$

holds with a constant c independent of h. Finally, in Section 4.5 we give a discretization error bound for anisotropic tetrahedral $\mathcal{P}_1$-elements based on the Céa-lemma and the previously discussed interpolation results. We emphasize that (for three-dimensional problems) there is a need for other (better) interpolation operators for anisotropic finite element spaces to obtain a discretization error bound of the form

$$\|u - u_h\|_{H^1(\Omega)} \leq ch\,|u|_{H^2(\Omega)}, \quad \forall u \in H^2(\Omega),$$

with c independent of h (compare Section 5.4).

4.2 Numerical experiments

In this section we present MATLAB simulations to illustrate the results for two-dimensional anisotropic finite elements discussed above. We consider the discretization of the elliptic problem

$$\begin{cases} -\Delta u = -2 & \text{in } \Omega := (-1,1)^2, \\ u = x_1^2 & \text{on } \partial\Omega, \end{cases} \tag{4.7}$$

in the x_1, x_2-plane using different types of grids and linear triangular finite elements. The exact solution of problem (4.7) is given by $u(x_1, x_2) = x_1^2$, $(x_1, x_2) \in \bar{\Omega}$. In the following we describe the single elements that form the different grids.

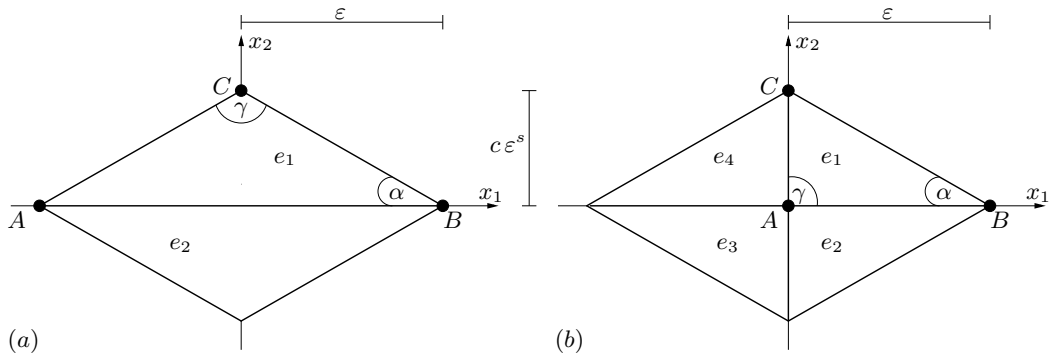


Figure 4.1: Anisotropic elements of (a) type A and (b) type B

Element type A: maximum angle condition (mac) violated

The first element type is pictured in Fig. 4.1 (a). The nodes of the triangle e_1 are $A = (-\varepsilon, 0)$, $B = (\varepsilon, 0)$ and $C = (0, c\varepsilon^s)$, for some c, $s \in \mathbb{N}$ and a real parameter $\varepsilon < 1$. Then $u(A) = u(B) = \varepsilon^2$ and $u(C) = 0$. When we choose $s > 1$ then the aspect ratio σ_{e_1} is unbounded, i.e. e_1 is an anisotropic element in the sense of (4.1), and the maximum angle condition is violated. Due to Theorem 3.13 (with the parameters $p = 2$ and $m = 0$) the L_2-error of standard nodal interpolation is of second order. For the gradient of the interpolation error we consider

$$\begin{aligned} \frac{\partial}{\partial x_1}(u - I_{e_1}u) &= 2x_1 - \frac{u(B) - u(A)}{2\varepsilon} = 2x_1, \\ \frac{\partial}{\partial x_2}(u - I_{e_1}u) &= -\frac{u(C) - (u(A) + u(B))/2}{c\varepsilon^s} = c^{-1}\varepsilon^{2-s}. \end{aligned}$$

While the exact solution $u(x_1, x_2) = x_1^2$ does not depend on x_2, the partial derivative of $I_{e_1}u$ in x_2-direction becomes arbitrarily large when the triangulations get finer, if we use anisotropic elements of type A with $s > 2$. Elementary calculus leads to

$$|u - I_{e_1}u|^2_{H^1(e_1)} = \int_{e_1} (2x_1)^2 + (c^{-1}\varepsilon^{2-s})^2 \, de_1 = \left(\frac{2}{3}\varepsilon^2 + (c^{-1}\varepsilon^{2-s})^2\right) \int_{e_1} 1 \, de_1. \tag{4.8}$$

The minimum and the maximum angles which are denoted by α and γ in Fig. 4.1 are referred to in the numerical experiments below.

Element type B: mac fulfilled & minimum angle condition (mic) violated

For the second grid we consider the elements of type A and insert additional nodes, such that the elements become right-angled as shown in Fig. 4.1 (b). Now the element e_1 is determined by the nodes $A = (0, 0)$, $B = (\varepsilon, 0)$ and $C = (0, c\varepsilon^s)$ with $u(A) = u(C) = 0$ and $u(B) = \varepsilon^2$. For $s > 1$ the aspect ratio σ_{e_1} again is unbounded but this time we do not expect the first order derivatives to cause problems due to the bounded maximum angle. In fact, the calculation of the partial derivatives yields

$$\begin{aligned} \frac{\partial}{\partial x_1}(u - I_{e_1}u) &= 2x_1 - \frac{u(B) - u(A)}{\varepsilon} = 2x_1 - \varepsilon, \\ \frac{\partial}{\partial x_2}(u - I_{e_1}u) &= -\frac{u(C) - u(A)}{c\varepsilon^s} = 0. \end{aligned}$$

For the gradient of the interpolation error on a single element we get

$$|u - I_{e_1}u|^2_{H^1(e_1)} = \int_{e_1} (2x_1 - \varepsilon)^2 \, de_1 = \frac{10}{3}\varepsilon^2 \int_{e_1} 1 \, de_1. \tag{4.9}$$

In the following numerical experiments we study different error norms for standard nodal interpolation and the finite element solution of (4.7) with $\mathcal{P}_1$-elements on different grids. We choose $c = 2$ and $s = 2$, which means that the grid refinement (from $h \to h/2$) in x_2-direction is twice as high as in x_1-direction.

Experiment I: mac violated, divergent finite element (FE) approximation

The hierarchy of grids used in this experiment is composed of the anisotropic elements of type A (see Fig. 4.1 (a)). For the discretization parameters $h = 1$ and $h = 0.5$ (i.e. $\varepsilon = 0.5$ and $\varepsilon = 0.25$) the triangulations are shown in Fig. 4.2. Clearly, the maximum and thus also the minimum angle condition are violated. The corresponding standard linear interpolant of $u(x_1, x_2) = x_1^2$ and the discrete finite element solution of problem (4.7) with $\mathcal{P}_1$-elements are shown in Fig. 4.3. A careful study of Fig. 4.3 (a) confirms the presented analytical results. For standard nodal interpolation the function value approximation is accurate. Note that

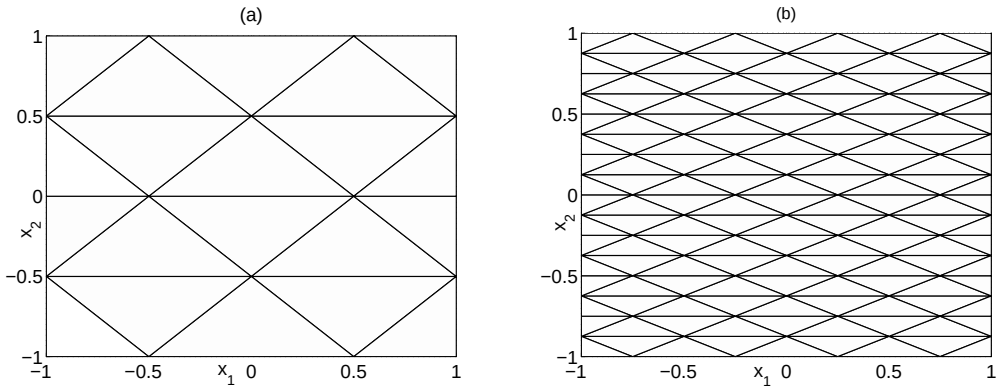


Figure 4.2: Triangulations for $h = 1$ (a) and $h = 0.5$ (b)

for $m = 0$ in Theorem 3.13 the factor ρ_e^{-m} is eliminated from the upper bound. However, the gradient interpolation may become arbitrarily poor for $h \to 0$. In fact, we observe that the interpolant has a kind of "zigzag"-shape with large gradients in x_2-direction.

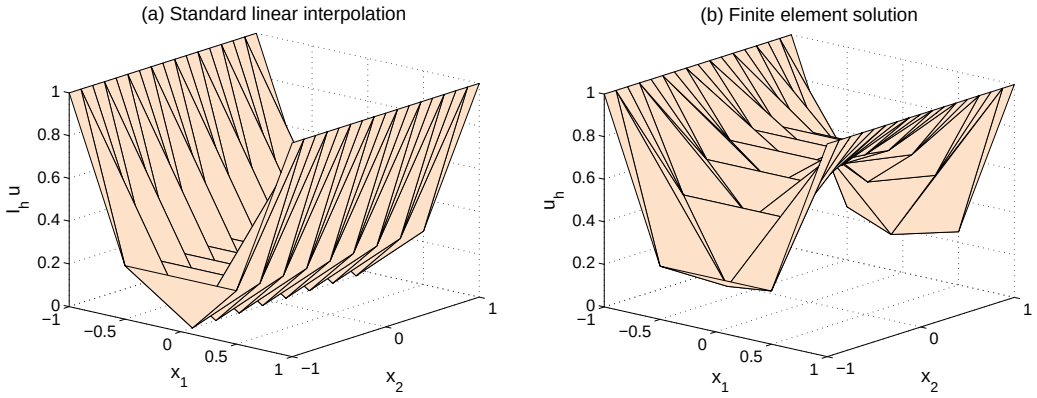


Figure 4.3: Linear interpolant (a) and finite element solution (b)

The simulation results are documented in Tab. 4.1. The minimum angle α tends to zero and the maximum angle γ tends to π as the grid refinement increases.

h	α	γ	$\|u - I_h u\|_{L_2}$	$\|u - I_h u\|_{H^1}$	$\|u - u_h\|_{L_2}$	$\|u - u_h\|_{H^1}$
1	0.2450	2.6516	0.2415	3.5473	0.8539	2.6142
0.5	0.1244	2.8929	0.0645	3.7625	0.7660	2.6328
0.25	0.0624	3.0168	0.0166	3.8782	0.7415	2.6447
0.125	0.0312	3.0791	0.0042	3.9383	0.7353	2.6488

Table 4.1: Global error norms for nodal interpolation and the finite element solution

The error reduction in the L_2-norm is of second order while the H^1-seminorm of the interpolation error is almost constant which is consistent with the bound in (4.8). The numerical finite element solution which is shown in Fig. 4.3 (b) exhibits large gradients as well but additionally does not yield an appropriate function value approximation. Note that for the L_2-error estimate of the finite element solution in Theorem 3.17 the duality argument of Aubin-Nitsche is applied. Since the interpolation error bound (3.22) with $m = 1$ is used for

this argument in (3.28) and (3.29), both orders of accuracy may be lost. Due to the computational time we only consider grid refinements up to approximately 30 thousand unknowns.

Experiment II: mac fulfilled & mic violated, convergent FE approximation

In this experiment we study the anisotropic elements of type B (see Fig. 4.1 (b)). For two different refinement levels the grids are illustrated in Fig. 4.4. The maximum angle condition

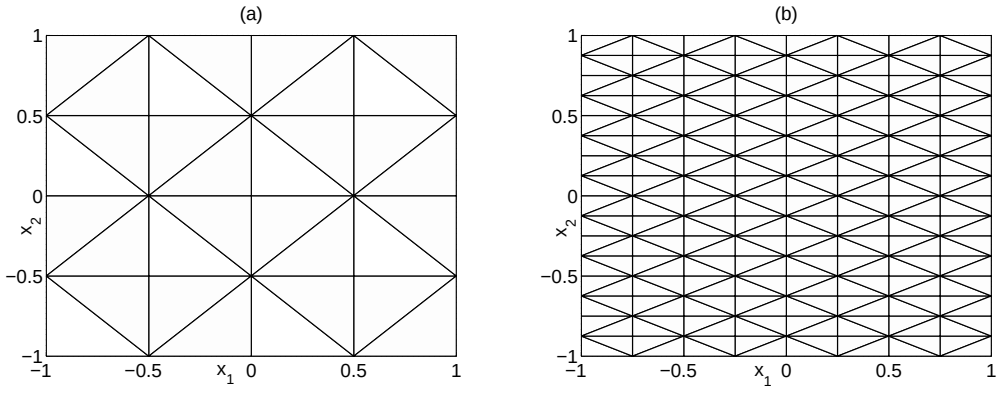


Figure 4.4: Triangulations for $\varepsilon = 0.5$ (a) and $\varepsilon = 0.25$ (b)

is fulfilled while one angle of the triangles tends to zero for $h \to 0$. This time the linear interpolant shown in Fig. 4.5 (a) seems to be accurate with respect to both, the function value and the gradient approximation. In particular the "zigzag"-shape described in experiment I has disappeared leading to smooth gradients in x_2-direction. Furthermore, the finite element solution in Fig. 4.5 (b) does not exhibit any visible differences in comparison to the linear interpolant.

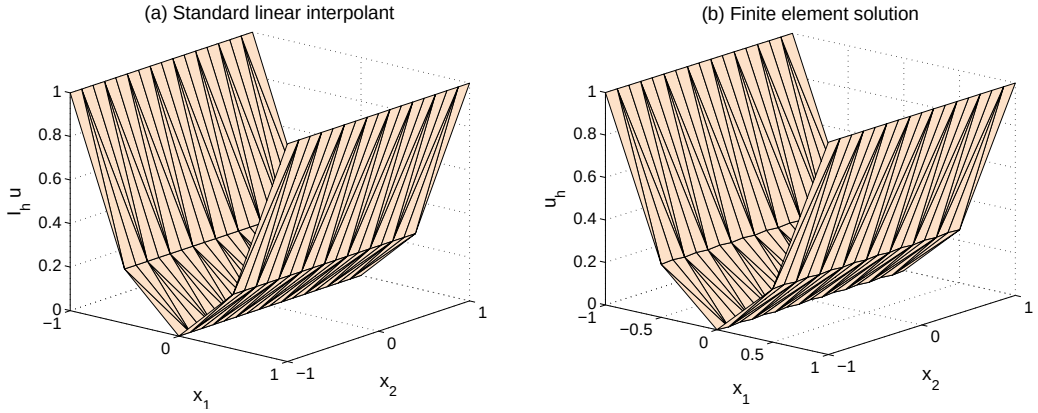


Figure 4.5: Linear interpolant (a) and finite element solution (b)

These first impressions are confirmed by the experimental results given in Tab. 4.2. The L_2- and H^1-errors are of second and first order, respectively. For the H^1-seminorm of the interpolation error this is consistent with the bound in (4.9).

Experiment III: mac violated, convergent FE approximation

The example meshes used in the numerical experiments I and II are introduced in the work of Babuška and Aziz [BA76]. As we have seen, the use of arbitrarily chosen anisotropic finite elements may cause serious problems with respect to the quality of the discretization. On the other hand, if the maximum angle condition holds, one can prove that the linear interpo-

h	α	γ	$\|u - I_h u\|_{L_2}$	$\|u - I_h u\|_{H^1}$	$\|u - u_h\|_{L_2}$	$\|u - u_h\|_{H^1}$
0.5154	0.2450	1.5708	0.0913	0.5774	0.0905	0.5731
0.2519	0.1244	1.5708	0.0228	0.2887	0.0227	0.2880
0.1252	0.0624	1.5708	0.0057	0.1443	0.0057	0.1442
0.0625	0.0312	1.5708	0.0014	0.0722	0.0014	0.0722

Table 4.2: Global error norms for nodal interpolation and the finite element solution

lation error in the H^1-norm is of first order (in two dimensions). Due to the Céa-lemma one can show that the H^1-error reduction of the finite element solution of (4.7) with $\mathcal{P}_1$-elements is of first order, too. In the last experiment of this section a further question in this context is answered. Using the anisotropic elements proposed by Apel and Dobrowolski in [AD92] we treat the question whether the maximum angle condition is a necessary condition to gain a first order H^1-error reduction of the finite element solution.

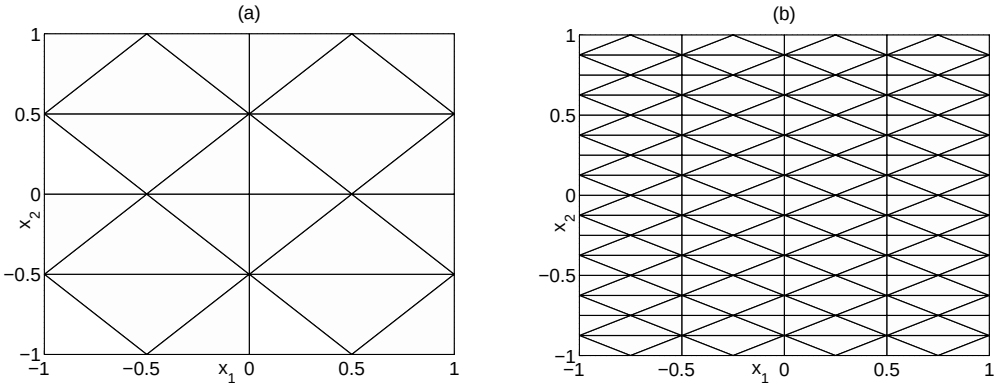


Figure 4.6: Triangulations for $\varepsilon = 0.5$ (a) and $\varepsilon = 0.25$ (b)

For two refinement levels the grids which (in contrast to so-called *semiregular* triangulations introduced in the next section) combine the anisotropic elements of type A and type B (see Fig. 4.1) are shown in Fig. 4.6.

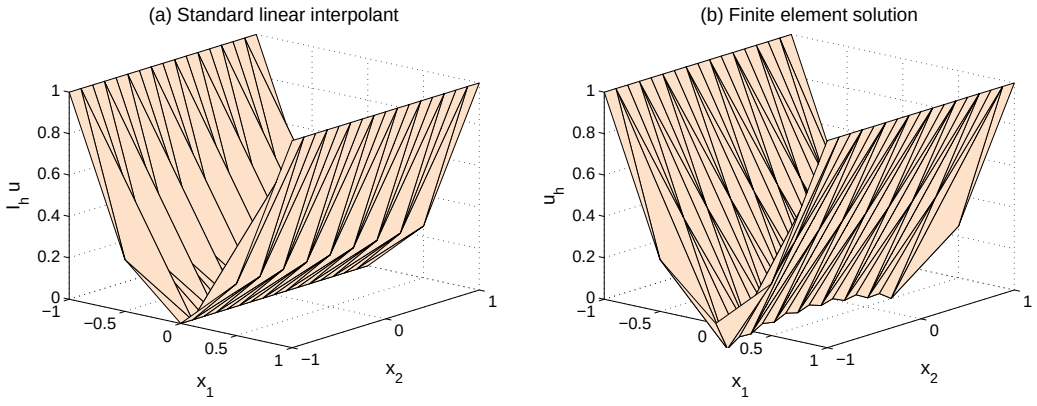


Figure 4.7: Linear interpolant (a) and finite element solution (b)

The corresponding nodal interpolant and the finite element solution are given in Fig. 4.7. We clearly recover the "zigzag"-behaviour in x_2-direction for some elements of the triangulation in Fig. 4.7 (a). This also seems to be the case in Fig. 4.7 (b) but this time the situation is

different (to the one in experiment I) as Tab. 4.3 clarifies. While the interpolation error in the H^1-seminorm is almost constant, the L_2- and H^1-error reduction of the finite element solution is of second and first order, respectively. We remark that in [AD92] the convergence of a modified interpolant is discussed.

h	α	γ	$\|u - I_h u\|_{L_2}$	$\|u - I_h u\|_{H^1}$	$\|u - u_h\|_{L_2}$	$\|u - u_h\|_{H^1}$
0.5154	0.2450	2.6516	0.2041	2.9155	0.2641	1.3712
0.2519	0.1244	2.8929	0.0510	2.8504	0.0664	0.6878
0.1252	0.0624	3.0168	0.0128	2.8339	0.0165	0.3420
0.0625	0.0312	3.0791	0.0032	2.8298	0.0041	0.1702

Table 4.3: Global error norms for nodal interpolation and the finite element solution

Remark 4.2 In the next section we give a proof of nodal interpolation error bounds for so-called *semiregular* tetrahedral elements. The basic idea in this proof concerns a geometrical argument that we can already explain in two dimensions. The geometrical property, which distinguishes the "bad" elements of type A from the "good" ones of type B (see Fig. 4.1) is the following: while for the elements of type B we can always find two vectors along the edges of the elements which are linearly independent uniformly in the discretization parameter h, this is not the case for the elements of type A.

4.3 Special interpolation error bounds

This section is dedicated to the study of Křížek's paper [Kri92], in which so-called *semiregular* finite elements are introduced and the three-dimensional analogon of (4.5) with $p = \infty$ is proven with a technique quite different from the approach used by Babuška and Aziz in two dimensions.

Maximum angle condition (in 3D):

Let there exist a constant $\bar{\gamma} < \pi$ such that

$$\gamma_e \leq \bar{\gamma} \quad \text{for all } e \in \mathcal{T}_h, \tag{4.10}$$

$$\varphi_e \leq \bar{\gamma} \quad \text{for all } e \in \mathcal{T}_h, \tag{4.11}$$

where γ_e denotes the maximum angle of all triangular faces of the tetrahedron e and φ_e denotes the maximum angle between the faces of e.

Definition 4.3 *A family $\mathcal{F} = \{\mathcal{T}_h\}$ of tetrahedral finite elements is said to be semiregular if for any $\mathcal{T}_h \in \mathcal{F}$ the inequalities* (4.10) *and* (4.11) *hold.*

Remark 4.4 *For local estimates in the next section the introduced maximum angle condition is used w.r.t. a single element in the following sense: an element e is said to fulfill the maximum angle condition* (4.10)-(4.11) *if both bounds hold uniformly in h_e.*

Note that any regular family of finite elements is also semiregular whereas the reverse inclusion does not hold in general. The following basic lemma concerning the geometry of tetrahedra is given without a proof (for more details see [Kri92]).

Lemma 4.5 *Let e be an arbitrary tetrahedron.*

(a) Let $\alpha \leq \beta \leq \gamma$ denote the angles of an arbitrary face of e. If (4.10) *holds then $\gamma \geq \frac{\pi}{3}$ and*

$$\beta, \gamma \in \left[\frac{\pi - \bar{\gamma}}{2}, \bar{\gamma} \right]. \tag{4.12}$$

(b) For an arbitrary vertex A of e let $\chi \le \psi \le \varphi$ denote the angles between the faces passing through A. If (4.11) *holds then $\varphi > \frac{\pi}{3}$ and*

$$\psi, \varphi \in \left(\frac{\pi - \bar{\gamma}}{2}, \bar{\gamma} \right]. \tag{4.13}$$

Based on this lemma we are able to give Křížek's extension of (4.5) for a semiregular family of triangulations. The result is deduced in the W^1_∞-norm, i.e. for the parameter $p = \infty$, and subsequently generalized using embedding properties of Sobolev spaces. We introduce the following notation. Let $\alpha := (\alpha_1, \dots, \alpha_n)$, $\alpha_i \in \mathbb{N}$, be a multi-index, $(x_1, \dots, x_n)^T$ the vector of coordinates and $\mathbf{y} \in \mathbb{R}^n$. Then we define

$$|\alpha| := \sum_{i=1}^{n} \alpha_i, \quad \mathbf{y}^\alpha := y_1^{\alpha_1} \cdots y_n^{\alpha_n}, \quad D^\alpha := \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}}. \tag{4.14}$$

Theorem 4.6 *Let $\mathcal{F} = \{\mathcal{T}_h\}$ be a semiregular family of decompositions of a polyhedron into tetrahedra. Then there exists a constant c such that for any $\mathcal{T}_h \in \mathcal{F}$ and any $e \in \mathcal{T}_h$ we have*

$$|u - I_e u|_{W^m_\infty(e)} \le c h_e^{2-m} |u|_{W^2_\infty(e)}, \quad m \in \{0, 1\}, \quad \forall u \in W^2_\infty(e).$$

Proof: Let $e \in \mathcal{T}_h \in \mathcal{F}$ be an arbitrary tetrahedron. Setting $p = \infty$ and $m = 0$ in Theorem 3.13 (given in Section 3.1.2 for an affine family of finite elements) we obtain

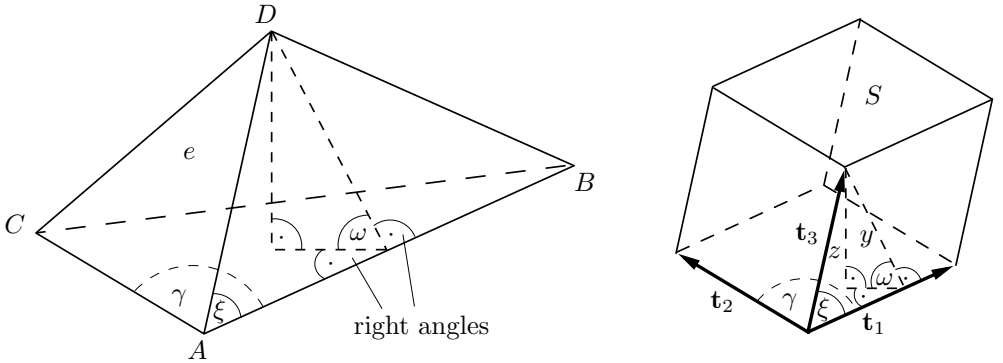
$$|u - I_e u|_{W^0_\infty(e)} \le c h_e^2 |u|_{W^2_\infty(e)}.$$

Note that this interpolation error bound is valid without posing any regularity assumptions upon the family $\mathcal{F} = \{\mathcal{T}_h\}$ of triangulations. Thus, it remains to derive the estimate

$$|u - I_e u|_{W^1_\infty(e)} \le c h_e |u|_{W^2_\infty(e)}, \quad \forall u \in W^2_\infty(e). \tag{4.15}$$

In order to show that (4.15) holds, the geometrical assumption of bounded maximum angles γ_e and φ_e is crucial.

The semiregularity of $\mathcal{F} = \{\mathcal{T}_h\}$ allows us to determine vectors $\mathbf{t}_i$ with $\|\mathbf{t}_i\|_2 = 1, i = 1, 2, 3$, along the edges of e such that the volume of the parallelepiped generated by these unit vectors is uniformly bounded away from zero, i.e. $\mathbf{t}_1, \mathbf{t}_2$ and $\mathbf{t}_3$ are linearly independent. For an appropriate choice of these vectors we consider the arbitrary tetrahedron e given in Fig. 4.8 (a). Let ABC be an arbitrary face of e with the opposite vertex D and the



(a) (b)

Figure 4.8: Tetrahedron (a) and linearly independent vectors (b) along its edges

maximum angle γ, which is without loss of generality assumed to be at the vertex A. From (4.12) we can conclude that

$$m := \min\left\{\sin(\frac{\pi-\bar{\gamma}}{2}), \sin\bar{\gamma}\right\} \leq \sin\gamma. \tag{4.16}$$

Let $\mathbf{t}_1$ and $\mathbf{t}_2$ be the unit vectors along the edges AB and AC, respectively. For the determination of the third unit vector we assume the angle ω between the faces ABC and ABD to be not less than the angle between the faces ABC and ACD (otherwise we exchange the notation of the vertices B and C). Then (4.13) leads to

$$m \leq \sin\omega. \tag{4.17}$$

Now we consider the triangle ABD and choose the vertex $K \in \{A, B\}$ such that the angle ξ at this vertex is not less than the angle at the other vertex. From (4.12) we get the bound

$$m \leq \sin\xi. \tag{4.18}$$

Let $\mathbf{t}_3$ be the unit vector along the edge KD. For the case $K = A$ the corresponding parallelepiped S is given in Fig. 4.8 (b). Denote by z the length of its spatial altitude perpendicular to the plane given by $\mathbf{t}_1$ and $\mathbf{t}_2$. Let y denote the length of its altitude perpendicular to $\mathbf{t}_1$, which lies in the plane given by $\mathbf{t}_1$ and $\mathbf{t}_3$. For the volume of S we get by (4.16)-(4.18)

$$\begin{aligned} |S| &= \|\mathbf{t}_1\|_2(\|\mathbf{t}_2\|_2\sin(\pi-\gamma))z = \sin\gamma(y\sin\omega) = \sin\gamma(\|\mathbf{t}_3\|_2\sin\xi)\sin\omega \\ &\geq m^3 > 0, \end{aligned} \tag{4.19}$$

independently of the discretization parameter h_e.

Let $u \in C^2(e)$ and P be an arbitrary fixed point in the interior of e. Since the unit vectors $\mathbf{t}_i$, $i = 1, 2, 3$, are linearly independent we can write

$$\nabla(u - I_e u)(P) = \sum_{j=1}^{3} c_j \mathbf{t}_j, \quad c_j \in \mathbb{R}, \tag{4.20}$$

with appropriate coefficients c_j. In this representation $\nabla(u - I_e u)(P)$ denotes the gradient vector of the function $u - I_e u$ evaluated in P. If we consider the orthogonal projection of this gradient vector on the edges of e along the unit vectors $\mathbf{t}_i$ we see that the coefficients c_j fulfill the equations

$$\mathbf{t}_i^T \nabla(u - I_e u)(P) = \sum_{j=1}^{3} c_j \mathbf{t}_i^T \mathbf{t}_j, \quad i = 1, 2, 3.$$

This leads to the matrix-vector-notation

$$\mathbf{D}\mathbf{c} = \mathbf{b}, \tag{4.21}$$

with the symmetric matrix $\mathbf{D} = (\mathbf{t}_i^T \mathbf{t}_j)_{i,j=1}^3$, the coefficient vector $\mathbf{c} = (c_1, c_2, c_3)^T$, and the right-hand side $\mathbf{b} = (\mathbf{t}_i^T \nabla(u - I_e u)(P))_{i=1}^3$. Since $\mathbf{D}$ is quadratic and nonsingular we can write

$$\mathbf{D}^{-1} = \frac{1}{\det(\mathbf{D})}\mathbf{D}^*, \tag{4.22}$$

where $\mathbf{D}^*$ is the matrix of cofactors of the entries of $\mathbf{D}$, i.e. $\mathbf{D}^* = (\det_{ij})_{i,j=1}^3$ with $\det_{ij}$ denoting the determinant of the submatrix that is obtained by eliminating the ith row and the jth column of $\mathbf{D}$. Since $|\mathbf{t}_i^T \mathbf{t}_j| = \|\mathbf{t}_i\|_2 \|\mathbf{t}_j\|_2 |\cos(\sphericalangle(\mathbf{t}_i, \mathbf{t}_j))| \leq 1$ we get $|\det_{ij}| \leq 2$ resulting in

$$\|\mathbf{D}^*\|_\infty \leq 6. \tag{4.23}$$

Finally, we can give the following bound for the gradient of the interpolation error using (4.20)-(4.23)

$$\begin{aligned}\|\nabla(u - I_e u)(P)\|_\infty &= \|\sum_{j=1}^{3} c_j \mathbf{t}_j\|_\infty \leq \sum_{j=1}^{3} |c_j| \leq 3\|\mathbf{c}\|_\infty \leq 3\|\mathbf{D}^{-1}\|_\infty \|\mathbf{b}\|_\infty \\ &\leq 3\frac{6}{|\det(\mathbf{D})|} \max_i |\mathbf{t}_i^T \nabla(u - I_e u)(P)|. \end{aligned} \tag{4.24}$$

We note that $|\det(\mathbf{N})|$ with $\mathbf{N}$ denoting the matrix whose columns consist of the unit vectors $\mathbf{t}_i$ is equal to the volume $|S|$ of the parallelepiped S in (4.19). Since the equality $\det(\mathbf{D}) = \det(\mathbf{N}^T\mathbf{N}) = \det^2(\mathbf{N})$ holds, we can conclude

$$\det(\mathbf{D}) \geq m^6 > 0,$$

independently of h_e. Thus, the proof is complete if we succeed in showing that the orthogonal projection of $\nabla(u - I_e u)(P)$ on any of the three unit vectors $\mathbf{t}_i$ cannot exceed constant times $h_e\,|u|_{W^2_\infty(e)}$.

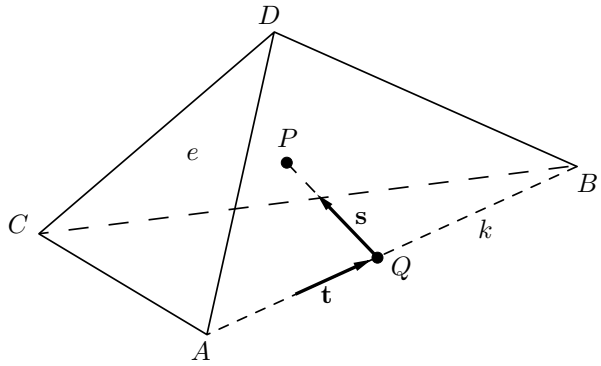


Figure 4.9: Unit vectors $\mathbf{t}$ and $\mathbf{s}$ along k and $\overline{QP}$

Let k be an arbitrary edge of e (see Fig. 4.9) and define $z := \mathbf{t}^T\nabla(u - I_e u)$ where $\mathbf{t}$ is a unit vector along k. Since $u - I_e u = 0$ at all the vertices of e ($I_e u$ is the nodal interpolant of u over e), by Rolle's theorem there exists a point $Q \in k$ such that

$$z(Q) = \mathbf{t}^T\nabla(u - I_e u)(Q) = 0.$$

If P is the above chosen arbitrary point in the interior of e and the mapping

$$\begin{aligned} g : [0,1] &\rightarrow \mathbb{R} \\ r &\mapsto (1-r)Q + rP \end{aligned}$$

parameterizes the way from Q to P we can write

$$z(P) - z(Q) = \int_0^1 (\nabla z \circ g)(r)|g'(r)|\,dr.$$

With $z(Q) = 0$ and $\mathbf{s}$ representing a unit vector along $\overline{QP}$ (compare Fig. 4.9) this is equivalent to

$$z(P) = \int_Q^P \mathbf{s}^T\nabla z(\sigma)\,d\sigma = \int_Q^P \mathbf{s}^T\Delta(u - I_e u)(\sigma)\mathbf{t}\,d\sigma,$$

where $\Delta(u - I_e u)(\sigma)$ is the Hessian matrix of $u - I_e u$ evaluated at σ. Thus, we conclude

$$\begin{aligned} |z(P)| &\leq \int_Q^P |\mathbf{s}^T \Delta(u - I_e u)(\sigma)\mathbf{t}|\, d\sigma \\ &\leq c \int_Q^P \sum_{|\alpha|=2} |D^\alpha (u - I_e u)(\sigma)|\, d\sigma \\ &\leq c \int_Q^P \operatorname{ess\,sup}_{x\in e} \Big(\sum_{|\alpha|=2} |D^\alpha (u - I_e u)(x)| \Big)\, d\sigma \\ &\leq c h_e \, |u - I_e u|_{W^2_\infty(e)} = c h_e \, |u|_{W^2_\infty(e)}, \end{aligned}$$

with ess sup denoting the essential supremum. If we use (4.24) and the density $W^2_\infty(e) = \overline{C^2(e)}$, we finally get

$$|u - I_e u|_{W^1_\infty(e)} \leq c h_e \, |u|_{W^2_\infty(e)}, \quad \forall u \in W^2_\infty(e).$$

◇

On the basis of this local result we can prove the following global interpolation error bounds for semiregular families of linear tetrahedral elements.

Theorem 4.7 (interpolation error bound) *Let $\mathcal{F} = \{\mathcal{T}_h\}$ be a semiregular family of tetrahedral triangulations of the polyhedron Ω whose boundary is Lipschitz. For $p \in [1, \infty]$ we have*

$$\begin{aligned} \|u - I_h u\|_{W^m_p(\Omega)} &\leq c_p \, \|u - I_h u\|_{W^m_\infty(\Omega)} \\ &\leq c h^{2-m} \, |u|_{W^2_\infty(\Omega)}, \quad m \in \{0, 1\}, \quad \forall u \in W^2_\infty(\Omega). \end{aligned} \tag{4.25}$$

Here, the constant c_p characterizes the topological embedding $W^m_\infty(\Omega) \hookrightarrow W^m_p(\Omega)$.

Note that estimate (4.25) in particular holds for $p = 2$. Since the standard nodal interpolant $I_h u$ belongs to the space X^1_h of $\mathcal{P}_1$-elements we can deduce the following approximation error bound.

Corollary 4.8 ($\mathcal{P}_1$-element approximation error bound) *For the approximation error of semiregular tetrahedral $\mathcal{P}_1$-elements we have*

$$\inf_{v_h \in X^1_h} \|u - v_h\|_{W^m_p(\Omega)} \leq c h^{2-m} \, |u|_{W^2_\infty(\Omega)}, \quad m \in \{0, 1\}, \quad \forall u \in W^2_\infty(\Omega), \tag{4.26}$$

and $p \in [1, \infty]$.

For the analysis of a robust multigrid method for anisotropic reaction-diffusion problems (see Chapter 5) we need a bound of the form (4.26) with $p = 2$ on the left- *and* right-hand side of the inequality. We remark that such a result holds (in the three-dimensional case, too). In the next section we illustrate that for the nodal Lagrangian interpolation operator a uniform bound of the form

$$\|u - I_h u\|_{H^1(\Omega)} \leq c h \, |u|_{H^2(\Omega)}, \quad \forall u \in H^2(\Omega),$$

with c independent of h does *not* hold in the anisotropic case. Thus, there is a need for other (better) interpolation operators for anisotropic finite element spaces (compare Section 5.4). However, using a more general approach to anisotropic estimates, we can give sharper bounds for the nodal Lagrangian interpolation error.

4.4 A general approach to anisotropic estimates

In the book of Apel [Ape99] estimates of the form

$$|u - I_h u|_{W_q^m(e)} \leq c\,(\mathrm{meas}_n\, e)^{\frac{1}{q}-\frac{1}{p}} h_e^{l-m} |u|_{W_p^l(e)},$$

with $u \in W_p^l(e)$, $\mathrm{meas}_n\, e$ denoting the area/volume of the n-dimensional finite element e, and admissible ranges of the parameters are called *isotropic estimates*. Apel states that these types of local interpolation error results are only rarely exploited for anisotropic finite element error estimates, since they do not extract the advantage of using elements with independent length scales. For this reason in [Ape99] so-called *anisotropic estimates* of the form

$$|u - I_h u|_{W_q^m(e)} \leq c\,(\mathrm{meas}_n\, e)^{\frac{1}{q}-\frac{1}{p}} \sum_{\substack{\alpha_1+\cdots+\alpha_n=l-m \\ \alpha_1,\ldots,\alpha_n \geq 0}} h_{1,e}^{\alpha_1} \cdots h_{n,e}^{\alpha_n} \left| \frac{\partial^{l-m} u}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}} \right|_{W_p^m(e)} \tag{4.27}$$

with different, suitably defined element sizes $h_{1,e}, \ldots, h_{n,e}$, are investigated. In this representation the idea of using small grid scales in the directions where the solution has large derivatives and relatively large grid scales in those directions where the derivatives are small becomes obvious. The main strategy in [Ape99] to achieve results of the form (4.27) still remains the same as for isotropic elements (compare Section 3.1.2), i.e. the derivation of an appropriate estimate on the reference element and the coordinate transformation. However, Apel shows that one has to deduce sharper estimates on the reference element and that the transformation has to be investigated very carefully. In this section we present the main ideas for triangular and tetrahedral elements given in [Ape99].

4.4.1 Estimates on the reference element

When the lemma of Deny and Lions (Lemma 3.12) is applied on the reference element, it leads to error bounds of the form

$$\|\hat{u} - \hat{I}_{\hat{e}} \hat{u}\|_{W_p^2(\hat{e})} \leq c\,|\hat{u}|_{W_p^2(\hat{e})}, \quad \forall \hat{u} \in W_p^2(\hat{e}), \tag{4.28}$$

for $p \in [1, \infty]$ and either $n/p < 2$ when $p > 1$ or $n = 2$ when $p = 1$ (see (3.20)). Using a transformation between an arbitrary element e and the reference element $\hat{e}$ and (4.28) an interpolation error bound for e can be shown (compare Theorem 3.13). However, for *anisotropic* elements the terms in the upper bound of the interpolation error that result from the transformation may have bad asymptotics as we discussed in the corresponding example of Section 3.1.2. In the following we reconsider this example (which is taken from Example 2.1 in [Ape99]) in order to make clear that there is a need for sharper estimates on the reference element to obtain satisfactory estimates of type (4.27) for anisotropic finite elements.

We consider the special triangular element

$$e = \{(x_1, x_2)^T \mid 0 < x_1 < h_1,\, 0 < x_2 < h_2(1 - x_1/h_1)\}, \tag{4.29}$$

which is shown in Fig. 3.2 from Section 3.1.2 and the corresponding reference element $\hat{e}$. We define $w := u - I_e u$ and $\hat{w} := w \circ F$, where I_e denotes the linear interpolation operator on e and $F : \hat{e} \to e$ the affine transformation. Then inequality (4.28) in particular leads to

$$\|\hat{D}^{(0,1)} \hat{w}\|_{L_p(\hat{e})} \leq c\,|\hat{u}|_{W_p^2(\hat{e})}, \quad \forall \hat{u} \in W_p^2(\hat{e}), \quad p \in [1, \infty]. \tag{4.30}$$

Let $\mathbf{h} := (h_1, h_2)^T$ and $\alpha = (\alpha_1, \alpha_2)$ be a multi-index. Then for the special element e in (4.29) the identity $\hat{D}^\alpha \hat{w} = \mathbf{h}^\alpha D^\alpha w$ holds. Thus, we get the following estimate for the partial

derivative in x_2-direction:

$$\begin{aligned}
\|D^{(0,1)}w\|^p_{L_p(e)} &= (h_1h_2)h_2^{-p}\,\|\hat{D}^{(0,1)}\hat{w}\|^p_{L_p(\hat{e})} \\
&\overset{(4.30)}{\leq} c(h_1h_2)h_2^{-p}\,|\hat{u}|^p_{W_p^2(\hat{e})} \\
&= c(h_1h_2)h_2^{-p}((h_1h_2)^{-1}\sum_{|\alpha|=2}\mathbf{h}^{\alpha p}\,\|D^\alpha u\|^p_{L_p(e)}) \\
&= ch_1^{2p}h_2^{-p}\,\|D^{(2,0)}u\|^p_{L_p(e)} + c\sum_{|\alpha|=1}\mathbf{h}^{\alpha p}\|D^{\alpha+(0,1)}u\|^p_{L_p(e)}. \quad (4.31)
\end{aligned}$$

If $h_2 = o(h_1)$ then the first term in (4.31) has bad asymptotics $h_1^2h_2^{-1} \sim h_e^2/\rho_e$. However, we can avoid this term if (4.30) can be improved to the sharper estimate

$$\|\hat{D}^{(0,1)}\hat{w}\|_{L_p(\hat{e})} \leq c\,|\hat{D}^{(0,1)}\hat{u}|_{W_p^1(\hat{e})}, \quad p \in [1,\infty].$$

Let $\hat{u} \in W_p^l(\hat{e})$ for some $l \leq k+1$. Here, k denotes the degree of polynomials contained in the space $\hat{\mathcal{P}}_{k,\hat{e}}$ of shape functions on the element $\hat{e}$. For various element types and suitably chosen parameter ranges Apel succeeds in showing inequalities of the form

$$\|\hat{D}^\alpha(\hat{u} - \hat{I}_{\hat{e}}\hat{u})\|_{L_q(\hat{e})} \leq c\,|\hat{D}^\alpha\hat{u}|_{W_p^{l-m}(\hat{e})}, \quad \forall\alpha,\ |\alpha| = m. \tag{4.32}$$

In this thesis we are interested in the choices $q = p$, $l = 2$ and $m \in \{0,1\}$. Since the resulting admissible ranges of the parameter p are different for two- and three-dimensional elements these are treated separately.

Triangular elements

From Lemma 2.4 in [Ape99] we get the estimate

$$\|\hat{D}^\alpha(\hat{u} - \hat{I}_{\hat{e}}\hat{u})\|_{L_p(\hat{e})} \leq c\,|\hat{D}^\alpha\hat{u}|_{W_p^{2-m}(\hat{e})}, \quad \forall\alpha,\ |\alpha| = m, \quad m \in \{0,1\},$$

for $\hat{u} \in C(\bar{\hat{e}})$ with $\hat{D}^\alpha\hat{u} \in W_p^{2-m}(\hat{e})$ and $p \in [1,\infty]$.

Tetrahedral elements

For tetrahedral elements the situation is slightly different. A crucial ingredient in the proof of (4.32) is the embedding $W_p^{l-m}(\hat{e}) \hookrightarrow L_1(G)$ for some domain $G \subset \hat{e}$. This embedding holds for triangular elements $\hat{e}$ without any restrictions but for tetrahedral elements $\hat{e}$ and one-dimensional G (i.e. two dimensions less than the one of $\hat{e}$) only if $p > 2$ in the case $l-m = 1$. Thus, for certain choices of the parameters the methods of proof in two dimensions cannot be applied in three dimensions since the embedding theorems depend on the space dimension. However, Lemma 2.6 in [Ape99] yields the estimate

$$\|\hat{D}^\alpha(\hat{u} - \hat{I}_{\hat{e}}\hat{u})\|_{L_p(\hat{e})} \leq c\,|\hat{D}^\alpha\hat{u}|_{W_p^{2-m}(\hat{e})}, \quad \forall\alpha,\ |\alpha| = m, \quad m \in \{0,1\}, \tag{4.33}$$

for $\hat{u} \in C(\bar{\hat{e}})$ with $\hat{D}^\alpha\hat{u} \in W_p^{2-m}(\hat{e})$ provided that

$$\begin{cases} p > 3/2 & \text{if } m = 0, \\ p > 2 & \text{if } m = 1. \end{cases}$$

Thus, for triangular elements we can choose $p = 2$ while this case is not covered for tetrahedral elements if $m = 1$. Example 2.6 in [Ape99] shows that for $m = 1$ the condition $p > 2$ in (4.33) is necessary. The following variant of this example is also given in [AD92].

Example:

Let $\hat{e}$ be the unit tetrahedron in $\mathbb{R}^3$ and $1 \le p \le 2$. We define the multi-index $\alpha := (1,0,0)$. For $\varepsilon > 0$ we introduce the function

$$\hat{u}_\varepsilon(\mathbf{x}) := (1 - \min\{1, \varepsilon \ln|\ln(r/\mathrm{e})|\})\, x_1,$$

with $r = r(x_2, x_3) := (x_2^2 + x_3^2)^{1/2}$ and Euler's constant e. From $\hat{u}_\varepsilon(x_1, 0, 0) = 0$ it follows that $\hat{D}^\alpha \hat{I}_{\hat{e}} \hat{u}_\varepsilon(\mathbf{x}) = 0$. Furthermore,

$$\hat{D}^\alpha \hat{u}_\varepsilon(\mathbf{x}) = (1 - \min\{1, \varepsilon \ln|\ln(r/\mathrm{e})|\}) \to 1, \quad \varepsilon \downarrow 0,$$

pointwise in $\hat{e}$. Thus, for $\varepsilon \downarrow 0$, we get

$$\|\hat{D}^\alpha(\hat{u}_\varepsilon - \hat{I}_{\hat{e}}\hat{u}_\varepsilon)\|^p_{L_p(\hat{e})} = \int_{\hat{e}} |\hat{D}^\alpha(\hat{u}_\varepsilon - \hat{I}_{\hat{e}}\hat{u}_\varepsilon)|^p \, d\mathbf{x} \to |\hat{e}| = \frac{1}{6},$$

while

$$\begin{aligned} |\hat{D}^\alpha \hat{u}_\varepsilon|^p_{W^1_p(\hat{e})} &= |1 - \min\{1, \varepsilon \ln|\ln(r/\mathrm{e})|\}|^p_{W^1_p(\hat{e})} \\ &\le c\varepsilon \int_0^1 r^{-p+1} |\ln(r/\mathrm{e})|^{-p} \, dr \\ &\le c\varepsilon \int_0^1 r^{-1} |\ln(r/\mathrm{e})|^{-2} \, dr \to 0. \end{aligned} \tag{4.34}$$

In [Ape99] Apel states that $|\hat{D}^\alpha \hat{u}_\varepsilon|^p_{W^1_p(\hat{e})} \to \infty$ for $\varepsilon \downarrow 0$ in the case $p > 2$, which means that the integral in (4.34) is unbounded for $p > 2$. For the further considerations we restrict ourselves to three dimensions.

4.4.2 The coordinate transformation

In the preceding section we discussed sharper estimates on the reference element. Now we turn to the second essential task for the derivation of anisotropic estimates, i.e. we consider properties of the transformation matrix (compare Lemma 3.11 in Section 3.1.2). We first introduce some notation.

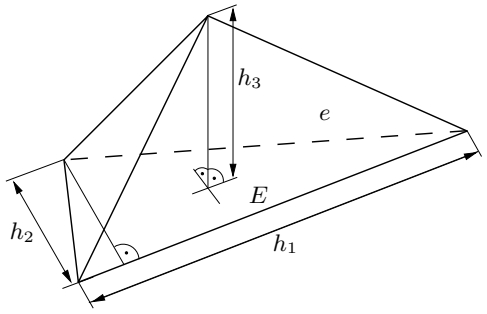


Figure 4.10: Element sizes of the tetrahedron e

Let E be the longest edge of the element e, and let Γ_E be the larger of the two faces of e with $E \subset \overline{\Gamma_E}$. Then we define the element sizes

$$h_1 := |E|, \quad h_2 := 2|\Gamma_E|/h_1, \quad h_3 := 6|e|/(h_1 h_2),$$

as shown in Fig. 4.10 (with $|\cdot|$ denoting the associated measures). On the basis of these definitions in [Ape99] an element related Cartesian coordinate system $\mathbf{x}_e = (x_{1,e}, x_{2,e}, x_{3,e})$ such that one of the vertices of e is at the origin is introduced and the following condition is formulated.

Coordinate system condition (in 3D):

The transformation of the element related coordinate system $(x_{1,e}, x_{2,e}, x_{3,e})$ to the discretization independent system (x_1, x_2, x_3) can be determined as a translation and three rotations around the $x_{j,e}$-axes by angles ϑ_j, $j = 1, 2, 3$, where

$$|\sin\vartheta_1| \lesssim \frac{h_3}{h_2}, \quad |\sin\vartheta_2| \lesssim \frac{h_3}{h_1}, \quad |\sin\vartheta_3| \lesssim \frac{h_2}{h_1}. \tag{4.35}$$

Here, the notation $\lesssim$ means smaller than up to a constant. We give an illustration of this condition in Fig. 4.11.

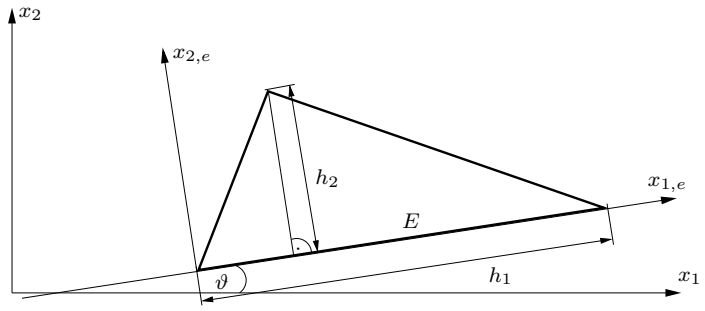


Figure 4.11: Illustration of the coordinate system condition

In $\mathbb{R}^2$ we only have the condition $|\sin\vartheta| \lesssim h_2/h_1$, which means that for $h_2/h_1 \ll 1$ the angle ϑ between the longest side E of the element e and the x_1-axis has to be small. The coordinate system condition and the maximum angle condition formulated in Section 4.3 (see also Remark 4.4) yield the following properties of the transformation matrix (compare [Ape99]). These are sufficient to prove an anisotropic estimate of the form (4.27).

Lemma 4.9 *Assume that the tetrahedron e satisfies the maximum angle condition* (4.10)-(4.11) *and the coordinate system condition* (4.35). *Then the entries of the transformation matrix* $\mathbf{B}$ *in* (3.14) *and of its inverse* $\mathbf{B}^{-1}$ *satisfy the conditions:*

$$\begin{aligned} |b_{i,j}| &\lesssim \min\{h_i, h_j\}, \quad i,j = 1,2,3, \\ |b^{-1}_{i,j}| &\lesssim \min\{h_i^{-1}, h_j^{-1}\}, \quad i,j = 1,2,3. \end{aligned}$$

Theorem 4.10 *Assume that the element e satisfies the maximum angle condition* (4.10)-(4.11) *and the coordinate system condition* (4.35). *Let* $\mathbf{h} = (h_{1,e}, h_{2,e}, h_{3,e})^T$ *be the vector that contains the different length scales of the element e. Then there exists a constant c independent of e such that the anisotropic interpolation error estimate*

$$|u - I_h u|_{W_p^m(e)} \le c \sum_{|\alpha|=2-m} \mathbf{h}^\alpha |D^\alpha u|_{W_p^m(e)}, \quad m \in \{0,1\}, \tag{4.36}$$

holds for $u \in W_p^2(e) \cap C(\bar{e})$ and $p \in (2, \infty]$.

We remark that the constant c in (4.36) depends on p, but it is independent of u. Theorem 4.10 trivially implies the *iso*tropic interpolation error estimate. We emphasize that in the following corollary the coordinate system condition is *not* needed anymore (cf. [Ape99]).

Corollary 4.11 *Assume that the element e satisfies the maximum angle condition* (4.10)-(4.11). *Then there exists a constant c independent of e such that the isotropic interpolation error estimate*

$$|u - I_h u|_{W_p^m(e)} \le c h_e^{2-m} |u|_{W_p^2(e)}, \quad m \in \{0,1\},$$

holds for $u \in W_p^2(e) \cap C(\bar{e})$ and $p \in (2, \infty]$.

Proof: Suppose that the coordinate system condition (4.35) holds. Then the result follows immediately from Theorem 4.10. Since the seminorms remain equivalent during a rotation of the coordinate system the corollary is proved. ◇

Standard arguments (compare Section 4.3) yield the following global result.

Theorem 4.12 (interpolation error bound) *Let $\mathcal{F} = \{\mathcal{T}_h\}$ be a semiregular family of tetrahedral triangulations of the polyhedron Ω whose boundary is Lipschitz. Then there exists a constant c independent of h such that*

$$\|u - I_h u\|_{W_p^m(\Omega)} \leq c h^{2-m} |u|_{W_p^2(\Omega)}, \quad m \in \{0,1\}, \quad \forall u \in W_p^2(\Omega), \tag{4.37}$$

and $p \in (2, \infty]$.

Note that from the general approach of Apel the special result of Křížek in (4.25) directly follows. For $p \in [1, \infty]$, $q \in (2, \infty]$ with $q \geq p$ and all $u \in W_\infty^2(\Omega)$, using (4.37), we can write

$$\begin{aligned} \|u - I_h u\|_{W_p^m(\Omega)} &\leq c \|u - I_h u\|_{W_q^m(\Omega)} \\ &\leq c h^{2-m} |u|_{W_q^2(\Omega)} \\ &\leq c h^{2-m} |u|_{W_\infty^2(\Omega)}, \quad m \in \{0,1\}, \end{aligned}$$

due to the topological embeddings $W_\infty^l(\Omega) \hookrightarrow W_q^l(\Omega) \hookrightarrow W_p^l(\Omega)$, $l \in \mathbb{N}$.

Corollary 4.13 ($\mathcal{P}_1$-element approximation error bound) *For the approximation error of semiregular tetrahedral $\mathcal{P}_1$-elements we have*

$$\inf_{v_h \in X_h^1} \|u - v_h\|_{W_p^m(\Omega)} \leq c h^{2-m} |u|_{W_p^2(\Omega)}, \quad m \in \{0,1\}, \quad \forall u \in W_p^2(\Omega), \tag{4.38}$$

and $p \in (2, \infty]$.

4.5 Finite element discretization error bound

Now we consider the reaction-diffusion problem (3.1) and its finite element discretization with semiregular tetrahedral $\mathcal{P}_1$-elements. To obtain a bound for the discretization error we use the Céa-lemma and the error bounds for standard nodal interpolation discussed in the previous section.

Theorem 4.14 *Let u be the solution of the continuous problem* (3.3) *and u_h the solution of the discrete problem* (3.8) *with the space $V_h = X_h^1$ of semiregular $\mathcal{P}_1$-elements. If $u \in W_p^2(\Omega)$ for some $p \in (2, \infty]$ then the error estimate*

$$\|u - u_h\|_{H^1(\Omega)} \leq c h |u|_{W_p^2(\Omega)}$$

holds, with a constant c independent of h and u.

Proof: Let $p \in (2, \infty]$ and assume $u \in W_p^2(\Omega)$. From the Céa-lemma 3.6 it follows that the continuous and the discrete problems have unique solutions. Since $W_p^1(\Omega) \hookrightarrow H^1(\Omega)$, using the approximation error bound (4.38), we get

$$\begin{aligned} \|u - u_h\|_{H^1(\Omega)} &\leq c \inf_{v_h \in X_h^1} \|u - v_h\|_{H^1(\Omega)} \\ &\leq c \inf_{v_h \in X_h^1} \|u - v_h\|_{W_p^1(\Omega)} \\ &\leq c h |u|_{W_p^2(\Omega)}. \end{aligned}$$

◇

From this result we can see that semiregular tetrahedral $\mathcal{P}_1$-elements are satisfactory for the finite element discretization of our problem class. Note that we are not able to give a bound for the L_2-error in analogy to Theorem 3.17 since the duality argument of Aubin-Nitsche does not lead to the desired estimate. Due to the restriction $p > 2$ in Theorem 4.14, for the analysis of a robust multigrid method in the next chapter, we study a modified interpolation operator which is also introduced by Apel in [Ape99].

Chapter 5

A robust multigrid method for anisotropic reaction-diffusion problems

In this chapter we consider a linear reaction-diffusion boundary value problem in a three-dimensional *thin* domain. The very different length scales in the geometry result in an anisotropy effect. Our study is motivated by the inverse heat conduction problem studied in this thesis [GSM+05]. In the solution method (explained in Chapter 2) direct heat conduction problems in a thin foil have to be solved leading to such linear anisotropic reaction-diffusion problems in each time step of an implicit time integration method (compare Remark 3.19). The reaction-diffusion problem contains two important parameters, namely $\varepsilon > 0$ which parameterizes the thickness of the domain and $\mu > 0$ denoting the measure for the size of the reaction term relative to that of the diffusion term. We analyze the convergence of a multigrid method with a robust (line) smoother [RS07]. Both, for the W- and the V-cycle method we derive contraction number bounds smaller than one uniform with respect to the mesh size and the parameters ε and μ.

5.1 An anisotropic reaction-diffusion model problem

We study a reaction-diffusion boundary value problem on the domain $\Omega_\varepsilon := [0,1]^2 \times [0,\varepsilon]$ with $0 < \varepsilon \leq 1$. For the standard scalar product and norm in $L_2(\Omega_\varepsilon)$ we use the notation $(\cdot,\cdot)_0$ and $\|\cdot\|_0$. The scalar products and corresponding norms in $H^k(\Omega_\varepsilon)$, $k = 1, 2$, are denoted by $(\cdot,\cdot)_k$ and $\|\cdot\|_k$, respectively. Let the Dirichlet and Neumann boundaries of Ω_ε be denoted by

$$\begin{aligned} \Gamma_D &:= \{(x,y,z) \,|\, (x,y) \in [0,1]^2, z \in \{0,\varepsilon\}\}, \\ \Gamma_N &:= \{(x,y,z) \,|\, (x,y) \in \{0,1\}^2, z \in [0,\varepsilon]\}, \end{aligned} \tag{5.1}$$

and define $U_\varepsilon := \{v \in H^1(\Omega_\varepsilon) \,|\, v = 0 \text{ on } \Gamma_D\}$.

Remark 5.1 For the robustness analysis of the z-line smoother in Section 5.6 we assume Dirichlet boundary conditions on the upper and lower faces of the domain Ω_ε. These boundary conditions are conserved by the modified interpolation operator that is introduced below. As far as we know there is no variant of this operator that handles the problem with pure Dirichlet boundary conditions. Therefore, we restrict our considerations to the combination of Dirichlet and Neumann boundaries as in (5.1). Note that we have to treat reaction-diffusion problems with *pure* Neumann boundary conditions for the solution of the inverse heat conduction problem (see Section 2.3). However, the numerical experiments of Chap-

ter 6 show, that the suggested robust multigrid method of this chapter seems to work quite well for this case, too.

For $\mu > 0$ we introduce the bilinear form

$$a(u,v) := (\nabla u, \nabla v)_0 + \mu(u,v)_0 \quad \text{for all} \;\; u, v \in U_\varepsilon.$$

This bilinear form is continuous and elliptic on U_ε. For given $f \in L_2(\Omega_\varepsilon)$ we consider the following problem: find $u \in U_\varepsilon$ such that

$$a(u,v) = (f,v)_0 \quad \text{for all} \;\; v \in U_\varepsilon. \tag{5.2}$$

For the discretization of this problem we use standard linear conforming finite elements on a nested family of uniform tetrahedral grids. To obtain a bound for the discretization error we use the Céa-lemma and a suitable interpolation operator. In a *two*-dimensional domain one can apply the standard Lagrangian interpolation operator even for such anisotropic problems. For the *three*-dimensional case, however, this operator is not satisfactory due to the fact that the case $p = 2$ is not covered in Theorem 4.12 (compare also Theorem 4.10 in Chapter 4). Instead we use a modified Scott-Zhang interpolation operator which is introduced in [Ape99]. Based on the modified Scott-Zhang interpolation operator we derive a finite element discretization error bound in which the dependence on the parameters ε and μ is explicit. To solve the discrete problem we consider a multigrid method with a symmetric z-line Gauss-Seidel smoother. The main topic of this chapter is a convergence analysis of this method. For the multigrid W-cycle we analyze the convergence in the framework of the approximation and smoothing property. We prove robustness of the multigrid W-cycle method in the sense that (for sufficiently many smoothing iterations) its contraction number in the Euclidean norm is bounded by a constant smaller than one independent of all the parameters. On the basis of [Ste93b] and [Ste94] we also prove a robustness result for the V-cycle multigrid method. Finally, we present numerical experiments that illustrate these robustness properties.

In the literature the convergence of multigrid methods for anisotropic pure diffusion problems, i.e. a problem as in (5.2) with $\mu = 0$, has been studied in [BZ00, Ste93a, Ste93b, Ste94] and [Wit89a, Wit89b]. In the latter papers the robustness of smoothers is studied, whereas in the former the convergence of W-cycle and V-cycle algorithms is analyzed. These convergence analyses of multigrid methods are based on the standard Lagrangian interpolation operator and (thus) are restricted to the *two*-dimensional case. In [AS02] a finite element method for the Poisson problem on three-dimensional domains with anisotropic mesh refinement is studied. For a multigrid scheme in which semicoarsening and line smoothers are combined, robust convergence of the V-cycle is shown. In all these analyses only the case $\mu = 0$ is considered. In this thesis we treat the *three*-dimensional case and consider the *additional (reaction) parameter* $\mu > 0$ [RS07].

5.2 Finite element discretization

From the Lax-Milgram Lemma 3.3 it follows that problem (5.2) has a unique solution. Note, that the very different length scales in the (x,y)- and the z-direction (for $\varepsilon \ll 1$) result in an anisotropy effect.

Remark 5.2 Instead of (5.2) we could also consider the weak formulation of the following anisotropic reaction-diffusion problem on the unit cube $\Omega := [0,1]^3$:

$$\begin{aligned} -u_{xx} - u_{yy} - \lambda u_{zz} + \mu u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma_D, \\ \frac{\partial u}{\partial \mathbf{n}} &= 0 \quad \text{on } \Gamma_N, \end{aligned} \tag{5.3}$$

with a parameter $\lambda = 1/\varepsilon^2 \geq 1$. The discrete versions of both formulations (5.2) and (5.3) lead to operators that have very similar anisotropy properties. In this thesis we consider (5.2) because our research is motivated by the heat conduction problem in a thin foil, which is a domain of the form Ω_ε with $\varepsilon \ll 1$ (cf. Chapter 2 and [GSM+05]). An implicit time integration method applied to this parabolic problem leads to a problem of the form (5.2) in each time step.

For the discretization we apply a standard finite element method based on a *uniform* family of nested triangulations. The uniform subdivision of the domain is based on Kuhn's triangulation, as illustrated in Fig. 5.1.

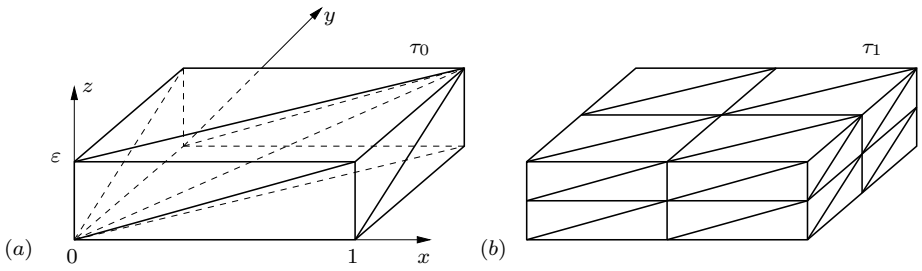


Figure 5.1: (a) Initial Kuhn triangulation $\mathcal{T}_0$ and (b) first refinement $\mathcal{T}_1$.

A stable regular (red) refinement strategy results in a family of consistent, nested triangulations (see [Bey98]), which is denoted by $\{\mathcal{T}_k\}_{k\geq 0}$. With $\mathcal{T}_k$ we associate the mesh size parameter $h_k = (\frac{1}{2})^k$. Note that for all $T \in \mathcal{T}_k$ we have

$$\begin{aligned} h_T &= \operatorname{diam}(T) = \sqrt{2}\, h_k, \\ \rho_T &= \sup\{\operatorname{diam}(S) \,|\, S \text{ is a ball contained in } T\} \sim \varepsilon h_k. \end{aligned}$$

Thus, this family of triangulations is regular in the sense that

$$\sigma := \sup_{k\in\mathbb{N}} \sup_{T\in\mathcal{T}_k} \frac{h_T}{\rho_T} < \infty$$

holds. However, $\sigma \sim \varepsilon^{-1}$ and thus $\sigma \to \infty$ for $\varepsilon \downarrow 0$. In Fig. 5.2 we show one particular hexahedron on level k and a typical tetrahedron $T \in \mathcal{T}_k$. Due to the degeneracy of the given domain Ω_ε, inside the elements T arbitrarily small angles appear for $\varepsilon \downarrow 0$. On the other hand the maximum angles that occur are right angles meaning that a *maximum* angle condition is satisfied uniformly w.r.t. k and ε.

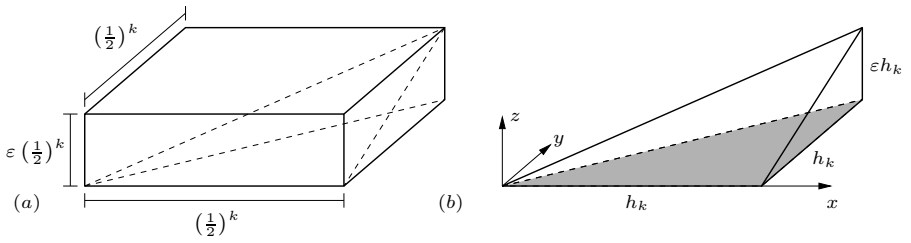


Figure 5.2: (a) Hexahedron on level k and (b) typical tetrahedron $T \in \mathcal{T}_k$.

For the discretization of (5.2) we use conforming finite elements and piecewise linear functions ($\mathcal{P}_1$-elements) with respect to the sequence of nested triangulations $\{\mathcal{T}_k\}_{k\geq 0}$. This results in a hierarchy of nested finite element spaces

$$U_{\varepsilon,0} \subset U_{\varepsilon,1} \subset \ldots \subset U_\varepsilon.$$

The discrete problem on level k is: find $u_k \in U_{\varepsilon,k}$ such that

$$a(u_k, v_k) = (f, v_k)_0 \quad \text{for all} \quad v_k \in U_{\varepsilon,k}. \tag{5.4}$$

Due to the fact that a maximum angle condition is satisfied the spaces $U_{\varepsilon,k}$ of dimension n_k are suitable for the spatial discretization of the parabolic problem from which (after implicit time integration) problem (5.2) originates. For fixed $\mu > 0$, this follows from Theorem 4.14.

5.3 Fourier analysis of the discrete model problem

In this section we study the discrete reaction-diffusion problem that is obtained by a *finite difference* discretization of problem (5.2) on the uniform family of triangulations $\{\mathcal{T}_k\}_{k\geq 0}$ by means of Fourier analysis. This is possible since the differential equation has constant coefficients. We analyze the spectrum of the corresponding stiffness matrix $\mathbf{A}_{FD,k}$ on level k in dependence of the anisotropy parameter ε. Further, we derive an upper bound for the spectral condition number $\text{cond}_2(\mathbf{A}_{FD,k})$ uniformly in ε. Note, that the matrix $\mathbf{A}_{FD,k}$ is slightly different from the one obtained by the finite element discretization (5.4) due to the boundary conditions. However, the structure of the eigenvalue distributions of both matrices is comparable. The gained insights play an important role with respect to the convergence of the (preconditioned) CG algorithm when applied to problem (5.4).

We consider the pure Poisson problem on Ω_ε ($\mu = 0$) with mixed boundaries Γ_D and Γ_N as in (5.1). For the discretization (including the Neumann boundary) we use central second-order finite differences. In an interior grid point the corresponding difference stencil is the same as for the discretization based on linear tetrahedral finite elements (compare Fig. 5.6). However, on the Neumann boundary Γ_N the situation is slightly different. In Fig. 5.3 we give the difference stencils for two points on the left boundary ($x = 0$) of Ω_ε.

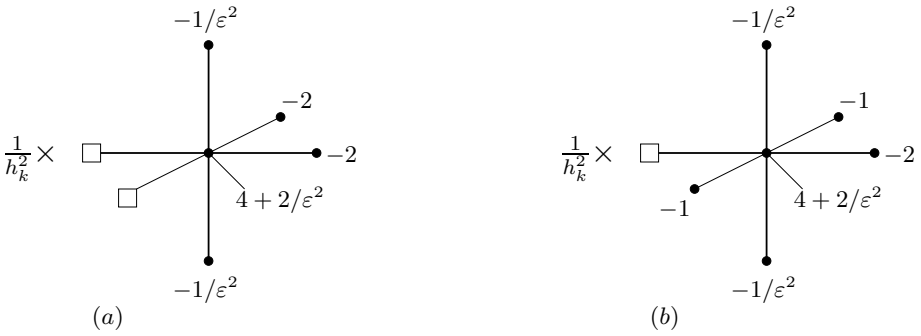


Figure 5.3: 3D difference stencil in (a) the left front edge and (b) the interior of the left boundary; (□) marking the auxiliary points.

As described in [TOS01] for the discretization of Γ_N an extended grid with auxiliary points (□) outside Ω_ε is introduced. These auxiliary points are also called *ghost points*. In this way, assuming that the right-hand side of the Poisson equation can be extended to Γ_N, the discrete Laplace operator can also be applied on the Neumann boundary. Then the discrete Poisson equation is used to provide a new approximation for the unknowns on Γ_N whereas the discrete Neumann boundary condition is used to eliminate the unknowns associated with the ghost points. In that way we get the couplings given in Fig. 5.3.

The corresponding stiffness matrix $\mathbf{A}_{FD,k}$ is symmetric and thus there exists an orthogonal basis of eigenvectors of this matrix. For the construction of this basis we introduce the continuous *Fourier modes*

$$e^{\nu\eta\xi}(x, y, z) = \cos(\nu\pi x)\cos(\eta\pi y)\sin\left(\xi\frac{\pi}{\varepsilon}z\right), \quad \nu, \eta, \xi \in \mathbb{N}.$$

Note that the Fourier modes which satisfy the mixed Dirichlet and Neumann boundary conditions are eigenfunctions of the Laplace operator

$$-\Delta e^{\nu\eta\xi} = ((\nu\pi)^2 + (\eta\pi)^2 + (\xi\pi/\varepsilon)^2)e^{\nu\eta\xi}.$$

Let $N_k := 2^k - 1$. The corresponding discrete Fourier modes on the grid $\mathcal{T}_k$ are defined by

$$e_k^{\nu\eta\xi}(\mathbf{x}) = e^{\nu\eta\xi}(\mathbf{x}), \quad \text{for } \mathbf{x} \in \mathcal{T}_k, \quad 0 \le \nu, \eta \le N_k + 1, \quad 1 \le \xi \le N_k.$$

If we use the same ordering of the grid points as for setting up the stiffness matrix to each Fourier mode $e_k^{\nu\eta\xi}$ there corresponds a unique vector which is denoted by $\mathbf{e}_k^{\nu\eta\xi}$. These vectors are eigenvectors of the stiffness matrix

$$\begin{aligned} \mathbf{A}_{FD,k}\mathbf{e}_k^{\nu\eta\xi} &= \lambda_{\nu\eta\xi}\mathbf{e}_k^{\nu\eta\xi}, \\ \lambda_{\nu\eta\xi} &= 4h_k^{-2}(\cos^2(\nu\pi h_k/2) + \cos^2(\eta\pi h_k/2) + \varepsilon^{-2}\sin^2(\xi\pi h_k/2)), \\ &\quad 0 \le \nu, \eta \le N_k + 1, \quad 1 \le \xi \le N_k. \end{aligned} \tag{5.5}$$

From this we are able to deduce the following result.

Lemma 5.3 *For the spectral condition number of the stiffness matrix $\mathbf{A}_{FD,k}$ on level k that corresponds to the second-order central difference discretization of the Poisson equation on Ω_ε with mixed boundaries* (5.1) *we have*

$$\mathrm{cond}_2(\mathbf{A}_{FD,k}) \le ch_k^{-2}, \quad k = 0, 1, \ldots$$

with c independent of ε and k.

Proof: Let $\lambda_{\max}(\mathbf{A}_{FD,k})$ and $\lambda_{\min}(\mathbf{A}_{FD,k})$ denote the in absolute value largest and smallest eigenvalues of $\mathbf{A}_{FD,k}$, respectively. Using (5.5) we get

$$\lambda_{\max}(\mathbf{A}_{FD,k}) \le 4h_k^{-2}(1 + 1 + \varepsilon^{-2}), \tag{5.6}$$

$$\lambda_{\min}(\mathbf{A}_{FD,k}) \ge 4h_k^{-2}(0 + 0 + \varepsilon^{-2}\sin^2(\pi h_k/2)) \ge 4h_k^{-2}\varepsilon^{-2}(\pi h_k/2)^2. \tag{5.7}$$

Thus, using $0 < \varepsilon \le 1$, we obtain

$$\mathrm{cond}_2(\mathbf{A}_{FD,k}) = \frac{\lambda_{\max}(\mathbf{A}_{FD,k})}{\lambda_{\min}(\mathbf{A}_{FD,k})} \le \frac{2 + \varepsilon^{-2}}{\varepsilon^{-2}(\pi h_k/2)^2} \le \frac{12}{\pi^2}h_k^{-2}.$$

◇

From Lemma 5.3 it follows that we can give an upper bound for the condition number of $\mathbf{A}_{FD,k}$ uniformly in the anisotropy parameter ε. Further, we can conclude that for decreasing ε clusters of eigenvalues develop since for $\varepsilon \downarrow 0$ the last summand in (5.5) gets dominant. This effect is shown in Fig. 5.4. While the spectrum of $\mathbf{A}_{FD,3}$ is almost continuous on the unit cube (a) there are $N_3 = 7$ eigenvalue clusters on the degenerate domain (b). The spectrum of the stiffness matrix $\mathbf{A}_k$ that corresponds to the space discretization with linear tetrahedral finite elements on level k is also given in Fig. 5.4 for $k = 3$. As already mentioned, there are some differences between the finite difference and the finite element discretization concerning the unknowns on the Neumann boundary Γ_N. These differences are related to the spacial orientation of the tetrahedra and consequently the varying supports of the finite element basis functions on Γ_N, resulting e.g. in varying diagonal entries of $\mathbf{A}_3$. However, the plots in Fig. 5.4 show comparable eigenvalue distributions. Note, that the cluster development for $\varepsilon \downarrow 0$ in general leads to a convergence improvement of the CG method when used as a solver for the discrete problem (see Lemma 3.23 in Section 3.3.2).

Note that the above considerations also hold for the reaction-diffusion problem (5.4) with $\mu > 0$. In this case the spectrum of the stiffness matrix is shifted to the right on the positive real axes due to the mass matrix term. The above analysis is also applicable for *pure* Dirichlet or Neumann boundary conditions in (5.4).

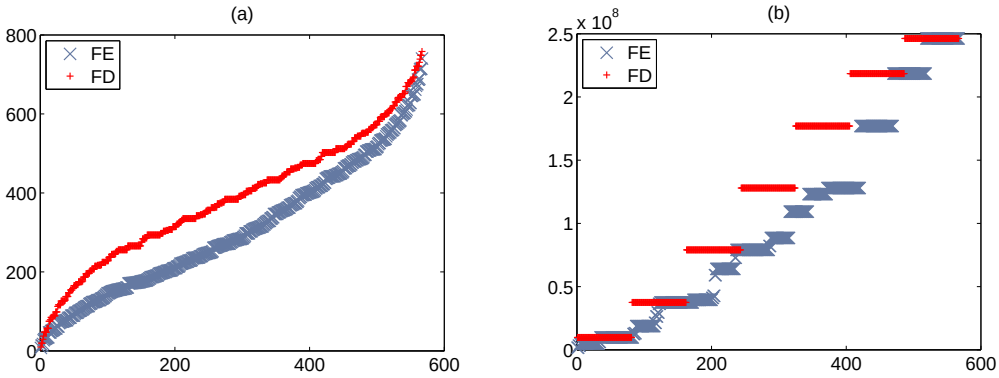


Figure 5.4: Spectrum of the stiffness matrix for (a) $\varepsilon = 1$ and (b) $\varepsilon = 10^{-3}$ based on finite differences (FD) and finite elements (FE) on level $k = 3$ with h_3=1/8.

Remark 5.4 In Lemma 5.11 from Section 5.6 we consider the scaling of the stiffness matrix $\mathbf{A}_k$ and show that

$$\lambda_{\max}(\mathbf{A}_k) \sim \left(\frac{1}{\varepsilon^2 h_k^2} + \mu \right) \tag{5.8}$$

holds, i.e. the largest eigenvalue of $\mathbf{A}_k$ tends to infinity for $\varepsilon \downarrow 0$. This bound has the same form as the one in (5.6). For the smallest eigenvalue we obtain

$$\lambda_{\min}(\mathbf{A}_k) = \min_{v \in U_{\varepsilon,k}} \frac{|v|_1^2 + \mu \|v\|_0^2}{\|v\|_0^2} \geq c + \mu, \tag{5.9}$$

with a constant $c > 0$, if we use the Poincaré-Friedrichs inequality

$$\|v\|_0 \leq c(\Omega_\varepsilon)\, |v|_1 \quad \text{for all } v \in U_\varepsilon,$$

in which the constant only depends on the diameter of Ω_ε. This bound is worse than the one in (5.7). Combination of (5.8) and (5.9) leads to

$$\mathrm{cond}_2(\mathbf{A}_k) = \mathcal{O}\left(\frac{1}{\varepsilon^2 h_k^2} \right),$$

which is worse than the corresponding result for $\mathbf{A}_{FD,k}$ in Lemma 5.3. Numerical experiments confirm that $\mathrm{cond}_2(\mathbf{A}_k)$ is bounded uniformly in ε, too.

5.4 Interpolation error bounds

In our convergence analysis of the multigrid method we need finite element discretization error bounds. If we apply the standard approach based on the Céa-lemma then a key ingredient for obtaining such bounds is a suitable (quasi-)interpolation operator

$$I_k : H^2(\Omega_\varepsilon) \to U_{\varepsilon,k}, \quad k = 0, 1, \ldots$$

This operator I_k should be such that for all $u \in H^2(\Omega_\varepsilon)$ the following error bounds hold

$$\|u - I_k u\|_0 \leq c_1 h_k^2\, |u|_2, \tag{5.10}$$
$$\|u - I_k u\|_1 \leq c_2 h_k\, |u|_2, \tag{5.11}$$

with constants c_1, c_2 *independent of* ε (and, as usual, also of k, u). We refer to [Ape99] for an extensive treatment of interpolation operators for anisotropic finite element spaces.

Here we briefly discuss a few issues that are relevant for the multigrid convergence analysis in this chapter. For the *two*-dimensional case uniform bounds as in (5.10)-(5.11) hold for the standard nodal Lagrangian interpolation operator (cf. Section 4.4 and Corollary 2.1 in [Ape99]). For *three*-dimensional problems the nodal Lagrangian interpolation operator still satisfies the uniform bound (5.10), but a result as in (5.11) does *not* hold in the anisotropic case (see Section 4.4). However, for this operator the result in (5.11) "almost" holds, in the following sense. For any $p > 2$ an estimate of the form

$$\|u - I_k u\|_{W_p^1(\Omega_\varepsilon)} \leq c_p h_k \, |u|_{W_p^2(\Omega_\varepsilon)} \quad \text{for all} \;\; u \in W_p^2(\Omega_\varepsilon),$$

holds, with c_p independent of ε (cf. Theorem 4.12). For the Clément and Scott-Zhang interpolation operators the uniform bound (5.10) holds but again (5.11) does not hold (cf. Example 3.1 and further comments in [Ape99]). Thus, there is a need for other (better) interpolation operators for anisotropic finite element spaces. Such operators are presented in [Ape99]. In particular a modification of the original Scott-Zhang operator is introduced which can be shown to satisfy, for the *three*-dimensional case, both uniform bounds (5.10) and (5.11). This operator is needed in the finite element discretization error analysis in the next section and therefore we describe how this operator is defined. A detailed discussion of this operator and its properties can be found in [Ape99] (Section 3.4).

For the description of the modified Scott-Zhang operator we need some additional notation. Let $\{X_i\}_{1 \leq i \leq \tilde{n}_k}$ denote the set of vertices of the triangulation $\mathcal{T}_k$ including those on the entire boundary (i.e. in particular on the Dirichlet boundary) and $\{\phi_i\}_{1 \leq i \leq \tilde{n}_k}$ the corresponding standard nodal basis which generates the finite element space $V_{\varepsilon,k}$. Note that we consider $\tilde{n}_k$ vertices while n_k denotes the dimension of the original finite element space $U_{\varepsilon,k}$. For an element $T \in \mathcal{T}_k$ we introduce the patch of surrounding elements

$$S_T := \bigcup \{T' \in \mathcal{T}_k \mid \bar{T}' \cap \bar{T} \neq \emptyset\}.$$

To each node X_i we associate a planar subdomain $\sigma_i \subset \Omega_\varepsilon$ with the following properties:

(P1) σ_i is parallel to the x, y-plane.

(P2) $X_i \in \bar{\sigma}_i$.

(P3) There exists a face E of some element $T \in \mathcal{T}_k$ such that the projection of E on the x, y-plane is identical to the projection of σ_i.

(P4) If the projections of any two points X_i and X_j on the x, y-plane coincide then so do the projections of σ_i and σ_j.

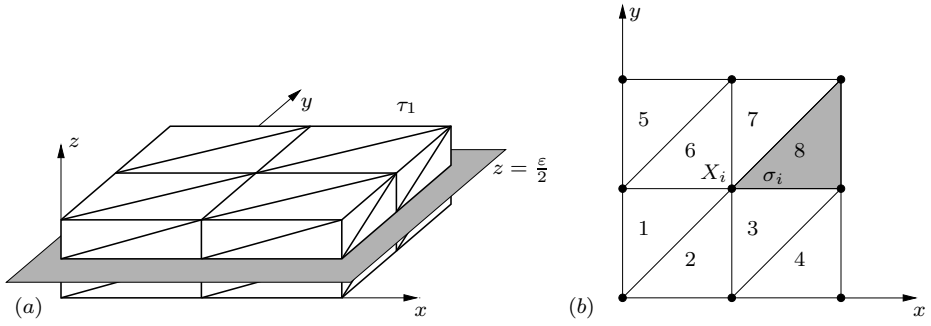


Figure 5.5: (a) $\mathcal{T}_1$ with plane $z = \frac{\varepsilon}{2}$ and (b) corresponding subdivision into faces.

In the triangulation $\mathcal{T}_k$ all the vertices are contained in planes $z = l\,\varepsilon(1/2)^k$, $l = 0, \ldots, 2^k$, which are subdivided into faces (triangles). Such a subdivision and the corresponding degrees

of freedom are shown in Fig. 5.5 for the case $k = 1$, $l = 1$.

One possibility to select the subdomains σ_i is to assume a lexicographical numbering of the faces (see Fig. 5.5 (b)) which should be the same in all the planes and to associate to each node X_i the face with maximum number. In this way the properties (P1)–(P3) are obviously satisfied (cf. shaded face in Fig. 5.2 (b) for (P3)). Due to the refinement strategy used, on each fixed level k the corresponding subdivisions into faces are identical for all the planes and thus (P4) is fulfilled, too. Given these subdomains σ_i the modified Scott-Zhang type interpolation operator $\mathcal{L}_k : H^1(\Omega_\varepsilon) \to V_{\varepsilon,k}$ from [Ape99] is based on the local L_2-orthogonal projections on σ_i:

$$\|u - \Pi_{\sigma_i} u\|_{L_2(\sigma_i)} = \min_{v \in \mathcal{P}_1} \|u - v\|_{L_2(\sigma_i)}, \quad i = 1, \ldots, \tilde{n}_k, \quad \text{for } \; u \in H^1(\Omega_\varepsilon).$$

For $u \in H^1(\Omega_\varepsilon)$ we define

$$\mathcal{L}_k u := \sum_{i=1}^{\tilde{n}_k} a_i \phi_i, \quad \text{with } \; a_i := (\Pi_{\sigma_i} u)(X_i). \tag{5.12}$$

In Theorem 3.3 from [Ape99] the following local stability and approximation property of the operator $\mathcal{L}_k$ is given.

Theorem 5.5 *The modified Scott-Zhang operator $\mathcal{L}_k$ defined in* (5.12) *satisfies the following estimates for all $T \in \mathcal{T}_k$ and all $u \in W_p^l(S_T)$:*

$$\begin{aligned} |\mathcal{L}_k u|_{W_q^m(T)} &\leq c(\mathrm{meas}_3\, T)^{1/q-1/p}\, |u|_{W_p^m(S_T)}, \\ |u - \mathcal{L}_k u|_{W_q^m(T)} &\leq c(\mathrm{meas}_3\, T)^{1/q-1/p} \sum_{|\alpha|=l-m} h_k^{|\alpha|} \varepsilon^{\alpha_3} |D^\alpha u|_{W_p^m(S_T)}, \end{aligned} \tag{5.13}$$

with $0 \leq m \leq l$, $l = 1, 2$. For (5.13) *the numbers p, $q \in [1, \infty]$ must be such that $W_p^l(T) \hookrightarrow W_q^m(T)$. The constant c is independent of k and ε.*

Here $|\cdot|_{W_p^m(T)}$ denotes the seminorm in the Sobolev space $W_p^m(T)$. Note that the term $h_k^{|\alpha|}\varepsilon^{\alpha_3}$ represents the product of the length scales of the edges of T in the three coordinate directions. The estimate (5.13) is an example of an *anisotropic estimate*. Theorem 3.3 in [Ape99] is more general than the result formulated in Theorem 5.5. The former gives a similar result for more general (for example, higher order,) polynomial finite elements in d-dimensional spaces, with $d = 2, 3$.

Corollary 5.6 *The modified Scott-Zhang operator $\mathcal{L}_k$ defined in* (5.12) *satisfies the following estimates*

$$\|u - \mathcal{L}_k u\|_0 \leq c_1 h_k^2\, |u|_2, \tag{5.14}$$

$$|u - \mathcal{L}_k u|_1 \leq c_2 h_k\, |u|_2, \tag{5.15}$$

for all $u \in H^2(\Omega_\varepsilon) \cap U_\varepsilon$ and with constants c_1, c_2 independent of k and ε. Moreover, $\mathcal{L}_k : H^2(\Omega_\varepsilon) \cap U_\varepsilon \to U_{\varepsilon,k}$ holds.

Proof: If in (5.13) we take $p = q = 2$, $l = 2$ and $m \in \{0, 1\}$, and sum over all $T \in \mathcal{T}_k$ we obtain (using $\varepsilon \leq 1$) the results in (5.14) and (5.15). Take $u \in H^2(\Omega_\varepsilon) \cap U_\varepsilon$, i.e. $u_{|\Gamma_D} = 0$. From the construction of the modified Scott-Zhang operator $\mathcal{L}_k : H^1(\Omega_\varepsilon) \to V_{\varepsilon,k}$ in (5.12) it follows that $a_i = (\Pi_{\sigma_i} u)(X_i) = 0$ if X_i is a vertex on the Dirichlet boundary Γ_D. Thus $(\mathcal{L}_k u)_{|\Gamma_D} = 0$ holds. Using $U_{\varepsilon,k} = \{v \in V_{\varepsilon,k} \mid v_{|\Gamma_D} = 0\}$ we conclude that $\mathcal{L}_k u \in U_{\varepsilon,k}$. $\diamond$

5.5 Finite element discretization error bound

In this section, using a standard approach, we derive a finite element discretization error bound that is uniform w.r.t. the parameters μ and ε.

Remark 5.7 Due to the special geometry of the considered domain Ω_ε one can use the so-called *Schwarz reflection principle* (see [Fol76] page 143) to show that the solution u of (5.2) lies in $H^2(\Omega_\varepsilon)$ taking into account the mixed boundary conditions. The idea in applying this principle is to get the original problem in a larger domain $\tilde{\Omega}_\varepsilon \supset\supset \Omega_\varepsilon$ with an extended right-hand side $\tilde{f}$ that belongs to $L_2(\tilde{\Omega}_\varepsilon)$. This can be realized by even extension over the Neumann boundaries and odd extension over the Dirichlet boundaries, respectively. Using the inner regularity of the extended problem solution finally leads to the desired result $u \in H^2(\Omega_\varepsilon)$. Similar results for cases with pure Neumann or Dirichlet boundary conditions can be found e.g. in [Gri85].

We proceed with an elementary lemma.

Lemma 5.8 *For all $u \in H^2(\Omega_\varepsilon) \cap U_\varepsilon$ the identity*

$$|u|_2 = \|\Delta u\|_0 \tag{5.16}$$

holds.

Proof: It is sufficient to prove (5.16) in the dense subset $C^\infty(\Omega_\varepsilon) \cap U_\varepsilon$. For u from this space we have

$$|u|_2^2 = \int_{\Omega_\varepsilon} u_{xx}^2 + u_{yy}^2 + u_{zz}^2 + 2u_{xy}^2 + 2u_{xz}^2 + 2u_{yz}^2 \, dx\, dy\, dz. \tag{5.17}$$

The unit outward pointing normal on the boundary $\Gamma_\varepsilon := \partial\Omega_\varepsilon$ is denoted by $\mathbf{n} = (n_x, n_y, n_z)^T$. On the sides of Ω_ε with normal $\mathbf{n} = (0, 0, \pm 1)^T$, i.e. the Dirichlet boundary, we have the identities $u_x = u_{xx} = 0$ and $u_y = u_{yy} = 0$. On the sides with normal $\mathbf{n} = (0, \pm 1, 0)^T$ we have $u_y = 0$ and thus $u_{xy} = 0$. Similar relations hold on the remaining Neumann boundaries. Integration by parts yields

$$\begin{aligned}
\int_{\Omega_\varepsilon} u_{xy}^2 \, dx\, dy\, dz &= \int_{\Gamma_\varepsilon} \underbrace{u_x u_{xy} n_y}_{=0} \, d\Gamma_\varepsilon - \int_{\Omega_\varepsilon} u_x u_{xyy} \, dx\, dy\, dz \\
&= -\int_{\Gamma_\varepsilon} \underbrace{u_x u_{yy} n_x}_{=0} \, d\Gamma_\varepsilon + \int_{\Omega_\varepsilon} u_{xx} u_{yy} \, dx\, dy\, dz.
\end{aligned}$$

Similar expressions can be derived for the mixed derivatives u_{xz} and u_{yz} in (5.17). This yields

$$\begin{aligned}
|u|_2^2 &= \int_{\Omega_\varepsilon} u_{xx}^2 + u_{yy}^2 + u_{zz}^2 + 2u_{xx}u_{yy} + 2u_{xx}u_{zz} + 2u_{yy}u_{zz} \, dx\, dy\, dz \\
&= \int_{\Omega_\varepsilon} (u_{xx} + u_{yy} + u_{zz})^2 \, dx\, dy\, dz = \int_{\Omega_\varepsilon} (\Delta u)^2 \, dx\, dy\, dz = \|\Delta u\|_0^2
\end{aligned}$$

and thus the lemma is proved. ◇

Lemma 5.9 *Let u be the solution of the continuous problem* (5.2)*. Then the inequalities*

$$\|u\|_0 \leq \frac{1}{\mu}\|f\|_0, \tag{5.18}$$

$$|u|_2 \leq 2\|f\|_0, \tag{5.19}$$

hold.

Proof: Setting $v = u$ in (5.2) gives

$$\mu\|u\|_0^2 \le a(u,u) = (f,u)_0 \le \|f\|_0\|u\|_0,$$

and thus (5.18) holds. Since the solution u of (5.2) lies in $H^2(\Omega_\varepsilon)$ (see Remark 5.7) we can write $-\Delta u = f - \mu u$. Using Lemma 5.8 we get

$$|u|_2 = \|\Delta u\|_0 = \|f - \mu u\|_0 \le \|f\|_0 + \mu\|u\|_0 \le 2\|f\|_0,$$

which proves the result in (5.19). ◇

Theorem 5.10 *Let u be the solution of the continuous problem* (5.2) *and u_k the solution of the discrete problem* (5.4)*. Then*

$$\|u - u_k\|_0 \le c \min\left\{\frac{1}{\mu}, h_k^2\right\}\|f\|_0$$

holds, with a constant c independent of f, ε, μ and k.

Proof: We use the notation $e_k = u - u_k$. Since $a(e_k, v_k) = 0$ for all $v_k \in U_{\varepsilon,k}$ we get

$$\mu\|e_k\|_0^2 \le a(e_k,e_k) = a(u,e_k) = (f,e_k)_0 \le \|f\|_0\|e_k\|_0$$

and thus

$$\|e_k\|_0 \le \frac{1}{\mu}\|f\|_0. \tag{5.20}$$

Now we apply Nitsche's duality argument and use the interpolation results from Corollary 5.6. Let $w \in U_\varepsilon$ be such that $a(w,v) = (e_k,v)_0$ for all $v \in U_\varepsilon$. From Lemma 5.9 we get

$$|w|_2 \le 2\|e_k\|_0. \tag{5.21}$$

We also have (due to the Céa-lemma)

$$\begin{aligned} |e_k|_1^2 \le a(e_k,e_k) &\le |u - \mathcal{L}_k u|_1^2 + \mu\|u - \mathcal{L}_k u\|_0^2 \\ &\le (c_2^2 h_k^2 + \mu c_1^2 h_k^4)|u|_2^2 \\ &\le c(1+\mu h_k^2)h_k^2\,\|f\|_0^2. \end{aligned} \tag{5.22}$$

Thus we get (with a constant c independent of k, ε and μ)

$$\begin{aligned} \|e_k\|_0^2 = a(w,e_k) &= a(w - \mathcal{L}_k w, e_k) \\ &\le |w - \mathcal{L}_k w|_1|e_k|_1 + \mu\|w - \mathcal{L}_k w\|_0\|e_k\|_0 \\ &\le (c_2 h_k|e_k|_1 + \mu c_1 h_k^2\|e_k\|_0)|w|_2 \\ &\overset{(5.20),(5.21)}{\le} c(h_k|e_k|_1 + h_k^2\|f\|_0)\|e_k\|_0 \\ &\overset{(5.22)}{\le} c\left([1+\mu h_k^2]^{\frac{1}{2}} h_k^2\|f\|_0 + h_k^2\|f\|_0\right)\|e_k\|_0 \\ &= c\left([1+\mu h_k^2]^{\frac{1}{2}} + 1\right)h_k^2\|f\|_0\|e_k\|_0. \end{aligned}$$

Hence, for $h_k^2 \le \frac{1}{\mu}$ we obtain

$$\|e_k\|_0 \le c h_k^2\,\|f\|_0.$$

Combining this estimate with the one in (5.20) proves the theorem. ◇

5.6 Multigrid convergence analysis

In this section we investigate the convergence behaviour of a multigrid method applied to the discrete problem (5.4). We use the approach introduced by Hackbusch (see [Hac85]) based on the approximation and smoothing property (compare Section 3.3.3). It is well-known that the anisotropy in the discrete problem (5.4) causes standard pointwise relaxation methods, such as the damped Jacobi method or the (symmetric) Gauss-Seidel method, to smooth the error only in the direction corresponding to the strong couplings. This causes a (strong) deterioration in the rate of convergence of a multigrid method with such smoothers for $\varepsilon \downarrow 0$. One possibility to deal with this problem is to keep pointwise relaxation but to adapt the strategy of grid coarsening e.g. by doubling the mesh size only in the directions in which the error is smooth. An alternative approach, which is used in this thesis, is to modify the smoothing procedure from pointwise relaxation to *linewise relaxation* meaning that the unknowns belonging to a line are updated simultaneously. Hackbusch (cf. [Hac85]) introduced the notion of a "robust smoother" for anisotropic problems. Such a smoother should be a fast iterative (or even direct) solver for the discrete problem in the limit case $\varepsilon \downarrow 0$. In the model reaction-diffusion problem that we consider we do not only have the anisotropy parameter ε but also the parameter μ in front of the reaction term.

In this section, using a fairly standard approach, we derive an approximation property, Theorem 5.12, and a smoothing property for the symmetric z-line Gauss-Seidel method, Theorem 5.15, in which the dependence of the bounds on ε, μ and k is explicit. Combination of these results immediately yields a uniform bound (< 1 for sufficiently many smoothing iterations) for the contraction number of the two-grid method and of the W-cycle iteration.

We introduce the isomorphism

$$P_k : X_k := \mathbb{R}^{n_k} \to U_{\varepsilon,k}, \quad P_k \mathbf{x} = \sum_{i=1}^{n_k} x_i \phi_i.$$

In order to establish the norm equivalence

$$c^{-1} \|\mathbf{x}\|_\varepsilon \leq \|P_k \mathbf{x}\|_0 \leq c \|\mathbf{x}\|_\varepsilon \quad \text{for all } \mathbf{x} \in X_k, \tag{5.23}$$

with a constant c independent of ε and k we use the scaled Euclidean scalar product

$$\langle \mathbf{x}, \mathbf{y} \rangle_\varepsilon := \varepsilon h_k^3 \sum_{i=1}^{n_k} x_i y_i, \qquad \|\cdot\|_\varepsilon^2 := \langle \cdot, \cdot \rangle_\varepsilon.$$

Standard arguments yield that for this scaled norm indeed the uniform norm equivalence (5.23) holds. Let the corresponding matrix norm (which is independent of ε) be denoted by $\|\cdot\|$. Note that the adjoint $P_k^* : U_{\varepsilon,k} \to X_k$ satisfies $(P_k \mathbf{x}, v)_0 = \langle \mathbf{x}, P_k^* v \rangle_\varepsilon$ for all $\mathbf{x} \in X_k$, $v \in U_{\varepsilon,k}$. The stiffness matrix $\mathbf{A}_k$ on level k is defined by

$$\langle \mathbf{A}_k \mathbf{x}, \mathbf{y} \rangle_\varepsilon = a(P_k \mathbf{x}, P_k \mathbf{y}) \quad \text{for all } \mathbf{x}, \mathbf{y} \in X_k. \tag{5.24}$$

In an interior grid point the discrete problem has the stencil given in Fig. 5.6. Note that for a point on the Neumann boundary only certain off-diagonal stencil entries in the x, y-plane change which does not affect the further analysis. For the prolongation and restriction in the multigrid method the canonical choice

$$\mathbf{p}_k \quad : \quad X_{k-1} \to X_k, \quad \mathbf{p}_k = P_k^{-1} P_{k-1} \tag{5.25}$$

$$\mathbf{r}_k \quad : \quad X_k \to X_{k-1}, \quad \mathbf{r}_k = P_{k-1}^* (P_k^*)^{-1} = \left(\frac{h_k}{h_{k-1}} \right)^3 \mathbf{p}_k^T, \tag{5.26}$$

is used. We use stationary linear iterative methods as smoothers and thus these are of the form

$$\mathbf{x}^{new} = \mathbf{x}^{old} - \mathbf{W}_k^{-1} (\mathbf{A}_k \mathbf{x}^{old} - \mathbf{b}).$$

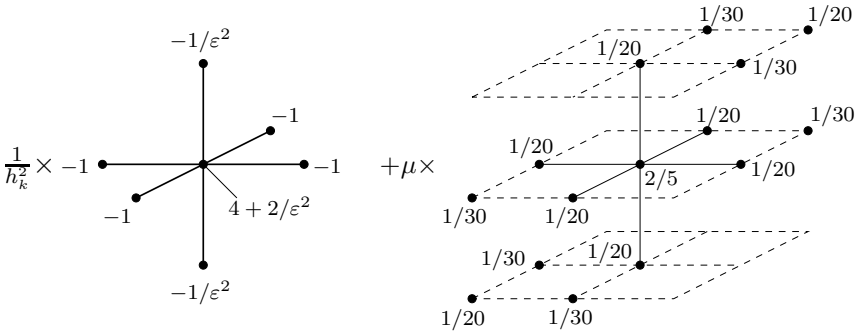


Figure 5.6: 3D difference stencil.

The corresponding iteration matrix is denoted by

$$\mathbf{S}_k = \mathbf{I} - \mathbf{W}_k^{-1}\mathbf{A}_k. \tag{5.27}$$

We consider the damped z-line Jacobi method and the symmetric z-line Gauss-Seidel method as smoothers. For the matrix representation of the discrete operator (see the stencil notation in Fig. 5.6) we assume a z-line ordering of the grid points, i.e. within each line of unknowns in z-direction (the z-lines) the vertices are numbered from bottom to top while the z-lines themselves are ordered lexicographically in the x, y-plane. Let $N_k := 2^k - 1$. The stiffness matrix can be decomposed as

$$\mathbf{A}_k = \mathbf{D}_k - \mathbf{L}_k - \mathbf{L}_k^T,$$

with the block-diagonal matrix $\mathbf{D}_k = \text{blockdiag}(\hat{\mathbf{D}}_k) \in \mathbb{R}^{n_k \times n_k}$ and diagonal blocks

$$\hat{\mathbf{D}}_k = \frac{1}{\varepsilon^2 h_k^2}\,\text{tridiag}(-1, 2, -1) + \frac{4}{h_k^2}\mathbf{I}_k + \frac{\mu}{20}\,\text{tridiag}(1, 8, 1) \in \mathbb{R}^{N_k \times N_k}. \tag{5.28}$$

Note that the upper and lower faces of Ω_ε are Dirichlet boundaries. The matrix $\mathbf{L}_k$ is strictly lower block-triangular. The choice

$$\mathbf{W}_k := \omega^{-1}\mathbf{D}_k, \quad \omega \in (0, 1],$$

defines a (damped) line Jacobi smoother, and

$$\mathbf{W}_k := (\mathbf{D}_k - \mathbf{L}_k)\mathbf{D}_k^{-1}(\mathbf{D}_k - \mathbf{L}_k^T) \tag{5.29}$$

yields the symmetric line Gauss-Seidel method. In the convergence analysis below we only consider the symmetric line Gauss-Seidel method. Similar results, however, can be obtained for the damped line Jacobi method if we use a damping factor $\omega = \omega(\varepsilon, \mu, k)$ such that $\omega^{-1}\mathbf{D}_k \geq \mathbf{A}_k$ holds.

Based on these components a standard multigrid algorithm with ν_1 pre- and ν_2 post-smoothing iterations can be formulated (see [Hac94]) with an iteration matrix that satisfies the recursion

$$\begin{aligned} &\mathbf{M}_0(\nu_2, \nu_1) = 0, \\ &\mathbf{M}_k(\nu_2, \nu_1) = \mathbf{S}_k^{\nu_2}(\mathbf{I} - \mathbf{p}_k(\mathbf{I} - \mathbf{M}_{k-1}^{\gamma})\mathbf{A}_{k-1}^{-1}\mathbf{r}_k\mathbf{A}_k)\mathbf{S}_k^{\nu_1}, \quad k = 1, 2, \ldots \end{aligned} \tag{5.30}$$

The choices $\gamma = 1$ and $\gamma = 2$ correspond to the V- and W-cycle, respectively. Results of numerical experiments with this method are presented in Section 5.7.

We now turn to the convergence analysis of this multigrid method. All constants (denoted by c or c_i) that appear in the analysis are independent of ε, μ and k. The following lemma gives a result on the scaling of the stiffness matrix.

Lemma 5.11 *Let $\mathbf{A}_k$ be the stiffness matrix from* (5.24)*. Then the inequalities*

$$c_1 \left(\frac{1}{\varepsilon^2 h_k^2} + \mu \right) \leq \|\mathbf{A}_k\| \leq c_2 \left(\frac{1}{\varepsilon^2 h_k^2} + \mu \right), \tag{5.31}$$

hold with constants $c_1 > 0$, c_2 independent of ε, μ and k.

Proof: Let $\mathbf{e}_i$ denote the ith basis vector in $\mathbb{R}^{n_k}$ and $S_i := \mathrm{supp}(\phi_i)$ the support of the nodal basis function ϕ_i. Then we have

$$\begin{aligned}
(\mathbf{A}_k)_{ii} &= \frac{\langle \mathbf{A}_k \mathbf{e}_i, \mathbf{e}_i \rangle_\varepsilon}{\langle \mathbf{e}_i, \mathbf{e}_i \rangle_\varepsilon} = \varepsilon^{-1} h_k^{-3} a(\phi_i, \phi_i) \\
&= \varepsilon^{-1} h_k^{-3} (|\phi_i|_1^2 + \mu \|\phi_i\|_0^2) \\
&= \varepsilon^{-1} h_k^{-3} \left(\int_{\Omega_\varepsilon} (\nabla \phi_i)^2 d\Omega_\varepsilon + \mu \int_{\Omega_\varepsilon} \phi_i^2 d\Omega_\varepsilon \right) \\
&\sim \varepsilon^{-1} h_k^{-3} \left(\int_{S_i} \left(\frac{1}{\varepsilon h_k} \right)^2 d\Omega_\varepsilon + \mu \int_{S_i} 1 \, d\Omega_\varepsilon \right) \\
&\sim \varepsilon^{-1} h_k^{-3} \left(\varepsilon h_k^3 \frac{1}{\varepsilon^2 h_k^2} + \mu \varepsilon h_k^3 \right) = \frac{1}{\varepsilon^2 h_k^2} + \mu.
\end{aligned}$$

Thus, we have $\|\mathbf{A}_k\| \geq (\mathbf{A}_k)_{ii} \geq c_1 \, (1/(\varepsilon^2 h_k^2) + \mu)$ with $c_1 > 0$, yielding the left inequality in (5.31). Using the inverse inequality

$$|P_k \mathbf{x}|_1^2 \leq c \varepsilon^{-2} h_k^{-2} \, \|P_k \mathbf{x}\|_0^2$$

we get

$$\begin{aligned}
\langle \mathbf{A}_k \mathbf{x}, \mathbf{x} \rangle_\varepsilon &= a(P_k \mathbf{x}, P_k \mathbf{x}) = |P_k \mathbf{x}|_1^2 + \mu \|P_k \mathbf{x}\|_0^2 \\
&\leq c \left(\frac{1}{\varepsilon^2 h_k^2} + \mu \right) \|P_k \mathbf{x}\|_0^2 \leq c_2 \left(\frac{1}{\varepsilon^2 h_k^2} + \mu \right) \|\mathbf{x}\|_\varepsilon^2.
\end{aligned}$$

Finally, we note that all the constants used in this proof are independent of ε, μ and k. $\diamond$

Theorem 5.12 (Approximation property) *Let $\mathbf{A}_k$ be the stiffness matrix from* (5.24) *and $\mathbf{p}_k$, $\mathbf{r}_k$ the prolongation and restriction from* (5.25) *and* (5.26)*, respectively. Then the approximation property*

$$\|\mathbf{A}_k^{-1} - \mathbf{p}_k \mathbf{A}_{k-1}^{-1} \mathbf{r}_k\| \leq c \frac{1 + \varepsilon^2 \mu h_k^2}{\varepsilon^2} \min \left\{ \frac{1}{\mu h_k^2}, 1 \right\} \|\mathbf{A}_k\|^{-1} \tag{5.32}$$

holds with a constant c independent of ε, μ and k.

Proof: We consider an arbitrary $\mathbf{y}_k \in X_k$. Let $w \in U_\varepsilon$, $w_k \in U_{\varepsilon,k}$ and $w_{k-1} \in U_{\varepsilon,k-1}$ be such that

$$\begin{aligned}
a(w, v) &= ((P_k^*)^{-1} \mathbf{y}_k, v)_0 \quad \text{for all } v \in U_\varepsilon, \\
a(w_k, v) &= ((P_k^*)^{-1} \mathbf{y}_k, v)_0 \quad \text{for all } v \in U_{\varepsilon,k}, \\
a(w_{k-1}, v) &= ((P_k^*)^{-1} \mathbf{y}_k, v)_0 \quad \text{for all } v \in U_{\varepsilon,k-1}.
\end{aligned} \tag{5.33}$$

Setting $v = P_k \mathbf{y}_k \in U_{\varepsilon,k}$ in (5.33) we get the identity

$$\begin{aligned}
\langle \mathbf{A}_k P_k^{-1} w_k, \mathbf{y}_k \rangle_\varepsilon &\overset{(5.24)}{=} a(w_k, P_k \mathbf{y}_k) \\
&= ((P_k^*)^{-1} \mathbf{y}_k, P_k \mathbf{y}_k)_0 = \langle \mathbf{y}_k, \mathbf{y}_k \rangle_\varepsilon.
\end{aligned}$$

Thus we obtain $w_k = P_k \mathbf{A}_k^{-1} \mathbf{y}_k$. Using the same line of argumentation it follows that $w_{k-1} = P_{k-1} \mathbf{A}_{k-1}^{-1} \mathbf{r}_k \mathbf{y}_k$. We use Theorem 5.10 with $f := (P_k^*)^{-1} \mathbf{y}_k \in L_2(\Omega_\varepsilon)$ and obtain

$$\|w - w_l\|_0 \leq c \min\left\{\frac{1}{\mu}, h_l^2\right\} \|(P_k^*)^{-1} \mathbf{y}_k\|_0 \quad \text{for } l \in \{k-1, k\}.$$

Using a triangle inequality and $h_{k-1} = 2h_k$ we get

$$\|w_k - w_{k-1}\|_0 \leq c \min\left\{\frac{1}{\mu}, h_k^2\right\} \|(P_k^*)^{-1} \mathbf{y}_k\|_0.$$

Due to the norm equivalence (5.23) we have

$$\begin{aligned} \|(\mathbf{A}_k^{-1} - \mathbf{p}_k \mathbf{A}_{k-1}^{-1} \mathbf{r}_k) \mathbf{y}_k\|_\varepsilon &\leq c \|P_k \mathbf{A}_k^{-1} \mathbf{y}_k - P_k \mathbf{p}_k \mathbf{A}_{k-1}^{-1} \mathbf{r}_k \mathbf{y}_k\|_0 \\ &= c \|w_k - w_{k-1}\|_0 \\ &\leq c \min\left\{\frac{1}{\mu}, h_k^2\right\} \|(P_k^*)^{-1} \mathbf{y}_k\|_0 \\ &\leq c \min\left\{\frac{1}{\mu}, h_k^2\right\} \|\mathbf{y}_k\|_\varepsilon, \end{aligned}$$

and thus

$$\|\mathbf{A}_k^{-1} - \mathbf{p}_k \mathbf{A}_{k-1}^{-1} \mathbf{r}_k\| \leq c \min\left\{\frac{1}{\mu}, h_k^2\right\} \tag{5.34}$$

holds. Using the scaling property from Lemma 5.11 one easily derives the bound in (5.32). ◇

We now turn to the smoothing property of the symmetric line Gauss-Seidel method for which the matrix $\mathbf{W}_k$ in (5.29) is symmetric positive definite. We start with an elementary lemma.

Lemma 5.13 *Let $\mathbf{D}_k = \text{blockdiag}(\hat{\mathbf{D}}_k)$ be the block-diagonal part of $\mathbf{A}_k$. The smallest eigenvalue of $\mathbf{D}_k$ is bounded from below by*

$$\lambda_{\min}(\mathbf{D}_k) \geq c_1 \left(\frac{1}{\varepsilon^2} + \frac{1}{h_k^2} + \mu\right). \tag{5.35}$$

For the lower block-triangular part $\mathbf{L}_k$ of $\mathbf{A}_k$ we have

$$\|\mathbf{L}_k\| \leq c_2 \left(\frac{1}{h_k^2} + \mu\right). \tag{5.36}$$

The constants $c_1 > 0$ and c_2 are independent of ε, μ and k.

Proof: Consider the diagonal blocks $\hat{\mathbf{D}}_k$ of $\mathbf{D}_k$ in (5.28). For the third term in this representation we have

$$\frac{\mu}{20} \lambda_{\min}\big(\text{tridiag}(1, 8, 1)\big) \geq \frac{3}{10}\,\mu.$$

Furthermore, for the smallest eigenvalue of the first summand in (5.28) we have

$$\begin{aligned} \varepsilon^{-2} h_k^{-2} \lambda_{\min}\big(\text{tridiag}(-1, 2, -1)\big) &= \varepsilon^{-2} h_k^{-2} 4 \sin^2\left(\frac{\pi}{2}(N_k + 1)^{-1}\right) \\ &= 4\varepsilon^{-2} h_k^{-2} \sin^2(\pi 2^{-k-1}) \geq c\,\varepsilon^{-2}, \end{aligned}$$

with $c > 0$ independent of ε and k. Since all three terms in (5.28) are symmetric positive definite we get the lower bound in (5.35). We now prove (5.36). From Fig. 5.6 we see that the matrix $\mathbf{L}_k$ does not contain any entries depending on ε and that $\|\mathbf{L}_k\|_p \leq c\,(h_k^{-2} + \mu)$, for $p = 1, \infty$, holds with c independent of k and μ. Thus we obtain

$$\|\mathbf{L}_k\|^2 \leq \|\mathbf{L}_k\|_1 \|\mathbf{L}_k\|_\infty \leq c\,(h_k^{-2} + \mu)^2,$$

which completes the proof. ◇

For the symmetric line Gauss-Seidel method we have

$$\mathbf{W}_k = \mathbf{A}_k + \mathbf{L}_k\mathbf{D}_k^{-1}\mathbf{L}_k^T \geq \mathbf{A}_k. \tag{5.37}$$

The following result can be found in [Hac94].

Lemma 5.14 *For all symmetric matrices* $\mathbf{B}$ *with* $0 \leq \mathbf{B} \leq \mathbf{I}$*, the inequality*

$$\|\mathbf{B}(\mathbf{I}-\mathbf{B})^\nu\|_2 \leq \eta_0(\nu) \quad (\nu \geq 0)$$

holds, where the function $\eta_0(\nu)$ *is defined by*

$$\eta_0(\nu) := \nu^\nu/(\nu+1)^{\nu+1} \leq \frac{1}{\mathrm{e}\nu+1}.$$

Theorem 5.15 (Smoothing property) *For the symmetric z-line Gauss-Seidel smoother* $\mathbf{S}_k$ *defined by* (5.27) *and* (5.37) *the following property holds*

$$\|\mathbf{A}_k\mathbf{S}_k^\nu\| \leq c\,\frac{\varepsilon^4}{\nu}\frac{(1+\mu h_k^2)^2}{\big(h_k^2+\varepsilon^2(1+\mu h_k^2)\big)(1+\varepsilon^2\mu h_k^2)}\,\|\mathbf{A}_k\|, \quad \nu = 1,2,\ldots \tag{5.38}$$

with a constant c *independent of* ε, k, μ *and* ν.

Proof: The symmetric block Gauss-Seidel method corresponds to the splitting $\mathbf{A}_k = \mathbf{W}_k - \mathbf{R}_k$ with $\mathbf{R}_k := \mathbf{L}_k\mathbf{D}_k^{-1}\mathbf{L}_k^T$. Note that $\mathbf{R}_k = \mathbf{R}_k^T$, $\mathbf{W}_k = \mathbf{W}_k^T > 0$ and $\mathbf{W}_k > \mathbf{R}_k \geq 0$ holds, and thus $\sigma(\mathbf{W}_k^{-1}\mathbf{R}_k) \subset [0,1)$. Moreover, $\mathbf{R}_k^{\frac{1}{2}}$ is well-defined and

$$0 \leq \mathbf{C}_k := \mathbf{R}_k^{\frac{1}{2}}\mathbf{W}_k^{-1}\mathbf{R}_k^{\frac{1}{2}} < \mathbf{I}.$$

Using the identity (for $\nu \geq 1$)

$$\mathbf{A}_k\mathbf{S}_k^\nu = (\mathbf{W}_k - \mathbf{R}_k)(\mathbf{W}_k^{-1}\mathbf{R}_k)^\nu = \mathbf{R}_k^{\frac{1}{2}}\mathbf{C}_k^{\nu-1}(\mathbf{I}-\mathbf{C}_k)\mathbf{R}_k^{\frac{1}{2}}$$

and Lemma 5.14 with $\mathbf{B} = \mathbf{I} - \mathbf{C}_k$ we obtain

$$\|\mathbf{A}_k\mathbf{S}_k^\nu\| \leq \|\mathbf{R}_k\|\eta_0(\nu-1) \leq \frac{c}{\nu}\|\mathbf{R}_k\|$$

with c being independent of all the parameters. Using the result in Lemma 5.13 we get

$$\begin{aligned}\|\mathbf{R}_k\| &\leq \|\mathbf{L}_k\|\|\mathbf{D}_k^{-1}\|\|\mathbf{L}_k^T\| = \|\mathbf{L}_k\|^2\lambda_{\min}(\mathbf{D}_k)^{-1} \\ &\leq c\left(\frac{1}{h_k^2}+\mu\right)^2\left(\frac{1}{\varepsilon^2}+\frac{1}{h_k^2}+\mu\right)^{-1} \\ &= c\frac{\varepsilon^2(1+\mu h_k^2)^2}{h_k^2(\varepsilon^2+h_k^2+\varepsilon^2\mu h_k^2)}.\end{aligned} \tag{5.39}$$

In combination with the scaling property of $\|\mathbf{A}_k\|$ in Lemma 5.11 we obtain the bound in (5.38). ◇

Remark 5.16 Note that for fixed k and μ the symmetric z-line Gauss-Seidel smoother becomes an exact solver for $\varepsilon \downarrow 0$ while for the standard point relaxation method this is not the case. Due to the strong couplings in $\mathbf{L}_k$ for the symmetric point Gauss-Seidel smoother we get

$$\begin{aligned}\|\mathbf{R}_k\| \geq (\mathbf{R}_k)_{ii} &\geq c\left(\frac{1}{\varepsilon^2h_k^2}+\mu\right)^2\left(\frac{1}{\varepsilon^2h_k^2}+\frac{1}{h_k^2}+\mu\right)^{-1} \\ &= c\frac{(1+\mu\varepsilon^2h_k^2)^2}{\varepsilon^2h_k^2(1+\varepsilon^2+\mu\varepsilon^2h_k^2)}.\end{aligned}$$

As a direct consequence of the approximation and smoothing property we obtain the following main result.

Theorem 5.17 *For the two-grid iteration matrix with $\nu_1 = \nu$ pre- and $\nu_2 = 0$ post-smoothing iterations with the symmetric z-line Gauss-Seidel method we have*

$$\|(\mathbf{I} - \mathbf{p}_k \mathbf{A}_{k-1}^{-1} \mathbf{r}_k \mathbf{A}_k)\mathbf{S}_k^\nu\| \leq \frac{C_T}{\nu} \frac{\varepsilon^2(1+\mu h_k^2)}{h_k^2 + \varepsilon^2(1+\mu h_k^2)} \leq \frac{C_T}{\nu}, \quad \nu = 1, 2, \ldots \tag{5.40}$$

with C_T independent of ε, μ, k, and ν.

Proof: From Theorem 5.12 and Theorem 5.15 we obtain

$$\begin{aligned} & \|(\mathbf{I} - \mathbf{p}_k \mathbf{A}_{k-1}^{-1} \mathbf{r}_k \mathbf{A}_k)\mathbf{S}_k^\nu\| \\ \leq\ & \|\mathbf{A}_k^{-1} - \mathbf{p}_k \mathbf{A}_{k-1}^{-1} \mathbf{r}_k\| \|\mathbf{A}_k \mathbf{S}_k^\nu\| \\ \leq\ & c \frac{1+\varepsilon^2 \mu h_k^2}{\varepsilon^2} \min\left\{\frac{1}{\mu h_k^2}, 1\right\} \frac{\varepsilon^4}{\nu} \frac{(1+\mu h_k^2)^2}{\big(h_k^2 + \varepsilon^2(1+\mu h_k^2)\big)(1+\varepsilon^2 \mu h_k^2)} \\ =\ & \frac{c}{\nu} \frac{\varepsilon^2(1+\mu h_k^2)}{h_k^2 + \varepsilon^2(1+\mu h_k^2)} (1+\mu h_k^2) \min\left\{\frac{1}{\mu h_k^2}, 1\right\}. \end{aligned}$$

Finally, note that $(1+x)\min\{\frac{1}{x}, 1\} \leq 2$ for all $x > 0$. ◇

For the multigrid W-cycle we can apply Theorem 3.29 (compare Theorem 10.6.25 from [Hac94]) and thus obtain the following result.

Theorem 5.18 *For every fixed $\psi \in (0,1)$ there exists $\nu_0 > 0$ independent of ε, μ and k such that for the iteration matrix $\mathbf{M}_k$ of the multigrid W-cycle ($\gamma = 2$ in (5.30)) with symmetric z-line Gauss-Seidel smoothing we have*

$$\|\mathbf{M}_k(0,\nu)\| \leq \psi \quad \textit{for all } \nu \geq \nu_0.$$

From the first bound in (5.40) we see that for fixed k and μ the norm of the two-grid iteration matrix tends to zero for $\varepsilon \downarrow 0$. The same holds for the iteration matrix of the multigrid W-cycle. Thus we expect very fast convergence of the multigrid method for $\varepsilon \ll 1$. This is confirmed by numerical experiments in the next section.

We now derive a convergence result for the multigrid V-cycle, based on the analysis given in [Ste94]. We use the energy norm $\|\mathbf{B}\|_A := \|\mathbf{A}_k^{\frac{1}{2}} \mathbf{B} \mathbf{A}_k^{-\frac{1}{2}}\|$ for $\mathbf{B} \in \mathbb{R}^{n_k \times n_k}$. Note that this norm depends on the parameters k, μ and ε.

Theorem 5.19 *Let $\mathbf{M}_k = \mathbf{M}_k(\nu,\nu)$ be the iteration matrix of the multigrid V-cycle ($\gamma = 1$ in (5.30)) with symmetric z-line Gauss-Seidel smoothing. There exists a constant c independent of ε, μ and k such that*

$$\|\mathbf{M}_k\|_A \leq \frac{c}{c+\nu} \quad \textit{for all } \ \nu \geq 1.$$

Proof: From Theorem 2.1 in [Ste94], with $\alpha = 1$, $\gamma = 1$, it follows that

$$\begin{aligned} \|\mathbf{M}_k\|_A &\leq \frac{\delta}{1+\delta} \quad \text{with} \\ \delta &:= \rho(\mathbf{A}_k^{-1} - \mathbf{p}_k \mathbf{A}_{k-1}^{-1} \mathbf{r}_k)\, \rho\big(\mathbf{A}_k[(\mathbf{I} - \mathbf{S}_k^{2\nu})^{-1} - \mathbf{I}]\big). \end{aligned} \tag{5.41}$$

Theorem 2.3 in [Ste94] yields

$$\rho\big(\mathbf{A}_k[(\mathbf{I} - \mathbf{S}_k^{2\nu})^{-1} - \mathbf{I}]\big) \leq \frac{\rho(\mathbf{R}_k)}{2\nu}.$$

Using the bounds in (5.34) and (5.39) we obtain

$$\begin{aligned}\delta &\leq \frac{1}{2\nu}\|\mathbf{A}_k^{-1} - \mathbf{p}_k\mathbf{A}_{k-1}^{-1}\mathbf{r}_k\|\,\|\mathbf{R}_k\| \\ &\leq \frac{c}{2\nu}\min\left\{\frac{1}{\mu}, h_k^2\right\}\frac{\varepsilon^2(1+\mu h_k^2)^2}{h_k^2(\varepsilon^2+h_k^2+\varepsilon^2\mu h_k^2)} \\ &\leq \frac{c}{2\nu}\min\left\{\frac{1}{\mu h_k^2}, 1\right\}(1+\mu h_k^2) \leq \frac{c}{2\nu}\max_{x>0}\left[\min\left\{\frac{1}{x}, 1\right\}(1+x)\right] = \frac{c}{\nu}.\end{aligned}$$

Substitution of this result in (5.41) yields

$$\|\mathbf{M}_k\|_A \leq \frac{c/\nu}{1+c/\nu} = \frac{c}{c+\nu},$$

and thus the theorem is proved. ◇

A very similar result holds for the V-cycle method with ν_1 pre- and ν_2 post-smoothing iterations (with the symmetric z-line Gauss-Seidel method), if $\nu_1+\nu_2 > 0$ but not necessarily $\nu_1 = \nu_2$ (cf. [Ste94]).

5.7 Numerical experiments

In this section we present results of some numerical experiments using the multigrid method with a symmetric z-line Gauss-Seidel smoother. The results confirm the robustness of the W- and the V-cycle multigrid algorithms w.r.t. variation in the discretization parameter h_k and the problem parameters ε and μ.

We consider a linear system with stiffness matrix as in (5.24) and a random right-hand side vector. The zero vector is used as starting vector. For the stopping criterion we take a reduction of the relative residual by a factor 10^6. The block problems arising within the smoother are solved with a tridiagonal LU decomposition.

First we study the smoother without coarse grid correction (Table 5.1 and Table 5.2). In these and all other tables the numbers between brackets give the average residual reduction factor per iteration.

	h_k			
ε	1/8	1/16	1/32	1/64
1	**82**(0.85)	**279**(0.95)	**988**(0.99)	**3514**(0.996)
1e-1	**3**(0.11e-1)	**6**(0.1)	**14**(0.43)	**42**(0.77)
1e-3	**1**(0.5e-10)	**1**(0.27e-9)	**1**(0.15e-8)	**1**(0.85e-8)

Table 5.1: Number of GS iterations and average reduction for $\mu = 1$.

	h_k			
μ	1/8	1/16	1/32	1/64
1e-3	**3**(0.11e-1)	**6**(0.1)	**14**(0.43)	**42**(0.77)
1e+3	**3**(0.52e-2)	**3**(0.11e-1)	**8**(0.2)	**23**(0.61)
1e+6	**8**(0.18)	**8**(0.16)	**7**(0.14)	**7**(0.12)

Table 5.2: Number of GS iterations and average reduction for $\varepsilon = 0.1$.

In Table 5.1 and Table 5.2 we observe convergence phenomena as expected for the symmetric z-line Gauss-Seidel method (GS). For fixed parameters ε and μ (not too large) the rate of convergence decreases with increasing refinement level. For fixed values of μ and k the rate of convergence increases if ε decreases.

	h_k			
ε	1/8	1/16	1/32	1/64
1	**7**(0.13)	**7**(0.11)	**6**(0.11)	**6**(0.98e-1)
1e-1	**2**(0.54e-4)	**3**(0.21e-2)	**4**(0.17e-1)	**4**(0.27e-1)
1e-3	**1**(0.36e-15)	**1**(0.52e-15)	**1**(0.86e-15)	**1**(0.11e-13)

Table 5.3: Number of W-cycle iterations and average reduction for $\mu = 1$.

	h_k			
μ	1/8	1/16	1/32	1/64
1e-3	**2**(0.55e-4)	**3**(0.21e-2)	**4**(0.17e-1)	**4**(0.27e-1)
1e+3	**2**(0.22e-4)	**2**(0.72e-4)	**3**(0.45e-2)	**4**(0.21e-1)
1e+6	**4**(0.21e-1)	**4**(0.19e-1)	**3**(0.12e-1)	**3**(0.86e-2)

Table 5.4: Number of W-cycle iterations and average reduction for $\varepsilon = 0.1$.

We now turn to the W-cycle multigrid algorithm with $\nu_1 = 2$ pre- and $\nu_2 = 0$ post-smoothing iterations. Table 5.3 shows very fast convergence for $\varepsilon \ll 1$ which is consistent with the first bound in (5.40). Furthermore, we clearly observe a uniform upper bound < 1 for the reduction number w.r.t. variation in all three parameters. Finally, in Table 5.5 and Table 5.6 we show results for the V-cycle multigrid algorithm with $\nu_1 = 2$ pre- and $\nu_2 = 0$ post-smoothing iterations. These results show no significant differences compared to those for the W-cycle algorithm.

	h_k			
ε	1/8	1/16	1/32	1/64
1	**8**(0.15)	**8**(0.15)	**8**(0.15)	**8**(0.15)
1e-1	**2**(0.54e-4)	**3**(0.21e-2)	**4**(0.17e-1)	**4**(0.3e-1)
1e-3	**1**(0.36e-15)	**1**(0.52e-15)	**1**(0.86e-15)	**1**(0.11e-13)

Table 5.5: Number of V-cycle iterations and average reduction for $\mu = 1$.

	h_k			
μ	1/8	1/16	1/32	1/64
1e-3	**2**(0.55e-4)	**3**(0.21e-2)	**4**(0.17e-1)	**4**(0.3e-1)
1e+3	**2**(0.22e-4)	**2**(0.72e-4)	**3**(0.45e-2)	**4**(0.2e-1)
1e+6	**4**(0.22e-1)	**4**(0.19e-1)	**3**(0.12e-1)	**3**(0.86e-2)

Table 5.6: Number of V-cycle iterations and average reduction for $\varepsilon = 0.1$.

Finally, we consider the multigrid V-cycle algorithm with z-line Jacobi smoothing and $\nu_1 = 4$ pre- and $\nu_2 = 0$ post-smoothing iterations. In Table 5.7 we address the case with damping ($\omega = \frac{2}{3}$) in all the smoothing iterations. The more theoretical approach of using damping in all the smoothing steps except the last one is shown in Table 5.8. Further experiments which are not given in detail show that the choice of a suitable damping factor becomes more significant (w.r.t. the multigrid convergence) with decreasing number of smoothing steps.

	h_k			
ε	1/8	1/16	1/32	1/64
1	**16**(0.40)	**16**(0.41)	**16**(0.41)	**16**(0.40)
1e-1	**4**(0.18e-1)	**4**(0.29e-1)	**5**(0.62e-1)	**5**(0.79e-1)
1e-3	**4**(0.13e-1)	**4**(0.13e-1)	**4**(0.13e-1)	**4**(0.13e-1)

Table 5.7: V-cycle iterations for $\mu = 1$ and damping in all the smoothing steps.

	h_k			
ε	1/8	1/16	1/32	1/64
1	**14**(0.35)	**14**(0.36)	**14**(0.36)	**14**(0.36)
1e-1	**3**(0.40e-2)	**4**(0.18e-1)	**5**(0.51e-1)	**5**(0.73e-1)
1e-3	**1**(0.30e-6)	**1**(0.47e-6)	**1**(0.70e-6)	**2**(0.13e-5)

Table 5.8: V-cycle iterations for $\mu = 1$ and damping in all except the last smoothing step.

Chapter 6

Numerical solution of an inverse heat conduction problem

In this chapter the IHCP described in Section 2.2.1 is solved numerically. The CGNE algorithm in combination with iterative regularization and a suitable discretization based on *semiregular* finite elements is applied. In Section 6.1 the software components and some implementation details are discussed. For method and code validation we use simulated measurement data obtained from the solution of a direct heat conduction problem in which the heat flux on the estimation boundary is known. Both, simulation results with error-free and noisy measurement data are illustrated in Section 6.2. The falling film application with *realistic* high resolution temperature measurements is considered in Section 6.3. For the solution of the discretized anisotropic direct heat conduction problems two different solvers are applied. We discuss the efficiency of a preconditioned conjugate gradient and a robust multigrid method. Finally, we study the method of *nested* optimization. The idea of this approach is to start the entire optimization process on a relatively coarse level (space grid) in order to find a good start approximation of the estimation quantity on a next finer level.

6.1 Software realization

In the iterative optimization process illustrated in Fig. 2.6 anisotropic direct heat conduction problems have to be solved. All these problems, i.e. the direct, the sensitivity and the adjoint problem are of parabolic type. The solutions are computed by means of the software package DROPS[1] [GPRR02], which was extended and improved for that purpose. DROPS is based on multi-level nested grids and conforming finite element discretization methods. The CGNE algorithm (i.e. the optimization loop) is realized in MATLAB.

In Fig. 6.1 the software components and the data transfer via the mex-interface (**mex** stands for **m**atlab **ex**tension), which embeds C++ programs into the MATLAB software framework at run time is illustrated. Since the IHCP is affine in the searched for boundary heat flux, the direct problem has to be solved only once at the beginning of the iteration (see Section 2.3.2). In each of the further CGNE steps, the adjoint and the sensitivity problems have to be solved, i.e. DROPS is called twice per iteration. The corresponding geometry and discretization parameters as well as the initial temperature distributions and the time-dependent boundary data on the estimation or measurement boundaries are transferred. The remaining boundaries are always assumed to be adiabatic. In DROPS the instationary three-dimensional problems are treated on the basis of conforming finite elements on tetrahedral nested grids. For the resulting large sparse systems of equations various solvers are available in this software package. Note, that in the optimization procedure only the restrictions of the calculated time-dependent solutions to the estimation or measurement boundary of

[1] http://www.igpm.rwth-aachen.de/DROPS/

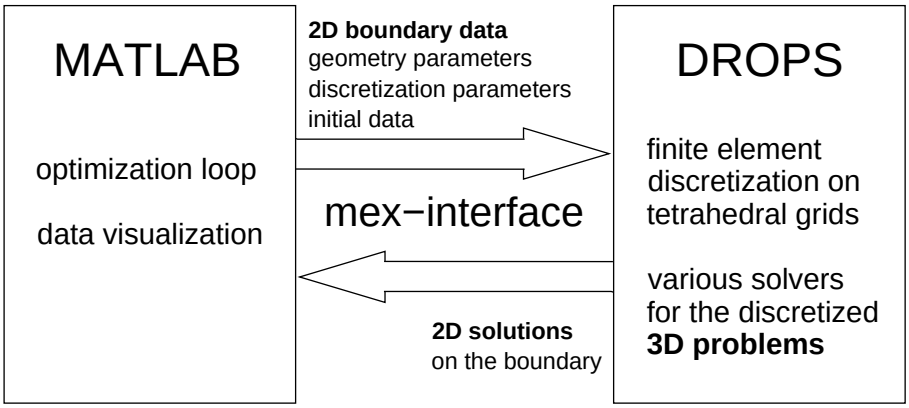


Figure 6.1: Data transfer between MATLAB and DROPS

the heat conductor are needed. Thus, the data transferred between MATLAB and DROPS are time-varying but only *two*-dimensional in space. The obtained results presented in this chapter are visualized with the graphic tools of MATLAB.

6.2 Simulation examples for method and code validation

In this section we consider some case studies for the numerical solution of the inverse heat conduction problem. Since we focus on method and code validation the material and geometry properties do not reflect the data in the realistic falling film experiment. We use simulated measurement data obtained from the solution of a direct heat conduction problem.

For the time discretization a standard one-step θ-method is applied. We use the Crank-Nicholson scheme ($\theta = 0.5$) while in [GSM$^+$05] simulation examples for the implicit Euler scheme ($\theta = 1$) can be found. For the space discretization linear finite elements on tetrahedral grids are employed. The resulting systems of linear equations are solved by a standard SSOR preconditioned (symmetric Gauss-Seidel) CG method. For the stopping criterion of the discrete problem solver we take a reduction factor of the initial residual by a factor 10^6.

6.2.1 Continuous and time-varying heat flux

In this first illustrative case study we consider the space domain $\Omega := 10 \times 40 \times 100\ mm^3$. Note that in this case there is no anisotropy effect. The material properties are lumped in the parameter $a = 10^{-4}\ \frac{m^2}{s}$. For the time discretization (Crank-Nicholson scheme) we use the time step size $\tau = 0.01\ s$ and apply 200 time steps. The initial and known boundary conditions consist of a constant temperature distribution $T_0(\mathbf{x}) = 20\ ^\circ C$, $\mathbf{x} \in \Omega$, a constant heat flux $\bar{q}(\mathbf{x}, t) = 2\ \frac{kW}{m^2}$, $(\mathbf{x}, t) \in S_m$, for heat addition and adiabatic boundaries S_r (cf. the notation of Section 2.2.1). For the initialization of the optimization procedure we choose $q^0(\mathbf{x}, t) = 0\ \frac{kW}{m^2}$, $(\mathbf{x}, t) \in S_e$ (compare Fig. 2.6). Let the given exact heat flux on the estimation boundary be denoted by $q^{ex}(\mathbf{x}, t)$, $(\mathbf{x}, t) \in S_e$. For this heat flux we consider a sinusoidal pattern in z-direction, which represents an intuitive approximation of the real quantity in the falling film experiment. The wavy structure is assumed to be time-dependent, such that the waves travel along the z-coordinate in the flow direction of the falling film

$$q^{ex}(x, y, z; t) = \sin\left(4\pi\left(\frac{z}{100} - \frac{t}{200}\right)\right)\ \frac{kW}{m^2}, \quad (\mathbf{x}, t) \in S_e. \tag{6.1}$$

In this example a uniform space discretization with 35937 unknowns is used, which corresponds to an initial grid triangulation with $32\times32\times32$ parallelepipeds and 196608 tetrahedra [GPRR02].

Estimation with error-free measurements

First, we present estimation results for the exact boundary heat flux in (6.1) with error-free measurements. As measurement data, we take the temperature T_m^{ex} on S_m obtained from the solution of the direct heat conduction problem with the chosen quantity q^{ex} as the corresponding boundary condition on S_e.

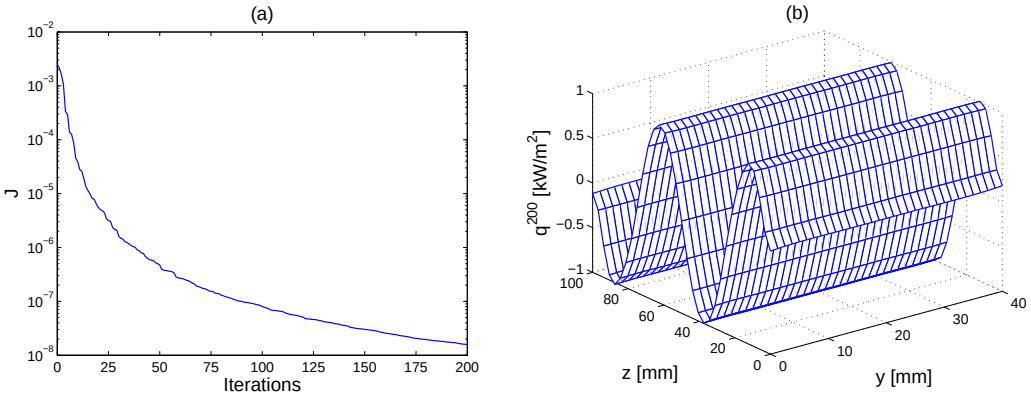


Figure 6.2: Objective functional (a) and estimated heat flux q^{200} (b)

In Fig. 6.2 (a), the objective functional is plotted over the number of optimization iterations, whereas a snapshot at one point in time of the estimated time-varying heat flux at the end of the optimization (after 200 iterations) is presented in Fig. 6.2 (b). We observe the expected convergence behaviour of the applied CGNE method in terms of a more rapid decrease of the objective functional at the beginning of the iterative process followed by stagnation at a certain level. Since the estimated heat flux is, like the exact quantity, constant in y-direction, in the following plots we consider a cross-section through the y-axis in order to look at the estimation quality in more detail. Both, exact and estimated heat fluxes are given in Fig. 6.3 over the z-direction for constant $y = 20\ mm$, i.e. in the middle of the heat conductor w.r.t. to its dimension in y-direction, and different numbers of time steps, which are denoted by n_t. Except for a region at final time the recovered heat flux is of high quality, since there are only little visible deviations compared to the exact quantity. At final time, the estimation quality decreases, due to the fact that the solution of the adjoint problem, i.e. the gradient of the objective functional, is zero and therefore yields no improvement of the start approximation. We already explained the effect that it is impossible to reconstruct the exact heat flux at $t = t_f$ from measurement data in the time interval $[t_0, t_f]$ in Section 1.2. In order to obtain a better reconstruction quality at $t = t_f$ one could use measurement data (if available) that exceed the time point t_f, e.g. in $[t_0, t_F]$ with $t_F > t_f$. However, for the concrete choice of t_F further investigations are needed. An alternative approach is based on certain (smoothness) assumptions, as e.g. the assumption that the searched for heat flux varies linearly in time. In that case the estimation result at an earlier time level could be used to obtain an extrapolated estimation at $t = t_f$. The effect of the iterative CGNE method is illustrated in Fig. 6.4. Here, the exact and estimated heat fluxes are shown at the fixed time level $n_t = 100$ and for different iterations of the optimization, which are denoted by n_{opt}. We clearly observe that the estimation quality increases with an increasing number of optimization steps, since we use simulated error-free temperature measurements. To stop the iteration, we consider the usual procedure and specify a small threshold parameter ϵ for the objective functional, i.e.

$$J(q^{n_{opt}}) \leq \epsilon. \tag{6.2}$$

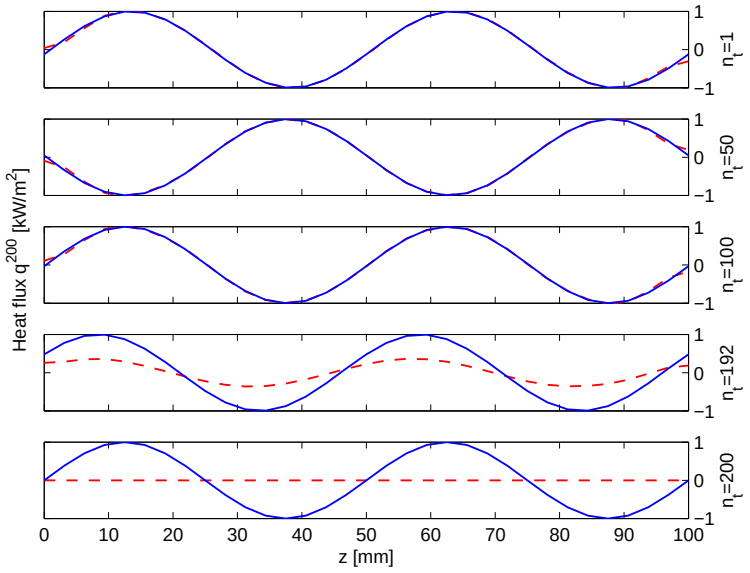


Figure 6.3: Exact heat flux q^{ex} (solid line) and estimated heat flux q^{200} (dashed line) for different time steps n_t

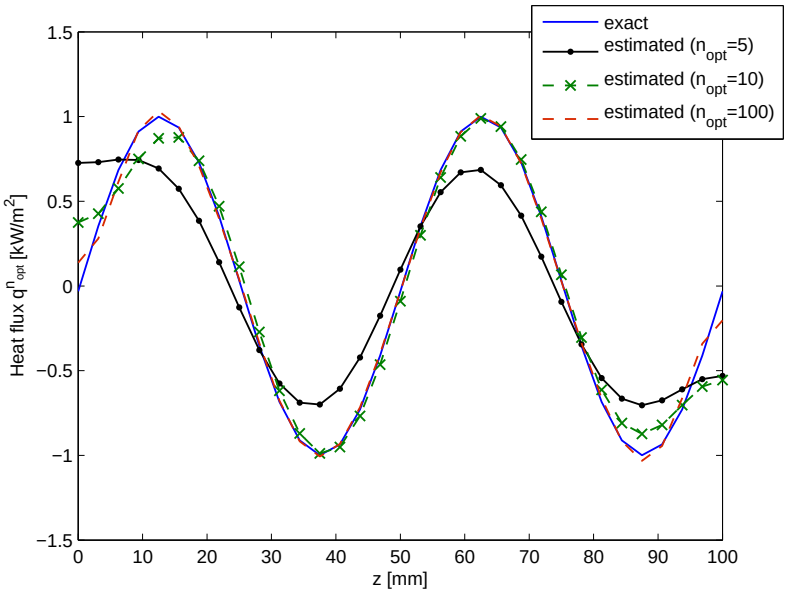


Figure 6.4: Estimated heat flux $q^{n_{opt}}$ for different optimization iterations n_{opt} and $n_t = 100$

From Fig. 6.2 (a) we see that about 90 iterations are needed until the stopping criterion in (6.2) with $\epsilon = 10^{-7}$ is fulfilled.

Estimation in the presence of measurement errors

Now we perturb the exact temperature data T_m^{ex}, obtained as described previously, with an

artificial measurement error ω. We assume that the perturbed temperature T_m is given by

$$T_m = T_m^{ex} + \delta\omega,$$

with δ being the standard deviation of the measurement error. The values of ω are generated from a zero mean normal distribution with variance one. The parameter δ is used to control the error amount added to the exact data. In the case of measurement errors, we cannot expect that the objective functional becomes arbitrarily small. In order to find an appropriate ϵ to stop the iteration, we use the parameter choice rules discussed in Section 2.3.3.

The discrepancy principle suggests that we stop the iteration when the residual approximately equals δ. If we set $|T(q)|_{S_m} - T_m| \approx \delta$ on S_m in the objective functional, we get

$$J(q) = \frac{1}{2}\int_{t_0}^{t_f}\int_{\Gamma_m} (T(q)|_{S_m} - T_m)^2 \, d\mathbf{x}\, dt \approx \frac{1}{2}(t_f - t_0)A_m\delta^2,$$

where A_m denotes the surface of the measurement boundary Γ_m. Thus, for the threshold parameter ϵ in (6.2) we choose

$$\epsilon = \tau\frac{1}{2}(t_f - t_0)A_m\delta^2,$$

with a parameter $\tau > 1$, which is introduced for theoretical considerations (compare (2.42)). Since the regularized solutions become "oversmoothed" if the value of this parameter is chosen too large, in the following simulations we used $\tau = 1.02$. For $\delta = 0.5$ some estimates at the fixed time level $n_t = 100$ are given in Fig. 6.5.

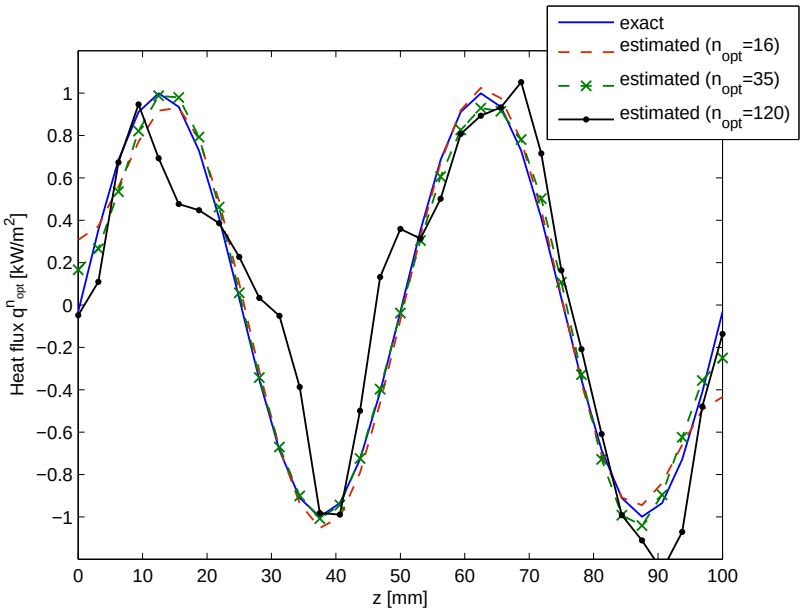


Figure 6.5: Estimated heat flux $q^{n_{opt}}$ for perturbed measurements with $\delta = 0.5$

The optimal result is obtained after $n_{opt} = 16$ iterations using the stopping rule based on the discrepancy principle as described above. In that case a good reconstruction of the exact heat flux is achieved. If too many iterations are performed, the estimated heat flux begins to oscillate and the estimation quality decreases (consider $n_{opt} = 120$). In Fig. 6.6 the estimated quantity q^{16} is compared to the exact heat flux q^{ex} for different time levels (and constant $y = 20\ mm$). The corresponding temperature distributions at one time level

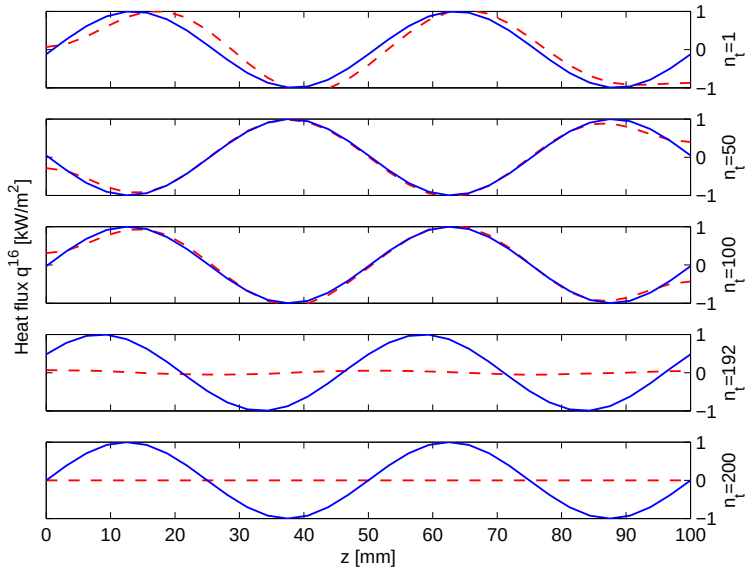


Figure 6.6: Exact heat flux q^{ex} (solid line) and estimated heat flux q^{16} (dashed line) for perturbed measurements with $\delta = 0.5$

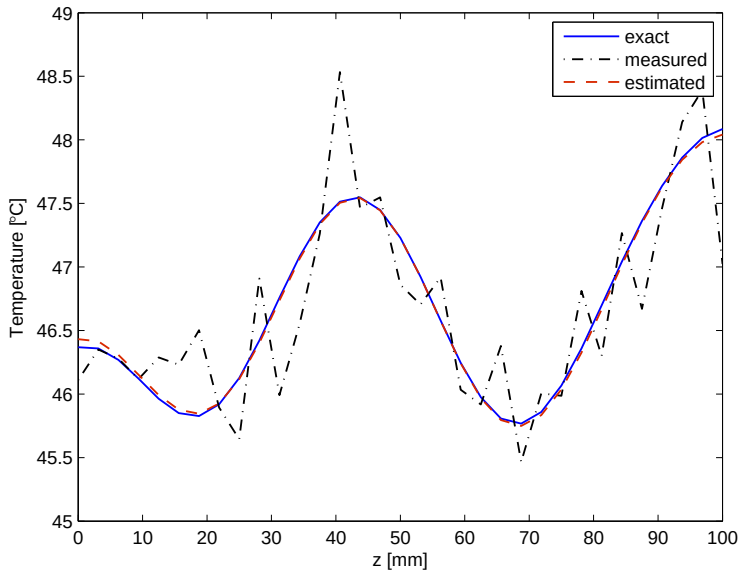


Figure 6.7: Exact, measured, and estimated temperatures

$n_t = 100$ are illustrated in Fig. 6.7. Similar results have been obtained for various values of the noise level δ (compare [GSM+05]).

The alternative heuristic stopping rule discussed in Section 2.3.3 is based on the L-curve, which is a parameterized plot of the residual against a solution norm. For two different solution norms, the L-curve is shown in Fig. 6.8. The best compromise between the approx-

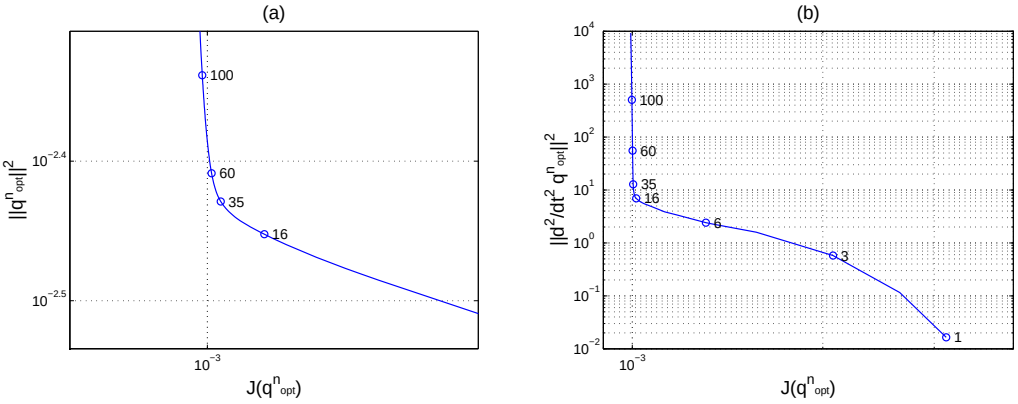


Figure 6.8: L-curve for perturbed measurements with $\delta = 0.5$; residual against the norm of (a) the solution and (b) the second order time derivative

imation and the propagated data errors is found at the point of the L-curve with maximum curvature. If we consider the plot of the residual against the standard discrete L_2-norm of the solution in Fig. 6.8 (a), the optimal value is about $n_{opt} = 35$ iterations. If we take into account certain smoothness properties of the solution e.g. by measuring the second order time derivative as shown in Fig. 6.8 (b), the optimal value is found at about $n_{opt} = 16$ iterations, i.e. the result is comparable to the one obtained by the discrepancy principle. Similar results are obtained by measuring the (space) gradient of the solution.

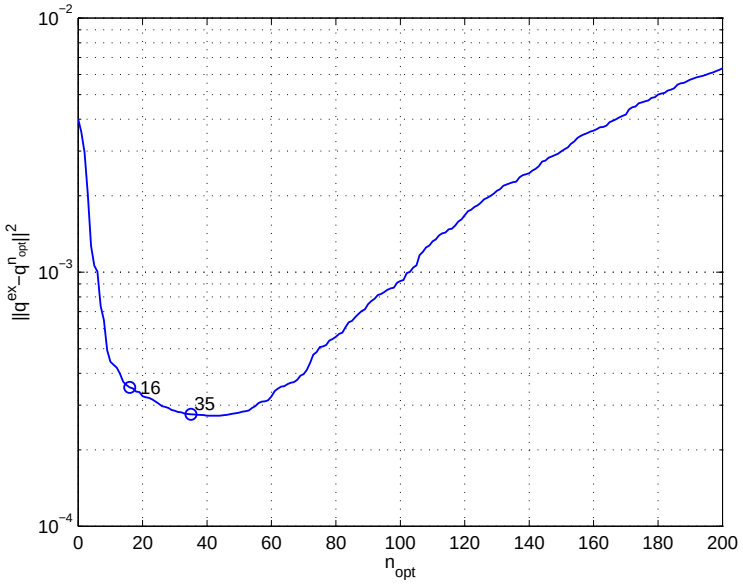


Figure 6.9: Approximation error over the number of optimization steps

The approximation error is shown in Fig. 6.9 for the whole range of optimization steps. Both solutions, at $n_{opt} = 16$ and $n_{opt} = 35$ iterations, seem to yield appropriate approximations of the exact quantity, which is confirmed by Fig. 6.5.

Multi-level optimization

We consider a multi-level approach for the optimization procedure. This concept is also

called method of nested iteration in Section 3.3.3 (see Fig 3.4). The idea of this approach is to start the entire optimization process on a relatively coarse level (space grid) in order to find a good start approximation of the estimation quantity, and in that way reduce the number of iterations needed, on the next finer level. For this reason we introduce a hierarchy of nested grids. Let l_k denote the level with $2^k(4 \times 4 \times 4)$, $k \geq 0$, parallelepipeds. On the finest level l_3 the measurement data are given. These are restricted subsequently to the coarser levels by standard injection, i.e. the values at the coarse grid points are identified with the corresponding values at the fine grid points. Now we reconsider the simulation example of this section, first for the case with error-free measurement data. We start the nested optimization process on the coarsest level l_0 with an initial approximation

$$q^0_{l_0}(\mathbf{x}, t) = 0 \ \frac{kW}{m^2}, \ (\mathbf{x}, t) \in S_e.$$

For the start approximation $q^0_{l_{k+1}}$ on level l_{k+1} we use the estimation result $q^{n_{opt}}_{l_k}$ from level l_k, $k = 0, 1, 2$, which is prolongated to the finer grid by standard linear interpolation. We take the threshold value $\epsilon = 10^{-6}$ in the stopping criterion (6.2) on each level. On the levels l_0 and l_1 the optimization is terminated after at most 10 iterations. The iteration numbers and the *initial* values of the objective functional on each level are documented in Table 6.1.

k	$J(q^0_{l_k})$	# iterations
0	$3.4 \cdot 10^{-3}$	10
1	$1.1 \cdot 10^{-3}$	10
2	$1.8 \cdot 10^{-4}$	20
3	$1.5 \cdot 10^{-5}$	21

Table 6.1: Iteration numbers for nested optimization with error-free data

From Fig. 6.2 (a) we see that 40 iterations are needed until the stopping criterion is fulfilled when the optimization is carried out completely on the finest level. In contrast to that if we use the method of nested optimization we only need about half the number of iterations on the finest level and half the number of iterations on the next coarser level. Since the computational time needed on the coarse grids l_0, l_1 and l_2 is negligible, the described multi-level optimization approach leads to a significant performance improvement. For the case with perturbed measurement data we consider the discrepancy principle as the stopping rule on each level. Again, we stop the iteration on the coarsest levels l_0 and l_1 after at most 10 iterations.

k	$J(q^0_{l_k})$	# iterations
0	$4.3 \cdot 10^{-3}$	10
1	$2.1 \cdot 10^{-3}$	10
2	$1.2 \cdot 10^{-3}$	9
3	$1.04 \cdot 10^{-3}$	6

Table 6.2: Iteration numbers for nested optimization with perturbed data

Instead of 16 iterations that are needed with the standard optimization performed exclusively on the finest level, from Table 6.2 we see that essentially 6 and 9 iterations on the levels l_3 and l_2, respectively, are needed with the method of nested optimization. The plot of the regularized solution $q^6_{l_3}$ in the middle of the foil over the time does not show any visible differences to the quantity q^{16} shown in Fig. 6.6. We remark that the multi-level approach is more involved when the L-curve method is used as the stopping criterion. In order to locate the point of maximum curvature more iterations than needed have to be performed on each level and an algorithm that determines this point automatically is desirable. However, in combination with the discrepancy principle, the multi-level approach is a suitable method in

order to reduce the computational time. Further investigations concerning a combination of the multi-level approach in space and time could yield an additional efficiency improvement but this idea is not carried out in this thesis.

6.2.2 Discontinuous and steady-state heat flux

In this section we present the simulation of an IHCP that is known to be "hard" since the searched for exact heat flux is discontinuous on Γ_e (compare [JL00] and the references therein). Examples of that form can be found at different places in the literature as a kind of benchmark problem. The time-independent exact heat flux has the representation

$$q^{ex}(x,y,z;t) = \begin{cases} 0 \ \frac{kW}{m^2} & \text{for } (x,y,z) \in \Gamma_0, \\ -3 \ \frac{kW}{m^2} & \text{else,} \end{cases}$$

with $\Gamma_0 := \{(x,y,z) \in \Gamma_e \mid y \in [10,30] \vee z \in [20,80]\}$ and is shown in Fig. 6.10 (a).

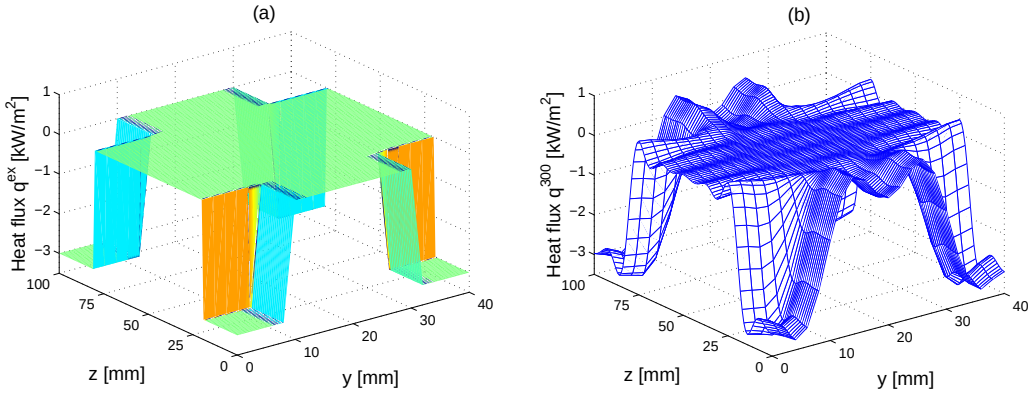


Figure 6.10: Exact (a) and estimated (b) heat flux

We reconsider the situation described in the first paragraph of Section 6.2.1, but this time we use a quasi-uniform space discretization with 36057 unknowns, which correspond to an initial triangulation with $16 \times 20 \times 100$ parallelepipeds w.r.t. the space coordinates. We take this triangulation in view of the realistic falling film application, where we choose a fine space resolution in the flow direction (i.e. the z-coordinate) and relatively coarse resolutions in the other directions (see Section 6.3).

The obtained estimation result is given in Fig. 6.10 (b). The objective functional shows the same typical behaviour as already described for the first simulation example and therefore we do not give the plot here. Both, exact and estimated solutions of the inverse problem are given in Fig. 6.11 and Fig. 6.12 for cross-sections through the y- and the z-axis, respectively (i.e. for $y = 0$ and $z = 0$). Again, the results are presented for different iteration numbers of the optimization procedure. Due to the discontinuities of the error-free boundary heat flux, oscillations appear in the piecewise linear approximations, when the number of optimization steps increases. A comparison of both plots, with 100 and 20 unknowns in the corresponding space directions, shows a better approximation quality for the case with a higher space resolution. For efficiency reasons the number of unknowns used to reach a desired accuracy should be kept as small as possible. Instead of uniform grid refinements, it may be better to use locally refined triangulations. Local grid refinement leads to good approximation properties, while at the same time the number of unknowns is severely reduced compared to the global refinement case. The multi-level refinement algorithm of DROPS [GPRR02] makes it easy to use such locally refined grids. However, the solution of the considered inverse heat conduction problem on such grids remains a challenging task for future investigations. In this context, suitable error estimators have to be developed that are based on the spatial

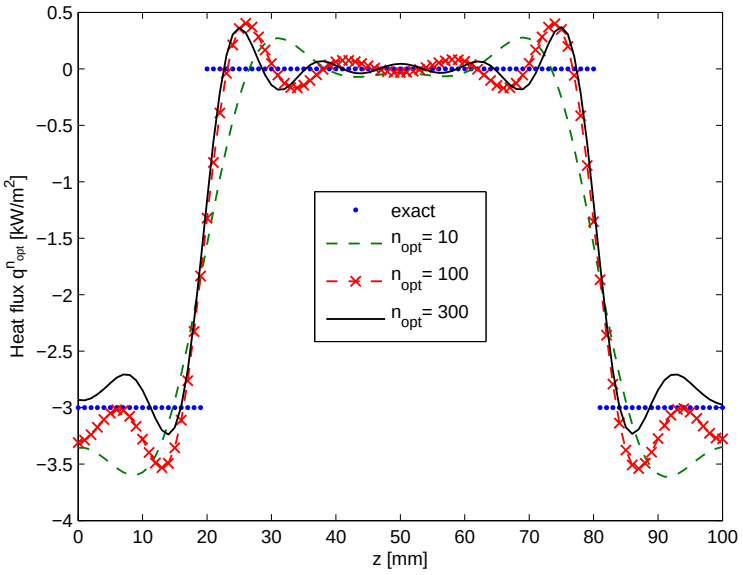


Figure 6.11: 2D plot of the heat fluxes for $y = 0$ at different optimization steps n_{opt}

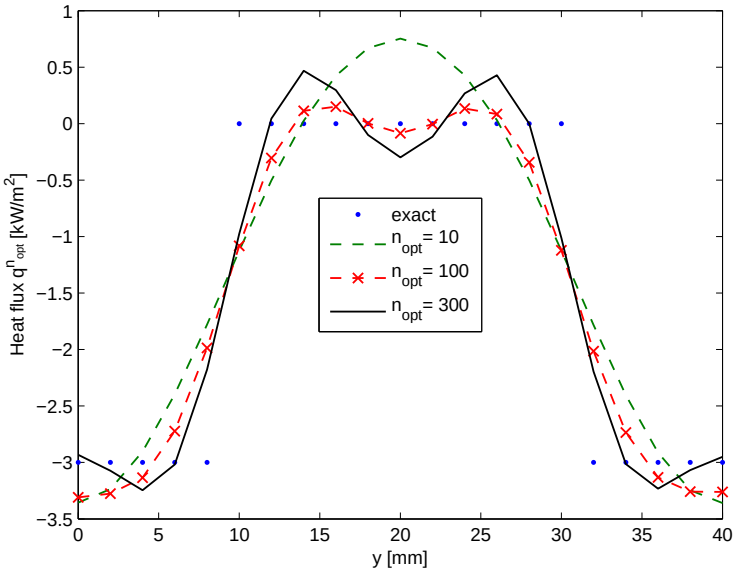


Figure 6.12: 2D plot of the heat fluxes for $z = 0$ at different optimization steps n_{opt}

behaviour of the estimation quantity. Thus, these error estimators do not have to take care of direct problem solutions (i.e. the temperature distributions) but rather of the boundary heat flux on Γ_e.

The temperature distributions which correspond to the estimated quantities are shown in Fig. 6.13 (a) and (b) together with the exact values for different optimization steps. The plots show that we obtain good approximations of the exact temperature distribution after

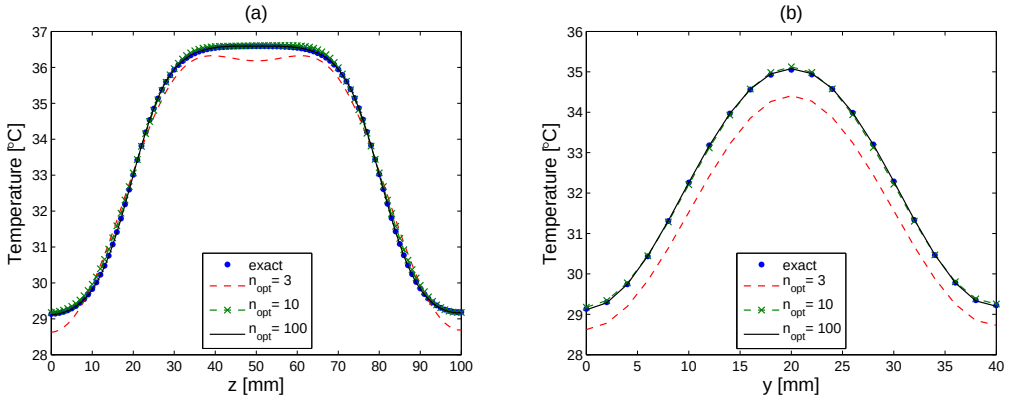


Figure 6.13: 2D plots of the temperatures for $y = 0$ (a) and $z = 0$ (b) at different optimization steps n_{opt}

only few optimization iterations in contrast to the corresponding heat fluxes. This example shows, that the value of the functional is in general not a good measure for the quality of the estimation. The fit of the temperatures may be almost perfect, even though the estimated heat flux is quite different from the exact one. This largely unavoidable effect is due to the ill-posedness of inverse heat conduction problems caused by the strong smoothing properties of the direct problem.

6.3 Estimation results with realistic measurement data

In this section, we present an estimation case study for the realistic falling film experiment (compare [GSM+05]) employing high resolution temperature measurements, which were performed at the Chair of Heat and Mass Transfer (RWTH Aachen University). These are taken with an infrared camera on the back side S_m of a constantan foil, which has a thickness of 25 μm (see Fig. 2.1). The main idea behind choosing a very thin heating foil consists of reaching a small temperature gradient across the foil thickness, i.e. the temperatures on both sides of the foil should be almost identical. In that case, the measured data provide a good estimate of the temperature distribution on the inaccessible film side of the foil. The measurement section has the dimension $19.5 \times 39\ mm^2$. Hence, we define the space domain $\Omega := 0.025 \times 19.5 \times 39\ mm^3$. Due to the very different length scales of the heat conductor we use anisotropic semiregular $\mathcal{P}_1$-elements (cf. Chapter 4).

The measurement data are taken with a sampling frequency of 500 Hz and a space resolution of $100 \times 200\ pixel$. These technical data translate to a time step size of $\tau = 2\ ms$ in the one-step θ-scheme and a space discretization of 100×200 unknowns in the y, z-plane in the case of a one-by-one allocation, i.e. if we consider the same resolution for the measurement data and the numerical simulation. For the space discretization in x-direction only 5 unknowns are used, which turned out to be an appropriate choice. To investigate the effect of the discretization in x-direction on the temperature profile, we solved the direct problem for 5 and 9 unknowns, respectively. Since no additional frequencies appeared using the finer space mesh, we conclude that already the coarser grid is appropriate for the resolution of the temperature changes in that direction. Altogether, we get a space discretization with 472824 tetrahedra. The final time of the experiment is $t_f = 0.3\ s$, which corresponds to 150 temperature frames that are taken with the infrared camera in order to observe the influence of some waves moving along with the laminar falling film. The electrical heating generates a constant heat flux $\bar{q}(\mathbf{x}, t) = 6.4\ \frac{kW}{m^2}$, $(\mathbf{x}, t) \in S_m$. For the initial approximation, we choose $q^0(\mathbf{x}, t) = \bar{q}(\mathbf{x}, t)$, $(\mathbf{x}, t) \in S_e$, because we expect q and $\bar{q}$ to have the same order of magnitude, due to the very thin heating foil. The other boundaries of the space domain

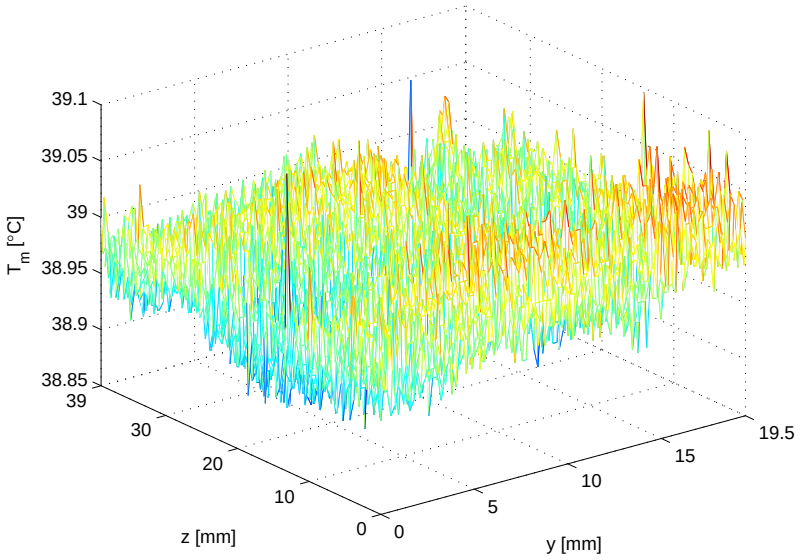


Figure 6.14: Realistic measurement data

are assumed to be adiabatic and the initial temperature distribution corresponds to the first temperature frame assumed to be constant across the foil thickness (in x-direction). The material properties of the foil are

$$\rho = 8900 \ \frac{kg}{m^3}, \quad c = 410 \ \frac{J}{kg \ K}, \quad \lambda = 23 \ \frac{W}{m \ K},$$

resulting in a thermal diffusivity of $a = 6.3 \cdot 10^{-6} \ \frac{m^2}{s}$.

In Fig. 6.14 the measured temperature distribution over the y, z-plane at one point in time is shown. The plot clearly shows that these data are perturbed by a relatively large amount of noise. Without going into more detail here, we mention the measurement preprocessing applied to the experimental data of the falling film. The convective heat transport in flow direction causes a constant rise of the film temperature of approximately $0.05 \ K/mm$, which overlays the local fluctuations caused by the waves. In order to remove this effect of the convective heat transport in the flow direction, a reference picture is subtracted from all the temperature frames taken with the infrared camera. This is the reason why we use adiabatic boundary conditions in the flow direction.

6.3.1 Single-level optimization

In this section the optimization is performed on one space grid which corresponds to the above described discretization parameters. Again, a CG method (symmetric GS preconditioned) with a reduction of the initial residual by 10^6 is applied to the resulting discrete heat conduction problems.

A typical observation in the context of inverse problems deals with the effect of noise in the given input data with respect to the number of optimization iterations. Fig. 6.15 shows the evolution of the corresponding objective functional, which decreases rapidly in the first iterations and flattens in the following steps. Although the temperature residual gets smaller, the quality of the corresponding estimated heat flux gets worse because of oscillations that appear with a rising number of optimization steps. This is an important reason why we

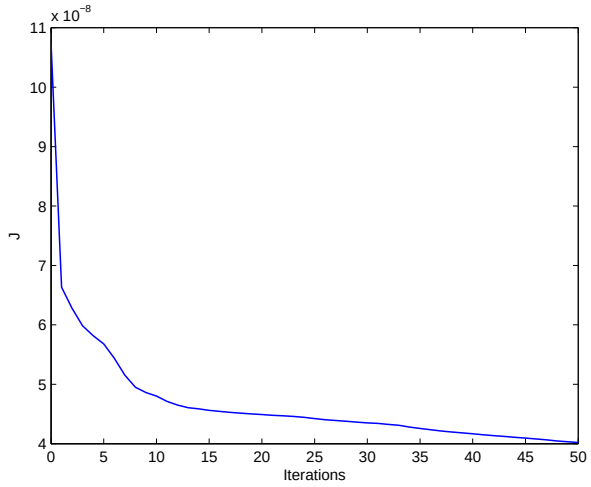


Figure 6.15: Objective functional over the number of iterations

have to investigate suitable regularization methods.

The discrepancy principle as described above with $\tau = 1.02$ has been used as the stopping rule resulting in $n_{opt} = 12$ iterations. The estimated standard deviation of the measurement error is $\delta = 0.02$. Two different L-curves for this case are shown in Fig. 6.16. In Fig. 6.16 (a) we use the standard discrete L_2-norm of the estimated heat flux for measuring its quality. The typical L-shape of the corresponding graph does not appear. If we use the norm of the second order time derivative instead, as shown in Fig. 6.16 (b), we see that the optimal estimate is obtained after 15 iterations. Similar results are obtained by measuring the (space) gradient of the iterates. However, the point of maximum curvature is not very much exposed. Again, the termination indices obtained by the discrepancy principle and the L-curve method are comparable.

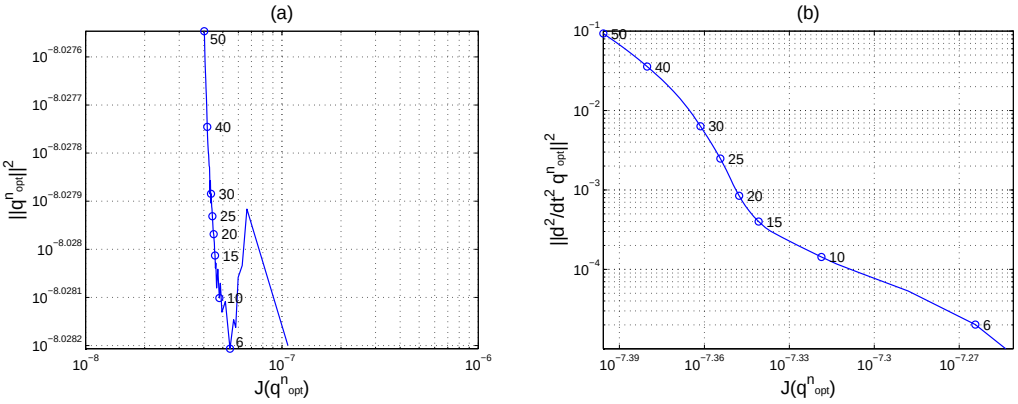


Figure 6.16: Different L-curves for the realistic measurement data

In Fig. 6.17 the estimated heat flux q^{15} at one point in time computed on the basis of realistic measured input data is presented. In Fig. 6.18 (a) this quantity is plotted for different time values in the middle of the foil over the flow direction (i.e. the z-coordinate). As a reference for the quality of the computed result the measured and the calculated temperatures are shown in Fig. 6.18 (b). Looking at the solution in time we observe that the estimated heat flux shows a wavy structure moving along in the flow direction of the falling film with the

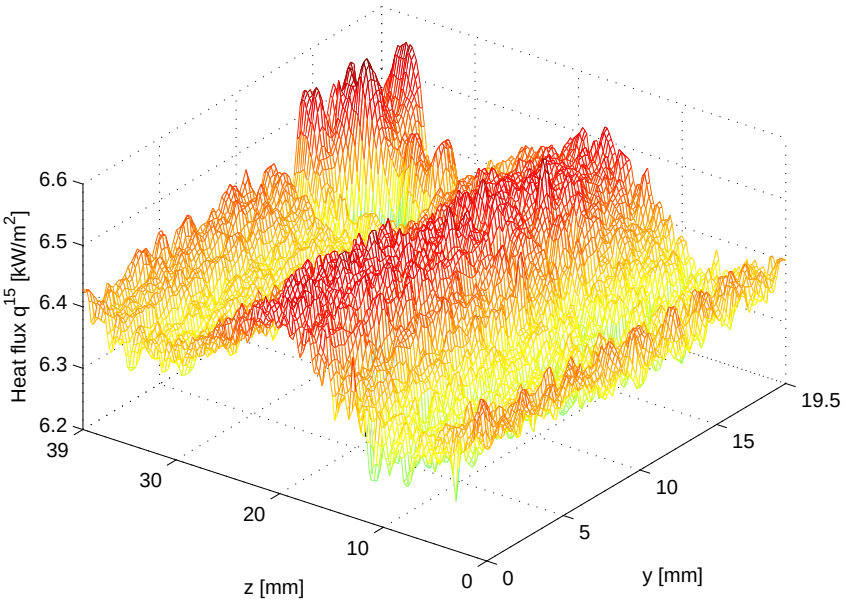


Figure 6.17: Heat flux estimate after 15 iterations

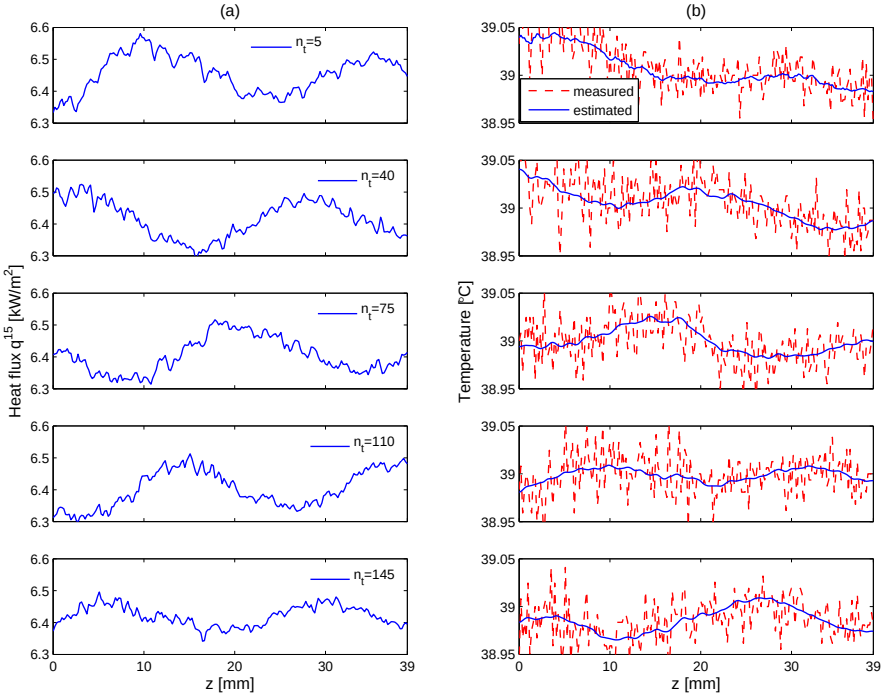


Figure 6.18: Estimation result after 15 iterations at different time steps: (a) heat flux estimate q^{15}; (b) measured (dashed line) and estimated (solid line) temperatures

same frequency (10 Hz) as the film waves. This can be traced back to the influence of the wavy film surface whose varying thickness effects the amount of heat that is transferred from the foil to the film. The quality of the approximation decreases at the end of time, since the solution of the adjoint problem is zero at final time and therefore yields no improvement of the corresponding iterative solution at $t = t_f$.

6.3.2 Multi-level optimization

In this section we consider a multi-level approach in two *different* ways. On the one hand, the up to now applied iterative solver (PCG) for the linear systems of equations is substituted by a robust multigrid method. On the other hand, we consider the method of nested optimization. Both methods require a hierarchy of (nested) grids. Let l_k denote the level with $2^k(1 \times 24 \times 49)$, $k \geq 0$, parallelepipeds. Since the number of grid points on the finest level l_2 (with $5 \times 97 \times 197$ unknowns) does not agree with the space resolution of the infrared camera, we only use the information of a cutout with 97×197 *pixel*. The temperature data in the remaining pixels are not used. For the dimension of the measurement section we still take 19.5×39 mm^2.

Robust multigrid

In the previous chapter a detailed convergence analysis of a multigrid method with a robust line smoother for anisotropic reaction-diffusion problems is given. This multigrid method is adapted to the falling film setup, in which the heating foil is very thin (25 μm) in x-direction relatively to the other directions. The very different length scales of the heat conductor result in an anisotropy effect. In each step of the implicit time integration method (Crank-Nicholson) an anisotropic reaction-diffusion problem has to be solved. We consider the multigrid V-cycle algorithm on the levels l_k, $k = 0, 1, 2$, with $\nu_1 = 2$ pre- and $\nu_2 = 0$ post-smoothing iterations. For the smoother we take the symmetric x-line Gauss-Seidel relaxation. The block problems arising within the smoother are solved with a tridiagonal LU decomposition. The zero vector is used as starting vector. Again, we take a reduction of the relative residual by 10^6.

On average (in time and the performed 50 optimization steps) we need 2 V-cycle iterations for the approximate solution of the corresponding discrete problems. Note, that in contrast to Section 5.6 for the numerical solution of the IHCP we have to treat linear reaction-diffusion problems (compare Remark 3.19) with *pure* Neumann boundary conditions. Thus, the suggested robust multigrid method of the previous chapter seems to work quite well for this case, too. However, since the eigenvalues of the corresponding system matrices are clustered due to the anisotropy (compare Section 5.3), for the PCG solver only 20 iterations have to be performed on average (and on the original grid with $5 \times 100 \times 200$ unknowns).

k	PCG	MG
0	13	-
1	16	2
2	20	2

Table 6.3: Iteration numbers for single-level optimization on level l_k

However, if the level number k on which the optimization is performed increases, also an increasing number of PCG iterations needed is expected. We emphasize that in contrast to that, for the robust multigrid method the rate of convergence seems to be uniform w.r.t. variation in the level number as Table 6.3 suggests.

Nested optimization

Finally, we consider the nested optimization approach on the levels l_k, $k = 0, 1, 2$. For the initial approximation on the coarsest level, we choose $q^0_{l_0}(\mathbf{x}, t) = \bar{q}(\mathbf{x}, t)$, $(\mathbf{x}, t) \in S_e$. The

initial temperature distributions on each level correspond to the first temperature frame assumed to be constant across the foil thickness (in x-direction). We use the discrepancy principle as the stopping rule on each level.

k	$J(q^0_{l_k})$	# iterations
0	$1.144 \cdot 10^{-7}$	14
1	$5.042 \cdot 10^{-8}$	8
2	$4.735 \cdot 10^{-8}$	3

Table 6.4: Iteration numbers for nested optimization with realistic measurement data

When the whole optimization is performed exclusively on level l_3, we have to apply 13 optimization steps. For the method of nested optimization the iteration numbers and the initial values of the objective functional are given in Table 6.4. Again, the plots of the regularized solutions $q^3_{l_3}$ and q^{13} are comparable (see Fig. 6.18). Thus, a significant reduction of the computational time can be achieved, if the optimization is performed on a hierarchy of nested space grids.

Chapter 7

Conclusions

In this thesis a three-dimensional anisotropic inverse heat conduction problem (IHCP) in falling films is solved numerically based on realistic high resolution temperature measurements that were performed with infrared thermography in a falling film experiment. The simulation studies show that a transient boundary heat flux on the (inaccessible) film side of a thin heating foil can be predicted adequately from the measurement data on the foil back side. While most numerical solution methods for inverse heat transfer problems that can be found in the literature are restricted to two-dimensional problems or, in the three-dimensional case, to a limited space-time resolution, we succeed in reconstructing the searched for quantity with high resolution in space and time. Besides high resolution measurement data, efficient and robust numerical methods are needed for this. For the solution of the ill-posed IHCP the conjugate gradient type method CGNE is used. This iterative optimization technique does not require an explicit representation of the operator in the inverse problem formulation, which (in general) is not available. From the theory of inverse problems it is known that the CGNE algorithm in combination with an appropriate stopping criterion is an order-optimal regularization method. Although a suitable problem discretization results in a kind of regularization, too, it is not sufficient for our problem. In the numerical experiments we clearly observe the typical behaviour of such a solution strategy, namely a limited data error amplification in the beginning of the iteration while the estimated quantity begins to oscillate, if too many iterations are applied. For the choice of the regularization parameter, i.e. the termination index of the CGNE iteration, we use the heuristic L-curve method which turned out to be satisfactory for our problem class. This parameter choice rule does not require any information about the noise level in the measurement data.

In the CGNE algorithm standard direct heat conduction problems have to be solved. However, the very different length scales of the heat conductor (the thin constantan foil) result in an anisotropy effect. In each step of an implicit time integration method a linear anisotropic reaction-diffusion problem has to be solved. For the numerical treatment we use a special space discretization based on linear anisotropic finite elements. In order to obtain a bound for the discretization error we use the Céa-lemma and error bounds for standard nodal Lagrangian interpolation. We give a uniform presentation of results for (triangular and) tetrahedral finite elements that can be found at different places in the literature and illustrate important phenomena by means of numerical experiments. A maximum angle condition is formulated which is the essential condition to deduce suitable bounds for the interpolation error. Different approaches for the derivation of these error bounds are considered. We discuss a special approach using a geometric argument and a more general technique based on adequate estimates on the reference element and the coordinate transformation. From these results we can conclude that anisotropic tetrahedral finite elements which satisfy a maximum angle condition are appropriate for the space discretization of our problem. In this thesis the finite element discretizations on multi-level nested grids were

computed by means of the software package DROPS[1], which was extended and improved for that purpose. The CGNE algorithm (i.e. the optimization) was implemented in MATLAB and an interface for the data transfer between MATLAB and DROPS was built.

In each step of the implicit time integration method linear anisotropic reaction-diffusion problems have to be solved, which contain two important parameters: there is a parameter which describes the anisotropy of the domain and one that measures the size of the reaction term relative to that of the diffusion term. After discretization we have a third parameter, namely the mesh width. For the efficient solution of the corresponding discrete problems the convergence of a multigrid method with a special (symmetric line Gauss-Seidel) smoother is analyzed. In the analysis there is a need for other (better) interpolation operators than the standard Lagrangian interpolation operator. Thus, a modified Scott-Zhang interpolation operator is studied. Both, for the W- and the V-cycle method we derive contraction number bounds smaller than one uniform with respect to all three parameters. For the multigrid W-cycle the convergence is analyzed in the framework of the approximation and smoothing property. Numerical experiments illustrate these robustness properties.

The combination of these numerical methods is applied to the realistic high resolution measurement data obtained from the falling film experiment. For a further efficiency improvement a multi-level approach for the optimization procedure is considered. This concept is also called method of nested optimization. The idea of this approach is to start the entire optimization process on a relatively coarse level (space grid) in order to find a good start approximation of the estimation quantity on a next finer level. The simulation studies show that a significant reduction of the computational time can be achieved, if the optimization is performed on a hierarchy of nested grids.

The presented results in this thesis may serve as a basis for future investigations of more complex inverse transport problems, such as the heat transfer through the falling film, mass transfer from the film to the gas phase, or reaction processes inside the film. Moreover, using the optimization technique presented in this thesis, a method for optimal design of the experimental setup could be developed and conditions to enhance the information content of the measurement data could be determined.

[1] http://www.igpm.rwth-aachen.de/DROPS/

Appendix A

Inverse problems

A.1 Integral equations

The following definition of specific integral equations is taken from [Smi58].

Definition A.1 (Integral equations of the first and second kind) *The equations*

$$y(t) = \int_a^b k(t,s)\,x(s)\,ds, \quad a \le t \le b, \tag{A.1}$$

$$x(t) = y(t) + \int_a^b k(t,s)\,x(s)\,ds, \quad a \le t \le b, \tag{A.2}$$

in which $x(t)$ is the unknown function, and all other functions are regarded as given, are linear integral equations. The general equations (A.1) *and* (A.2) *are called Fredholm equations of the first and second kind, respectively. If $k(t,s) = 0$, $t < s$, holds for the kernel these equations can be written as*

$$y(t) = \int_a^t k(t,s)\,x(s)\,ds, \quad a \le t \le b, \tag{A.3}$$

$$x(t) = y(t) + \int_a^t k(t,s)\,x(s)\,ds, \quad a \le t \le b. \tag{A.4}$$

The equations (A.3) *and* (A.4) *are called Volterra equations of the first and second kind.*

A.2 Regularization theory

We consider operator equations of the form $\mathcal{K}f = g$, where $\mathcal{K}$ is a bounded linear operator between the Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$. The operator $\mathcal{K}$ has nullspace $\mathcal{N}(\mathcal{K})$ and range $\mathcal{R}(\mathcal{K})$. Its domain is denoted by $\mathcal{D}(\mathcal{K})$. We introduce some basic results from regularization theory using the notation of [EHN96].

Definition A.2 (best-approximate solution) *Let $\mathcal{K} : \mathcal{X} \to \mathcal{Y}$ be a bounded linear operator.*

i. $f \in \mathcal{X}$ is called least-squares solution of $\mathcal{K}f = g$ if

$$\|\mathcal{K}f - g\| = \inf\{\|\mathcal{K}z - g\| \mid z \in \mathcal{X}\}.$$

ii. $f \in \mathcal{X}$ is called best-approximate solution of $\mathcal{K}f = g$ if f is a least-squares solution of $\mathcal{K}f = g$ and

$$\|f\| = \inf\{\|z\| \mid z \text{ is least-squares solution of } \mathcal{K}f = g\}.$$

Definition A.3 (Moore-Penrose inverse) *The Moore-Penrose (generalized) inverse $\mathcal{K}^\dagger$ of a bounded linear operator $\mathcal{K} : \mathcal{X} \to \mathcal{Y}$ is defined as the unique linear extension of $\tilde{\mathcal{K}}^{-1}$ to*

$$\mathcal{D}(\mathcal{K}^\dagger) := \mathcal{R}(\mathcal{K}) \oplus \mathcal{R}(\mathcal{K})^\perp,$$

with

$$\mathcal{N}(\mathcal{K}^\dagger) = \mathcal{R}(\mathcal{K})^\perp,$$

where

$$\tilde{\mathcal{K}} := \mathcal{K}|_{\mathcal{N}(\mathcal{K})^\perp} : \mathcal{N}(\mathcal{K})^\perp \to \mathcal{R}(\mathcal{K}).$$

In this definition the superscript $^\perp$ denotes the orthogonal complement and the composition $\oplus$ denotes the direct sum. The generalized inverse is the solution operator mapping g onto the best-approximate solution of $\mathcal{K}f = g$, as the following theorem states.

Theorem A.4 *Let $g \in \mathcal{D}(\mathcal{K}^\dagger)$. Then $\mathcal{K}f = g$ has a unique best-approximate solution, which is given by*

$$f^\dagger := \mathcal{K}^\dagger g.$$

The set of all least-squares solutions is $f^\dagger + \mathcal{N}(\mathcal{K})$.

Proposition A.5 *The Moore-Penrose generalized inverse $\mathcal{K}^\dagger$ is bounded (i.e. continuous) if and only if $\mathcal{R}(\mathcal{K})$ is closed.*

The best-approximate solution can be characterized by the normal equation.

Theorem A.6 *Let $g \in \mathcal{D}(\mathcal{K}^\dagger)$. Then $f \in \mathcal{X}$ is a least-squares solution of $\mathcal{K}f = g$ if and only if the normal equation*

$$\mathcal{K}^*\mathcal{K}f = \mathcal{K}^*g$$

holds.

Thus, $\mathcal{K}^\dagger g$ is the solution of $\mathcal{K}^*\mathcal{K}f = \mathcal{K}^*g$ with minimum norm, i.e. $\mathcal{K}^\dagger = (\mathcal{K}^*\mathcal{K})^\dagger\mathcal{K}^*$.

Definition A.7 (regularization method) *Let $\mathcal{K} : \mathcal{X} \to \mathcal{Y}$ be a bounded linear operator between the Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$, $\alpha_0 \in (0, \infty]$. For every $\alpha \in (0, \alpha_0)$ let*

$$\mathcal{R}^\alpha : \mathcal{Y} \to \mathcal{X}$$

be a continuous (not necessarily linear) operator. The family $\{\mathcal{R}^\alpha\}_{\alpha>0}$ is called a regularization or a regularization operator (for $\mathcal{K}^\dagger$), if for all $g \in \mathcal{D}(\mathcal{K}^\dagger)$ there exists a parameter choice rule $\alpha = \alpha(\delta, g_\delta)$ such that

$$\lim_{\delta\to 0} \sup\{\|\mathcal{R}^{\alpha(\delta,g_\delta)}g_\delta - \mathcal{K}^\dagger g\| \,|\, g_\delta \in \mathcal{Y},\ \|g_\delta - g\| \le \delta\} = 0 \tag{A.5}$$

holds. Here, $\alpha : \mathbb{R}^+ \times \mathcal{Y} \to (0, \alpha_0)$ is such that

$$\lim_{\delta\to 0} \sup\{\alpha(\delta, g_\delta) \,|\, g_\delta \in \mathcal{Y},\ \|g_\delta - g\| \le \delta\} = 0. \tag{A.6}$$

For a specific $g \in \mathcal{D}(\mathcal{K}^\dagger)$, a pair $(\mathcal{R}^\alpha, \alpha)$ is called a (convergent) regularization method (for solving $\mathcal{K}f = g$) if (A.5) *and* (A.6) *hold.*

Proposition A.8 *Let $\mathcal{K}$ be compact with singular system $(\sigma_n; v_n, u_n)$. Then, for $\mu > 0$,*

$$\mathcal{K}^\dagger g \in \mathcal{R}((\mathcal{K}^*\mathcal{K})^{\mu/2}) \tag{A.7}$$

if and only if

$$\sum_{n=1}^{\infty} \frac{1}{\sigma_n^{2+2\mu}} |(g, u_n)|^2 < \infty.$$

Thus, (A.7) is a condition on the decay rate of the Fourier coefficients (g, u_n) relative to the singular values σ_n, which becomes more severe as μ becomes larger.

Appendix B

Variational formulation of PDEs

Let $\Omega \subset \mathbb{R}^n$, $n = 1, 2, 3$, be an open, bounded and connected domain with a (piecewise) smooth boundary $\partial\Omega$. We introduce some function spaces that are often referred to in this thesis.

B.1 Basic function spaces

Let $C^k(\Omega)$, $k \in \mathbb{N}$, denote the space of functions $f : \Omega \to \mathbb{R}$ for which all (partial) derivatives

$$D^\alpha f := \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}, \quad \alpha = (\alpha_1, \cdots, \alpha_n), \quad |\alpha| = \alpha_1 + \cdots + \alpha_n,$$

of order $|\alpha| \leq k$ are continuous functions on Ω. The space $C^k(\bar{\Omega})$, $k \in \mathbb{N}$, consists of all functions in $C^k(\Omega)$ for which all derivatives up to order $|\alpha| \leq k$ have continuous extensions to $\bar{\Omega}$. For $f : \Omega \to \mathbb{R}$ we define the support by

$$\text{supp}(f) := \overline{\{\mathbf{x} \in \Omega \,|\, f(\mathbf{x}) \neq 0\}},$$

which leads to the space $C_0^k(\bar{\Omega})$, $k \in \mathbb{N}$, of all functions in $C^k(\bar{\Omega})$ which have a compact support in Ω, i.e. $\text{supp}(f) \subset \Omega$.

The space of all measurable and to the power of $p \in [1, \infty)$ integrable functions is denoted by

$$L_p(\Omega) := \left\{ v \,|\, \int_\Omega |v|^p \, d\mathbf{x} < \infty \right\},$$

with the associated norm

$$\|v\|_{L_p(\Omega)} := \left(\int_\Omega |v|^p \, d\mathbf{x} \right)^{\frac{1}{p}}.$$

In extension of the above definitions we introduce the space

$$L_\infty(\Omega) := \{ v \,|\, \text{ess sup}_{\mathbf{x} \in \Omega} \, |v(\mathbf{x})| < \infty \},$$

with the norm

$$\|v\|_{L_\infty(\Omega)} := \text{ess sup}_{\mathbf{x} \in \Omega} \, |v(\mathbf{x})|.$$

Here, ess sup denotes the essential supremum, i.e. the smallest upper bound except for null sets.

B.2 Sobolev spaces

An introduction and properties of Sobolev spaces can be found e.g. in [BS02, Cia78].

Definition B.1 (weak derivative) *Consider $u \in L_2(\Omega)$ and $|\alpha| > 0$. If there exists $v \in L_2(\Omega)$ such that*

$$\int_\Omega v\phi \, d\mathbf{x} = (-1)^{|\alpha|} \int_\Omega u D^\alpha \phi \, d\mathbf{x}, \quad \forall \phi \in C_0^\infty(\Omega),$$

then v is called the αth weak derivative of u and is denoted by $D^\alpha u := v$.

For $p \in [1, \infty]$ the Sobolev space $W_p^m(\Omega)$ consists of those functions in $L_p(\Omega)$, for which all the αth weak derivatives with $|\alpha| \le m$ belong to the space $L_p(\Omega)$, i.e.

$$W_p^m(\Omega) := \{u \in L_p(\Omega) \,|\, D^\alpha u \in L_p(\Omega), \forall\, |\alpha| \le m\}, \quad m = 0, 1, \dots$$

Equipped with the norm

$$\|u\|_{W_p^m(\Omega)} := \begin{cases} \left(\sum\limits_{|\alpha| \le m} \int_\Omega |D^\alpha u|^p \, d\mathbf{x} \right)^{\frac{1}{p}}, & \text{if } 1 \le p < \infty, \\ \max\limits_{|\alpha| \le m} (\operatorname{ess\,sup}_{\mathbf{x} \in \Omega} |D^\alpha u(\mathbf{x})|), & \text{if } p = \infty, \end{cases} \tag{B.1}$$

the space $W_p^m(\Omega)$ is a Banach space. The corresponding seminorm $|u|_{W_p^m(\Omega)}$ is defined by (B.1) and $|\alpha| \le m$ substituted by $|\alpha| = m$. The Sobolev space $W_{p,0}^m(\Omega)$ is the closure of $C_0^\infty(\Omega)$ in the space $W_p^m(\Omega)$. For $p = 2$ the Sobolev spaces $W_p^m(\Omega)$ and $W_{p,0}^m(\Omega)$ together with the scalar product

$$(u, v)_m := \sum_{|\alpha| \le m} \int_\Omega D^\alpha u D^\alpha v \, d\mathbf{x}, \quad m = 0, 1, \dots$$

form Hilbert spaces, which are denoted by $H^m(\Omega)$ and $H_0^m(\Omega)$, respectively.

Lemma B.2 (Green's formula) *Let $\Omega \subset \mathbb{R}^n$, $n = 2, 3$, be a domain with a smooth boundary Γ. Then the equality*

$$\int_\Omega \Delta u v \, d\mathbf{x} = \int_\Gamma \frac{\partial u}{\partial \mathbf{n}} v \, ds - \int_\Omega \nabla u \nabla v \, d\mathbf{x}, \quad \forall u \in H^2(\Omega),\, v \in H^1(\Omega),$$

holds.

In this thesis we often refer to embeddings of two Banach spaces $\mathcal{X}$ and $\mathcal{Y}$. We denote by $\mathcal{X} \hookrightarrow \mathcal{Y}$ the continuous embedding of $\mathcal{X}$ into $\mathcal{Y}$, which means that $\mathcal{X} \subset \mathcal{Y}$ and

$$\exists c = c(\Omega) : \|u\|_{\mathcal{Y}} \le c \|u\|_{\mathcal{X}}, \quad \forall u \in \mathcal{X}.$$

The following embedding results can be found e.g. in [Cia78]:

$$\begin{array}{lcll} W_p^m(\Omega) & \hookrightarrow & C(\bar{\Omega}), & \text{if } \begin{cases} mp > n, & p > 1, \\ m \ge n, & p = 1, \end{cases} \\ W_p^m(\Omega) & \hookrightarrow & L_q(\Omega), & \text{if } \begin{cases} mp < n, & \frac{1}{q} = \frac{1}{p} - \frac{m}{n}, \\ mp = n, & 1 \le q < \infty. \end{cases} \\ W_p^m(\Omega) & \hookrightarrow & W_p^l(\Omega), & \text{if } m \ge l, \\ W_p^m(\Omega) & \hookrightarrow & W_q^m(\Omega), & \text{if } p \ge q. \end{array}$$

Bibliography

[AB84] O. Axelsson and V.A. Barker: *Finite Element Solution of Boundary Value Problems.* Academic Press New York, 1984.

[AD92] T. Apel and M. Dobrowolski: *Anisotropic interpolation with applications to the finite element method.* Computing, 47:277–293, 1992.

[Ali94] O.M. Alifanov: *Inverse Heat Transfer Problems.* Springer Berlin Heidelberg, 1994.

[ALR02] F. Al-Sibai, A. Leefken, and U. Renz: *Local and instantaneous distribution of heat transfer rates through wavy films.* Int. J. Therm. Sci., 41:658–663, 2002.

[Als05] F. Al-Sibai: *Experimentelle Untersuchungen der Strömungscharakteristik und des Wärmeübergangs bei welligen Rieselfilmen.* PhD thesis, RWTH Aachen, 2005.

[Ape99] T. Apel: *Anisotropic Finite Elements: Local Estimates and Applications.* Teubner Stuttgart Leipzig, 1999.

[AS02] T. Apel and J. Schöberl: *Multigrid methods for anisotropic edge refinement.* SIAM J. Numer. Anal., 40(5):1993–2006, 2002.

[BA76] I. Babuška and A.K. Aziz: *On the angle condition in the finite element method.* SIAM J. Numer. Anal., 13:214–226, 1976.

[BBH96] J.V. Beck, B. Blackwell, and A. Haji-Sheikh: *Comparison of some inverse heat conduction methods using experimental data.* Int. J. Heat Mass Transfer, 39(17):3649–3657, 1996.

[Bey98] J. Bey: *Finite-Volumen- und Mehrgitter-Verfahren für elliptische Randwertprobleme.* Teubner Stuttgart Leipzig, 1998.

[BM97] J. Blum and W. Marquardt: *An optimal solution to inverse heat conduction problems based on frequency domain interpretation and observers.* Numer. Heat Transfer: Part B: Fundam., 32:453–478, 1997.

[BS02] S.C. Brenner and L.R. Scott: *The Mathematical Theory of Finite Element Methods.* Springer New York, 2002.

[BZ00] J.H. Bramble and X. Zhang: *Uniform convergence of the multigrid V-cycle for an anisotropic problem.* Math. Comput., 70(234):453–470, 2000.

[Cha01] S. Chantasiriwan: *An algorithm for solving multidimensional inverse heat conduction problems.* Int. J. Heat Mass Transfer, 44:3823–3832, 2001.

[Cia78] P.G. Ciarlet: *The Finite Element Method for Elliptic Problems.* North-Holland Amsterdam, 1978.

[EHN96] H.W. Engl, M. Hanke, and A. Neubauer: *Regularization of Inverse Problems.* Kluwer Academic Publishers Dordrecht, 1996.

[Eld83] L. Eldén: *The numerical solution of a non-characteristic Cauchy problem for a parabolic equation.* In P. Deuflhard and E. Hairer, editors, Numerical treatment of inverse problems in differential and integral equations, Birkhäuser Boston, pages 246–268, 1983.

[Eva98] L.C. Evans: *Partial Differential Equations.* Appl. Math. Sci., Providence Rhode Island, 1998.

[Fol76] G.B. Folland: *Introduction to Partial Differential Equations.* Mathematical Notes, Princeton University Press and University of Tokyo Press, 1976.

[GPRR02] S. Groß, J. Peters, V. Reichelt, and A. Reusken: *The DROPS package for numerical simulations of incompressible flows using parallel adaptive multigrid techniques.* Technical report no. 211, 2002, http://www.igpm.rwth-aachen.de/reports/ps/IGPM211.ps.gz.

[GR94] C. Großmann and H.-G. Roos: *Numerik partieller Differentialgleichungen.* Teubner Stuttgart, 1994.

[Gri85] P. Grisvard: *Elliptic Problems in Nonsmooth Domains.* Pitman, 1985.

[GSM+05] S. Groß, M. Soemers, A. Mhamdi, F. Al-Sibai, A. Reusken, W. Marquardt, and U. Renz: *Identification of boundary heat fluxes in a falling film experiment using high resolution temperature measurements.* Int. J. Heat Mass Transfer, 48:5549–5562, 2005.

[Hac85] W. Hackbusch: *Multi-grid Methods and Applications.* Springer Berlin Heidelberg New York Tokyo, 1985.

[Hac94] W. Hackbusch: *Iterative Solution of Large Sparse Systems of Equations.* Appl. Math. Sci. 95, Springer New York, 1994.

[Had23] J. Hadamard: *Lectures on Cauchy's Problem in Linear Partial Differential Equations.* Yale University Press New Haven, 1923.

[Han92] P.C. Hansen: *Analysis of discrete ill-posed problems by means of the L-curve.* SIAM Review, 34(4):561–580, 1992.

[Han95] M. Hanke: *Conjugate Gradient Type Methods for Ill-posed Problems.* Pitman Research Notes in Mathematics Series 327, Longman Scientific & Technical, 1995.

[Hao98] D.H. Háo: *Methods for Inverse Heat Conduction Problems.* Peter Lang, Europäischer Verlag der Wissenschaften, 1998.

[HME+03] G. Hetsroni, D. Mewes, C. Enke, M. Gurevich, A. Mosyak, and R. Rozenblit: *Heat transfer to two-phase flow in inclined tubes.* Int. J. Multiphase Flow, 29(2):173–194, 2003.

[HMG+08] Y. Heng, A. Mhamdi, S. Groß, A. Reusken, M. Buchholz, H. Auracher, and W. Marquardt: *Reconstruction of local heat fluxes in pool boiling experiments along the entire boiling curve from high resolution transient temperature measurements.* Int. J. Heat Mass Transfer, accepted 2008.

[HO93] P.C. Hansen and D.P. O'Leary: *The use of the L-curve in the regularization of discrete ill-posed problems.* SIAM J. Sci. Comp., 14(6):1487–1503, 1993.

[HR94] G. Hetsroni and R. Rozenblit: *Heat transfer to a liquid-solid mixture in a flume.* Int. J. Multiphase Flow, 20:671–689, 1994.

[HR98] D.H. Háo and H.-J. Reinhardt: *Gradient methods for inverse heat conduction problems.* Inverse Problems in Engineering, 6:177–211, 1998.

[HW99] C.-H. Huang and S.-P. Wang: *A three-dimensional inverse heat conduction problem in estimating surface heat flux by conjugate gradient method.* Int. J. Heat Mass Transfer, 42:3387–3403, 1999.

[JL00] P. Jonas and A.K. Louis: *Approximate inverse for a one-dimensional inverse heat conduction problem.* Inverse Problems, 16:175–185, 2000.

[KGM+08] M. Karalashvili, S. Groß, A. Mhamdi, A. Reusken, and W. Marquardt: *Incremental identification of transport coefficients in convection-diffusion systems.* SIAM J. Sci. Comput., accepted 2008.

[Kir96] A. Kirsch: *An Introduction to the Mathematical Theory of Inverse Problems.* Appl. Math. Sci. 120, Springer, 1996.

[Kri92] M. Křížek: *On the maximum angle condition for linear tetrahedral elements.* SIAM J. Numer. Anal., 29:513–520, 1992.

[Lam00] P.K. Lamm: *A survey of regularization methods for first-kind Volterra equations.* In D. Colton, H.W. Engl, A. Louis, J.R. McLaughlin, and W. Rundell, editors, Surveys on solution methods for inverse problems, Springer Vienna New York, pages 53–82, 2000.

[LAR01] A. Leefken, F. Al-Sibai, and U. Renz: *Measurement of instantaneous heat transfer using a hot-foil infrared technique.* In Thermosense XXIII, Proceedings of SPIE, Orlando USA, 4630:21–29, April 16–19, 2001.

[LM91] T.H. Lyu and I. Mudawar: *Statistical investigation of the relationship between interfacial waviness and sensible heat transfer to a falling liquid film.* Int. J. Heat Mass Transfer, 34(5):1451–1464, 1991.

[LMFA99] S. Leuthner, A.H. Maun, S. Fiedler, and H. Auracher: *Heat and mass transfer in wavy falling films of binary mixtures.* Int. J. Therm. Sci., 38:937–943, 1999.

[LMM05] T. Lüttich, A. Mhamdi, and W. Marquardt: *Design, formulation and solution of multi-dimensional inverse heat conduction problems.* Numer. Heat Transfer: Part B: Fundamentals, 47(2):111–133, 2005.

[LSU68] O.A. Ladyzenskaja, V.A. Solonnikov, and N.N. Ural'ceva: *Linear and Quasilinear Equations of Parabolic Type.* Appl. Math. Sci., Translations of Mathematical Monographs 23, 1968.

[LT03] S. Larsson and V. Thomée: *Partial Differential Equations with Numerical Methods.* Springer Berlin Heidelberg New York, 2003.

[Mar05] W. Marquardt: *Model-based experimental analysis of kinetic phenomena in multi-phase reactive systems.* Chem. Eng. Res. Des., 83(A8):1–13, 2005.

[MM94] K.W. Morton and D.F. Mayers: *Numerical Solution of Partial Differential Equations.* Cambridge University Press Cambridge New York Melbourne, 1994.

[NW99] J. Nocedal and S.J. Wright: *Numerical Optimization.* Springer Berlin Heidelberg New York, 1999.

[QV94] A. Quarteroni and A. Valli: *Numerical Approximation of Partial Differential Equations.* Springer Berlin Heidelberg, 1994.

[Reg96] T. Regińska: *A regularization parameter in discrete ill-posed problems.* SIAM J. Sci. Comput., 17(3):740–749, 1996.

[Reu05] A. Reusken: *Numerical methods for elliptic partial differential equations.* Lecture notes, 2005.

[RS07] A. Reusken and M. Soemers: *On the robustness of a multigrid method for anisotropic reaction-diffusion problems.* Computing, 80:299–317, 2007.

[Saa03] Y. Saad: *Iterative Methods for Sparse Linear Systems – 2nd edition.* SIAM, 2003.

[Smi58] F. Smithies: *Integral Equations.* Cambridge University Press, 1958.

[Ste93a] R. Stevenson: *Robustness of multi-grid applied to anisotropic equations on convex domains with re-entrant corners.* Numer. Math., 66:373–398, 1993.

[Ste93b] R. Stevenson: *New estimates of the contraction number of V-cycle multi-grid with applications to anisotropic equations.* In W. Hackbusch and G. Wittum, editors, Incomplete decompositions, Proceedings of the Eighth GAMM Seminar, Notes on Numerical Fluid Mechanics, Vieweg Braunschweig, 41:159–167, 1993.

[Ste94] R. Stevenson: *Robust multi-grid with 7-point ILU smoothing.* In P.W. Hemker and P. Wesseling, editors, Multigrid methods IV, Birkhäuser, 116:295–308, 1994.

[Syn57] J.L. Synge: *The Hypercircle in Mathematical Physics.* Cambridge University Press New York, 1957.

[Tho97] V. Thomée: *Galerkin Finite Element Methods for Parabolic Problems.* Springer Berlin Heidelberg New York, 1997.

[TOS01] U. Trottenberg, C. Oosterlee, and A. Schüller: *Multigrid.* Academic Press London, 2001.

[vdV03] H.A. van der Vorst: *Iterative Krylov Methods for Large Linear Systems.* Cambridge University Press, 2003.

[Wit89a] G. Wittum: *Linear iterations as smoothers in multigrid methods: theory with applications to incomplete decompositions.* IMPACT Comput. Sci. Eng., 1:180–215, 1989.

[Wit89b] G. Wittum: *On the robustness of ILU smoothing.* SIAM J. Sci. Stat. Comput., 10(4):699–717, 1989.

[YC97] C. Yang and C. Chen: *Inverse estimation of the boundary condition in three-dimensional heat conduction.* J. Phys. D. Appl. Phys., 30(15):2209–2216, 1997.

[Zla68] M. Zlámal: *On the finite element method.* Numer. Math., 12:394–409, 1968.

Lebenslauf

Persönliche Daten

• Name	Marcus Soemers
• Geburtsdatum	15.4.1974
• Geburtsort	Mönchengladbach

Schulbildung

• 1980 – 1981	Gemeinschaftsgrundschule Keyenberg
• 1981 – 1984	Ev. Grundschule Pahlkestraße in Mönchengladbach
• 1984 – 1993	Städt. Gymnasium a.d. Gartenstraße in Mönchengladbach
• 5/1993	Abitur

Wehrdienst

• 7/1993 – 9/1994	Ersatzdienst

Studium

• 10/1994 – 9/1996	Maschinenbaustudium, *RWTH Aachen*
• 10/1996 – 8/2002	Mathematikstudium mit Nebenfach Informatik, *RWTH Aachen*
• 10/1997 – 12/1997	stud. Hilfskraft am Lehrst. D für Mathematik
• 4/1998 – 1/2000	stud. Hilfskraft am Lehrst. für Numerische Mathematik
• 3/1999	Vordiplom im Studiengang Maschinenbau
• 8/2002	Diplom im Studiengang Mathematik (Thema: Diskretisierung und iterative Lösung von Konvektions-Diffusions-Gleichungen auf Shishkin-Gittern)

Beruf

• seit 9/2002	Promotionsstudium am Lehrstuhl für Numerische Mathematik (LNM), *RWTH Aachen*
• 9/2002 – 8/2005	Förderung durch das Graduiertenkolleg „Hierarchie und Symmetrie in Mathematischen Modellen“
• seit 9/2005	Wissenschaftlicher Mitarbeiter am LNM
• seit 6/2007	Assoziiertes Mitglied der AICES Graduiertenschule, *RWTH Aachen*