Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

J. LANG

Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems. Theory, Algorithm, and Applications

Preprint SC 99–20 (September 1999)

to Susanne, Annemarie, and Elisabeth

Preface

This monograph has been written to illustrate the interlocking of theory, algorithm, and application in developing solution techniques for complex PDE systems. A deep theoretical understanding is necessary to produce a powerful idea leading to a successful algorithm. Efficient and robust implementation is the key to make the algorithm perform satisfactorily. The extra insight obtained by solving real–life problems brings out the structure of the method more clearly and often suggests ways to improve the numerical algorithm.

It is my intention to impart the beauty and complexity found in both the theoretical investigation of the adaptive algorithm proposed here, i.e., the coupling of Rosenbrock methods in time and multilevel finite elements in space, and its realization. I hope that this method will find many more interesting applications.

Berlin–Dahlem, Januar 1999 *Jens Lang*

Acknowledgements

I have looked forward to writing this section since it is a pleasure for me to thank all friends who made this work possible and provided valuable input.

I would like to express my gratitude to Peter Deuflhard for giving me the opportunity to work in the field of Scientific Computing. I have benefited immensely from his help to get the right perspectives, and from his continuous encouragement and support over several years. He certainly will forgive me the use of Rosenbrock methods rather than extrapolation methods to integrate in time.

My colleagues at the ZIB form a friendly and open–minded community. In particular, I thank Ulrich Nowak for discussing the never ending story about adaptive strategies and about many practical issues necessary to design an efficient and robust PDE solver. With admirable teaching skill Rainer Roitzsch introduced me to the subject of dynamical programming. From him I have learnt all I know about C and C++. He has been and still is a tremendous source of information. I also thank Bodo Erdmann for coding a three–dimensional version of KARDOS and sharing time with me to find the definitely last programming error.

Working in interdisciplinary areas I have enjoyed the joint works with Ulrich Maas and Jochen Fröhlich who taught me how to bring flames under control, with Wilhelm Ruppel who gave me an excellent insight into the real engineering world, with Wilhelm Merz who never lost hope in dealing with semiconductors, and with Martin Seebass who showed me the best way to look at a temperature distribution in a human body. Thank you.

I have plagued Gert Lube with several versions of the manuscript. His constructive criticism and stimulating discussions have greatly improved the quality of the final product. I also indebted to Sigrid Wacker, Uwe Pöhle, and Rainald Ehrig for carefully reading the manuscript. Special thanks to Regine Kossick for providing me quickly with all the literature I have ordered.

I would like to acknowledge my gratitude to Reinhard Lehmann who laid the foundations of my finite element thinking.

Finally, thanks to my beloved wife Susanne for her love and support.

Contents

I

Introduction

Diverse physical phenomena in such fields as biology, chemistry, metallurgy, medicine, and combustion are modelled by systems of nonlinear parabolic partial differential equations (PDEs). Nowadays there is an increasing activity in mathematics to analyse the properties of such models, including existence, uniqueness, and regularity of their solutions (e.g. AMANN [6], LUNARDI [97]). Due to the great complexity of such systems only little is known about true solutions. Furthermore, the permanent advance in computational capabilities allows the incorporation of more and more detailed physics into the models. Apart from a few situations, where mathematical analysis can actually be applied, the numerical analysis of PDEs is the central tool to assess the modelling process for large scale physical problems. In fact, a posteriori error estimates can be used to judge the quality of a numerical approximation and to determine an adaptive strategy to improve the accuracy where needed. In such a way numerical and modelling errors can be clearly distinguished with the effect that the reliability of the modelling process can be assessed. Moreover, successful adaptive methods lead to substantial savings in computational work for a given tolerance. They are now entering into real–life applications and starting to become a standard feature of modern software.

In this work, we concentrate on nonlinear parabolic systems which can be written as abstract Cauchy problems of the form

$$\partial_t u = F(t,u)\,, \quad u(0) = u_0\,, \quad 0 < t \leq T\,,$$

where the vector–valued solution is supposed to be unique and temporally smooth, at least after an initial transitional phase. Our main assumption which gives a parabolic character to this problem is that for each u and t the Fréchet derivative $A = -\partial_u F(t,u)$ is a negative infinitesimal generator of an analytic semigroup (see e.g. AMANN [5]). Because analytic semigroups are generated by sectorial operators, the eigenvalues of $A + \kappa I$, where $\kappa > 0$ is sufficiently large, belong to a sector $\{\lambda : |\arg(\lambda)| < \phi,\, \phi < \pi/2\}$ in the right complex half plane.

It is well–known that differential operators give rise to infinite stiffness. Therefore, often an implicit discretization method coupled with a Newton–like iteration is applied to integrate in time. Investigating the convergence of Newton's method in function space, DEUFLHARD [47] pointed out that one calculation of the Jacobian or an approximation of it per time step is sufficient to integrate stiff problems efficiently. In this work, we use linearly implicit methods of

Rosenbrock type which are constructed by working the exact Jacobian directly into the formula (ROSENBROCK [118]). These methods offer several advantages. They completely avoid the solution of nonlinear equations, which means that no Newton iteration has to be controlled. There is no problem to construct Rosenbrock methods with optimal linear stability properties for stiff equations. Because of their one–step nature, they allow a rapid change of step sizes. Last but not least, they are very easy to program – as simple as explicit methods. Most of the knowledge about linearly implicit methods can be found in the books of STREHMEL & WEINER [137], HAIRER & WANNER [72], and DEUFLHARD & BORNEMANN [48].

Our aim is to analyse and to design an adaptive algorithm for nonlinear parabolic systems with time– and solution–dependent operators, where linearly implicit methods in time are coupled with multilevel finite elements in space. The time steps and the mesh sizes are automatically chosen during the integration in order to control the discretization error with respect to a prescribed tolerance given by the user. Apart from a few results for semilinear and quasilinear equations, it seems that this has not been studied previously. There are, however, a number of different adaptive techniques which can be classified by the discretization sequence used. A posteriori error estimates for parabolic equations have been developed primarily within the classical method of lines approach (MOL). Discretizing in space first, the time–dependent PDE is transformed into an ODE–system which can be solved by an appropriate variable step–size time integrator. The accuracy in space is controlled by a posteriori error estimators constructed for stationary problems (e.g. BIETERMAN and BABUŠKA [27], ADJERID and FLAHERTY [2], TROMPERT and VERWER [141], MOORE [103], NOWAK [106], VANDE WOUWER, SAUCEZ, and SCHIESSER [143], BERZINS, CAPON, and JIMACK [25]). More recently, the reverse discretization sequence, first in time then in space, known as Rothe's method has been investigated. Interpreting the time–dependent PDE as an ODE in a Hilbert (or Banach) space, the temporal error can be estimated by classical ODE–procedures. The spatial discretization is considered as a perturbation of the time integration and can be assessed by standard error estimators for stationary problems (e.g. BORNEMANN [29, 30, 31], LANG and WALTER [88], LANG [84]). For a comparative study of the MOL and Rothe's approach we refer to DEUFLHARD, LANG, and NOWAK [49]. A third possibility is to discretize simultaneously in space and time employing a discontinuous Galerkin method and to apply coupled space–time estimators (e.g. ERIKSSON and JOHNSON [55], VERFÜRTH [146]). The method of Moving Finite Elements invented by MILLER and MILLER [102] is a further way of adaptively solving PDEs using mesh points which automatically move in the space–time domain. The physical PDE is replaced by an extended system of the PDE and the so-called moving mesh equation (e.g. BAINES [13], HUANG and RUSSELL [76], ZEGELING [153]).

In this work, we follow Rothe's approach. In Chapter II the nonlinear parabolic problem is introduced in a Hilbert space setting. We summarize known convergence results for Rosenbrock methods applied to partial differential equations.

It is a known fact that Rosenbrock methods suffer from order reduction, i.e., the classical order in general cannot be achieved. This phenomenon has also been seen for implicit Runge–Kutta and extrapolation methods. Sharp error estimates showing fractional orders of convergence were established in a sequence of papers by LUBICH, OSTERMANN, and ROCHE [107, 108, 94, 95]. Fortunately, there are conditions which imply higher order of convergence for Rosenbrock methods (LUBICH and OSTERMANN [95]).

In Chapter III, we investigate the approximation properties of finite elements applied to spatial discretization of Rosenbrock schemes. For this, we use a perturbation technique proposed by LUBICH and OSTERMANN [94] for Runge–Kutta approximations of quasilinear parabolic problems. Application of resolvent bounds and standard finite element techniques yields full spatial convergence order. Global error estimates are given in large generality, including discrete versions of the $C^0([0,T];L^2)$–norm, that is, the maximum L^2–norm in space taken over all time levels, or the $L^2(0,T;H^1)$–norm which measures also spatial derivatives of the error. The results remain valid for variable step sizes satisfying the condition of quasiuniformity.

A fundamental property of stable one–step integration methods is that the global error consists of propagated and accumulated local truncation errors. Thus, controlling the local errors of each individual time step with respect to a given tolerance leads to a control of the global error. So our objective is to construct efficient computational estimations of the local numerical errors arising in the temporal and spatial discretization.

In Chapter IV, we discuss our a posteriori error estimators which are based on the difference of higher and lower order solutions. The classical embedding technique for ODE integrators is employed to estimate the error in time. An automatic step size selection procedure ensures that the step size is as large as possible to guarantee the desired precision. It turns out that a combination of the standard controller and the PI–controller proposed by GUSTAFSSON et al. [70] works very well for a large class of problems with a great diversity in the dynamic behaviour.

To estimate the error in space we extend the hierarchical bases technique to Rosenbrock schemes. Hierarchical error estimators have been accepted to provide efficient and reliable assessment of spatial errors for stationary problems (DEUFLHARD, LEINEN, and YSERENTANT [50], ZIENKIEWICZ et al. [155], BANK and SMITH [18], BORNEMANN, ERDMANN, and KORNHUBER [34]). They can be computed locally by small element–by–element calculations. We study the robustness of our hierarchical error estimator with respect to a stepsize–dependent norm. By robustness we mean that the estimator yields upper and lower bounds on the error uniformly in the time step and in the mesh size. First, exact robustness is proven for one–stage Rosenbrock methods. For higher stage numbers the special structure of the Rosenbrock methods gives rise to a nonlinear spatial error transport which strongly influences the whole estimation process. In this case, the derived estimates are just nearly optimal. A closer discussion, however, shows that the occuring perturbation terms in general are negligible for practical computations.

In Chapter V, we discuss practical issues which are useful to implement our adaptive strategies. The spatial error estimator is used to construct an efficient and reliable adaptive strategy for an automatic mesh control. A sequence of improved spatial meshes is built up in order to get a mesh with as few degrees of freedom as possible such that the computed error is less than the prescribed tolerance. This approach directly corresponds to the multilevel finite element technique well established for the adaptive solution of stationary problems (e.g. DEUFLHARD, LEINEN, and YSERENTANT [50], BORNEMANN, ERDMANN, and KORNHUBER [33]). However, beginning with a time–fixed coarse grid at each time level and using the multilevel technique would be wasteful. For time–dependent problems, there is often considerable information which can be used from the optimum grid at the previous time to construct a first approximation of the desired grid at the advanced time. Thus, an efficient grid removal based on the same error estimators as used for refinement is applied to determine where degrees of freedom are no longer needed. Note, that this requires data structures that allow both grid refinement and robust coarsening.

In Chapter VI, we present illustrative numerical tests including three Rosenbrock solvers to demonstrate that the theoretical order predictions are indeed of interest for the numerical practice. For linear parabolic equations, similar tests can be found in OSTERMANN and ROCHE [108]. The observed temporal convergence rates nicely correspond to the theoretical values. Optimum convergence order is obtained for the finite element discretization. Furthermore, we assess the quality of the hierarchical error estimator in terms of the effectivity index and present some results for the performance of the whole adaptive algorithm.

The final Chapter VII is dedicated to a series of real–life applications that arise in today's chemical industry, semiconductor–device fabrication, and health care. Usually, apart from a few exceptions it takes ten or more years before the results of academic research become available in professional practice (BABUŠKA and SZABO [12]). One way to shorten this transfer is to create one's own software product and to demonstrate that the developed numerical algorithms work robustly and safely over a wide range of practically relevant problems. The program package KARDOS that is based on the stationary solver KASKADE [54] was coded along the adaptive principles proposed in the theoretical part of this work.

We have chosen problems that are on one hand of great importance to industry and on the other hand challenging for numerical solution because information is highly nonuniformly distributed in space and time. The goal is to impart the excitement and usefulness of the investigated adaptive approach as a tool in efficiently solving real–world problems.

II

The Continuous Problem and Its Discretization in Time

In this chapter, the nonlinear problem class of parabolic type which will be studied is introduced in a Hilbert space setting. The main assumptions are formulated in terms of sectorial operators which are negative infinitesimal generators of analytic semigroups. They are general enough to cover a huge class of important applications. Our setting is taken from Lubich and Ostermann [95], where a rigorous analysis of linearly implicit time discretizations applied to nonlinear parabolic equations is given. The assumptions are slightly extended to a family of operators $A(t,w(t))$, where $w(t)$ is varying in a neighbourhood of the solution. This allows later the study of spatial projection errors.

Some fundamental examples which fit into the chosen framework are described. We discuss the close connection to the theory of sesquilinear forms associated with elliptic operators. This approach provides sufficient conditions for our assumptions and therefore allows in general an easy check of them.

Working with resolvent bounds instead of Gårding's inequality yields a theoretical explanation of the noninteger temporal convergence orders observed for linearly implicit one–step methods [95]. We summarize known convergence results up to order three after reviewing the definition of Rosenbrock methods.

§1. Nonlinear Evolution Problem

We consider the nonlinear initial boundary value problem

$$\begin{aligned} \partial_t u(x,t) &= F(x,t,u(x,t)) && \text{in } \Omega\times(0,T], \\ B(x,t,u(x,t))u(x,t) &= g(x,t,u(x,t)) && \text{on } \partial\Omega\times(0,T], \\ u(x,0) &= u_0(x) && \text{on } \overline{\Omega}, \end{aligned} \tag{II.1}$$

where $\Omega\subset\mathbb{R}^d$, $d=1,2$ or 3, is a bounded open domain with smooth boundary $\partial\Omega$ lying locally on one side of Ω, and $T>0$. The boundary operator B stands for an appropriate system of boundary conditions and has to be interpreted in the sense of traces. The unknown u is allowed to be vector–valued.

These equations will be considered in a Hilbert space setting. Let

$$\mathcal{V}\hookrightarrow^{ds}\mathcal{H}\hookrightarrow^{ds}\mathcal{V}' \tag{II.2}$$

be a Gelfand triple of separable Hilbert spaces (see also Appendix A§1), where the antiduality between $\mathcal{V}$ and $\mathcal{V}'$ is denoted by $\langle\cdot,\cdot\rangle$. We introduce norms $\|\cdot\|$, $|\cdot|$ in $\mathcal{V}$ and $\mathcal{H}$, induced by the scalar products $((\cdot,\cdot))$ and $(\cdot,\cdot)$, respectively. The dual norm on $\mathcal{V}'$ is defined in the usual way by

$$\|\cdot\|_* = \sup_{v\in\mathcal{V},\|v\|=1} \langle\cdot,v\rangle. \tag{II.3}$$

We write (II.1) as an abstract Cauchy problem

$$\partial_t u = F(t,u(t)), \quad u(0)=u_0, \quad 0<t\le T, \tag{II.4}$$

and assume that this equation has a unique, temporally smooth solution $u(t)$ in $\mathcal{V}$. Here, we suppose that the mapping $F:(0,T]\times\mathcal{V}\to\mathcal{V}'$ is sufficiently differentiable. Setting

$$A(t,v) := -F_u(t,v(t)), \quad Q(t,v(t)) := F(t,v(t)) + A(t,v)v(t), \tag{II.5}$$

for all $v\in\mathcal{V}$, we derive from (II.4)

$$\partial_t u + A(t,u)u(t) = Q(t,u(t)), \quad u(0)=u_0, \quad 0<t\le T. \tag{II.6}$$

Equation (II.6) looks like a quasilinear Cauchy problem, but in general the nonlinear perturbation Q strongly depends on A making both of the same differential order, i.e., $Q(t,u)$ is only defined for $u\in\mathcal{V}$.

We will study this equation in the framework of analytic semigroups [110, 81]. Assume that we are given, for $t\ge 0$ and $w\in\mathcal{W}\subset\mathcal{V}$, a family of linear uniformly bounded sectorial operators $A(t,w):\mathcal{V}\to\mathcal{V}'$ having uniformly bounded inverses such that

$$\|A(t,w)\|_{\mathcal{L}(\mathcal{V},\mathcal{V}')} + \|A^{-1}(t,w)\|_{\mathcal{L}(\mathcal{V}',\mathcal{V})} \le C_A \tag{II.7}$$

with C_A independent of t and w. Since sectorial operators are negative infinitesimal generators of analytic semigroups (see Appendix A§4), the Cauchy problem (II.6) is said to be *parabolic*. If we assume, in addition, that a whole neighborhood of zero is contained in $\rho(-A)$, then there are time–independent constants $M>0$ and $\phi<\pi/2$ such that for all $w\in\mathcal{W}$ and for all complex λ with $|\arg(\lambda)|\le\pi-\phi$

$$\|(\lambda I + A(t,w))^{-1}\|_{\mathcal{L}(\mathcal{V})} \le \frac{M}{1+|\lambda|}. \tag{II.8}$$

This condition can be interpreted as an abstract ellipticity assumption on the operator A. It is very general and often satisfied by elliptic partial differential operators for different function spaces $\mathcal{V}$ and $\mathcal{W}$.

To illustrate this, we shall discuss sufficient conditions for (II.7) and (II.8) in terms of sesquilinear forms which are very common in the theory of elliptic operators. The operators $A(t,w) : \mathcal{V} \to \mathcal{V}'$ are associated with the sesquilinear form $a(t,w;v_1,v_2)$ on $\mathcal{V} \times \mathcal{V}$ defined by the identity

$$a(t,w;v_1,v_2) = \langle A(t,w)v_1, v_2 \rangle\,, \quad v_1, v_2 \in \mathcal{V}\,. \tag{II.9}$$

Let us assume the time– and solution–dependent sesquilinear form a to satisfy uniformly in $t \in [0,T]$ and $w \in \mathcal{W}$ the continuity condition

$$|a(t,w;v_1,v_2)| \le M_a \|v_1\| \|v_2\|\,, \quad v_1, v_2 \in \mathcal{V}\,, \tag{II.10}$$

and the Gårding–type inequality

$$a(t,w;v_1,v_1) \ge \mu_a \|v_1\|^2 - \kappa_0 |v_1|^2\,, \quad v_1 \in \mathcal{V}\,, \tag{II.11}$$

with constants M_a, $\mu_a > 0$, and κ_0 independent of $t \ge 0$, w, v_1, and v_2. Since we work with finite time intervals, we may consider $A + \kappa_0 I$ instead of A employing the classical Gårding transformation $\hat{u}(t) := e^{-\kappa_0 t} u(t)$. Therefore, we can assume $\mathcal{V}$–ellipticity

$$a(t,w;v_1,v_1) \ge \mu_a \|v_1\|^2\,, \quad v_1 \in \mathcal{V}\,, \tag{II.12}$$

without any loss of generality, hereafter.

The conditions (II.10) and (II.12) are strongly related to our assumptions (II.7) and (II.8). In fact, the continuity property is equivalent to the uniform boundedness of the operator A, i.e., $\|A(t,w)\|_{\mathcal{L}(\mathcal{V},\mathcal{V}')} \le M_a$. Lax–Milgram's theorem reveals further that A is an isomorphism of $\mathcal{V}$ onto $\mathcal{V}'$, showing A^{-1} is also uniformly bounded. We have $\|A^{-1}(t,w)\|_{\mathcal{L}(\mathcal{V}',\mathcal{V})} \le 1/\mu_a$, and therefore $C_A = M_a + 1/\mu_a$. Furthermore, the $\mathcal{V}$-ellipticity of the sesquilinear form a implies the resolvent bound (II.8) (see [139] §8, [110] §7.2, [90, 41]).

For our analysis including the study of discretization errors, it is suitable to impose the Lipschitz continuity of $t \to A(t,w(t))$ in the $\mathcal{L}(\mathcal{V},\mathcal{V}')$-norm. We assume that

$$\|A(t_2,w(t_2)) - A(t_1,w(t_1))\|_{\mathcal{L}(\mathcal{V},\mathcal{V}')} \le L\,|t_2 - t_1|, \quad t_1, t_2 \in [0,T]\,. \tag{II.13}$$

It is well–known that (II.13) and (II.8) with $u(t) \in \mathcal{W}$ guarantees the existence and uniqueness of the solution of the Cauchy problem (II.6) in the homogeneous case [110, 6].

We make the following regularity assumptions on the second derivatives of the nonlinear function $F : (0,T] \times \mathcal{V} \to \mathcal{V}'$:

$$\begin{aligned} \|F_{tu}(t,v)v_1\|_* &\le C\,\|v_1\| \quad \text{for all } v_1 \in \mathcal{V}, &\text{(II.14)}\\ \|F_{uu}(t,v)[v_1,v_2]\|_* &\le C\,\|v_1\|\|v_2\| \quad \text{for all } v_1, v_2 \in \mathcal{V}, &\text{(II.15)} \end{aligned}$$

with v varying in bounded subsets of $\mathcal{V}$.

Remark 1. In many practical situations, the operator $A(t,w)$ with domain $D(A(t,w))$ is given by a continuous sesquilinear form as unbounded operator in a Hilbert space $\mathcal{H}$. Supposing that $D(A^{1/2})=D(A^{*1/2})=\mathcal{V}\hookrightarrow\mathcal{H}$ where $\mathcal{V}$ does not depend on t and w, $A(t,w)$ can be extended to a bounded sesquilinear form on $\mathcal{V}\times\mathcal{V}$. Then Lax–Milgram's theorem [Theorem A.3] shows that $A(t,w)$ is an isomorphism from $\mathcal{V}$ to its conjugate linear dual $\mathcal{V}'$.

Example 1. *Linear Scalar Equations.* Let Ω be a bounded domain in $\mathbb{R}^d$ with a sufficiently smooth boundary $\partial\Omega=\Gamma_D\cup\Gamma_C$, where $\Gamma_D\cap\Gamma_C=\emptyset$. We consider a second–order differential operator $A(t)$ defined by

$$A(t)u := -\sum_{i,j}\partial_i(a_{ij}(\cdot,t)\partial_j u) + \sum_i a_i(\cdot,t)\partial_i u + a_0(\cdot,t)u$$

where $\partial_i=\partial/\partial x_i$, $a_{ij}=a_{ji}$, a_i, $a_0\in L^\infty(\Omega\times(0,T))$, and

$$\sum_{i,j} a_{ij}(x,t)\xi_i\xi_j \ge \alpha\,|\xi|^2 \quad \text{for all } (x,t)\in\Omega\times(0,T),\ \xi\in\mathbb{R}^d\setminus\{0\}, \tag{II.16}$$

with a constant $\alpha>0$ independent of x and t. We impose homogeneous boundary conditions, setting

$$0 = B(t)u := \begin{cases} u & \text{on } \Gamma_D\,, \\ \partial_{\nu_{A(t)}}u + b_0(\cdot,t)u & \text{on } \Gamma_C\,, \end{cases}$$

where $\partial_{\nu_{A(t)}} = \sum_{i,j} n_i a_{ij}\partial_j$ is the outer conormal with respect to the matrix $(a_{ij}(\cdot,t))$, and $b_0\in L^\infty(\Gamma_C\times(0,T))$. Finally, we assume that either $\Gamma_D\neq\emptyset$ or $a_0\neq 0$, if $b_0=0$. The inhomogeneous case can be handled as usual by transformation (cf. [43], p. 247).
We consider the operator $A(t)$ as unbounded on $\mathcal{H}=L^2(\Omega)$. We have

$$D(A(t)) = \{v\in H^2(\Omega)\ |\ B(t)u=0 \text{ on } \partial\Omega\}\,.$$

Let $\mathcal{V}=\{v\in H^1(\Omega)\ |\ v=0 \text{ on } \Gamma_D\}$. Then we can associate the operator $A(t)$ with the sesquilinear form

$$a(t;v,w) = \int_\Omega \left(\sum_{i,j} a_{ij}\partial_i v\partial_j w + \sum_i a_i\partial_i v w + a_0 v w\right)dx + \int_{\Gamma_C} b_0 v w\, d\sigma\,,$$

where $v,w\in\mathcal{V}$. Here, we recall that from the trace theorem (cf. [92]) follows the inequality

$$\|v\,|_{\partial\Omega}\,\|_{L^2(\partial\Omega)} \le \varepsilon\|v\|_{H^1(\Omega)} + c_\varepsilon\|v\|_{L^2(\Omega)}\,,\quad v\in H^1(\Omega)\,,$$

for each $\varepsilon > 0$ with a constant $c_\varepsilon > 0$. The uniform ellipticity (II.16) and the boundedness of all coefficients imply the continuity of $a(t;\cdot,\cdot)$ on $\mathcal{V} \times \mathcal{V}$. Straightforward calculation shows also that for each $\mu_a > 0$ there exists a $\kappa_0 \in \mathbb{R}$ such that (II.11) is valid. Therefore, the operators $A_\kappa = A + \kappa I$, where $\kappa \geq \kappa_0$, satisfy conditions (II.10) and (II.12) showing that they fit into our framework of analytic semigroups. Including C^1–regularity in t of all problem coefficients, we get (II.13) and (II.14) without any difficulties.
Now a standard linear parabolic problem can be defined in a weak formulation as follows: find $u \in C^0([0,T];\mathcal{H}) \cap L^2(0,T;\mathcal{V})$ such that

$$\begin{aligned} \langle \partial_t u, v \rangle + \langle A(t)u, v \rangle &= \langle f(\cdot,t), v \rangle \quad \forall v \in \mathcal{V}, \\ (u(0),\phi) &= (u_0,\phi) \quad \forall \phi \in \mathcal{H}, \end{aligned}$$

where $f \in L^2(0,T;\mathcal{V}')$ and $u_0 \in L^2(\Omega)$.

Example 2. *Systems of Quasilinear Equations* [4, 5, 51]. For the applications we have in mind, we discuss shortly a system of second order partial differential equations of quasilinear form

$$\begin{aligned} \partial_t u + \mathcal{A}(\cdot,u)u &= f(\cdot,u) \quad \text{in } \Omega \times (0,T], \\ B(\cdot,u)u &= g(\cdot,u) \quad \text{on } \partial\Omega \times (0,T], \\ u(0) &= u_0 \quad \text{on } \overline{\Omega}, \end{aligned} \tag{II.17}$$

where

$$\mathcal{A}(\cdot,u)v := -\sum_{i,j} \partial_i(a_{ij}(\cdot,u)\partial_j v) + \sum_i a_i(\cdot,u)\partial_i v$$

and

$$B(\cdot,u)v := \delta\,(\partial_{\nu_\mathcal{A}} v + b_0(\cdot,u)v) + (1-\delta)v$$

acting on $\mathbb{R}^N$–valued functions $u = (u_1,\ldots,u_N)^T$. The functions f and g are N–vectors. The $(N \times N)$–coefficient matrices

$$a_{ij} = \left(a_{ij}^{rs}(x,u)\right)_{1 \leq r,s \leq N}, \; a_i = (a_i^{rs}(x,u))_{1 \leq r,s \leq N}, 1 \leq i,j \leq d,$$

are assumed to be measurable in $L^\infty(\Omega \times \mathbb{R}^N, \mathcal{L}(\mathbb{R}^N))$ such that a uniform Legendre condition is satisfied, i.e.,

$$\sum_{r,s,i,j} a_{ij}^{rs}(x,v)\,\xi_r^j \xi_s^i > 0 \quad \text{for all } (x,v) \in \Omega \times \mathbb{R}^N, \xi \in \mathbb{R}^{dN}.$$

Moreover, we suppose that their dependencies on the solution vector u are smooth.
To describe the boundary conditions, we let

$$\delta := \operatorname{diag}(\delta_1, \ldots, \delta_N) : \partial\Omega \to \mathcal{L}(\mathbb{R}^N)$$

such that $\delta_r \in C^0(\partial\Omega, \{0,1\}), 1 \le r \le N$. Consequently, $\delta_r = 0$ recovers Dirichlet boundary conditions, whereas $\delta_r = 1$ leads to mixed boundary conditions.
We have $F(t,u) = -\mathcal{A}(u)u + f(u)$, and put

$$A(u)v = \mathcal{A}(u)v + \partial_u \mathcal{A}(u)[v,u] - \partial_u f(u)v$$

equipped with the boundary conditions defined by $B(u)$.
The well–posedness of system (II.17) has been investigated by AMANN [5] in terms of analytic semigroups. Using the so–called theory of extrapolation spaces, the nonlinearity $g(x,t,u)$ can be handled separately.
Due to the great complexity of these problems it is not always possible to find an appropriate set of Hilbert spaces $\mathcal{V} \hookrightarrow \mathcal{H} \hookrightarrow \mathcal{V}'$ independent of t such that our assumptions are satisfied. The situation is often more favourable for linear Dirichlet boundary conditions, for which the operator B is strongly simplified.
Let us consider problem (II.17) with $a_i = 0$, $g = 0$, and $\delta_r = 0$ for all components. We set $\mathcal{H} = [L^2(\Omega)]^N$, $\mathcal{V} = [H^1_0(\Omega)]^N$, and thus $\mathcal{V}' = [H^{-1}(\Omega)]^N$. The weak formulation of the operator $A(w)v$ is given by

$$\begin{aligned}\langle A(w)v, \psi \rangle \;&=\; \int_\Omega \sum_{i,j} a_{ij}(x,w)\partial_i v \partial_j \psi + \sum_{i,j} \partial_u a_{ij}(x,w) v \partial_i w \partial_j \psi \\ &\qquad -\partial_u f(w) v \psi \, dx.\end{aligned}$$

If $d = 1$, then we have Sobolev's embedding $H^1_0(\Omega) \hookrightarrow C^0(\overline{\Omega})$. It is a well–known fact that the solutions of parabolic evolution equations have even better regularity properties if the data are more regular. For sufficient regularity, we can suppose $u \in C^1_t(\mathcal{V})$ (see e.g. [110], Theorem 3.1). Setting $\mathcal{W} = \{w \in C^1_t(\mathcal{V}) : \|w - u\| \le s,\ s > 0 \text{ fixed}\}$, the operators $A(w)$ make sense as linear operators from $\mathcal{V}$ to $\mathcal{V}'$. Providing suitable hypotheses on the regularity and boundedness of the coefficients, the nonlinearity f, and their derivatives, we are able to derive estimates (II.7), (II.8), and (II.13). The constants depend on $\max_{0\le t\le T}\|u(t)\|$, s, and corresponding bounds of the data. In particular, (II.13) follows from

$$\|A(w(t_2)) - A(w(t_1))\|_{\mathcal{L}(\mathcal{V},\mathcal{V}')} \le C_1 \|w(t_2) - w(t_1)\| \le C\,|t_2 - t_1|$$

where Lipschitz constants of a_{ij}, $\partial_u a_{ij}$, and $\partial_u f$ with respect to u determine the value of C_1. The second inequality is a direct consequence of $w \in C^1_t(\mathcal{V})$. Clearly, conditions (II.14) and (II.15) can also be satisfied.
The above choice of $\mathcal{H}$ and $\mathcal{V}$ is no longer possible for higher spatial dimensions. However, we can consider the operator $A(w)$ as unbounded on $\mathcal{H} = [H^s_0(\Omega)]^N$,

with $0<s<1$ for $d=2$, and $1/2<s<1$ for $d=3$. Setting then $\mathcal{V}=[H^{s+1}(\Omega)\cap H_0^1(\Omega)]^N$, the embedding $\mathcal{V}\hookrightarrow[C^0(\overline{\Omega})]^N$ is still valid. The same comment applies to operators including first–order terms and the time variable t.
The situation is completely changed if Neumann boundary conditions have to be imposed. Because the conormal derivative may depend on the solution a variable domain of $A(w)$ may result (cf. Example 3). Thus, it is not always possible to define an appropriate set of spaces $\mathcal{V}$ and $\mathcal{H}$ (see [94] for a more thorough discussion).
If the coefficients a_{ij} do not depend on the solution, we get the famous class of semilinear equations. These problems have been extensively studied in the monograph by HENRY [74] (see also [110], Chap. 6), which has been the basis for numerous investigations of reaction–diffusion equations. There can be found a rich collection of examples. Surprisingly, the Navier–Stokes equations fit also in our framework utilizing suitable divergence–free function spaces ([62], [74], Example 3.8). In LUBICH and OSTERMANN ([95], Appendix), appropriate spaces and norms for stiff reaction–diffusion equations are derived to ensure that the constants in the assumptions are independent of the possibly very large reaction coefficients.

§2. Rosenbrock Methods and Basic Results

We are interested in approximating the nonlinear Cauchy problem (II.4) in time. For this, we use linearly implicit one–step methods proposed by ROSENBROCK [118] to achieve higher order methods for stiff problems by working the Jacobian matrix into the integration formula. Applied to the initial–value problem (II.4) with step size $\tau>0$ a so–called s–stage Rosenbrock method has the recursive form

$$\begin{aligned} u_{n+1} &= u_n+\tau\sum_{i=1}^{s} b_i K'_{ni}\,, \\ K'_{ni} &= F(t_n+\alpha_i\tau, K_{ni})-\tau A(t_n,u_n)\sum_{j=1}^{i}\gamma_{ij}K'_{nj} \\ &\quad + \tau\gamma_i F_t(t_n,u_n), \end{aligned} \tag{II.18}$$

with the intermediate values

$$K_{ni}=u_n+\tau\sum_{j=1}^{i-1}\alpha_{ij}K'_{nj}\,,\quad 1\le i\le s, \tag{II.19}$$

and

$$\alpha_i=\sum_{j=1}^{i-1}\alpha_{ij},\quad \gamma_i=\sum_{j=1}^{i}\gamma_{ij}\,.$$

Here, u_n denotes an approximation of $u(t_n)$ at $t_n = n\tau$. The coefficients b_i, α_{ij}, and γ_{ij} are suitable chosen to obtain a desired order of consistency and stability for stiff problems (see also Appendix B§1). We always assume that $\gamma_{ii} = \gamma > 0$, and $\alpha_i \in [0,1]$ for all i.
It was the fundamental idea of ROSENBROCK that only linear systems with the operators $I + \tau\gamma_{ii}A(t_n, u_n)$ have to be solved successively one after the other. An iterative Newton method as known from implicit Runge–Kutta methods is no longer required.
For convenience, we set $\alpha_{ij} = 0$ for $j \geq i$, $\gamma_{ij} = 0$ for $j > i$, and use the notation

$$\beta_{ij} = \alpha_{ij} + \gamma_{ij}, \quad c_i = \alpha_i + \gamma_i, \quad \mathcal{B} = (\beta_{ij})_{i,j=1}^{s} \ ,$$

$$b = (b_1, \ldots, b_s)^T, \quad \alpha^k = (\alpha_1^k, \ldots, \alpha_s^k)^T, \quad \mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^s \ .$$

The Rosenbrock method applied to ordinary differential equations (ODEs) with sufficiently differentiable right–hand side has (classical) order p if the global error satisfies

$$\epsilon_n := u_n - u(t_n) = O(\tau^p) \quad \text{ as } \tau \to 0,$$

uniformly on bounded time intervals. The method is called $\mathcal{A}(\Theta)$*–stable*, if its stability function

$$R(z) = 1 + zb^T(I - z\mathcal{B})^{-1}\mathbf{1}$$

is bounded in modulus by 1 for $|\arg(z)| \geq \pi - \Theta$. If additionally the absolute limit of the stability function at infinity $|R(\infty)|$ is strictly smaller than 1, we call the method *strongly* $\mathcal{A}(\Theta)$*–stable*.

Convergence results for one–step methods of Rosenbrock type applied to partial differential equations were derived in [108, 95]. It turns out that the classical order of ODE–integrators can in general not be achieved. This phenomenon is known as order reduction and has been first investigated for Runge–Kutta schemes by several authors [124, 123, 107, 137]. Nowadays it is much better understood than before why (lower) fractional orders occur. This reduction is not induced by lack of smoothness of the solution $u(t)$ but rather by the presence of powers of the operators $A(t)$ in the local truncation error.

For Rosenbrock methods of order $p \geq 3$ which are strongly $\mathcal{A}(\Theta)$–stable with $\Theta > \phi$, the following result has been established by LUBICH and OSTERMANN ([95], Theorem 4.3). If $u(t) \in \mathcal{W}$, then the error bound

$$\left(\tau \sum_{n=0}^{N} \|\, \epsilon_n \|^2 \right)^{1/2} + \max_{0 \leq n \leq N} |\epsilon_n| \ \leq \ C\tau^{2+\beta} \, \|A^\beta \partial_t^2 u(t)\|_{L_t^2(\mathcal{V})} \tag{II.20}$$

holds for $N\tau \leq T$ and $\beta \in [0,1]$ such that the range of $A(t,u(t))^{-\beta} : \mathcal{V} \to \mathcal{V}$ is independent of t. This result shows that the temporal order of convergence is influenced by spatial regularity. The attainable value of β depends on the domains of the fractional powers of $A(t)$. Note there is no order reduction for $p \leq 2$.

Example 3. For second–order strongly elliptic differential operators taken as unbounded in $\mathcal{H} = L^2(\Omega)$ corresponding values of β can be found in [63, 68], see also the discussion in [94]. We recall some results for $\mathcal{V} = D(A^{1/2}(t))$, taking up $A(t)$ from Example 1. For homogeneous Dirichlet boundary conditions, we have $D(A^{\alpha}(t)) = H^{2\alpha}(\Omega) \cap H_0^1(\Omega)$, $1/2 \leq \alpha < 5/4$. Considering the condition $A^{\beta}\partial_t^2 u(t) \in \mathcal{V}$ or equivalently $\partial_t^2 u(t) \in D(A^{1/2+\beta}(t))$, we get $\beta = 3/4 - \varepsilon$ with arbitrary small $\varepsilon > 0$. In the case of homogeneous Neumann boundary conditions, $D(A^{\alpha}(t)) = H^{2\alpha}(\Omega)$ for $\alpha < 3/4$. Thus, $\beta = 1/4 - \varepsilon$ can be used. The better value $\beta = 5/4 - \varepsilon$ is obtained if the boundary conditions are time–independent. Otherwise, $\beta = 1/4 - \varepsilon$ in general for inhomogeneous Neumann boundary conditions. Finally, order reduction is more severe for nonhomogeneous Dirichlet boundary conditions due to $\partial_t^2 u(t) \in D(A^{\alpha}(t))$ with $\alpha < 1/4$.

Fortunately, there are conditions which imply also a higher order of convergence. To obtain full order 3 independent of the spatial regularity we have to fulfill in the above situation

$$b^T \mathcal{B}^j (2\mathcal{B}^2 \mathbf{1} - \alpha^2) = 0 \quad \text{for } p-2 \leq j \leq s-1. \tag{II.21}$$

These conditions were found to be necessary to improve the order for the linear time–invariant case in [108], and yield also improved order for the nonlinear case [95]. Analogous conditions for the stiff ODE case related to B–convergence properties were announced previously in [136, 127].

Remark 2. At a first glance, it should be attractive to use an approximate Jacobian to reduce computational costs and to be more robust with respect to perturbations caused by spatial discretization errors. Unfortunately, there are some well–known disadvantages: significant increase of order conditions (see W–methods [134]), lack of conservation properties in general, loss of accuracy long before stability is affected, and more severe order reduction compared with Rosenbrock methods [95]. Nevertheless, it seems to be sometimes worthwhile to apply lower–order Rosenbrock methods that also satisfy the conditions of W–methods [148].

III

Convergence of the Discretization in Time and Space

We study spatial approximations of Rosenbrock schemes by means of finite elements coupled with a Galerkin method. The error is evaluated in a discrete $L^2_t(\mathcal{V})\cap C^0_t(\mathcal{H})$–norm under the usual assumption that the solution is temporally smooth. Since we are mainly interested in studying Rosenbrock methods of order $p\geq 2$, we need $H^q_t(\mathcal{V})$–regularity with $q\geq 3$. The parabolic nature of our equations often yields smooth solutions, at least after an initial transitional phase. The obtained convergence results show a natural separation of temporal and spatial error terms, which simplifies their control in an adaptive solution process. Keeping the spatial discretization error below a prescribed tolerance would nearly result in a time integration procedure similar to the unperturbed case. Variable step sizes are also allowed, but the relation between them must remain bounded (quasiuniform meshes).

In a first step we put our parabolic Cauchy problem into a discrete framework, introducing a new Gelfand triple $(\mathcal{V}_h,\mathcal{H}_h,\mathcal{V}'_h)$. An analogous approach was used by Savaré [125] who investigated fully discretized $A(\Theta)$–stable multistep methods for linear problems. Unlike higher regularity assumptions in time, we are working with minimal spatial regularity conditions. The finite element subspaces are supposed to possess only the approximation property without fixing the order. The technique of quasi–interpolation can be used to construct quasi–optimal projection operators $\Pi_h:\mathcal{V}\to\mathcal{V}_h$ that are bounded uniformly in the meshwidth h.

To derive error estimates of our fully discretized scheme, we make use of the splitting

$$u-u_{h,n}=(u-\Pi_h u)+(\Pi_h u-u_{h,n})\,,$$

where $u_{h,n}\in\mathcal{V}_h$ is an approximation of the solution $u(t)$ at $t=t_n$. The first term represents the spatial projection error, while the second can be treated in terms of spatial truncation errors $d_h(t)\in\mathcal{V}'_h$ defined by the perturbed PDE

$$P_h\partial_t\Pi_h u=P_hF(t,\Pi_h u)+d_h(t)\,,$$

where $P_h:\mathcal{V}'\to\mathcal{V}'_h$ denotes a restriction operator. This technique was carried out in Lubich and Ostermann [94] for Runge–Kutta approximations of quasilinear parabolic equations. We extend their result to Rosenbrock schemes in Lemma III.2. It remains to estimate the spatial truncation error and its temporal derivative at certain time points. The corresponding result is given in Lemma III.3. We use a specially defined interpolation operator in time to handle summation over all intermediate integration points. Here, Lemma III.1 which is based on a related result by SAVARÉ [125] comes into the play.

§1. Finite Element Discretization in Space

To get a fully discretized scheme we introduce a family of finite element subspaces $\{\mathcal{V}_h\}_{h\in(0,1)}$ of $\mathcal{V}$ with the approximation property

$$\inf_{\psi\in\mathcal{V}_h} \|v-\psi\| \;\longrightarrow\; 0 \quad \text{for } h\to 0, \tag{III.1}$$

for all $v\in\mathcal{V}$. In general there is no difficulty in constructing such approximation spaces [39, 35]. If Ω has a curved boundary special techniques have been developed, as the use of isoparametric finite elements [91].

We construct a new Gelfand triple denoting by $\mathcal{V}_h'$ the antidual of $\mathcal{V}_h$, and introducing $\mathcal{H}_h$ as closure of $\mathcal{V}_h$ in the $\mathcal{H}$-norm. Thus, we take the $\mathcal{V}$- and $\mathcal{H}$-norm to measure functions in $\mathcal{V}_h$ and $\mathcal{H}_h$, respectively. The duality pairing is denoted by $\langle\cdot,\cdot\rangle_h$. Let $P_h\in\mathcal{L}(\mathcal{V}',\mathcal{V}_h')$ be a restriction operator defined by

$$\langle P_h w,\psi\rangle_h \;=\; \langle w,\psi\rangle \qquad \text{for all } \psi\in\mathcal{V}_h, w\in\mathcal{V}', \tag{III.2}$$

which directly implies the contraction property

$$\|P_h w\|_{\mathcal{V}_h'} \le \|w\|_* \quad \text{for all } w\in\mathcal{V}'\,. \tag{III.3}$$

It can be established easily that $\partial_t P_h = P_h\partial_t$, using (III.2) and properties of sesquilinear forms.

Let $\Pi_h:\mathcal{V}\to\mathcal{V}_h$ be a projection operator such that for all $v\in\mathcal{V}$

$$\|\Pi_h v\| \;\le\; C\,\|v\|\,, \tag{III.4}$$

$$\|v-\Pi_h v\| \;\le\; C\inf_{\psi\in\mathcal{V}_h}\|v-\psi\|\,, \tag{III.5}$$

with constants C independent of v and h. That means, $\Pi_h v$ is quasi–optimal with respect to the best approximation. Assuming the usual property $\Pi_h\partial_t = \partial_t\Pi_h$, this implies that for a function $u\in H_t^3(\mathcal{V})\hookrightarrow C_t^2(\mathcal{V})$

$$\|u-\Pi_h u\|_{H_t^2(\mathcal{V})} \;\longrightarrow\; 0 \quad \text{for } h\to 0\,. \tag{III.6}$$

It should be useful to consider an example demonstrating that our assumptions can be satisfied in general situations.

Example 1. Let Ω be a bounded, open, connected domain in $\mathbb{R}^d$, $d \geq 2$, with polyhedral Lipschitz continuous boundary $\partial\Omega$, for simplicity. We consider a simplicial subdivision $\mathcal{T}_h$ of Ω such that any two simplicies share a complete smooth submanifold of their boundaries if they are not disjoint. For each d-simplex $K \in \mathcal{T}_h$, let h_K be the diameter of K, ρ_K be the radius of the largest closed ball contained in $\overline{K}$, and set $h := \max_{K\in\mathcal{T}_h} h_K$. Clearly, the spatial variables can be normalized such that $h \in (0,1)$. We assume $\max_{K\in\mathcal{T}_h} h_K/\rho_K$ to be bounded from above independently of h. This guarantees shape regularity of the subdivision and allows locally refined meshes.
Let $H_0^1(\Omega) \subset \mathcal{V} \subset H^1(\Omega)$, $H = L^2(\Omega)$, which represents our standard case. We set

$$\mathcal{V}_h := \{v \in C^0(\Omega) \,|\, v|_K \in P_m \text{ for all } K \in \mathcal{T}_h\,, m \geq 1\}$$

consisting of all continuous piecewise polynomials of degree m or less. It is well–known that $\mathcal{V}_h$ fulfills the approximation property (III.1) whenever $v \in H^l(\Omega), l > 1$.
An appropriate way to define a projection operator $\Pi_h : \mathcal{V} \to \mathcal{V}_h$ under minimal regularity conditions (note $\mathcal{V} \not\hookrightarrow C^0(\Omega)$ for $d \geq 2$) is the use of quasi–interpolation operators [128, 40]. We follow an idea by SCOTT and ZHANG [128] and define

$$\Pi_h v(x) = \sum_{i=1}^{dim(\mathcal{V}_h)} \phi_i(x) \int_{\sigma_i} \psi_i(\xi) v(\xi)\, d\xi\,,$$

where $\{\phi_i\}$ denotes the nodal basis in $\mathcal{V}_h$, σ_i is either a d– or $(d-1)$–simplex according to the type of the node i, and $\{\psi_i\}$ is a dual basis satisfying

$$\int_{\sigma_i} \psi_i(\xi)\phi_j(\xi)\, d\xi = \delta_{ij}\,, i,j = 1,\ldots, dim(\mathcal{V}_h)\,,$$

where δ_{ij} denotes the Kronecker symbol. One simply proves $\Pi_h v = v$ for all $v \in \mathcal{V}_h$ showing that Π_h is a projection. Now our assumptions (III.4) and (III.5) follow directly from Theorem 4.1 and Corollary 4.1 of [128], where more details can be found too. The operator Π_h can be extended to the interval $(0,T]$, replacing $v(\cdot)$ by $v(\cdot,t)$. Obviously, ∂_t and Π_h commute.

For $t \in [0,T]$ and $v \in \mathcal{V}$, we define operators A_h and F_h mapping from $\mathcal{V}$ to $\mathcal{V}_h'$ by

$$A_h(t,v) := P_h A(t,v)\,, \quad F_h(t,v) := P_h F(t,v). \tag{III.7}$$

It is a direct consequence of the contraction property (III.3) that the family of linear operators $A_h(t,w) : \mathcal{V} \to \mathcal{V}_h'$ remains uniformly bounded

$$\|A_h(t,w)v\|_{\mathcal{V}_h'} \leq \|A(t,w)v\|_* \leq C_A \|v\| \,,\; v\in\mathcal{V},\, w\in\mathcal{W}\,. \tag{III.8}$$

In order to guarantee the solvability of the discrete equations, we will assume that there exist uniformly bounded inverses of $A_h(t,w)$ such that altogether

$$\|A_h(t,w)\|_{\mathcal{L}(\mathcal{V},\mathcal{V}_h')} + \|A_h^{-1}(t,w)\|_{\mathcal{L}(\mathcal{V}_h',\mathcal{V})} \leq C \tag{III.9}$$

with a constant C independent of t, w, and h. We further suppose the discrete version of (II.8), i.e., there are time– and mesh–independent constants $M'\geq 0$ and $\phi'<\pi/2$ such that for all complex λ with $|\arg(\lambda)|\leq\pi-\phi'$ and for all $w\in\mathcal{W}$

$$\|(\lambda I + A_h(t,w))^{-1}\|_{\mathcal{L}(\mathcal{V}_h)} \leq \frac{M'}{1+|\lambda|}\,. \tag{III.10}$$

The classical way to derive such resolvent bounds for finite element applications is to start from the $\mathcal{V}_h$–ellipticity of the operator A_h which directly follows from the $\mathcal{V}$–ellipticity of A whenever $\mathcal{V}_h\subset\mathcal{V}$ ([139],§8, see also [90, 41]). In this case the constants remain the same, $M'=M$ and $\phi'=\phi$.

Using once again (III.3), we get directly

$$\|A_h(t_2,w(t_2)) - A_h(t_1,w(t_1))\|_{\mathcal{L}(\mathcal{V},\mathcal{V}_h')} \leq L\,|t_2-t_1| \tag{III.11}$$

for all $w\in\mathcal{W}$ and $t_1,t_2\in[0,T]$, and

$$\|P_hF_{tu}(t,v)v_1,\|_{\mathcal{V}_h'} \leq C\,\|v_1\|\,,\; v_1\in\mathcal{V}\,, \tag{III.12}$$

$$\|P_hF_{uu}(t,v)[v_1,v_2]\|_{\mathcal{V}_h'} \leq C\,\|v_1\|\|v_2\|\,,\; v_1,v_2\in\mathcal{V}\,, \tag{III.13}$$

with v varying in bounded subsets of $\mathcal{V}$. That means, for every $r>0$ there exist constants $C=C(r)$ such that

$$\|P_hF_t(t,v_1) - P_hF_t(t,v_2)\|_{\mathcal{V}_h'} \leq C\,\|v_1-v_2\|\,, \tag{III.14}$$

$$\|P_hF_u(t,v_1) - P_hF_u(t,v_2)\|_{\mathcal{L}(\mathcal{V},\mathcal{V}_h')} \leq C\,\|v_1-v_2\|\,, \tag{III.15}$$

for all $v_1,v_2\in\mathcal{V}$ with $\|v_1\|+\|v_2\|\leq r$.

Now we turn to the finite element approximation of our Rosenbrock scheme. Employing the standard Galerkin principle, the fully discretized problem to (II.18)-(II.19) consists in the sequence of linear equations in the unknowns $K_{h,ni}'\in\mathcal{V}_h$, $u_{h,n}\in\mathcal{V}_h$

$$\left\langle K'_{h,ni}, \psi \right\rangle = \left\langle F(t_n+\alpha_i\tau, K_{h,ni}) - \tau A(t_n, u_{h,n}) \sum_{j=1}^{i} \gamma_{ij} K'_{h,nj} \right. \tag{III.16}$$
$$\left. + \tau\gamma_i F_t(t_n, u_{h,n}), \psi \right\rangle \quad \text{for all } \psi \in \mathcal{V}_h ,$$

or equivalently

$$P_h K'_{h,ni} = F_h(t_n+\alpha_i\tau, K_{h,ni}) - \tau A_h(t_n, u_{h,n}) \sum_{j=1}^{i} \gamma_{ij} K'_{h,nj} + \tau\gamma_i P_h F_t(t_n, u_{h,n}) ,$$

taken as an equation in $\mathcal{V}'_h$ with the intermediate values

$$K_{h,ni} = u_{h,n} + \tau \sum_{j=1}^{i-1} \alpha_{ij} K'_{h,nj} , \quad 1 \le i \le s. \tag{III.17}$$

Choosing an initial value $u_{h,0} = \Pi_h u_0 \in \mathcal{V}_h$ we construct an approximation $u_{h,n+1}$ of the solution $u(t_{n+1})$ by the summation

$$u_{h,n+1} = u_{h,n} + \tau \sum_{i=1}^{s} b_i K'_{h,ni} . \tag{III.18}$$

Let us define the spatial projection error on the initial values by

$$\epsilon_{h0}[u] = \sqrt{\tau}\, \|u(0) - \Pi_h u(0)\| + |u(0) - \Pi_h u(0)| . \tag{III.19}$$

Then we have the following convergence results.

Theorem 1. *Let the assumptions (III.9)–(III.15) be satisfied with $\sigma u + (1-\sigma)\Pi_h u \in \mathcal{W}$ for all $\sigma \in [0,1]$. The Rosenbrock method used has order $p \ge 2$ and is strongly $A(\Theta)$-stable with $\Theta > \phi'$. Let $u \in H_t^3(\mathcal{V})$. Then the error of the fully discretized equations (III.16)-(III.18) is bounded for sufficiently small τ and $N\tau \le T$ by*

$$\left(\tau \sum_{n=0}^{N} \| u(t_n) - u_{h,n} \|^2 \right)^{1/2} + \max_{0 \le n \le N} | u(t_n) - u_{h,n} |$$
$$\le \epsilon_{h0}[u] + C \left(\tau^2 + \|u - \Pi_h u\|_{H_t^2(\mathcal{V})} \right) .$$

The constant C depends on the constants in (III.9)–(III.15), on the $H_t^3(\mathcal{V})$–norm of the solution, on the coefficients of the Rosenbrock method, and on T.

Theorem 2. *Let the assumptions (III.9)–(III.15) be satisfied with* $\sigma u + (1-\sigma)\Pi_h u \in \mathcal{W}$ *for all* $\sigma \in [0,1]$. *The Rosenbrock method used has order* $p \geq 3$, *is strongly* $A(\Theta)$*-stable with* $\Theta > \phi'$, *and satisfies the condition (II.21). Let* $u \in H_t^4(\mathcal{V})$. *Then the error of the fully discretized equations (III.16)-(III.18) is bounded for sufficiently small* τ *and* $N\tau \leq T$ *by*

$$\left(\tau \sum_{n=0}^{N} \| u(t_n) - u_{h,n}\|^2\right)^{1/2} + \max_{0\leq n\leq N} | u(t_n) - u_{h,n}|$$
$$\leq \epsilon_{h0}[u] + C\left(\tau^3 + \|u - \Pi_h u\|_{H_t^2(\mathcal{V})}\right) .$$

The constant C *depends on the constants in (III.9)–(III.15), on the* $H_t^4(\mathcal{V})$*–norm of the solution, on the coefficients of the Rosenbrock method, and on* T.

Let us mention that there is a natural way to extend the above results to variable step sizes in time ([94], Theorem 5.1). The theorems remain valid for sequences $\{\tau_n\}$ satisfying

$$\sum_{n=1}^{N} \left| \frac{\tau_{n+1}}{\tau_n} - 1 \right| \leq C_1\,, \quad C_2\tau \leq \tau_n \leq \tau\,, \quad 1 \leq n \leq N\,, \tag{III.20}$$

with a fixed $\tau > 0$ and $C_1, C_2 > 0$. In [94] the authors argue that the whole interval $[0,T]$ can be subdivided into appropriate subintervals on which step sizes of different scales can be used. Therefore, (III.20) is no severe restriction in practice. The recent papers [14, 109] which are concerned to the approximation of holomorphic semigroups by variable stepsize rational functions are also of interest.

Remark 1. Theorem 2 has a straightforward extension to the case of Rosenbrock methods of order $p \geq 3$ which do not satisfy condition (II.21). In this situation, the attainable order of convergence in general is fractional and corresponds to the value stated in (II.20).

§2. Proof of the Convergence Results

We first introduce a preliminary definition to handle summation in time. Let $l_N^2(\mathcal{V})$ the space of $\mathcal{V}$–valued vectors $\underline{v} = \{v_n\}_{n=0,\ldots,N-1}$ equipped with the norm

$$\|\underline{v}\|_{l_N^2(\mathcal{V})} = \left(\tau \sum_{n=0}^{N-1} \|v_n\|_{\mathcal{V}}^2\right)^{1/2} .$$

We define an interpolation operator in time

$$\Pi_t^\alpha : H_t^q(\mathcal{V}) \to l_N^2(\mathcal{V})\,, \quad \alpha \in [0,1]\,,\ q \geq 1\,,$$

by

$$\Pi_t^\alpha v = \{v(t_j + \alpha\tau)\}_{j=0,\dots,N-1}\,. \tag{III.21}$$

Π_t^α interpolates at one fixed intermediate integration point in time. The following result is a modification of [125], Lemma 3.1 and Corollary 3.2. The proof of our result is based on ideas given in [126].

Lemma 1. *There exists a constant $C>0$ such that for all $v \in H_t^q(\mathcal{V})$, $q \geq 1$,*

$$\|\Pi_t^\alpha(v-\Pi_h v)\|_{l_N^2(\mathcal{V})} \leq C\left(\tau^q\|\partial_t^q(v-\Pi_h v)\|_{L_t^2(\mathcal{V})} + \|v-\Pi_h v\|_{L_t^2(\mathcal{V})}\right). \tag{III.22}$$

Proof. First let $v \in H^q(0,1;\mathcal{V})$ with $q \geq 1$. Then v can be identified with a continuous function lying in $C^0([0,T];\mathcal{V})$ (see Appendix A§5). That means, there is a constant C for which the inequality

$$\|v(\alpha)\| \leq C\,\|v\|_{H^q(0,1;\mathcal{V})}$$

is satisfied for arbitrary $\alpha \in [0,1]$. For the well–known equivalence result for Sobolev spaces of EHRLING, NIRENBERG, and GAGLIARDO ([1], Theorem 4.14, see also Corollary 4.10), we get further

$$\|v(\alpha)\| \leq C\,\left(\,\|\partial_t^q v\|_{L^2(0,1;\mathcal{V})} + \|v\|_{L^2(0,1;\mathcal{V})}\,\right), \quad 0\leq\alpha\leq 1\,. \tag{III.23}$$

We consider now $v \in H^q(0,\tau;\mathcal{V})$ and define a function $v_\tau \in H^q(0,1;\mathcal{V})$ by

$$v_\tau(t) := v(\tau t), \quad t \in [0,1]\,.$$

Using (III.23) and the transformation $t^* = \tau t$, we deduce

$$\begin{aligned}
&\tau\,\|v(\alpha\tau)\|^2 = \tau\,\|v_\tau(\alpha)\|^2 \\
\leq\;& C\tau\int_0^1\left(\|\partial_t^q v_\tau(t)\|^2 + \|v_\tau(t)\|^2\right)dt = C\int_0^\tau\left(\tau^{2q}\|\partial_t^q v(t^*)\|^2 + \|v(t^*)\|^2\right)dt^* \\
\leq\;& C\left(\,\tau^{2q}\,\|\partial_t^q v\|^2_{L^2(0,\tau;\mathcal{V})} + \|v\|^2_{L^2(0,\tau;\mathcal{V})}\,\right)
\end{aligned}$$

If $v \in H_t^q(\mathcal{V})$, then the summation over all intervalls $[t_n,t_{n+1}], n=0,\dots,N-1$, leads to

$$\tau\sum_{n=0}^{N-1}\|v(t_n+\alpha\tau)\|^2 \leq C\left(\tau^{2q}\|\partial_t^q v\|^2_{L_t^2(\mathcal{V})} + \|v\|^2_{L_t^2(\mathcal{V})}\right)\,.$$

Setting $v := v - \Pi_h v \in H_t^q(\mathcal{V})$, we get the desired result. □

We next derive some preliminary equations to be used in our further analysis. Applying the restriction operator P_h to (II.4) yields

$$P_h \partial_t u(t) = F_h(t, u(t)) \qquad \text{in } \mathcal{V}_h' . \tag{III.24}$$

The projected solution $\Pi_h u$ satisfies the perturbed equation

$$P_h \partial_t \Pi_h u(t) = F_h(t, \Pi_h u(t)) + d_h(t) \qquad \text{in } \mathcal{V}_h' , \tag{III.25}$$

where $d_h(t) \in \mathcal{V}_h'$ is the spatial truncation error. We equip this equation with the approximate initial value

$$\Pi_h u(0) = \Pi_h u_0 \in \mathcal{V}_h . \tag{III.26}$$

Differentiating (II.4) and restricting afterwards gives

$$P_h \partial_{tt} u(t) = P_h F_t(t, u(t)) - A_h(t, u(t)) u(t) \qquad \text{in } \mathcal{V}_h' . \tag{III.27}$$

We insert $\Pi_h u(t)$ and define a spatial defect $d_h'(t)$ by

$$P_h \partial_{tt} \Pi_h u(t) = P_h F_t(t, \Pi_h u(t)) - A_h(t, \Pi_h u(t)) \Pi_h u(t) + d_h'(t) \ \text{ in } \mathcal{V}_h' . \tag{III.28}$$

Since $\partial_t P_h = P_h \partial_t$, we can also differentiate (III.25) to get this equation. Hence, $d_h'(t) = \partial_t d_h(t)$.

We shall now compare the finite element solutions $u_{h,n}$ defined in (III.18) with the projections $\Pi_h u(t_n)$. We follow the theory of perturbed Rosenbrock methods established in [95] and extend the results to our full discretization scheme.

We have the following convergence estimate.

Lemma 2. *Let $d_h(t)$ be as defined in (III.25). Then for sufficiently small τ and $N\tau \le T$*

$$\begin{aligned} &\left(\tau \sum_{n=1}^{N} \| \Pi_h u(t_n) - u_{h,n} \|^2 \right)^{1/2} + \max_{1 \le n \le N} | \Pi_h u(t_n) - u_{h,n} | \\ &\le C\tau^\nu + C \left(\tau \sum_{n=0}^{N-1} \sum_{i=1}^{s} \| d_h(t_n + \alpha_i \tau) + \tau \gamma_i d_h'(t_n) \|_{\mathcal{V}_h'}^2 \right)^{1/2} , \end{aligned} \tag{III.29}$$

with

$\nu = 2$ if the assumptions of Theorem III.1 are satisfied, and
$\nu = 3$ if the assumptions of Theorem III.2 are satisfied.

The constants C depend on the constants in (III.9)–(III.15), on the $H_t^{\nu+1}(\mathcal{V})$–norm of the solution, on the coefficients of the Rosenbrock method, and on T.

Proof. We consider the perturbed Rosenbrock scheme

$$\begin{aligned}
P_h\tilde{K}'_{ni} &= F_h(t_n+\alpha_i\tau,\tilde{K}_{ni}) - \tau A_h(t_n,\tilde{u}_n)\sum_{j=1}^{i}\gamma_{ij}\tilde{K}'_{nj} \\
&\quad + \tau\gamma_i P_h F_t(t_n,\tilde{u}_n) + R'_{ni} \\
\tilde{u}_{n+1} &= \tilde{u}_n + \tau\sum_{i=1}^{s} b_i\tilde{K}'_{ni} + r_{n+1}\ ,
\end{aligned}$$

with perturbations $R'_{ni}\in\mathcal{V}'_h$ and $r_{n+1}\in\mathcal{V}_h$.
a) Let $p\geq 2$. Setting

$$\begin{aligned}
\tilde{K}'_{ni} &= \partial_t(\Pi_h u)(t_n+c_i\tau)\ , \\
\tilde{K}_{ni} &= \Pi_h u(t_n+\alpha_i\tau)\ , \quad \tilde{u}_n = \Pi_h u(t_n),
\end{aligned}$$

and taking into account the consistency conditions for Rosenbrock methods with order $p\geq 2$ [Appendix B, (B.1)], we derive by Taylor expansion

$$\begin{aligned}
R'_{ni} &= \tau\int_{t_n}^{t_{n+1}}\kappa_i\left(\tfrac{t-t_n}{\tau}\right)A_h(t_n,\Pi_h u(t_n))\partial_t^2(\Pi_h u)(t)\,dt \\
&\quad + \tau\int_{t_n}^{t_{n+1}}\hat{\kappa}_i\left(\tfrac{t-t_n}{\tau}\right)P_h\partial_t^3(\Pi_h u)(t)\,dt \\
&\quad + d_h(t_n+\alpha_i\tau) + \tau\gamma_i d'_h(t_n)\ , \\
r_{n+1} &= \tau^2\int_{t_n}^{t_{n+1}}\kappa\left(\tfrac{t-t_n}{\tau}\right)\partial_t^3(\Pi_h u)(t)\,dt\ .
\end{aligned} \tag{III.30}$$

Here, κ_i, $\hat{\kappa}_i$, and κ denote bounded Peano kernels.
b) Let $p\geq 3$. Setting now

$$\begin{aligned}
\tilde{K}'_{ni} &= \partial_t(\Pi_h u)(t_n+\alpha_i\tau) + \tau\gamma_i\partial_t^2(\Pi_h u)(t_n)\ , \\
\tilde{K}_{ni} &= \Pi_h u(t_n+\alpha_i\tau)\ , \quad \tilde{u}_n = \Pi_h u(t_n),
\end{aligned}$$

and using the corresponding consistency conditions, we get once again by Taylor expansion

$$
\begin{aligned}
R'_{ni} &= \tau^2 \sum_{j=1}^{i} \gamma_{ij} c_j A_h(t_n, \Pi_h u(t_n)) \partial_t^2 (\Pi_h u)(t_n) \\
&\quad + \tau^2 \int_{t_n}^{t_{n+1}} \kappa_i \left(\tfrac{t-t_n}{\tau}\right) A_h(t_n, \Pi_h u(t_n)) \partial_t^3 (\Pi_h u)(t)\, dt \\
&\quad + d_h(t_n + \alpha_i \tau) + \tau \gamma_i d'_h(t_n) , \\
r_{n+1} &= \tau^3 \int_{t_n}^{t_{n+1}} \kappa \left(\tfrac{t-t_n}{\tau}\right) \partial_t^4 (\Pi_h u)(t)\, dt ,
\end{aligned}
\tag{III.31}
$$

where κ_i and κ are bounded Peano kernels.

c) From (III.30) and (III.31) it is clear that the difference to the approach in [95] consists in the terms related to the spatial truncation error only. By the way we note that $e_0 = 0$ due to (III.26) and $u_{h,0} = \Pi_h u_0$. Repeating literally the proofs of Theorem 5.1 and Theorem 5.3 of [95], respectively, we obtain the desired result. □

Next we estimate the spatial truncation error.

Lemma 3. *Let $u \in H_t^l(\mathcal{V})$ with $l \geq 2$. Then for sufficiently small τ and $N\tau \leq T$*

$$
\tau \sum_{n=0}^{N-1} \sum_{i=1}^{s} \| d_h(t_n + \alpha_i \tau) + \tau \gamma_i d'_h(t_n) \|_{\mathcal{V}'_h}^2 \tag{III.32}
$$

$$
\leq C \left(\tau^{2(l-1)} \| \partial_t^l (u - \Pi_h u) \|_{L_t^2(\mathcal{V})}^2 + \| u - \Pi_h u \|_{H_t^1(\mathcal{V})}^2 + \tau^2 \| u - \Pi_h u \|_{H_t^2(\mathcal{V})}^2 \right) .
$$

The constant C depends on the constants in (III.9), (III.12)–(III.15), on the coefficients of the Rosenbrock method, and on the $H_t^2(\mathcal{V})$–norm of the solution.

Proof. Since $H_t^l(\mathcal{V}) \hookrightarrow C_t^1(\mathcal{V})$ for $l \geq 2$ there exists a constant C_0 such that $\|u\|_{C_t^1(\mathcal{V})} < C_0 \|u\|_{H_t^2(\mathcal{V})}$. Thus, in the following the constants C can be allowed to depend on the $C_t^1(\mathcal{V})$–norm of the solution. For brevity we shall use sometimes u_t instead of $\partial_t u$ in the notation above, and often omit the variable t and simply write u for $u(t)$, u_t for $u_t(t)$, etc.

a) First we estimate the residual error $d_h(t)$ for arbitrary but fixed $t \in (0, T]$. From equation (III.25) we subtract (III.24) to get

$$
d_h(t) = P_h \Pi_h u_t - P_h u_t + F_h(t, u) - F_h(t, \Pi_h u) .
$$

Using $A_h = -P_h F_u$, the mean–value theorem shows

$$
\| F_h(t, u) - F_h(t, \Pi_h u) \|_{\mathcal{V}'_h} = \left\| \int_0^1 A_h(t, \sigma u + (1-\sigma) \Pi_h u)(u - \Pi_h u)\, d\sigma \right\|_{\mathcal{V}'_h} .
$$

Since $w_\sigma := \sigma u + (1-\sigma)\Pi_h u \in \mathcal{W}$ for all $\sigma \in [0,1]$, the operators $A_h(t, w_\sigma)$ are uniformly bounded from $\mathcal{V}$ to $\mathcal{V}_h'$. Thus, we finally estimate

$$\|F_h(t,u) - F_h(t,\Pi_h u)\|_{\mathcal{V}_h'} \le C\,\|u - \Pi_h u\|\,.$$

Applying (III.3) and $\|v\|_* \le C\|v\|$ for all $v \in \mathcal{V}$, we conclude that

$$\|d_h(t)\|_{\mathcal{V}_h'} \le C\;(\,\|u_t(t) - \Pi_h u_t(t)\| + \|u(t) - \Pi_h u(t)\|\,)\;. \tag{III.33}$$

b) Now we estimate the first derivative $d_h'(t)$ of the spatial truncation error for arbitrary but fixed $t \in (0,T]$. We combine equations (III.27) and (III.28) to obtain the estimate

$$\begin{aligned} \|d_h'(t)\|_{\mathcal{V}_h'} \;\; &\le \;\; \|P_h\Pi_h u_{tt} - P_h u_{tt}\|_{\mathcal{V}_h'} + \|A_h(t,\Pi_h u)\Pi_h u_t - A_h(t,u)u_t\|_{\mathcal{V}_h'} \\ &\qquad + \|P_h F_t(t,u) - P_h F_t(t,\Pi_h u)\|_{\mathcal{V}_h'} =: I + II + III\,. \end{aligned}$$

It follows from (III.3) that

$$I \le \|u_{tt} - \Pi_h u_{tt}\|_*\,.$$

To estimate II, we use the uniform boundedness of $A_h(t,\Pi_h u)$ as an operator from $\mathcal{V}$ to $\mathcal{V}_h'$ and inequality (III.15):

$$\begin{aligned} II \;\; &\le \;\; \|A_h(t,\Pi_h u)\Pi_h u_t - A_h(t,\Pi_h u)u_t + A_h(t,\Pi_h u)u_t - A_h(t,u)u_t\|_{\mathcal{V}_h'} \\ &\le \;\; \|A_h(t,\Pi_h u)\Pi_h u_t - A_h(t,\Pi_h u)u_t\|_{\mathcal{V}_h'} + \|A_h(t,\Pi_h u)u_t - A_h(t,u)u_t\|_{\mathcal{V}_h'} \\ &\le \;\; C\,\|u_t - \Pi_h u_t\| + \|A_h(t,\Pi_h u) - A_h(t,u)\|_{\mathcal{L}(\mathcal{V},\mathcal{V}_h')}\,\|u_t\| \\ &\le \;\; C\;(\,\|u_t - \Pi_h u_t\| + \;\|u - \Pi_h u\|\,)\;. \end{aligned}$$

Next we use (III.14) to get directly

$$III \le C\,\|u - \Pi_h u\|\,.$$

Putting together the different contributions and using $\|v\|_* \le C\|v\|$ for all $v \in \mathcal{V}$, we conclude that

$$\|d_h'(t)\|_{\mathcal{V}_h'} \le C\,(\,\|u_{tt} - \Pi_h u_{tt}\| + \|u_t - \Pi_h u_t\| + \|u - \Pi_h u\|\,)\;. \tag{III.34}$$

c) In order to show (III.32), we observe

$$
\begin{aligned}
&\tau \sum_{n=0}^{N-1} \sum_{i=1}^{s} \|d_h(t_n+\alpha_i\tau) + \tau\gamma_i d_h'(t_n)\|^2_{\mathcal{V}_h'} \\
\leq\;& C\,\tau \sum_{n=0}^{N-1} \sum_{i=1}^{s} \Big(\|d_h(t_n+\alpha_i\tau)\|^2_{\mathcal{V}_h'} + \tau^2\gamma_i^2 \|d_h'(t_n)\|^2_{\mathcal{V}_h'} \Big) \,.
\end{aligned}
\tag{III.35}
$$

We begin with the first term related to d_h. Using (III.33) and recalling the definition of the interpolation operator Π_t^α, $\alpha \in [0,1]$, in (III.21), we find

$$
\begin{aligned}
I_d \;&:=\; \tau \sum_{n=0}^{N-1} \sum_{i=1}^{s} \|d_h(t_n+\alpha_i\tau)\|^2_{\mathcal{V}_h'} \\
&\leq\; C \sum_{i=1}^{s} \tau \sum_{n=0}^{N-1} \big(\|(u_t - \Pi_h u_t)(t_n+\alpha_i\tau)\|^2 + \|(u - \Pi_h u)(t_n+\alpha_i\tau)\|^2 \big) \\
&=\; C \sum_{i=1}^{s} \Big(\|\Pi_t^{\alpha_i}(u_t - \Pi_h u_t)\|^2_{l^2_N(\mathcal{V})} + \|\Pi_t^{\alpha_i}(u - \Pi_h u)\|^2_{l^2_N(\mathcal{V})} \Big) \,.
\end{aligned}
$$

To estimate the expressions on the right–hand side, we apply Lemma 1 for $v = u_t$, $q = l-1$, and $v = u$, $q = l$. Hence, with $\alpha = \alpha_i$

$$
\begin{aligned}
I_d \;\leq\; C \sum_{i=1}^{s} \Big(& \tau^{2(l-1)} \|\partial_t^{l-1}(u_t - \Pi_h u_t)\|^2_{L^2_t(\mathcal{V})} + \|u_t - \Pi_h u_t\|^2_{L^2_t(\mathcal{V})} \\
+\; & \tau^{2l} \|\partial_t^{l}(u - \Pi_h u)\|^2_{L^2_t(\mathcal{V})} + \|u - \Pi_h u\|^2_{L^2_t(\mathcal{V})} \Big) \,.
\end{aligned}
$$

Since the terms are now independent of the stage counter i, we deduce for sufficiently small τ

$$
I_d \leq C \,\Big(\tau^{2(l-1)} \|\partial_t^{l}(u - \Pi_h u)\|^2_{L^2_t(\mathcal{V})} + \|u - \Pi_h u\|^2_{H^1_t(\mathcal{V})} \Big) \,.
$$

In a similar way, we estimate the second term in (III.35). Using now (III.34) and $\Pi_t^{\alpha_i}$ with $\alpha_i = 0$, we get

$$
\begin{aligned}
II_d \;&:=\; \tau \sum_{n=0}^{N-1} \sum_{i=1}^{s} \tau^2\gamma_i^2 \|d_h'(t_n)\|^2_{\mathcal{V}_h'} \\
&\leq\; C \sum_{i=1}^{s} \tau^2\gamma_i^2\, \tau \sum_{n=0}^{N-1} \sum_{j=0}^{2} \|(\partial_t^j u - \Pi_h \partial_t^j u)(t_n)\|^2
\end{aligned}
$$

$$\leq \quad C \sum_{i=1}^{s} \gamma_i^2 \tau^2 \sum_{j=0}^{2} \|\Pi_t^0(\partial_t^j u - \Pi_h \partial_t^j u)\|^2_{l^2_N(\mathcal{V})} .$$

Successive use of Lemma 1 with $\alpha = 0$, $v = u, u_t, u_{tt}$, and $q = l, l-1, l-2$, respectively, yields

$$\begin{aligned} II_d \quad \leq \quad & C \sum_{i=1}^{s} \gamma_i^2 \tau^2 \, \Big(\tau^{2l} \|\partial_t^l (u - \Pi_h u)\|^2_{L^2_t(\mathcal{V})} \\ & + \tau^{2(l-1)} \|\partial_t^{l-1}(u_t - \Pi_h u_t)\|^2_{L^2_t(\mathcal{V})} + \tau^{2(l-2)} \|\partial_t^{l-2}(u_{tt} - \Pi_h u_{tt})\|^2_{L^2_t(\mathcal{V})} \\ & + \|u - \Pi_h u\|^2_{H^2_t(\mathcal{V})} \Big) \, . \end{aligned}$$

Taking into account that the terms are now independent of i, we conclude for sufficiently small τ

$$II_d \leq C \, \Big(\tau^{2(l-1)} \|\partial_t^l (u - \Pi_h u)\|^2_{L^2_t(\mathcal{V})} + \tau^2 \|u - \Pi_h u\|^2_{H^2_t(\mathcal{V})} \Big) \, .$$

The desired estimate (III.32) follows by combining the inequalities for I_d and II_d. ∎

Lemma 4. *Let $d_h(t)$ be as defined in (III.25). Then for sufficiently small τ and $N\tau \leq T$*

$$\begin{aligned} & \left(\tau \sum_{n=1}^{N} \| \Pi_h u(t_n) - u_{h,n}\|^2 \right)^{1/2} + \max_{1\leq n\leq N} |\Pi_h u(t_n) - u_{h,n}| \\ & \leq \; C \, \Big(\tau^\nu + \|u - \Pi_h u\|_{H^2_t(\mathcal{V})} \Big) \end{aligned} \tag{III.36}$$

with

$\nu = 2$ if the assumptions of Theorem III.1 are satisfied, and
$\nu = 3$ if the assumptions of Theorem III.2 are satisfied.

The constant C depends on the same quantities as in Lemma 2 and Lemma 3.

Proof. Apply Lemma 2 and Lemma 3 with $l = 3$ and $l = 4$. ∎

Now we are ready to prove our main convergence theorems.

Proof of Theorem III.1.

In the following we make permanent use of the splitting

$$u - u_{h,n} = (u - \Pi_h u) + (\Pi_h u - u_{h,n}) \, .$$

Since $|v| \leq C\|v\|$ for all $v \in \mathcal{V}$ (Appendix A§1), we have the embedding $H^l_t(\mathcal{V}) \hookrightarrow C^0_t(\mathcal{H})$ for $l \geq 1$, i.e., there exists in particular a constant $C > 0$ such that

$$\|v\|_{C_t^0(\mathcal{H})} \leq C\,\|v\|_{H_t^2(\mathcal{V})} \quad \text{for all } v \in H_t^2(\mathcal{V})\,. \tag{III.37}$$

By the definition of Π_t^1 we get

$$\begin{aligned}
\tau \sum_{n=1}^{N} \|u(t_n) - u_{h,n}\|^2 &\leq C\tau \sum_{n=1}^{N} \left(\|(u-\Pi_h u)(t_n)\|^2 + \|\Pi_h u(t_n) - u_{h,n}\|^2 \right) \\
&\leq C \left(\|\Pi_t^1 (u - \Pi_h u)\|^2_{l_N^2(\mathcal{V})} + \tau \sum_{n=1}^{N} \|\Pi_h u(t_n) - u_{h,n}\|^2 \right) .
\end{aligned}$$

Applying Lemma 1 for $v = u$, $q = 2$, $\alpha = 1$, and Lemma 4, we obtain

$$\begin{aligned}
&\tau \sum_{n=1}^{N} \|u(t_n) - u_{h,n}\|^2 \\
\leq\ & C \left(\tau^4 \|\partial_t^2 (u - \Pi_h u)\|^2_{L_t^2(\mathcal{V})} + \|u - \Pi_h u\|^2_{L_t^2(\mathcal{V})} + \tau^4 + \|u - \Pi_h u\|^2_{H_t^2(\mathcal{V})} \right) \\
\leq\ & C \left(\tau^4 + \|u - \Pi_h u\|^2_{H_t^2(\mathcal{V})} \right) .
\end{aligned}$$

Using Lemma 4 and the embedding (III.37), we deduce for the discrete maximum error in the $\mathcal{H}$–norm

$$\begin{aligned}
&\max_{1 \leq n \leq N} |u(t_n) - u_{h,n}| \\
\leq\ & \max_{1 \leq n \leq N} |u(t_n) - \Pi_h u(t_n)| + \max_{1 \leq n \leq N} |\Pi_h u(t_n) - u_{h,n}| \\
\leq\ & \|u - \Pi_h u\|_{C_t^0(\mathcal{H})} + C \left(\tau^2 + \|u - \Pi_h u\|_{H_t^2(\mathcal{V})} \right) \\
\leq\ & C \left(\tau^2 + \|u - \Pi_h u\|_{H_t^2(\mathcal{V})} \right) .
\end{aligned}$$

Putting all contributions together and taking into account the definition of the initial error (III.19), we conclude that

$$\begin{aligned}
&\left(\tau \sum_{n=0}^{N} \| u(t_n) - u_{h,n}\|^2 \right)^{1/2} + \max_{0 \leq n \leq N} | u(t_n) - u_{h,n}| \\
\leq\ & \epsilon_{h0}[u] + C \left(\tau^2 + \|u - \Pi_h u\|_{H_t^2(\mathcal{V})} \right) ,
\end{aligned}$$

which proves the theorem. □

Proof of Theorem III.2.
We proceed as in the proof of Theorem III.1. Using Lemma 1 for $q = 3$ and Lemma 4, we obtain order 3 in time. □

IV

Computational Error Estimation

The strategy for choosing time steps and mesh sizes is dictated by the a posteriori nature of the global bounds given in Theorem III.1 and Theorem III.2 above. Ideally, an adaptive method should keep the global error below a prescribed tolerance. But global errors are difficult to estimate. Thus, a standard approach is to adjust the discretization parameters during the integration in order to restrict the local truncation error. One would hope that smaller local errors lead also to a decrease of the global error - a property which is known as tolerance proportionality in the pure ODE case [138].

In the following a posteriori theory is used to estimate the local error of computations that are sufficiently accurate.

§1. Control of Time Steps

Step size control is an important and necessary means to increase the efficiency of a time integration method. In fact, a constant time step is often not adequate to reach a given accuracy, since it would require a huge number of small steps.

The discretization sequence, first time then space, permits us to consider naturally the spatial discretization as a perturbation of the time integration process. We assume for the moment that the spatial perturbation is always kept below a certain level and does not affect mainly the step size selection procedure. Thus, the generated time step sequence $\{\tau_j\}_{j=0,1,\dots}$ is supposed to be nearly the same as in the case of no perturbation.

The local truncation error $\delta_\tau(t)$ is defined as the error after a single step of length τ starting with the exact local solution $u(t)$. Using the short notation $u_{n+1}=\Phi(u_n)$ for the Rosenbrock method (II.18)-(II.19), we have at $t=t_n$ with time step $\tau_n=t_{n+1}-t_n$

$$\delta_{\tau_n}(t_n) = \Phi(u(t_n)) - u(t_n+\tau_n)\,. \tag{IV.1}$$

The asymptotic behaviour of the local error for an order p method can be described by

$$\delta_{\tau_n}(t_n) = \phi(t_n)\,\tau_n^{p+1} + o(\tau_n^{p+2})\,,\quad \tau_n\to 0\,. \tag{IV.2}$$

Assuming appropriate temporal regularity of the mapping $F(t,u(t))$ in (II.4), the coefficient vector $\phi(t)$ is a smooth function of t.

The global error $e_{n+1} := u_{n+1} - u(t_{n+1})$ at the forward time level $t = t_{n+1}$ can be seen to satisfy

$$e_{n+1} = \Phi(e_n + u(t_n)) - \Phi(u(t_n)) + \delta_{\tau_n}(t_n)\ . \tag{IV.3}$$

Consequently, this error is the sum of the local error and the difference of the actual Rosenbrock step $\Phi(u_n)$ and the hypothetical step $\Phi(u(t_n))$ taken from the exact solution $u(t_n)$. To measure the errors we introduce an appropriate norm $|||\cdot|||$ which is often a mixed absolute–relative norm in practical computations for reason of robustness (see Chapter V.§3). It is now a fundamental property of a stable one–step integration method that

$$|||\, e_{n+1}\, ||| \leq |||\, e_n\, ||| + |||\, \delta_{\tau_n}(t_n)\, ||| \leq |||\, e_0\, ||| + \sum_{j=0}^{n} |||\, \delta_{\tau_j}(t_j)\, |||\ , \tag{IV.4}$$

showing that the global error consists of propagated and accumulated local truncation errors. Estimating and controlling the latter within an automatic step size selection procedure ensure that the step sizes τ_j are chosen sufficiently small to have the desired precision, but they have to be also sufficiently large to avoid unnecessary computational work.

Unfortunately, the local error (IV.1) is not computable directly, but there are different ways to estimate it. Two classical devices are usually applied: Richardson extrapolation and embedding (see, e.g., [71], §II.4). For stiff ODEs a comparison of these techniques involving various Rosenbrock methods was published in [79]. It turned out that for low tolerances embedding yields satisfactory results while Richardson extrapolation becomes superior at more stringent tolerances. In the context of PDEs desired tolerances ranging approximately from 5% to 0.1% are usually required. Thus, an error estimation based on embedding should be a good choice here.

A pair of embedded Rosenbrock methods consists of two different methods. Replacing the coefficients b_i in (II.18) by different coefficients $\hat{b}_i$ a second solution $\hat{u}_{n+1}$ of inferior order, say $p-1$, can be obtained. For this, an enlarged set of consistency conditions has to be fulfilled. The difference of both solutions satisfactorily estimates the local truncation error of $\hat{u}_{n+1}$, and we define the scalar estimate

$$r_{n+1} := |||\, u_{n+1} - \hat{u}_{n+1}\, |||\ . \tag{IV.5}$$

The asymptotic behaviour of r_{n+1} is given by

$$r_{n+1} = \phi_n\, \tau_n^p\ , \quad \phi_n = |||\, \phi(t_n) + o(\tau_n)\, |||\ . \tag{IV.6}$$

Assuming $\phi_{n+1} \approx \phi_n$, i.e., the coefficient vector is varying a little only or constant, and setting $r_{n+2} = TOL_t$, the next time step can be chosen by

$$\tau_{n+1} = \rho \left(\frac{TOL_t}{r_{n+1}} \right)^{1/p} \tau_n \; , \tag{IV.7}$$

where ρ denotes a safety factor. Typically, one sets $\rho = 0.95$ to reduce the risk for a rejected step. If $TOL_t < r_{n+1}$ the step is rejected, and a new try is performed with $\tau_n := \tau_{n+1}$. Otherwise, if the computation was successful we step forward with τ_{n+1}.

Although the rule (IV.7) is standard and commonly used for ODE integrators, the traditional asymptotic model (IV.2) fails sometimes to describe correctly the relation between step size and local error. Often a nonsmooth behaviour of the time integration process can be observed. For instance, after a drastic step size reduction the corresponding error r_{n+1} becomes very small. The proposed new time step will be too optimistic leading to repeated rejections. Such oscillations yield unacceptable performance of any integration method. There is also a limit on how much the step size may decrease after an accepted step. According to $TOL_t \geq r_{n+1}$ we get $\tau_{n+1} \geq \rho \tau_n$. Consequently, the standard controller is unable to reduce drastically the time step without rejections.

Example 1. The nonlinear one–dimensional flame propagation problem [113, 60]

$$\begin{aligned}
\partial_t u_1 - \partial_{xx} u_1 &= f(u_1, u_2) \; , \\
\partial_t u_2 - \frac{1}{Le} \partial_{xx} u_2 &= -f(u_1, u_2) \; , \\
f(u_1, u_2) &:= \frac{\beta^2}{2\,Le} u_2 \exp\left(\frac{-\beta(1-u_1)}{1-\alpha(1-u_1)} \right)
\end{aligned}$$

is solved for $\alpha = 0.8$, $\beta = 20$, and $Le = 2$. The boundary and initial conditions are

$$\begin{aligned}
&u_1(-\infty, t) = 0, \\
&\partial_x u_1(\infty, t) = 0,
\end{aligned} \qquad u_1(x,0) = \begin{cases} \exp(x) & \text{for} \quad x \leq 0 \\ 1 & \text{for} \quad x > 0 \; , \end{cases}$$

$$\begin{aligned}
&u_2(-\infty, t) = 1, \\
&\partial_x u_2(\infty, t) = 0,
\end{aligned} \qquad u_2(x,0) = \begin{cases} 1 - \exp(Le \cdot x) & \text{for} \quad x \leq 0 \\ 0 & \text{for} \quad x > 0 \; . \end{cases}$$

For the chosen Lewis number Le the system describes an unregularly oscillating propagation of a flame changing its shape and velocity in time. We take a sufficiently large computational domain to ensure that the flame propagation is not affected by the boundary conditions. Spatial errors are kept always below an appropriate level. The standard controller applied for a fixed tolerance TOL_t reduces and increases the time step with respect to the varying flame speed. In Fig. IV.1 the employed values over a long time interval are plotted. The time step varies over more than two orders of magnitude. A special zoom shows that many computed solutions have to be rejected for reason of precision. In this phase of quickly changing dynamics the controller fails to reduce the time step in a smooth way. Computational effort is wasted and the integration performance becomes unacceptable.

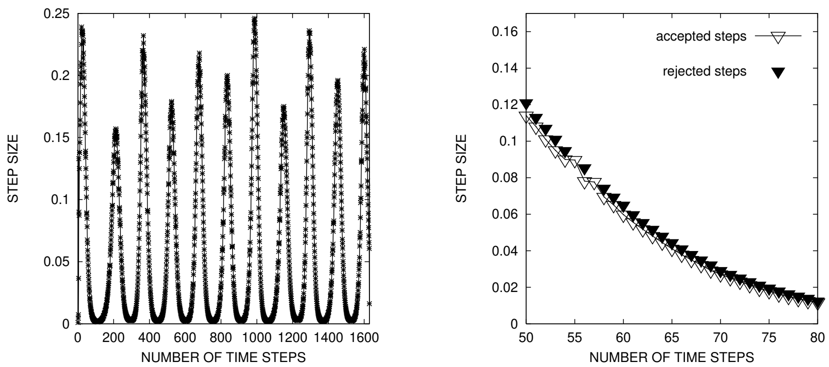


Figure IV.1: Standard controller for one–dimensional oscillating flame: Selected time steps τ_j versus j for $t \in [0, 100]$ (left) and for a critical phase of acceleration $t \in [8.10, 9.02]$ (right). The black triangles indicate time steps that had to be rejected, the white triangles correspond to accepted steps. The standard controller is unable to reduce drastically the time step without rejections.

A good step size control algorithm must work well for a large class of problems with a great diversity in the dynamic behaviour. The standard controller works normally quite well, but it does not have an entirely satisfactory performance. The basic assumptions that ϕ varies slowly and higher order error terms are negligible seem to be questionable in some cases.

In the pioneering work of GUSTAFSSON et al. [70] the step size control has been viewed as an automatic control problem. They proposed a new control algorithm employing a discrete PI (proportional integral) controller. The standard controller (IV.7) is based on the key equations

$$r_{n+2} = \phi_{n+1}\, \tau_{n+1}^{p} \;, \quad \phi_{n+1} = \phi_n \;, \tag{IV.8}$$

where a constant model for ϕ is used. To include significant changes of ϕ an improved approximation is necessary. Several models have been compared in [69]. One beneficial choice is to consider multiplicative changes by

$$\log \phi_{n+1} = \log \phi_n + (\log \phi_n - \log \phi_{n-1}) \,. \tag{IV.9}$$

The difference on the right–hand side measures the last change of $\log\phi$. A straightforward calculation yields the new step size selection rule

$$\tau_{n+1} = \frac{\tau_n}{\tau_{n-1}} \left(\frac{TOL_t\, r_n}{r_{n+1}\, r_{n+1}} \right)^{1/p} \tau_n \,, \tag{IV.10}$$

showing that now more data from previous steps are used to decide on the new step size. Optionally, one can involve once again a safety factor ρ.

Additionally, we have experienced that the following two modifications are useful. After successive rejections the standard controller with an approximate exponent

$$p \approx p_{n+1} = \frac{\log r_{n+1} - \log r_n}{\log \tau_n - \log \tau_{n-1}}$$

should be applied. Since the step proposed by the new PI controller is often too optimistic when steps have to be increased, we suggest to take in this case the standard proposal if it is smaller.

Remark 1. Applying the new strategy to the above flame problem with the same tolerance, we get a smooth integration behaviour where no computed solutions have been rejected. The resulting reduction in CPU time is about 20% for the entire run. Although the observed saving is not extremely large it should be recalled that it can be obtained with merely no additional cost. Moreover, the PI controller makes the computation more robust in situations with sudden changes in time.

Remark 2. Although the estimate of the local truncation error is only correct for the lower order solution $\hat{u}_{n+1}$, usually the higher order approximation u_{n+1} is used to proceed in time. For this, DEUFLHARD and BORNEMANN [48] give a theoretical justification based on arguments from control theory and from simple error relations. The authors present also a PID controller which includes further information.

§2. Estimation of Spatial Errors

Since the pioneering works of BABUŠKA and RHEINBOLDT [10, 11] quite a lot of a posteriori error estimates have been developed for mastering finite element calculations. Now they are widely used in the mesh–controlled solution of partial differential equations. A good survey is given in [9] and more recently in [144], where also a substantial bibliography on the subject can be found.

We deal with a posteriori error estimators based on the use of hierarchical basis functions. Such estimators have been accepted to provide efficient and reliable assessment of spatial errors and to form a basis of adaptive local mesh refinement [155, 50, 18, 34]. Our aim is the extension of the hierarchical bases technique to time–dependent nonlinear problems within the setting of linearly implicit time integrators. The crucial point herein is the construction of robust estimators for the fully discretized equations (III.16) which are singularly perturbed by the presence of the (variable) time step. A robust estimator has to yield upper and lower bounds on the error uniformly in the time step $\tau \geq 0$. In general, a straightforward application of standard adaptive finite element solvers runs into troubles in the limit case $\tau \to 0$. For selfadjoint scalar problems robust estimators were constructed in [31, 145]. Our analysis takes into account the abstract framework of [18].

Analogously to Chapter IV.§1 we are interested in analyzing the local error behaviour. The spatial discretization is considered as a perturbation of the time integration process. Starting with the Rosenbrock solution u_n at $t = t_n$, we will estimate the error $u_{n+1} - u_{h,n+1}$ caused by the spatial approximation of all stage values $K'_{h,ni} \in \mathcal{V}_h$. Although system (III.16) is linear the nonlinearity on the right–hand side gives rise to a nonlinear spatial error transport.

Let us consider a hierarchical decomposition

$$\bar{\mathcal{V}}_h = \mathcal{V}_h \oplus \mathcal{Z}_h \ , \tag{IV.11}$$

where $\mathcal{Z}_h$ is the subspace that corresponds to the span of all additional basis functions needed to enrich the space $\mathcal{V}_h$ to higher order. Consequently, any function $\bar{v} \in \bar{\mathcal{V}}_h$ has the unique decomposition $\bar{v} = v + z$, where $v \in \mathcal{V}_h$ and $z \in \mathcal{Z}_h$. The hierarchical basis error estimator tries to bound the spatial error by evaluating its components in the space $\mathcal{Z}_h$, i.e.,

$$C_1 |||E_{h,n+1}||| \leq |||u_{n+1} - u_{h,n+1}||| \leq C_2 |||E_{h,n+1}||| \ , \tag{IV.12}$$

where $E_{h,n+1} \in \mathcal{Z}_h$ is the computed a posteriori error estimate. This is justified by the fact that applying $\bar{\mathcal{V}}_h$ to approximate the solution, the component lying in $\mathcal{V}_h$ will change generally very little from the previous computation based on $\mathcal{V}_h$. One key point of our analysis is to choose the norm $|||\cdot|||$ in such a way that the constants C_1 and C_2 are independent of the mesh size and the time step, ensuring the robustness of the estimator.

In accordance with our main convergence Theorems III.1 and III.2 we introduce a τ–dependent error norm

$$||v||_\tau^2 := \tau ||v||^2 + |v|^2 \ , \quad v \in \mathcal{V} \ , \tag{IV.13}$$

and its associated sesquilinear form

$$a_\tau(v,w) = \tau((v,w)) + (v,w) \ , \quad v,w \in \mathcal{V}. \tag{IV.14}$$

To study the solution process of the Rosenbrock method at $t = t_n$ we define another sesquilinear form

$$b_n(v_1,v_2) := \langle v_1, v_2 \rangle + \gamma\tau \, a(t_n, u_n; v_1, v_2) \tag{IV.15}$$

on $\mathcal{V} \times \mathcal{V}$ and use the weak formulation of (II.18) for all $K'_{ni} \in \mathcal{V}$

$$b_n(K'_{ni}, \phi) = \langle r_{ni}(K'_n), \phi \rangle , \quad \phi \in \mathcal{V} , \tag{IV.16}$$

where

$$\begin{aligned} K'_n &= (K'_{n1}, \ldots, K'_{ns})^T , \\ r_{ni}(\underline{w}) &= F(t_n + \alpha_i \tau, u_n + \tau \sum_{j=1}^{i-1} \alpha_{ij} w_j) - \tau A(t_n, u_n) \sum_{j=1}^{i-1} \gamma_{ij} w_j \\ &\quad + \tau \gamma_i F_t(t_n, u_n) , \quad \underline{w} = (w_1, \ldots, w_s)^T \in \mathcal{V}^s . \end{aligned} \tag{IV.17}$$

The new solution at $t = t_{n+1}$ is calculated by

$$u_{n+1} = u_n + \tau \sum_{i=1}^{s} b_i K'_{ni} . \tag{IV.18}$$

In the following we assume that the sesquilinear form $a(t_n, u_n; \cdot, \cdot)$ satisfies the stronger conditions (II.10) and (II.12), i.e.,

$$\text{(A1)} \qquad |a(t_n, u_n; v_1, v_2)| \leq M_a \|v_1\| \|v_2\| , \quad v_1, v_2 \in \mathcal{V} , \tag{IV.19}$$

$$\text{(A2)} \qquad a(t_n, u_n; v_1, v_1) \geq \mu_a \|v_1\|^2 , \quad v_1 \in \mathcal{V} , \tag{IV.20}$$

with constants M_a, $\mu_a > 0$ independent of t_n, u_n, v_1, and v_2.
Now we can prove the following estimate.

Lemma 1. *Let b_n be defined as above and assume that (A1) and (A2) hold. Then there exist positive constants M_b and μ_b independent of τ such that*

$$\text{(B1)} \qquad |b_n(v_1, v_2)| \leq M_b \|v_1\|_\tau \|v_2\|_\tau , \tag{IV.21}$$

$$\text{(B2)} \qquad b_n(v_1, v_1) \geq \mu_b \|v_1\|_\tau^2 , \tag{IV.22}$$

for all functions $v_1, v_2 \in \mathcal{V}$.

Proof. Taking into account the embedding identity $\langle v', v \rangle = (v', v)$, $v' \in \mathcal{H}$, $v \in \mathcal{V}$, and applying the Cauchy–Schwarz–Bunyakovskii inequality, we deduce

$$\begin{aligned} |b_n(v_1, v_2)| &\leq |v_1||v_2| + \tau\gamma M_a \|v_1\| \|v_2\| \\ &\leq (\tau\gamma M_a \|v_1\|^2 + |v_1|^2)^{1/2} (\tau\gamma M_a \|v_2\|^2 + |v_2|^2)^{1/2} \\ &\leq \max(1, \gamma M_a) \|v_1\|_\tau \|v_2\|_\tau , \end{aligned}$$

and

$$b_n(v_1, v_1) \geq |v_1|^2 + \tau\gamma\mu_a \|v_1\|^2 \geq \min\,(1, \gamma\mu_a)\|v_1\|_\tau^2 .$$

With $M_b := \max\,(1, \gamma M_a)$ and $\mu_b := \min\,(1, \gamma\mu_a)$ we obtain the desired result. □

Because of (B1) and (B2) the linear systems (IV.16) are uniquely solvable. Thus, the τ–dependent norm (IV.13) is appropriate to describe properties of (IV.16). The same is true when $\mathcal{V}$ is replaced by $\mathcal{V}_h$. Let $K'_{h,n} = (K'_{h,n1}, \ldots, K'_{h,ns})^T$ be the vector of finite element solutions $K'_{h,ni} \in \mathcal{V}_h \subset \mathcal{V}$ satisfying

$$b_n(K'_{h,ni}, \phi) = \langle r_{ni}(K'_{h,n}), \phi \rangle\,, \quad \phi \in \mathcal{V}_h\,, \tag{IV.23}$$

and set

$$u_{h,n+1} = u_{h,n} + \tau \sum_{i=1}^{s} b_i K'_{h,ni} \tag{IV.24}$$

with $u_{h,n} \in \mathcal{V}_h$ chosen such that

$$b_n(u_n - u_{h,n}, \phi) = 0\,, \quad \phi \in \mathcal{V}_h\,. \tag{IV.25}$$

Remark 3. We note that (IV.23) differs slightly from (III.16), where the approximate solution $u_{h,n}$ is also used to calculate the terms on the right–hand side. Here, the Rosenbrock solution u_n is projected onto $\mathcal{V}_h$ to ensure $u_{h,n+1} \in \mathcal{V}_h$.

Let us recall $\bar{\mathcal{V}}_h = \mathcal{V}_h \oplus \mathcal{Z}_h$ as defined in (IV.11). For the theoretical analysis of an a posteriori error estimate for $u_{n+1} - u_{h,n+1}$, we need also the finite element solution vector $\bar{K}'_{h,n} = (\bar{K}'_{h,n1}, \ldots, \bar{K}'_{h,ns})^T$, with $\bar{K}'_{h,ni} \in \bar{\mathcal{V}}_h \subset \mathcal{V}$ defined by

$$b_n(\bar{K}'_{h,ni}, \phi) = \langle r_{ni}(\bar{K}'_{h,n}), \phi \rangle\,, \quad \phi \in \bar{\mathcal{V}}_h\,. \tag{IV.26}$$

With these stage values, we have an approximate solution $\bar{u}_{h,n+1} \in \bar{\mathcal{V}}_h$ by the summation

$$\bar{u}_{h,n+1} = \bar{u}_{h,n} + \tau \sum_{i=1}^{s} b_i \bar{K}'_{h,ni}\,, \tag{IV.27}$$

with $\bar{u}_{h,n}$ chosen such that

$$b_n(u_n - \bar{u}_{h,n}, \phi) = 0\,, \quad \phi \in \bar{\mathcal{V}}_h\,. \tag{IV.28}$$

Obviously, one expects that $\bar{u}_{h,n+1} \in \bar{\mathcal{V}}_h$ is a better approximation to the solution u_{n+1} than $u_{h,n+1} \in \mathcal{V}_h$. This is a very natural property of a higher order scheme expressed in terms of the *saturation assumption*

(A3) $$\|u_{n+1} - \bar{u}_{h,n+1}\|_\tau \leq \beta\, \|u_{n+1} - u_{h,n+1}\|_\tau\,, \tag{IV.29}$$

where $\beta < 1$ independent of h and τ. In general, one can anticipate $\beta \to 0$ as $h \to 0$.

We add also a *strengthened Cauchy–Schwarz–Bunyakovskii inequality* for the decomposition (IV.11) to the list of our assumptions, i.e., there exists a constant $\delta \in [0,1)$ independent of h and τ such that

$$\text{(A4)} \qquad |a_\tau(v,z)| \le \delta \, \|v\|_\tau \|z\|_\tau \quad v \in \mathcal{V}_h,\ z \in \mathcal{Z}_h\,. \tag{IV.30}$$

Remark 4. Typically, inequality (A4) is a direct consequence of $\mathcal{V}_h \cap \mathcal{Z}_h = \{0\}$ ([53], Theorem 1). The calculation of δ is usually straightforward, employing standard transformations to reference elements (cf. [36, 98, 16]). There is also an interesting equivalence result concerning certain interpolation operators widely used in the analysis of multilevel iterative methods ([16], Lemma 3).

A direct consequence of the saturation assumption and the triangle inequality is the estimate

$$\frac{1}{1+\beta}\|\bar{u}_{h,n+1} - u_{h,n+1}\|_\tau \le \|u_{n+1} - u_{h,n+1}\|_\tau \le \frac{1}{1-\beta}\|\bar{u}_{h,n+1} - u_{h,n+1}\|_\tau\,, \tag{IV.31}$$

showing that $\|\bar{u}_{h,n+1} - u_{h,n+1}\|_\tau$ is a robust estimator of the local spatial error. Clearly, an exact solution of (IV.26) to calculate $\bar{u}_{h,n+1}$ would be far too expensive.

We define our a posteriori error estimator $E_{h,n+1} \in \mathcal{Z}_h$ by

$$E_{h,n+1} = E_{h,n0} + \tau \sum_{i=1}^{s} b_i E_{h,ni}\,, \tag{IV.32}$$

where $E_{h,n0}$ approximates the projection error of the initial value in $\mathcal{Z}_h$, i.e,

$$b_n(E_{h,n0},\phi) = b_n(u_n - u_{h,n},\phi)\,, \quad \phi \in \mathcal{Z}_h\,, \tag{IV.33}$$

and the vector of all stage error estimators $E_{h,n} = (E_{h,n1},\ldots,E_{h,ns})^T \in \mathcal{Z}_h^s$ satisfies

$$b_n(E_{h,ni},\phi) = \big\langle r_{ni}(K'_{h,n} + E_{h,n}),\phi\big\rangle - b_n(K'_{h,ni},\phi)\,, \quad \phi \in \mathcal{Z}_h\,. \tag{IV.34}$$

According to the definition (IV.17), the computation of the error estimator $E_{h,n+1}$ requires the solution of a linear system. The stage error estimates $E_{h,ni}$ are used successively to improve the approximation of the nonlinear terms r_{ni}.

Our goal is to prove that the function $E_{h,n+1}$ which can be computed more easily than the difference $\bar{u}_{h,n+1} - u_{h,n+1}$ yields a good approximation of the local spatial error. This is shown in the following theorems.

Theorem 1. *Assume $s=1$, (A1)–(A4) hold, and $E_{h,n+1}$ is as defined above. Then*

$$\frac{\mu_b}{M_b}\|E_{h,n+1}\|_\tau \le \|u_{n+1}-u_{h,n+1}\|_\tau \le \frac{M_b}{(1-\beta)\mu_b\sqrt{1-\delta^2}}\|E_{h,n+1}\|_\tau\,. \qquad \text{(IV.35)}$$

Theorem 2. *Assume $s>1$, (A1)–(A4) hold, and $E_{h,n+1}$ is as defined above. Let $\bar u_{h,n+1}=\hat u_{h,n+1}+\hat E_{h,n+1}$, where $\hat u_{h,n+1}\in\mathcal{V}_h$ and $\hat E_{h,n+1}\in\mathcal{Z}_h$. Then*

$$\frac{\mu_b}{M_b}\|E_{h,n+1}\|_\tau \quad\le\quad \|u_{n+1}-u_{h,n+1}\|_\tau + D_n\,, \qquad \text{(IV.36)}$$

$$\|u_{n+1}-u_{h,n+1}\|_\tau \quad\le\quad \frac{M_b}{(1-\beta)\mu_b\sqrt{1-\delta^2}}\|E_{h,n+1}\|_\tau + \bar D_n\,, \qquad \text{(IV.37)}$$

where

$$\begin{aligned} D_n &= \frac{\sqrt{\tau}}{M_b}\left\|\sum_{i=1}^s b_i(r_{ni}(K'_{h,n}+E_{h,n})-r_{ni}(K'_n))\right\|_*\,,\\ \bar D_n &= \frac{M_b}{(1-\beta)\mu_b}\,\|\hat u_{h,n+1}-u_{h,n+1}\|_\tau\\ &\quad+\frac{\sqrt{\tau}}{(1-\beta)\mu_b\sqrt{1-\delta^2}}\left\|\sum_{i=1}^s b_i(r_{ni}(K'_{h,n}+E_{h,n})-r_{ni}(\bar K'_{h,n}))\right\|_*\,. \end{aligned}$$

We will shortly discuss the above results. In the case $s=1$, Theorem IV.1 shows that the estimator $E_{h,n+1}$ is robust in the sense (IV.12). With increasing s, the situation becomes less favourable since the optimal estimate is disturbed by the terms D_n and $\bar D_n$. However, we will see in a moment that in general these perturbations are negligible for practical computations making some natural assumptions.

If we resolve our elliptic problems in $\bar{\mathcal{V}}_h$ to compute $\bar u_{h,n+1}=\hat u_{h,n+1}+\hat E_{h,n+1}$, where $\hat u_{h,n+1}\in\mathcal{V}_h$ and $\hat E_{h,n+1}\in\mathcal{Z}_h$, we intuitively expect that

$$\|\hat u_{h,n+1}-u_{h,n+1}\|_\tau \ll \|\hat E_{h,n+1}\|_\tau\,.$$

In fact, this is the reason why hierarchical error estimators work well in practice. Applying Lemma IV.2 given in Chapter IV.§3 with $\bar v=\bar u_{h,n+1}$ and using the left–hand inequality of (IV.31), we get

$$\|\hat u_{h,n+1}-u_{h,n+1}\|_\tau \ll \|u_{n+1}-u_{h,n+1}\|_\tau\,, \qquad \text{(IV.38)}$$

demonstrating that the first term of $\bar D_n$ is small with respect to the local spatial error.

Next we study the influence of the nonlinear error transport caused by the sources r_{ni}. Taking into account the uniform boundedness of A as operator from $\mathcal{V}$ to $\mathcal{V}'$ and (II.15), we derive by Taylor expansion for sufficiently small τ

$$\sqrt{\tau}\left\|\sum_{i=1}^{s} b_i(r_{ni}(K'_{h,n}+E_{h,n})-r_{ni}(K'_n))\right\|_* \leq C_1\tau\sum_{i=1}^{s-1}\|K'_{h,ni}+E_{h,ni}-K'_{ni}\|_\tau$$

and

$$\sqrt{\tau}\left\|\sum_{i=1}^{s} b_i(r_{ni}(K'_{h,n}+E_{h,n})-r_{ni}(\bar{K}'_{h,n}))\right\|_* \leq C_2\tau\sum_{i=1}^{s-1}\|K'_{h,ni}+E_{h,ni}-\bar{K}'_{h,ni}\|_\tau$$

with $C_1, C_2 > 0$ independent of τ. Neglecting for a moment the initial error $u_n - u_{h,n}$ which is assumed to be sufficiently small, the combination of (IV.18) and (IV.24) yields

$$u_{n+1} - u_{h,n+1} \approx \tau\sum_{i=1}^{s} b_i(K'_{ni} - K'_{h,ni}). \tag{IV.39}$$

Thus, the local spatial error is mainly a sum of weighted stage errors multiplied by τ. Utilizing errors already computed in (IV.34) to calculate following stage errors, we have the hope to improve the whole estimation process such that

$$\sum_{i=1}^{s-1}\|K'_{h,ni}+E_{h,ni}-K'_{ni}\|_\tau < c_1\sum_{i=1}^{s-1}\|K'_{h,ni}-K'_{ni}\|_\tau\,, \tag{IV.40}$$

$$\sum_{i=1}^{s-1}\|K'_{h,ni}+E_{h,ni}-\bar{K}'_{h,ni}\|_\tau < c_2\sum_{i=1}^{s-1}\|K'_{h,ni}-K'_{ni}\|_\tau\,, \tag{IV.41}$$

with sufficiently small $c_1, c_2 > 0$. In this case, the corresponding terms of D_n and $\bar{D}_n$ have moderate size compared to the local spatial error $\|u_{n+1} - u_{h,n+1}\|_\tau$.

Although we could not prove exactly the robustness of our hierarchical error estimator for $s > 1$, the above discussion provides some intuitive insight justifying its use also for the considered general nonlinear problem class.

The expense of the error estimation can be further decreased if the sesquilinear form b_n on the left–hand sides of (IV.33) and (IV.34) is replaced by an approximation $\tilde{b}_n$ that allows a more efficient solution of the arising linear systems. In this case, we compute approximations $\tilde{E}_{h,n0}$ and $\tilde{E}_{h,ni}$ in $\mathcal{Z}_h$ satisfying

$$\tilde{b}_n(\tilde{E}_{h,n0},\phi) \;=\; b_n(u_n - u_{h,n},\phi)\,, \tag{IV.42}$$

$$\tilde{b}_n(\tilde{E}_{h,ni},\phi) \;=\; \langle r_{ni}(K'_{h,n}+\tilde{E}_{h,n}),\phi\rangle - b_n(K'_{h,ni},\phi)\,, \quad \phi\in\mathcal{Z}_h\,, \tag{IV.43}$$

where $\tilde{E}_{h,n} = (\tilde{E}_{h,n1}, \dots, \tilde{E}_{h,ns})^T$. From these relations, we derive for

$$\tilde{E}_{h,n+1} = \tilde{E}_{h,n0} + \tau \sum_{i=1}^{s} b_i \tilde{E}_{h,ni}$$

the equality for all $\phi \in \mathcal{Z}_h$

$$\begin{aligned} \tilde{b}_n(\tilde{E}_{h,n+1}, \phi) \quad &= \quad b_n(E_{h,n+1}, \phi) \\ &\quad + \tau \sum_{i=1}^{s} b_i \langle r_{ni}(K'_{h,n} + \tilde{E}_{h,n}) - r_{ni}(K'_{h,n} + E_{h,n}), \phi \rangle . \end{aligned} \tag{IV.44}$$

We have the following estimate to assess the quality of $\tilde{E}_{h,n+1}$.

Theorem 3. *Let $\tilde{b}_n$ be a sesquilinear form satisfying for all z_1, $z_2 \in \mathcal{Z}_h$*

$$(\tilde{B}1) \qquad |\tilde{b}_n(z_1, z_2)| \quad \leq \quad \tilde{M}_b \, \|z_1\|_\tau \|z_2\|_\tau \, , \tag{IV.45}$$

$$(\tilde{B}2) \qquad \tilde{b}_n(z_1, z_1) \quad \geq \quad \tilde{\mu}_b \, \|z_1\|_\tau^2 \, , \tag{IV.46}$$

with positive constants $\tilde{M}_b$ and $\tilde{\mu}_b$. Then

$$\frac{\tilde{\mu}_b}{M_b} \, \|\tilde{E}_{h,n+1}\|_\tau \quad \leq \quad \|E_{h,n+1}\|_\tau + \frac{D_n}{M_b} \, , \tag{IV.47}$$

$$\|E_{h,n+1}\|_\tau \quad \leq \quad \frac{\tilde{M}_b}{\mu_b} \, \|\tilde{E}_{h,n+1}\|_\tau + \frac{D_n}{\mu_b} \, , \tag{IV.48}$$

where

$$D_n = \sqrt{\tau} \left\| \sum_{i=1}^{s} b_i (r_{ni}(K'_{h,n} + \tilde{E}_{h,n}) - r_{ni}(K'_{h,n} + E_{h,n})) \right\|_* .$$

Proof. The inequalities follow immediately from (IV.44) and the conditions (B1), (B2), ($\tilde{\text{B}}$1), and ($\tilde{\text{B}}$2) (see also [18], Theorem 3.2). □

Once again Taylor expansion shows that for sufficiently small τ

$$D_n \leq C\tau \sum_{i=1}^{s-1} \|\tilde{E}_{h,ni} - E_{h,ni}\|_\tau .$$

Thus, the perturbation term D_n is negligible providing appropriate approximations $\tilde{b}_n$ to the sesquilinear form b_n.

Remark 5. One usual choice of an approximate sesquilinear form $\tilde{b}_n$ is based on the diagonalization of b_n over $\mathcal{Z}_h$. Let $\{\phi_j\}$ are the basis functions of $\mathcal{Z}_h$, and $z_1, z_2 \in \mathcal{Z}_h$ with the representations

$$z_1 = \sum\nolimits_j z_{1j} \phi_j \quad \text{and} \quad z_2 = \sum\nolimits_j z_{2j} \phi_j \, .$$

Then we define

$$\tilde{b}_n(z_1, z_2) = \sum_j z_{1j} z_{2j}\, b_n(\phi_j, \phi_j)\,.$$

In the context of hierarchical basis functions, this leads to a very efficient algorithm for computing an a posteriori error estimate (see for instance [155],[50]). The global calculation of $E_{h,n+1}$ is reduced to small element–by–element calculations to compute $\tilde{E}_{h,n+1}$. An analogous approach for the solution of reaction–diffusion equations was proposed in [84]. A second possibility is to use nonconforming approximation spaces. The arising stiffness matrix is then block–diagonal. For more details we refer to [16, 3].

§3. Proof of the Error Estimates

In this section, we use the strengthened Cauchy–Schwarz–Bunyakowskii inequality to have the following estimate.

Lemma 2. *Let $\bar{\mathcal{V}}_h = \mathcal{V}_h \oplus \mathcal{Z}_h$ and $\bar{v} = \hat{v} + \hat{z}$, where $\hat{v} \in \mathcal{V}_h$ and $\hat{z} \in \mathcal{Z}_h$. Then (IV.30) implies*

$$\|\hat{z}\|_\tau \leq \frac{1}{\sqrt{1-\delta^2}}\, \|\bar{v} - v\|_\tau\,, \quad \text{for all } v \in \mathcal{V}_h\,. \tag{IV.49}$$

Proof. We use (IV.30) to obtain directly for all $v \in \mathcal{V}_h$

$$\begin{aligned} \|\bar{v} - v\|_\tau^2 &= a_\tau(\hat{v} - v + \bar{v} - \hat{v}, \hat{v} - v + \bar{v} - \hat{v}) \\ &\geq \|\hat{v} - v\|_\tau^2 + \|\hat{z}\|_\tau^2 - 2\delta\, \|\hat{v} - v\|_\tau \|\hat{z}\|_\tau \\ &= \left(\|\hat{v} - v\|_\tau - \delta\, \|\hat{z}\|_\tau\right)^2 + (1-\delta^2)\, \|\hat{z}\|_\tau^2 \\ &\geq (1-\delta^2)\, \|\hat{z}\|_\tau^2\,, \end{aligned}$$

showing that the statement is valid. □

Proof of Theorem IV.1

First we mention that r_{n1} is only a function of u_n. It does not depend on any stage value. From the definitions of $E_{h,n+1}$ and u_{n+1}, we derive after straightforward calculation

$$b_n(E_{h,n+1}, \phi) = b_n(u_{n+1} - u_{h,n+1}, \phi)\,, \quad \phi \in \mathcal{Z}_h\,. \tag{IV.50}$$

Let $\phi = E_{h,n+1}$. Then, using (IV.21) and (IV.22), we estimate

$$\begin{aligned} \mu_b\, \|E_{h,n+1}\|_\tau^2 &\leq b_n(u_{n+1} - u_{h,n+1}, E_{h,n+1}) \\ &\leq M_b\, \|u_{n+1} - u_{h,n+1}\|_\tau \|E_{h,n+1}\|_\tau\,. \end{aligned}$$

Thus, we have the first inequality in (IV.35).

Analogously to (IV.50), we get with the definition of $\bar{u}_{h,n+1}$

$$b_n(\bar{u}_{h,n+1} - u_{h,n+1}, \phi) = b_n(E_{h,n+1}, \phi)\,, \quad \phi \in \mathcal{Z}_h\,. \tag{IV.51}$$

Let $\bar{u}_{h,n+1} = \hat{u}_{h,n+1} + \hat{E}_{h,n+1}$, where $\hat{u}_{h,n+1} \in \mathcal{V}_h$ and $\hat{E}_{h,n+1} \in \mathcal{Z}_h$. Then, using (IV.51) with $\phi = \hat{E}_{h,n+1}$ and taking into account $b_n(\bar{u}_{h,n+1} - u_{h,n+1}, \phi) = 0$ for all $\phi \in \mathcal{V}_h$, we obtain with (IV.21) and (IV.22)

$$
\begin{aligned}
\mu_b \, \|\bar{u}_{h,n+1} - u_{h,n+1}\|_\tau^2 &\leq b_n(\bar{u}_{h,n+1} - u_{h,n+1}, \bar{u}_{h,n+1} - u_{h,n+1}) \\
&= b_n(\bar{u}_{h,n+1} - u_{h,n+1}, \hat{u}_{h,n+1} - u_{h,n+1} + \hat{E}_{h,n+1}) \\
&= b_n(E_{h,n+1}, \hat{E}_{h,n+1}) \leq M_b \, \|E_{h,n+1}\|_\tau \|\hat{E}_{h,n+1}\|_\tau \, .
\end{aligned}
$$

To complete our proof, we must estimate $\|\hat{E}_{h,n+1}\|_\tau$. Applying Lemma IV.2 with $\bar{v} = \bar{u}_{h,n+1}$, we have

$$
\|\hat{E}_{h,n+1}\|_\tau \leq \frac{1}{\sqrt{1-\delta^2}} \, \|\bar{u}_{h,n+1} - u_{h,n+1}\|_\tau \, .
$$

Hence altogether

$$
\mu_b \, \|\bar{u}_{h,n+1} - u_{h,n+1}\|_\tau \leq \frac{M_b}{\sqrt{1-\delta^2}} \, \|E_{h,n+1}\|_\tau \, .
$$

Finally, we apply the right–hand inequality in (IV.31) to get the second bound in (IV.35). □

Proof of Theorem IV.2

a) We first prove the lower bound. The definitions (IV.32) in conjunction with (IV.33), (IV.34), and (IV.24) yield for all $\phi \in \mathcal{Z}_h$

$$
\begin{aligned}
b_n(E_{h,n+1}, \phi) &= b_n(E_{h,n0} + \tau \sum_{i=1}^{s} b_i E_{h,ni}, \phi) \\
&= b_n(u_n - u_{h,n+1}, \phi) + \tau \sum_{i=1}^{s} b_i \big\langle r_{ni}(K'_{h,n} + E_{h,n}), \phi \big\rangle \, .
\end{aligned}
\tag{IV.52}
$$

From (IV.16) and (IV.18) we get

$$
\tau \sum_{i=1}^{s} b_i \big\langle r_{ni}(K'_n), \phi \big\rangle - b_n(u_{n+1} - u_n, \phi) = 0 \, , \quad \phi \in \mathcal{Z}_h \subset \mathcal{V} \, . \tag{IV.53}
$$

Add both equalities to obtain for all $\phi \in \mathcal{Z}_h$

$$
b_n(E_{h,n+1}, \phi) = b_n(u_{n+1} - u_{h,n+1}, \phi) + \tau \sum_{i=1}^{s} b_i \big\langle r_{ni}(K'_{h,n} + E_{h,n}) - r_{ni}(K'_n), \phi \big\rangle \, .
$$

Now let $\phi = E_{h,n+1}$. Then, using (IV.21) and (IV.22), we estimate

$$
\mu_b \, \|E_{h,n+1}\|_\tau^2 \leq b_n(E_{h,n+1}, E_{h,n+1})
$$

$$\begin{aligned}
&= b_n(u_{n+1} - u_{h,n+1}, E_{h,n+1}) \\
&\quad + \tau \sum_{i=1}^{s} b_i \big\langle r_{ni}(K'_{h,n} + E_{h,n}) - r_{ni}(K'_n), E_{h,n+1} \big\rangle \\
&\leq M_b \, \|u_{n+1} - u_{h,n+1}\|_\tau \|E_{h,n+1}\|_\tau \\
&\quad + \sqrt{\tau} \left\| \sum_{i=1}^{s} b_i \big(r_{ni}(K'_{h,n} + E_{h,n}) - r_{ni}(K'_n) \big) \right\|_* \|E_{h,n+1}\|_\tau \,.
\end{aligned}$$

This inequality yields the first statement of Theorem IV.2.

b) To show the upper bound, we take (IV.52) and (IV.26)–(IV.28) to obtain for all $\phi \in \mathcal{Z}_h$

$$b_n(\bar{u}_{h,n+1} - u_{h,n+1} - E_{h,n+1}, \phi) = \tau \sum_{i=1}^{s} b_i \big\langle r_{ni}(\bar{K}'_{h,n}) - r_{ni}(K'_{h,n} + E_{h,n}), \phi \big\rangle \,. \tag{IV.54}$$

Recalling $\bar{u}_{h,n+1} = \hat{u}_{h,n+1} + \hat{E}_{h,n+1}$, where $\hat{u}_{h,n+1} \in \mathcal{V}_h$ and $\hat{E}_{h,n+1} \in \mathcal{Z}_h$, and setting $\phi = \hat{E}_{h,n+1}$ in (IV.54), we estimate

$$\mu_b \, \|\bar{u}_{h,n+1} - u_{h,n+1}\|_\tau^2 \leq b_n(\bar{u}_{h,n+1} - u_{h,n+1}, \bar{u}_{h,n+1} - u_{h,n+1})$$

$$\begin{aligned}
&= b_n(\bar{u}_{h,n+1} - u_{h,n+1}, \hat{u}_{h,n+1} - u_{h,n+1} + \hat{E}_{h,n+1}) \\
&= b_n(\bar{u}_{h,n+1} - u_{h,n+1}, \hat{u}_{h,n+1} - u_{h,n+1}) + b_n(E_{h,n+1}, \hat{E}_{h,n+1}) \\
&\quad + \tau \sum_{i=1}^{s} b_i \big\langle r_{ni}(\bar{K}'_{h,n}) - r_{ni}(K'_{h,n} + E_{h,n}), \hat{E}_{h,n+1} \big\rangle \\
&\leq M_b \left(\|E_{h,n+1}\|_\tau \|\hat{E}_{h,n+1}\|_\tau + \|\hat{u}_{h,n+1} - u_{h,n+1}\|_\tau \|\bar{u}_{h,n+1} - u_{h,n+1}\|_\tau \right) \\
&\quad + \sqrt{\tau} \left\| \sum_{i=1}^{s} b_i \big(r_{ni}(\bar{K}'_{h,n}) - r_{ni}(K'_{h,n} + E_{h,n}) \big) \right\|_* \|\hat{E}_{h,n+1}\|_\tau \,.
\end{aligned}$$

Applying Lemma IV.2 with $\bar{v} = \bar{u}_{h,n+1}$, we have

$$\|\hat{E}_{h,n+1}\|_\tau \leq \frac{1}{\sqrt{1-\delta^2}} \, \|\bar{u}_{h,n+1} - u_{h,n+1}\|_\tau \,.$$

Hence

$$\begin{aligned}
\mu_b \, \|\bar{u}_{h,n+1} - u_{h,n+1}\|_\tau \quad &\leq \quad \frac{M_b}{\sqrt{1-\delta^2}} \, \|E_{h,n+1}\|_\tau + M_b \, \|\hat{u}_{h,n+1} - u_{h,n+1}\|_\tau + \\
&\qquad \frac{\sqrt{\tau}}{\sqrt{1-\delta^2}} \left\| \sum_{i=1}^{s} b_i \big(r_{ni}(\bar{K}'_{h,n}) - r_{ni}(K'_{h,n} + E_{h,n}) \big) \right\|_* .
\end{aligned}$$

Finally, we apply the right–hand inequality in (IV.31) to verify the upper bound.
$\square$

V

Towards an Effective Code: Practical Issues

In this chapter, we discuss some practical issues which are useful for the implementation of our adaptive strategies. Efficient coding of Rosenbrock methods and dynamic multilevel techniques in two and three dimensions are addressed. A code KARDOS has been developed which allows one to accurately solve time–dependent systems of partial differential equations.

§1. Implementation of Rosenbrock Methods

Usually, Rosenbrock methods are not implemented in the form (II.18). The matrix–vector multiplication $A(t_n, u_n)\sum_j \gamma_{ij} K'_{nj}$ can be avoided by a simple transformation as suggested by several authors [152, 80, 130]. Introducing new variables

$$U_{ni} = \tau \sum_{j=1}^{i} \gamma_{ij} K'_{nj}\,, \quad i = 1, \ldots, s\,,$$

and defining the matrix $\Gamma = (\gamma_{ij})_{i,j=1}^{s}$ we derive

$$\begin{aligned} \left(\frac{1}{\tau\gamma} I + A(t_n, u_n)\right) U_{ni} &= F(t_n + \alpha_i \tau, u_n + \sum_{j=1}^{i-1} a_{ij}\, U_{nj}) + \sum_{j=1}^{i-1} \frac{c_{ij}}{\tau}\, U_{nj} \\ &\quad + \tau\gamma_i F_t(t_n, u_n), \\ u_{n+1} &= u_n + \sum_{i=1}^{s} m_i\, U_{ni}\,, \end{aligned} \tag{V.1}$$

where

$$(a_{ij})_{i,j=1}^{s} = (\alpha_{ij})_{i,j=1}^{s}\, \Gamma^{-1}\,, \quad (c_{ij})_{i,j=1}^{s} = \mathrm{diag}(\gamma^{-1}, \ldots, \gamma^{-1}) - \Gamma^{-1}\,,$$

$$(m_1, \ldots, m_s) = (b_1, \ldots, b_s)\, \Gamma^{-1}\,.$$

Remember our assumption $\gamma_{ii} = \gamma > 0$ which guarantees that Γ is invertible.

Remark 1. Rosenbrock methods can also be applied to implicit PDEs of the form

$$H(t,u)\partial_t u = F(t,u)\,,$$

where the matrix $H(t,u)$ may be singular. Rewriting this system as

$$\partial_t u = z\,, \quad 0 = F(t,u) - H(t,u)z\,,$$

we formally get a differential–algebraic system which can be attacked by Rosenbrock schemes satisfying additional algebraic order conditions [116, 96, 72]. One of the most popular solvers within this class is Rodas [72] being also "stiffly accurate" (see Appendix B§3).

§2. Implementation of Multilevel Finite Element Methods

The Rosenbrock time integration scheme (V.1) generates a sequence of linear elliptic problems. In the spirit of full adaptivity these stationary problems are solved by a multilevel finite element method (MFEM) as implemented in the Kaskade-toolbox developed at the Konrad–Zuse–Centre in Berlin [54]. The general principle of the multilevel technique consists of replacing the solution space by a sequence of discrete spaces with successively increasing dimensions to improve the approximation property. Starting with an approximation $u_{h,n+1}^{(0)}$ of the solution at $t=t_{n+1}$, we construct a sequence of improved spatial meshes

$$\mathcal{T}_{n+1}^{0} \subset \mathcal{T}_{n+1}^{1} \subset \ldots \subset \mathcal{T}_{n+1}^{m_{n+1}}\,,$$

and of corresponding nested FE–spaces

$$\mathcal{V}_h^{(0)} \subset \mathcal{V}_h^{(1)} \subset \ldots \subset \mathcal{V}_h^{(m_{n+1})}\,,$$

until a prescribed tolerance, say TOL_x, is reached. More precisely, the multilevel process comes to an end if for a certain number m_{n+1} the approximate error $E_{h,n+1}^{m_{n+1}}$ of the finite element solution $u_{h,n+1}^{(m_{n+1})}$ fulfills

$$|||E_{h,n+1}^{m_{n+1}}||| \le TOL_x\,,$$

where $|||\cdot|||$ denotes a specially weighted norm which will be defined later. Such an MFEM requires the specification of four modules: the finite element assembly, the estimation technique for the error in space, the mesh refinement strategy, and last but not least the solver of the resulting linear equations.

In Kaskade conforming FE–discretizations without slave nodes on intervals (1D), triangular grids (2D), and tetrahedral grids (3D) are provided. Allowing the use of highly unstructured grids, it is a flexible tool to handle complex geometries in higher spatial dimensions.

Once the approximate stage values $K'_{h,ni} \in \mathcal{V}_h^{(l)}, i = 1, \ldots, s$, have been computed on $\mathcal{T}_{n+1}^l$ for some l, the a posteriori error estimator $E_{h,n+1}^l$ developed in Chapter IV.§2 can be used to give specific assessment of the error distribution. Clearly, new grid points should be placed in those regions only where the current precision is insufficient. For this procedure it is required that the spatial discretization error can be estimated locally. We define for any finite element $K \in \mathcal{T}_{n+1}^l$ the local quantity

$$\eta_K = |||E_{h,n+1}^l|||_K \,.$$

These quantities are then used to judge the quality of the underlying discretization in the element K. In a next step a set R_l of elements which have to be refined are selected. There are several selection rules available [144, 77], e.g., R_l consists of all elements having an error estimator η_K larger than a local error barrier η_{bar} usually defined by $\eta_{bar} := \gamma \max_K \eta_K, 0 < \gamma < 1$. After choosing R_l we are led to a robust and stable refinement strategy. There are no specific problems in 1D, an interval is divided into two subintervals. Sometimes further refinement known as mass balancing is necessary to avoid large ratios of neighbouring element sizes.

In 2D all triangles of R_l are usually refined into four congruent triangles (regular or "red" refinement). Applied in several iterations for $l = 0, \ldots, m_{n+1} - 1$, this technique equilibrates the local error over the whole mesh and improves the finite element solution. It is standard and used in the PLTMG–package [17] and in KASKADE as well. To ensure that the new triangulation $\mathcal{T}_{n+1}^{l+1}$ does not possess slave nodes, triangles with one refined edge are subdivided into two triangles (irregular or "green" refinement), and those with two or three refined edges are refined "red" (see Fig. V.1). Recall that the finite element discretization error grows when the maximum angle tends to π [8]. Moreover, since the condition number of the stiffness matrix increases like $1/\sin\alpha$ [58], where α is the minimum angle, it is important to bound the angles away from 0 and π. Pure "red" refinement guarantees this immediately as the angles in each $\mathcal{T}_{n+1}^l$ are congruent to those in the initial grid $\mathcal{T}_{n+1}^0$. In order to avoid degeneration due to repeated "green" refinement, the "green" closure of each triangulation is removed before refining. Consequently, the refinement process is stable in the sense that the ratio of the diameter $diam(K)$ and the radius of the largest interior ball $\rho(K)$ remains uniformly bounded, i.e.,

$$diam(K)/\rho(K) \leq C \qquad \text{for all } K \in \mathcal{T}_{n+1}^l, l = 0, \ldots, m_{n+1}, \tag{V.2}$$

where the positive constant C is independent of l.

The regular refinement has been successfully extended to 3D by various authors (cf. [154, 26, 67, 15]). Connecting the midpoints of the edges of a given tetrahedron, we get four new tetrahedra corresponding to the vertices and one octahedron which has to be further refined (Fig. V.1). Each choice of the interior diagonal of the remaining octahedron yields a regular refinement into four additional tetrahedra. Altogether we obtain eight new tetrahedra. Unfortunately, stability in the sense of (V.2) cannot always be guaranteed if the interior diagonal is not properly selected. For instance, the permanent choice of the shortest diagonal can fail if the triangular faces of the initial grid have obtuse angles [154]. More robust strategies based on the history of the refinement process have been proposed in [26, 67]. Different "green" closures are used to obtain grids without slave nodes. Analogously to the 2D case, they are skipped at the beginning of further refinement. The refinement strategies by ZHANG [154] and BEY [26] are used in KASKADE and have been proven to be robust in many applications [33, 85].

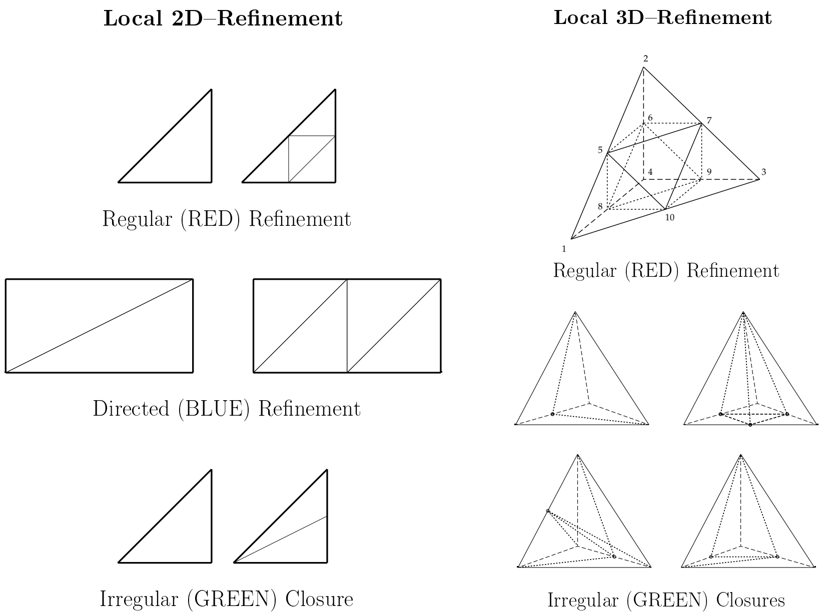


Figure V.1: Refinement in 2D and 3D.

An important question for the time integration process is how to choose $\mathcal{T}^0_{n+1}$, the starting grid for the computation of the new solution $u_{h,n+1}$. In the spirit of classical elliptic solvers often a time–fixed coarse grid $\mathcal{T}^0$ is taken. This approach avoids special coarsening algorithms which are in general more complicated to program than refinement strategies. However, in situations where the solution changes very slowly in time, the permanent start with a fixed mesh would be wasteful. But not only in this case one has the intuitive feeling that the new first mesh should be an approximation of the final mesh from the last time, i.e., $\mathcal{T}^0_{n+1} \sim \mathcal{T}^{m_n}_n$ (see Fig. V.2). Therefore, a considerably more efficient alternative is to remove degrees of freedom by analyzing the current solution to find regions where the errors are small. Of course, this requires data structures that allow grid enhancement and robust coarsening as well. Such a grid management is supported by the KASKADE–toolbox.

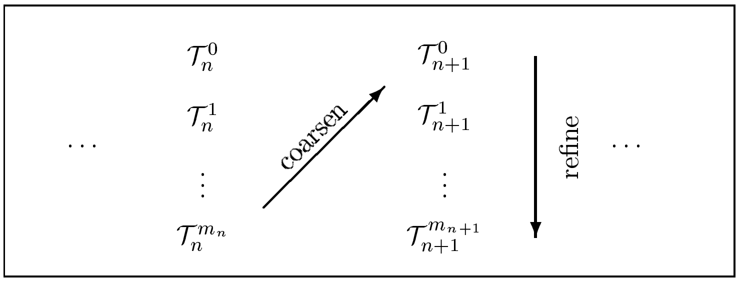


Figure V.2: Refinement and coarsening during time integration.

We first select candidates for coarsening employing the tree structure of the old mesh $\mathcal{T}_n^{m_n}$. An element is marked if it has a father which does not have a refined son. As usual a refined element is said to be the father of its subelements called sons. Fig. V.3 shows the result for a given simple one–dimensional tree structure. The extension of this strategy to more sophisticated tree structures in higher dimensions is straightforward. In a second step we take into account the local error behaviour of all marked elements. We identify regions of small errors by their η–values. Supposing an asymptotic behaviour of the form $\eta \sim ch^p$ as the characteristic mesh size h of the element K tends to zero, a simple prediction of the η–values after coarsening will be approximately

$$\eta_{pred} \sim c\,(\alpha h)^p \sim \alpha^p \eta\,.$$

Here, the value αh describes the characteristic mesh size after coarsening. For our refinement strategies we can use $\alpha = 2$. We remove the element K if η_{pred} does not exceed the local error barrier η_{bar} computed for the grid $\mathcal{T}_n^{m_n}$. Practical experiences have shown that the above described "trimming–tree" strategy works quite satisfactorily [84].

The linear systems arising from each of the grids $\mathcal{T}_{n+1}^l$ can be solved by direct or iterative methods (see [66] for a general overview). In 1D direct solvers are the method of choice. Band solvers such as DGBTRF/DGBTRS from LAPACK [7] and sparse solvers such as MA28 from HARWELL [52] benefit from the very regular sparse pattern of the system matrix. Since we have one and the same matrix for all stage problems (V.1), the matrix factorization has to be done only once. We have experienced that even in large one–dimensional applications the consumed CPU–time of a direct system solver amounts to only a very small part of the whole solution process. In higher dimensions the situation is completely different. Then iterative solvers perform considerably better with respect to CPU–time and memory requirements. Krylov subspace methods such as GMRES [121], CGS [133], and BICGSTAB [142], to name a few, are widely

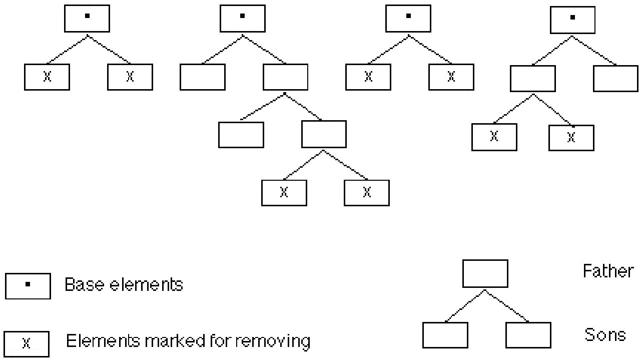


Figure V.3: One–dimensional tree structure used when removing elements.

used to solve the large linear systems which are in general non–symmetric and not positive definite. It is a well–known fact that differential operators and their discretizations give rise to "infinite" stiffness causing very ill–conditioned linear equations as pointed out by SHAMPINE ([130], p. 97). Thus, efficient preconditioning is often necessary to accelerate the convergence of the iterative solvers used. Since the best preconditioner is the matrix itself, various sorts of incomplete factorizations, e.g. ILU, are usually employed. It results in very few iterations, but on the other hand needs a large amount of memory to store the factors. When memory becomes more restrictive, one can use preconditioners working directly with the matrix, e.g., SSOR–preconditioning. Last but not least, the multilevel structure of the grid improvement provides good starting values for any refined grid. Thus, it speeds up the iterative process in a natural way. One can also look at the MFEM as a one–way multigrid method giving optimal convergence rates as demonstrated by BORNEMANN and DEUFLHARD [32]. Some practical observations can be found in Chapter VI.§3. Worthwhile to mention is that the door is also open to apply more sophisticated multigrid methods.

§3. KARDOS – an Accurate Adaptive PDE–Solver

As we started with coding our adaptive algorithm we have chosen the name KARDOS mnemotechnically for: KAskade Reaction–DiffusiOn System. Although nowadays KARDOS is used for a much larger problem class, we have decided not to change the name.

From the previous chapters one gets a first insight in the huge complexity required to code an adaptive method to solve PDEs efficiently. Working in the field of programming over a long period, it turns out that designing a code is sometimes not easier than developing an algorithm. Thus, it is proper for numerical analysis to address not only problems in approximation of functions and quality assessment of solutions, but also the ways in which numerical algorithms are implemented into a flexible and portable software structure on real hardware.

The idea of KARDOS is to combine the multilevel strategy of KASKADE with linearly implicit discretization methods in time. The modularization concept of KASKADE has proven to be flexible enough to allow the implementation of an outer time shell. Flexibility is always connected with a specification of interfaces. In [117] the need for low–level and high–level interface elements is discussed for the assembling procedure of KARDOS. There, a notification system and special dynamic construction of records are described to implement efficient mesh transfer operations between two different integration points in time.

A code is as flexible as it is possible to change one modul with an implementation of another technique. For example, it may be desirable to have more than one time integrator and one preconditioned iterative solver available. KARDOS consists of several exchangeable moduls: Rosenbrock solvers, direct and iterative methods, preconditioners, a posteriori error estimators, refinement strategies etc. We do not want to describe too many details of the implementation. For later use we shall focus on just two design aspects.

A problem that often turns up in practical computations is the different scales of the solution components in the PDE. Using a weighted root mean square norm,

$$|||\underline{v}|||_\omega = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{|||v_i|||^2}{\omega_i^2} \right)^{1/2}$$

for vector–valued functions $\underline{v} = (v_1, \ldots, v_n)^T$ and with weights

$$\omega_i = \mathrm{ATOL}_i + |||U_i||| \cdot \mathrm{RTOL}_i \,,$$

the components can be treated in an equally scaled way. Here, $U = (U_1, \ldots, U_n)^T$ should be a good approximation to the actual solution. The tolerances ATOL_i and RTOL_i have to be selected carefully to reflect accurately the scale of the problem (cf. [37], p. 131).

A second point is to balance the temporal and spatial discretization errors in order to keep the entire local error below a prescribed tolerance TOL. In [88] an adaptive strategy based on spectral information about the Jacobian is developed for 2nd order time integrators. In the general case, we set

$$TOL_t = TOL/2 \quad \text{and} \quad TOL_x = TOL/3 \,.$$

This heuristic of balancing the errors worked quite well for the problems we have solved with KARDOS.

VI

Illustrative Numerical Tests

We present some computational results to demonstrate that the theoretical order predictions given by Theorems III.1 and III.2 are indeed of interest for the numerical practice. The sharpness of the above convergence rates for the fully discretized schemes can be observed for two different nonlinear equations. The quality of the hierarchical error estimator in space is assessed in terms of the effectivity index. Our discussions in §1 and §2 are limited to the one–dimensional case where the separate study of temporal and spatial discretization errors is much easier than in higher dimensions. It allows also a direct comparison with similar studies given in [147, 103]. Finally, we look at the performance of our multilevel strategy which is applied to solve a combustion problem in different spatial dimensions. For this problem we observed multigrid complexity.

§1. Practical Convergence Observations

In this section we consider two nonlinear scalar parabolic equations in 1D which fit into our framework of Chapter III. For both equations smooth solutions are available for studying the error behaviour. The discussion includes three Rosenbrock methods listed in Table VI.1.

Method	stages	function calls	order	$A(\Theta)$–stable	$\lvert R(\infty)\rvert$	(II.21)	stiffly accurate
ROS2 Dekker, Verwer [44]	2	2	2	$\frac{\pi}{2}$	0	–	–
ROWDA3 Roche [116]	3	2	3	$\frac{\pi}{2}$	0	–	–
RODASP Steinebach [135]	6	6	4	$\frac{\pi}{2}$	0	+	+

Figure VI.1: List of Rosenbrock solvers.

All solvers are L–stable, a property which is strongly recommended for solving stiff problems. RODASP, a modification of the well–known RODAS [72], is the only one which satisfies condition (II.21). We note also that ROS2 may be used with an inexact Jacobian. Helpful information concerning the consistency and stability of Rosenbrock methods is summarized in Appendix B.

Example 1. Consider Burger's equation

$$\partial_t u - \nu \partial_x^2 u + u \partial_x u = 0\,,\; 0 < t \leq T = 1,\; \Omega = [0,1]\,, \tag{VI.1}$$

where initial and Dirichlet boundary conditions are chosen from the exact solution given by WHITHAM ([151], Chap. 4)

$$u(x,t) = \frac{0.1 r_1 + 0.5 r_2 + r_3}{r_1 + r_2 + r_3}\,,$$

with

$$r_1 = e^{-\frac{x-0.5}{20\nu} - \frac{99t}{400\nu}}\,,\; r_2 = e^{-\frac{x-0.5}{4\nu} - \frac{3t}{16\nu}}\,,\; r_3 = e^{-\frac{x-0.375}{2\nu}}\,.$$

This equation has also been used by VERWER [147] to study order reductions of diagonally implicit Runge–Kutta schemes. We choose the same value $\nu = 0.1$ to allow an additional comparison with results presented there.

Let us first focus on the temporal convergence behaviour of the chosen Rosenbrock solvers. To keep the spatial discretization error to an insignificant level we have used standard 4th order Lagrange finite elements and a uniform grid consisting of 8192 elements.

We get from (VI.1) by differentiation

$$A(u)v = -\nu \partial_x^2 v + u \partial_x v + \partial_x u v\,,$$

equipped with time–dependent Dirichlet boundary conditions. Hence, the domain $D(A^\alpha(t)) = D(A^\alpha(u(t)))$ is independent of t only for $\alpha < 1/4$ (see Example II.3). Therefore, the standard setting for second–order parabolic equations $\mathcal{H} = L^2(\Omega)$ and $\mathcal{V} = H^1(\Omega)$ is not suitable in this situation. Considering $A(t)$ with $D(A(t)) = H^1(\Omega)$ as unbounded on $\mathcal{H} = H^{-1}(\Omega)$, we get $\mathcal{V} = L^2(\Omega)$. Let $A_0(t)$ the same operator taken as unbounded on $L^2(\Omega)$. Then, $D(A^{1/2+\alpha}(t)) = D(A_0^\alpha(t))$ shows that we can use $\beta = 1/4 - \varepsilon$ to obtain from (II.20) an error estimate of the form

$$\left(\tau \sum_{n=0}^{N} \|\, \epsilon_n \|_{L^2(\Omega)}^2 \right)^{1/2} + \max_{0 \leq n \leq N} \|\epsilon_n\|_{H^{-1}(\Omega)} \;\leq\; O(\tau^{2.25-\varepsilon})\,,$$

neglecting the spatial discretization error. In Table VI.2 we have listed the global error $\|\underline{\epsilon}\|_{l^2_{N+1}(L^2(\Omega))}$, $\underline{\epsilon} = \{\epsilon_n\}_{n=0,\dots,N}$, $N\tau = T$, and the numerically observed temporal order of convergence

$$q_{num} = \log_2 \frac{\|\underline{\epsilon}\|_{l^2_{N+1}(L^2(\Omega))}}{\|\underline{\epsilon}\|_{l^2_{2N+1}(L^2(\Omega))}}\,.$$

The predicted order reduction of the 3rd order ROWDA3 to $p=2.25$ can be seen clearly. Recall that this method does not satisfy condition (II.21). By the way we mention that the additional algebraic order of ROWDA3 does not pay off here as one could expect for Dirichlet boundary conditions. The same convergence rate was observed for the implicit Runge–Kutta schemes SDIRK2 and SDIRK3 in [147] where $\|\epsilon_N\|_{L^2(\Omega)}$ was measured. In contrast, RODASP satisfies (II.21) and reaches $q_{num} \approx 3.75$. The last values in Table VI.2 and also in Table VI.3 were omitted for reason of insufficient accuracy even using double precision. Full order $p=4$ as observed in [108] is not attainable because the conditions for linear parabolic equations are not sufficient to raise the order further [95]. The values for ROS2 reveal $q_{num} \to 2$ for $\tau \to 0$ which corresponds to the theoretical value given in Theorem III.1.

Unless our theory is not applicable to obtain global H^1–errors it should be worthwhile to have a look at what happens for the error in this norm from the numerical point of view. The corresponding results are given in Table VI.3. The observed temporal orders are further decreased. The new order $q_{num} = 1.75$ of ROWDA3 clearly shows up and seems to be also true for ROS2. An interesting fact is that this value can be obtained by a formal application of (II.20) with $\mathcal{H} = L^2(\Omega)$ and $\mathcal{V} = H^1(\Omega)$. In fact, the discussion in Example II.3 shows $\beta = -1/4 - \varepsilon$ with arbitrary small $\varepsilon > 0$ which yields directly the observed order $p=1.75$.

Let us now turn to the spatial discretization error. To measure the convergence rates in space we have solved problem (VI.1) with standard linear ($q=1$) and quadratic ($q = 2$) Lagrange finite elements on different meshes providing a small time step $\tau = 10^{-4}$. Table VI.4 nicely shows that the observed orders of convergence correspond to the theoretical values well–known for the approximation property (III.1) of the finite element spaces. Due to the high spatial regularity of the solution, we get order $q + 1$ for the L^2–norm and order q for the H^1–norm. The results are nearly identical for all time integrators under consideration where, of course, we need a smaller time step for ROS2 to avoid dominating temporal errors.

Remark 1. The order reduction in the case of inhomogeneous Dirichlet boundary conditions becomes less severe if we reformulate (VI.1) in terms of

$$v(x,t) = u(x,t) - xu(1,t) - (1-x)u(0,t)\,.$$

The resulting equations for v are then solved with homogeneous Dirichlet boundary conditions. Since Rosenbrock methods are not invariant with respect to this transformation, we can use $\beta = 3/4 - \varepsilon$ in (II.20) employing $\mathcal{H} = L^2(\Omega)$

	ROS2		ROWDA3		RODASP	
τ	$\|\|\underline{\epsilon}\|\|_{l^2(L^2)}$	q_{num}	$\|\|\underline{\epsilon}\|\|_{l^2(L^2)}$	q_{num}	$\|\|\underline{\epsilon}\|\|_{l^2(L^2)}$	q_{num}
$\frac{1}{10}$	8.6_{10}^{-4}		2.7_{10}^{-5}		2.0_{10}^{-7}	
$\frac{1}{20}$	3.5_{10}^{-4}	1.31	5.2_{10}^{-6}	2.38	1.4_{10}^{-8}	3.79
$\frac{1}{40}$	1.2_{10}^{-4}	1.58	1.0_{10}^{-6}	2.33	1.1_{10}^{-9}	3.72
$\frac{1}{80}$	3.4_{10}^{-5}	1.76	2.1_{10}^{-7}	2.30		
$\frac{1}{160}$	9.4_{10}^{-6}	1.86	4.3_{10}^{-8}	2.27		
$\frac{1}{320}$	2.5_{10}^{-7}	1.92	9.0_{10}^{-9}	2.26		

Figure VI.2: Burger's equation with time–dependent Dirichlet boundary conditions. The observed temporal orders of convergence measured in the global L^2–norm correspond to the predicted values: $p = 2$ for ROS2, $p = 2.25$ for ROWDA3, and $p \geq 3$ for RODASP.

	ROS2		ROWDA3		RODASP	
τ	$\|\|\underline{\epsilon}\|\|_{l^2(H^1)}$	q_{num}	$\|\|\underline{\epsilon}\|\|_{l^2(H^1)}$	q_{num}	$\|\|\underline{\epsilon}\|\|_{l^2(H^1)}$	q_{num}
$\frac{1}{10}$	8.3_{10}^{-3}		4.1_{10}^{-4}		1.8_{10}^{-6}	
$\frac{1}{20}$	3.1_{10}^{-3}	1.44	1.2_{10}^{-4}	1.78	1.8_{10}^{-7}	3.32
$\frac{1}{40}$	1.0_{10}^{-3}	1.61	3.5_{10}^{-5}	1.76	2.0_{10}^{-8}	3.16
$\frac{1}{80}$	3.1_{10}^{-4}	1.71	1.0_{10}^{-5}	1.75		
$\frac{1}{160}$	9.1_{10}^{-5}	1.75	3.1_{10}^{-6}	1.75		
$\frac{1}{320}$	2.7_{10}^{-5}	1.76	9.3_{10}^{-7}	1.75		

Figure VI.3: Burger's equation with time–dependent Dirichlet boundary conditions. The observed temporal orders of convergence measured in the global H^1–norm reveal $p = 1.75$ for ROS2 and ROWDA3, and $p \geq 3$ for RODASP.

and $\mathcal{V} = H^1(\Omega)$. This results in a convergence rate improved by one. Unfortunately, the above transformation is not always practicable for more complicated problems.

	ROWDA3				RODASP			
h	$\|\|\underline{\varepsilon}\|\|_{l^2(L^2)}$	q_{num}	$\|\|\underline{\varepsilon}\|\|_{l^2(H^1)}$	q_{num}	$\|\|\underline{\varepsilon}\|\|_{l^2(L^2)}$	q_{num}	$\|\|\underline{\varepsilon}\|\|_{l^2(H^1)}$	q_{num}
	q=1							
$\frac{1}{64}$	1.4_{10}^{-5}		2.8_{10}^{-4}		1.4_{10}^{-5}		2.8_{10}^{-4}	
$\frac{1}{128}$	3.5_{10}^{-6}	2.00	1.4_{10}^{-4}	1.00	3.5_{10}^{-6}	2.00	1.4_{10}^{-5}	1.00
$\frac{1}{256}$	8.9_{10}^{-7}	2.00	7.0_{10}^{-5}	1.00	8.9_{10}^{-7}	2.00	7.0_{10}^{-6}	1.00
	q=2							
$\frac{1}{64}$	4.6_{10}^{-8}		1.9_{10}^{-5}		4.6_{10}^{-8}		1.9_{10}^{-7}	
$\frac{1}{128}$	5.8_{10}^{-9}	3.00	4.8_{10}^{-6}	2.00	5.8_{10}^{-9}	3.00	4.8_{10}^{-8}	2.00
$\frac{1}{256}$	7.3_{10}^{-10}	2.98	1.2_{10}^{-6}	1.99	7.3_{10}^{-10}	2.98	1.2_{10}^{-6}	2.00

Figure VI.4: Burger's equation with time–dependent Dirichlet boundary conditions. The observed spatial orders of convergence for linear elements ($q=1$) and quadratic elements ($q=2$) correspond to the theoretical values $q+1$ for the global L^2–norm and q for the global H^1–norm.

Example 2. Consider the reaction–diffusion equation

$$\partial_t u - \partial_x^2 u = p_3(1-u^2) + 2p_2^2(u-u^3)\,,\; 0<t,\; \Omega=[-3,3]\,, \qquad \text{(VI.2)}$$

where initial conditions are chosen from the exact solution

$$u(x,t) = \tanh\left(p_2(x-p_1)+p_3 t\right).$$

This travelling wave problem was considered by NOWAK [106] to study moving–mesh strategies. Here, we want to investigate the case of homogeneous Neumann boundary conditions. Setting $p_1=0.05$, $p_2=p_3=6.0$, the analytical solution satisfies

$$\partial_n u(-3,t) \simeq 0 \quad \text{and} \quad \partial_n u(3,t) \simeq 0$$

for sufficiently small t. We have integrated problem (VI.2) with $T=0.1$ resolving the travelling solution front by a highly adaptive spatial mesh. Once again 4–th order Lagrange finite elements were applied.

The discussion in Example II.3 shows that, in the case of time–independent homogeneous Neumann boundary conditions, $\beta=5/4-\varepsilon$ with arbitrarily small $\varepsilon>0$ can be used in (II.20) for the standard setting $\mathcal{H}=L^2(\Omega)$ and $\mathcal{V}=$

	ROS2		ROWDA3		RODASP	
τ	$\|\|\underline{\epsilon}\|\|_{l^\infty(L^2)}$	q_{num}	$\|\|\underline{\epsilon}\|\|_{l^\infty(L^2)}$	q_{num}	$\|\|\underline{\epsilon}\|\|_{l^\infty(L^2)}$	q_{num}
$\frac{1}{200}$	1.4_{10}^{-4}		6.2_{10}^{-6}		3.7_{10}^{-7}	
$\frac{1}{400}$	6.6_{10}^{-5}	1.48	8.4_{10}^{-7}	2.87	1.7_{10}^{-8}	4.46
$\frac{1}{800}$	2.4_{10}^{-5}	1.70	1.2_{10}^{-7}	2.86	8.5_{10}^{-10}	4.32
$\frac{1}{1600}$	7.3_{10}^{-6}	1.83	1.6_{10}^{-8}	2.90		
$\frac{1}{3200}$	2.0_{10}^{-6}	1.91	2.0_{10}^{-9}	2.96		
$\frac{1}{6400}$	5.4_{10}^{-7}	1.95	2.4_{10}^{-10}	3.08		

Figure VI.5: Travelling wave equation with homogeneous Neumann boundary conditions. The observed temporal orders of convergence in the $C_t^0(L^2)$–norm correspond to the classical orders.

	ROS2		ROWDA3		RODASP	
τ	$\|\|\underline{\epsilon}\|\|_{l^2(H^1)}$	q_{num}	$\|\|\underline{\epsilon}\|\|_{l^2(H^1)}$	q_{num}	$\|\|\underline{\epsilon}\|\|_{l^2(H^1)}$	q_{num}
$\frac{1}{200}$	2.9_{10}^{-4}		1.9_{10}^{-5}		1.6_{10}^{-6}	
$\frac{1}{400}$	1.5_{10}^{-4}	0.93	3.1_{10}^{-6}	2.61	1.1_{10}^{-7}	3.98
$\frac{1}{800}$	5.9_{10}^{-5}	1.38	4.6_{10}^{-7}	2.73	7.3_{10}^{-9}	3.91
$\frac{1}{1600}$	1.9_{10}^{-5}	1.63	6.5_{10}^{-8}	2.83		
$\frac{1}{3200}$	5.5_{10}^{-6}	1.79	8.7_{10}^{-9}	2.90		
$\frac{1}{6400}$	1.5_{10}^{-6}	1.88	1.1_{10}^{-9}	2.97		

Figure VI.6: Travelling wave equation with homogeneous Neumann boundary conditions. The observed temporal orders of convergence in the global H^1–norm correspond to the classical orders.

$H^1(\Omega)$. Indeed, the classical order is obtained by all methods. The errors and the observed order of convergence are displayed in Table VI.5 and Table VI.6. We observe that the L^2–convergence is slightly better than the global H^1–convergence.

Remark 2. We get the same results considering Dirichlet boundary conditions which are taken from the exact solution. At a first glance this could be surprising, but the solution is nearly constant in a neighbourhood of the boundary. Thus, the boundary conditions do not affect the solution process. This situation often occurs in practical computations when boundary conditions dominate the solution only at the very beginning.

Remark 3. We have also tested the 3rd order ROS3 and RODAS3 proposed by SANDU, VERWER et al. [122]. Both methods do not satisfy condition (II.21). Generally, we can state that their performance was very close to that of ROWDA3.

§2. Accuracy of the Spatial A Posteriori Error Estimate

In order to illustrate the accuracy of the spatial error estimate investigated in Chapter IV.§2, we consider once again problem (VI.2). This problem is solved for one appropriate time step with different meshes. We limit our discussion to linear finite elements. The subspace Z_h consists of all quadratic bump functions having support in one element only. Thus the proposed error estimator (IV.32) can be computed element–by–element by solving one linear equation. Let us consider the third–order ROWDA3. To ensure that the temporal error is dominated by the spatial error, we first perform one time step with ROS2 in a fully adaptive manner such that both errors are approximately balanced with respect to the L^2–norm. Employing different tolerances we get a sequence of time steps and the corresponding number of points shown in the first two columns of Table VI.7. Using then ROWDA3 which has higher order in time, we are able to compare estimated and exact spatial errors. If the time steps were too small, one would assess the projection error of the initial solution only.

The quality of estimators is usually measured by the effectivity index Θ defined by

$$\Theta = \frac{\text{Estimated Error}}{\text{Exact Error}} .$$

Table VI.7 shows the exact L^2– and H^1–errors together with the effectivity indices computed at $t=\tau$.

The results indicate that the hierarchical basis error estimator produces excellent estimates of the local error as long as the temporal error is dominated by the spatial error. Similar results are valid for the other Rosenbrock solvers.

§3. Performance of the Multilevel Strategy

We conclude this chapter by presenting some results for the performance of KARDOS measured on a SUN-SPARC ULTRA2. Let us consider the combustion problem

$$\begin{aligned} \partial_t C - \nabla^2 C &= -D\,C\,e^{-\delta/T}\,, \\ Le\,\partial_t T - \nabla^2 T &= \alpha D\,C\,e^{-\delta/T}\,, \quad t>0, \end{aligned}$$

τ	N	$\|\|e\|\|_{L^2}$	Θ_{L^2}	$\|\|e\|\|_{H^1}$	Θ_{H^1}
1.6_{10}^{-2}	31	2.1_{10}^{-3}	1.05	1.6_{10}^{-1}	1.01
4.1_{10}^{-3}	89	3.1_{10}^{-4}	0.98	5.2_{10}^{-2}	1.00
1.3_{10}^{-3}	238	2.9_{10}^{-5}	1.00	1.8_{10}^{-2}	1.00
5.6_{10}^{-4}	789	3.3_{10}^{-6}	1.00	5.4_{10}^{-3}	1.00
1.6_{10}^{-4}	3042	1.9_{10}^{-7}	1.00	1.3_{10}^{-3}	1.00

Figure VI.7: L^2– and H^1–errors and effectivity indices for Example 2 with time step τ and $N-1$ adaptively chosen linear elements.

where T is the dimensionless temperature and C the concentration of a reactant. We solve the system for different spatial dimensions on the domains

$$\Omega_d = \{x = (x_1, \ldots, x_d) \in \mathbb{R}^d,\ 0 < x_1, \ldots, x_d < 1\}\,, \quad d = 1, 2, 3\,,$$

taking the special combustion numbers $Le = 0.9$, $\delta = 20$, $\alpha = 1$, and $D = 5 \cdot e^{\delta}/(\alpha\delta)$. The initial and boundary conditions are

$$\begin{aligned} C(x,0) = T(x,0) = 1\,, &\qquad x \in \Omega_d\,, \\ \nabla C(x,t) \cdot n = \nabla T(x,t) \cdot n = 0\,, &\qquad \text{on } x_i = 0,\ i = 1, \ldots, d\,, \\ C(x,t) = T(x,t) = 1\,, &\qquad \text{on } x_i = 1,\ i = 1, \ldots, d\,. \end{aligned}$$

This system describes a one–step reaction in the presence of Arrhenius chemistry. The one–dimensional version of this problem was investigated via activation–energy asymptotics by KAPILA [78]. Initially, the temperature increases slowly during an induction period with relatively weak reaction. Induction is followed by an extremely rapid development and growth of a localized hot spot at $0 \in \mathbb{R}^d$. A sharply focused temperature region appears in which the concentration of the reactant is rapidly depleted. Then the reaction front propagates through the domain. The problem is complicated enough to study the performance of our adaptive multilevel solver KARDOS. We have chosen ROWDA3 and standard linear finite elements. The linear equations were solved by BICGSTAB preconditioned by ILU. Several runs with different tolerances were performed.

In Tables VI.8–VI.10 we can see the number of time steps, the total number of spatial discretization points over all time, N_{tot}, the corresponding averaged number of points, the computing time in seconds, and two different ratios which measure the complexity index of the algorithm. In the one–dimensional case an $O(N)$–complexity is visible. This corresponds to the observation that the main part of the computing time is consumed by the assembling routines. In two and three spatial dimension the algorithm tends to $O(N \log N)$–complexity. Thus, we get multigrid complexity of KARDOS. Considering the absolute values of CPU/N_{tot} for each dimension, we derive approximately a ratio 1 : 3 : 27, showing that the effort per degree of freedom grows rapidly when the dimension is increased.

RUN	♯ Timesteps	N_{tot}	$\frac{N_{tot}}{\sharp Timesteps}$	CPU[sec]	$\frac{10^4 \cdot \text{CPU}}{N_{tot}}$	$\frac{10^4 \cdot \text{CPU}}{N_{tot} \log N_{tot}}$
1	224	50 314	225	43	8.55	1.82
2	300	110 404	368	98	8.88	1.76
3	380	197 672	521	169	8.55	1.61
4	670	769 810	1 149	613	7.96	1.35

Figure VI.8: Performance of KARDOS for 1D–combustion with $t \in [0, 0.227]$.

RUN	♯ Timesteps	N_{tot}	$\frac{N_{tot}}{\sharp Timesteps}$	CPU[sec]	$\frac{10^3 \cdot \text{CPU}}{N_{tot}}$	$\frac{10^4 \cdot \text{CPU}}{N_{tot} \log N_{tot}}$
1	76	94 302	1 241	238	2.52	5.07
2	82	148 608	1 812	379	2.55	4.93
3	95	281 582	2 964	763	2.71	4.97
4	124	843 304	6 801	2 442	2.90	4.89

Figure VI.9: Performance of KARDOS for 2D–combustion with $t \in [0, 0.278]$.

RUN	♯ Timesteps	N_{tot}	$\frac{N_{tot}}{\sharp Timesteps}$	CPU[sec]	$\frac{10^2 \cdot \text{CPU}}{N_{tot}}$	$\frac{10^3 \cdot \text{CPU}}{N_{tot} \log N_{tot}}$
1	21	176 442	8 402	3 775	2.14	4.08
2	28	747 516	26 697	16 736	2.24	3.81
3	33	1 486 551	45 047	34 319	2.31	3.74
4	40	2 404 560	60 114	58 010	2.41	3.78

Figure VI.10: Performance of KARDOS for 3D–combustion with $t \in [0, 0.297]$.

VII

Applications from Computational Sciences

Having developed a rather complete theory for the efficient solution of nonlinear parabolic problems, we present a series of real–life applications in different spatial dimensions. Clearly, the acceptance of numerical algorithms can benefit from demonstrating that they work robustly and safely over a wide range of practically relevant problems. Starting with practical computations one will be quickly forced to realize that the understanding and creation of good models is just as challenging as proving a deep theorem. Therefore, it is not surprising that all of the following simulations were done in joint works with specialists of the considered areas over a period of a couple of years.

All models are explained and parameters are completely described, except for the bubble reactor where the data are protected by the BASF company. This should enable interested readers to make their own experiences. We mention that not all considered problems fit directly into our theoretical framework. Nevertheless, the proposed adaptive approach allows straightforward extensions for algebraic equations, pure ODEs, and equations without diffusion too.

The first problem is the synthesis process of two different chemicals in a bubble reactor. The investigations were done within the joint project "Development of simulation algorithms for reaction processes" of the BASF Research Institute and the Konrad–Zuse Centre. The two–film model used was developed by W. Ruppel [120].

The two–dimensional combustion problems which follow were jointly investigated with J. Fröhlich [59]. Propagating thin flame fronts demand dynamic spatial adaptivity when they have to be resolved numerically. Two reaction–diffusion systems are considered: laminar flames travelling through an obstacle and a solid–solid alloying reaction in a uniformly packed reactor. It should be noted that a large number of phenomena in biology, ecology, physics, and engineering are governed by equations of reaction–diffusion type.

The third problem modelling the diffusion step of semiconductor device fabrication was studied in a joint work with W. Merz [86, 87]. The simulation was based on an extensive model involving an enormous list of empirical parameters. Nevertheless, we succeeded in resolving the main effects of the anomalous diffusion of phosphorus known from the literature. More detailed information about the fundamentals behind the pair–diffusion model used and an analytical treatment can be found in [101].

In the fourth application the proposed adaptive method is incorporated in an optimization process specially designed for regional hyperthermia of deep seated tumors in order to achieve a desired steady–state temperature distribution. A nonlinear three–dimensional heat–transfer model based on temperature–dependent perfusion is applied to predict the temperature. This study was jointly done with B. ERDMANN and M. SEEBASS [85]. It is part of the substantial activities at the Konrad–Zuse Centre and the Klinikum Rudolf Virchow in Berlin to improve this cancer therapy (cf. [21]). The necessary electric field data were provided by R. BECK.

Last but not least we mention that all simulations were performed with the programming package KARDOS explained in Chapter V.§3. We used mainly the Rosenbrock solver ROWDA3 and solved the linear equations by the BICGSTAB–algorithm preconditioned by ILU or SSOR. Applying linear finite elements we measured the spatial errors in the space of quadratic functions. These ingredients worked quite well and gave satisfactory results.

§1. 1D: Two–Phase Bubble Reactor

Gas–fluid systems give rise to propagating phase boundaries changing their shape and size in time. In the following we consider a synthesis process of two gaseous chemicals A and B in a cylindrical bubble reactor filled with a catalytic fluid (see Fig. VII.1).

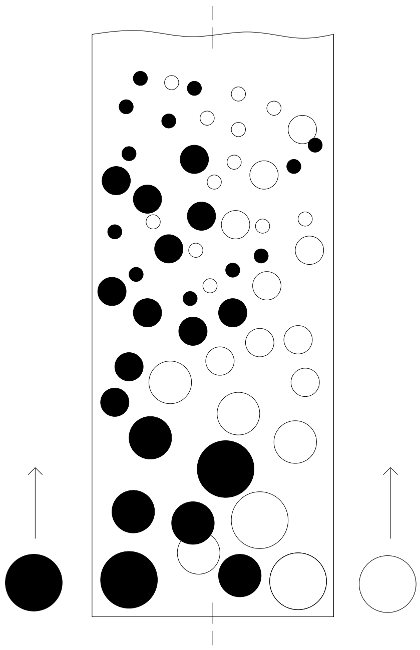
The bubbles stream in at the lower end of the reactor and rise to the top while dissolving and reacting with each other. The right proportions of such reactors depend, among other things, on the rising behaviour of the bubbles and specific reaction velocities. Therefore, modelling and simulation of the underlying two–phase system can provide engineers with useful knowledge necessary to construct economical plants.

A fully three–dimensional description of the synthesis process would become too complicated. We have used a one–dimensional two–film model developed by RUPPEL [120]. It is based mainly on the assumption that the interaction between the gas and the reactor fluid (bulk) takes place in very thin layers (films) with time–independent thickness (see Fig. VII.2). In the first film F_1 the chemical A dissolves into the bulk. From there it is transported very fast to the second film F_2 where reaction with chemical B takes place. As a result new chemicals C and D are produced causing further reactions.

Defining the assignment $(A,B,C,D,E,F,G) \to (u_1,\ldots,u_7)$ the model can be expressed by the following equations.

Diffusive process in F_1 only for the chemical A:

$$\begin{aligned} -D_1\frac{\partial^2 u_1}{\partial x^2} &= 0, \qquad x \in (\xi_1,\xi_2), \\ \frac{\beta_1}{\alpha_1 D_1}u_1(\xi_1) - \frac{\partial u_1}{\partial x}(\xi_1) &= c_2\frac{\beta_1}{D_1}, \qquad u_1(\xi_2^-) = u_1(\xi_2^+)\,. \end{aligned} \tag{VII.1}$$



$$
\begin{aligned}
A + B &\longrightarrow C + D \\
C + D &\longrightarrow E \\
C &\longrightarrow F + G
\end{aligned}
$$

Figure VII.1: Bubble reactor in section and reaction mechanism.

Transport of all the chemicals through the bulk:

$$
\begin{aligned}
&c_1 \frac{\partial u_i}{\partial t} = S_1(t) D_i \frac{\partial u_i}{\partial x}(\xi_2^-) + S_2(t) D_i \frac{\partial u_i}{\partial x}(0^+),\ x \in (\xi_2, 0), t > 0\,, \\
&u_1(\xi_2^+) = u_1(\xi_2^-),\ \ \frac{\partial u_i}{\partial x}(\xi_2^+) = 0,\ i \neq 1,\ \ u_i(0^-) = u_i(0^+)\,.
\end{aligned}
\tag{VII.2}
$$

Reaction and diffusion in F_2:

$$
\begin{aligned}
&-D_i \frac{\partial^2 u_i}{\partial x^2} = \sum_j k_{i,j} u_j u_i,\ x \in (0, \xi_3), t > 0. \\
&u_i(0^+) = u_i(0^-),\ \ \frac{\beta_2}{\alpha_2 D_2} u_2(\xi_3) + \frac{\partial u_2}{\partial x}(\xi_3) = c_3 \frac{\beta_2}{D_2}, \\
&\frac{\partial u_i}{\partial x}(\xi_3) = 0,\ i \neq 2,\ \ u_i(x, 0) = u_i^0.
\end{aligned}
\tag{VII.3}
$$

Here, D_i and β_i denote the diffusion and the coupling coefficient of the i–th component, and α_i represents the Henry coefficient. The specific exchange areas S_1 and S_2 depend nonlinearly on the decreasing bubble radii $r_1(t)$ and $r_2(t)$.

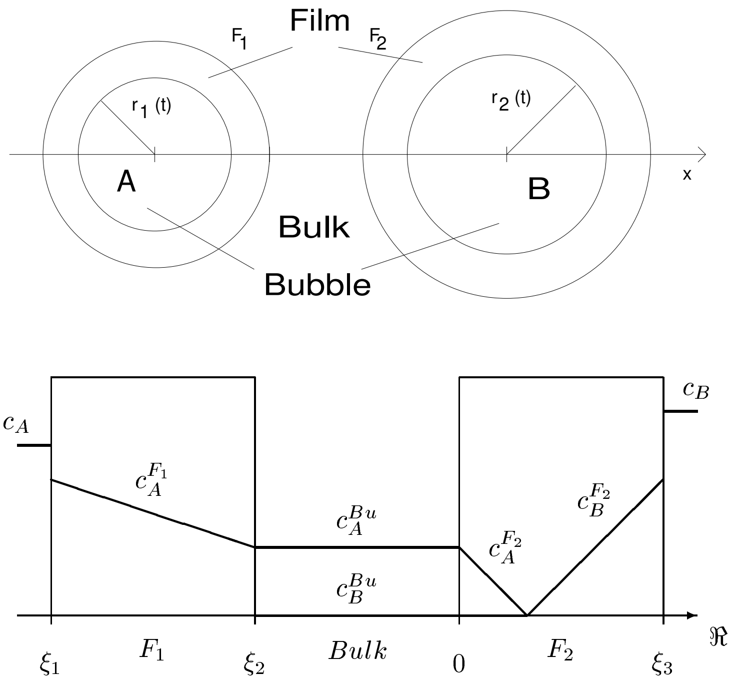


Figure VII.2: Two–film model based on interaction zones with constant thickness (top). Behaviour of the chemicals A and B on the computational domain (bottom).

As a consequence of the applied two–film model the dynamical synthesis process can be simulated with a fixed spatial domain involving the bulk and the film F_2. Equation (VII.1) is solved analytically. Clearly, the spatial discretization needs some adaptation due to the presence of internal boundary conditions between bulk and film. We refer to [83] for a more thorough discussion. Here we will report only on the temporal evolution of the grid used to resolve the reaction front in the film $F_2 = [0, 15]\mu m$. Fig. VII.3 shows that at the beginning the reaction front is travelling very fast from the outer to the inner boundary of the film where the chemical B enters permanently. During the time period $[0.1, 0.5]$ the reaction zone does not change its position which allows larger time steps. After that with decreasing concentration of the chemical A at the outer boundary the front travels back, but now with moderate speed. Obviously, the adaptively controlled discretization is able to follow automatically the dynamics of the problem.

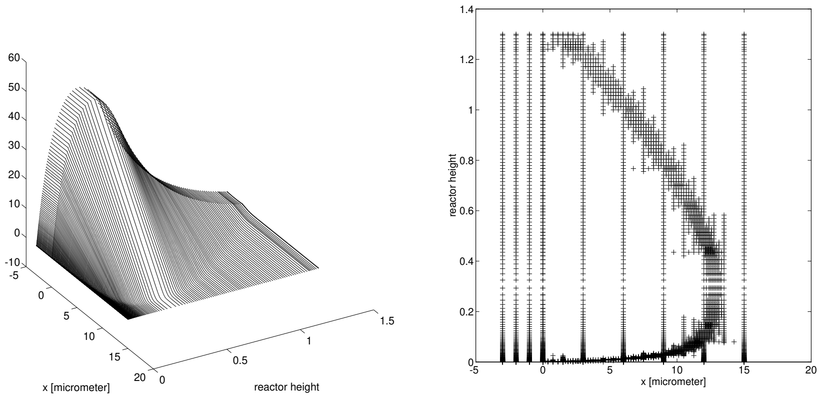


Figure VII.3: Evolution of the chemical component A (left) and of the grid (right) where the reactor height is taken as time axis.

§2. 2D: Propagation of Laminar Flames

Combustion problems are known to range among the most demanding for spatial adaptivity when the thin flame front is to be resolved numerically. This is often required as the inner structure of the flame determines global properties such as the flame speed, the formation of cellular patterns or even more important the mass fraction of reaction products (e.g. NO_x formation). A large part of numerical studies in this field is devoted to the different instabilities of such flames. The observed phenomena include cellular patterns, spiral waves, and transition to chaotic behaviour [113, 46, 61, 45, 24].

Apart from spatial adaptivity these problems can be solved generally with a constant time step. A reliable error control as available with the present method, however, is of great advantage. When dealing with ignition and extinction processes or complicated geometries and non–uniform material the relevant time scales can change by orders of magnitude. The proposed method then automatically adjusts the time step in accordance with the spatial tolerance so that the invested computational effort results in an optimal advancement of the calculation in time.

§2.1. Laminar Flames Through an Obstacle

The major part of gaseous combustion processes can adequately be described under the low Mach number hypothesis. This essentially amounts to eliminating the pressure dependence of the fluid density while retaining its temperature dependence. When the latter is not accounted for either, the motion of the fluid becomes independent from temperature and species concentration. Then the velocity field influences these quantities only via a convection term. Hence, one can solve the temperature and species equations alone specifying any solenoidal velocity field $u(x, y, t)$. In particular, $u = 0$, $u = u_0 = const.$, and $u = -u_f$ with $u_f(t)$ being the velocity of the flame front are important cases. Introducing the dimensionless temperature $\theta = (T - T_{unburnt})/(T_{burnt} - T_{unburnt})$, denoting by Y the species concentration, and assuming constant diffusion coefficients yields [112]

$$\partial_t \theta - \nabla^2 \theta = \omega \,, \tag{VII.4}$$

$$\partial_t Y - \frac{1}{Le} \nabla^2 Y = -\omega \,, \tag{VII.5}$$

where the Lewis number Le is the ratio of diffusivity of heat and diffusivity of mass. The time has been nondimensionalized with the heat conduction time scale, and the heat release parameter has entered in the reaction term through the definition of θ. We use a simple one-species reaction mechanism governed by an Arrhenius law

$$\omega = \frac{\beta^2}{2Le} \, Y \, e^{\frac{-\beta(1-\theta)}{1-\alpha(1-\theta)}} \,, \tag{VII.6}$$

in which an approximation for large activation energy has been employed [38]. The temperature ratio $\alpha = (T_{burnt} - T_{unburnt})/T_{burnt}$ is the quantity that determines the gas expansion in non-constant density flows so that the above thermo-diffusive model is exact for $\alpha = 0$. The extension of (VII.4), (VII.5) to a complex reaction scheme is straightforward by adding similar equations for additional species and modifying the reaction terms. We have also performed computations with an additional convection term. This can either describe the response of a thermodiffusive flame to a given velocity field under fixed boundary conditions or it can be used to furnish a moving reference frame (equivalent to dynamic regridding) in which a propagating flame front may become stationary. In the latter case the spatially uniform velocity is chosen proportional to the instantaneous integral of the reaction rate.

Here we consider a freely propagating laminar flame described by eqs. (VII.4), (VII.5) and its response to a heat absorbing obstacle, a set of cooled parallel rods with rectangular cross section. The computational domain has width $H = 16$ and length $L = 60$. The obstacle covers half of the width and has length $L/4$. The absorption of heat is modelled by the boundary condition

$$\partial_n \theta = - \, k \, (\theta - \theta_{ref}) \,, \tag{VII.7}$$

where k is a heat loss parameter and where the reference temperature is chosen as $\theta_{ref} = \theta_{unburnt} = 0$. On the left boundary of the domain (cf. Fig. VII.4) Dirichlet conditions corresponding to the burnt state are prescribed while the remaining boundary conditions are of homogeneous Neuman type. The initial condition is the analytical solution of a one–dimensional right–travelling flame in the limit $\beta \to \infty$ located left of the obstacle:

$$\theta(x,y,0) = \begin{cases} 1 & \text{for} \quad x \le x_0\,, \\ e^{x_0 - x} & \text{for} \quad x > x_0\,, \end{cases}$$

$$Y(x,y,0) = \begin{cases} 0 & \text{for} \quad x \le x_0\,, \\ 1 - e^{Le\,(x_0 - x)} & \text{for} \quad x > x_0\,, \end{cases}$$

where we have used $x_0 = 9$ in our computations. Two different situations may occur in this experiment according to the value of k. For small k the flame becomes curved and is slowed down in the interior of the channel but manages to pass through. For stronger heat loss the flame is extinguished. Computations of this phenomenon in a simple channel geometry have been done in [23]. In the present setting the extinction limit is a function of many parameters: length and width of the obstacle, its geometry, the Lewis number, the type of boundary condition and the amount of heat loss. We therefore do not aim here at determining precise thresholds but rather show a sample computation for $k = 0.1$ choosing $Le = 1$, $\beta = 10$, $\alpha = 0.8$. In this case the heat loss is below the critical value. We mention that the flame is extinguished in the obstacle for $k = 0.2$ [59].

Fig. VII.4 shows the propagation of the reaction front with the help of the reaction rate ω. In our experience the reaction rate is by far the best quantity to judge for adequate spatial resolution of such a computation, as it is related to the smallest spatial scales of the problem. The present results show (backed by additional verifications) that ω is indeed well–resolved although being controlled only indirectly through the adaptation process based on θ and Y. Fig. VII.5 displays the corresponding grids to give an impression of these as well.

Fig. VII.6 depicts total reaction rate, time step, and degrees of freedom during the propagation. When the flame passes through the channel the total reaction rate diminishes not only due to the smaller width of the front but also due to heat loss. The peak near the end of the graph results from the increased flame area after the front leaves the obstacle. According to the temporally reduced flame velocity the time step automatically increases about one order of magnitude. This results in essential savings of computation time compared to a constant time step. The latter would furthermore have to be adjusted by hand safely below the limit for stability and precision. The peaks in this diagram pointing to small values are related to the fact that a plot at a prescribed time t has been requested. This generally imposes one very small time step to exactly reach this point. Handling this is not straightforward with the estimation procedure for the timestep described in section IV.1. We therefore restarted with a given time step safely below the required one (a modified estimation scheme might be used as well). The figure thus permits one to appreciate the rapidity and robustness of the temporal adaptation which returns to the optimal value in about three steps.

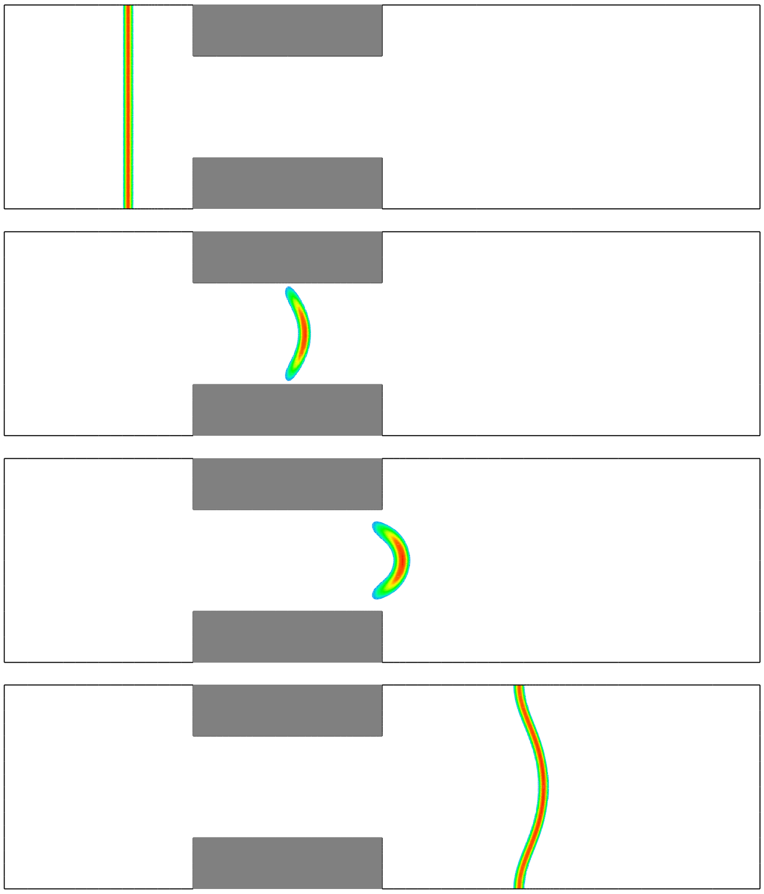


Figure VII.4: Flame through cooled grid, $Le = 1$, $k = 0.1$. Reaction rate ω at $t = 1$, 20, 40, 60.

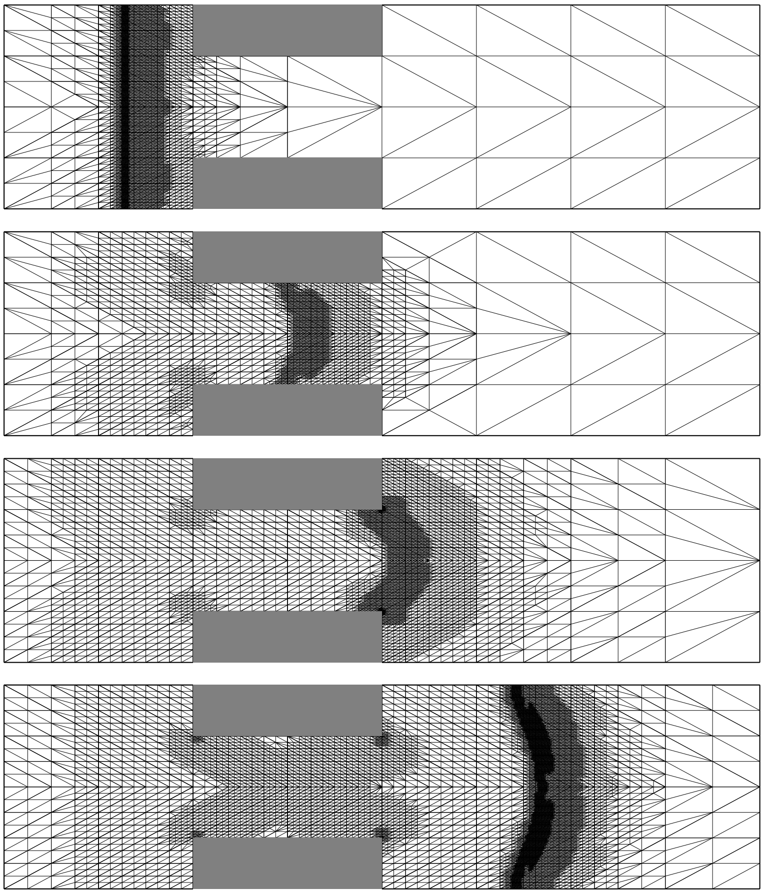


Figure VII.5: Flame through cooled grid, $Le = 1$, $k = 0.1$. Spatial discretization at $t = 1$, 20, 40, 60.

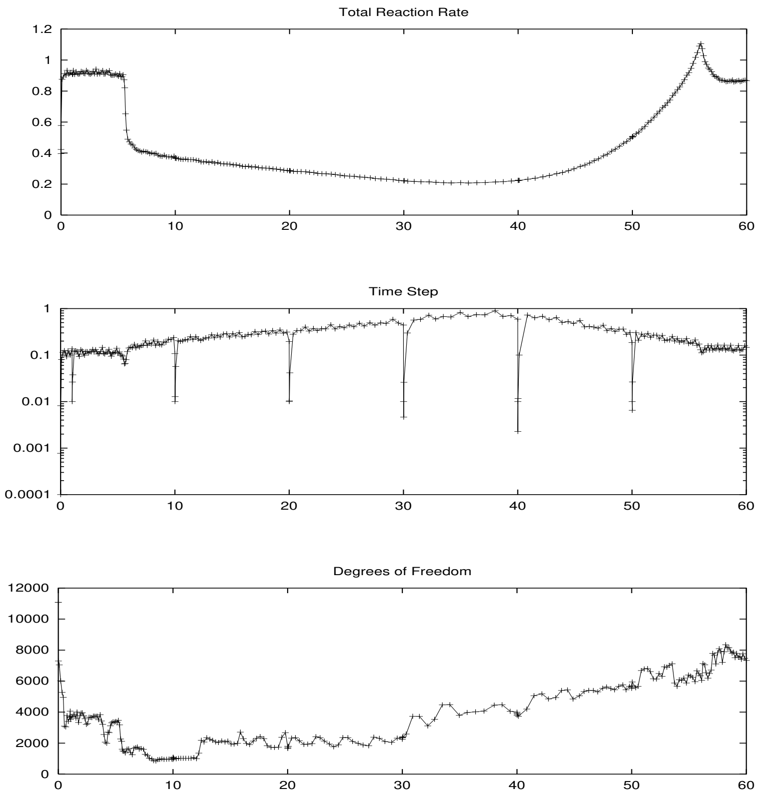


Figure VII.6: Flame through cooled grid, $Le = 1$, $k = 0.1$. Temporal evolution of the scaled total reaction rate $\int \omega d\Omega / H$, the time step τ, and the number of nodes N for $TOL = 2.0\ 10^{-3}$.

§2.2. Reaction Front in a Solid

The second class of problems that we will deal with refers to solid-solid combustion. The particularity of such a process is that convection is impossible and that the macroscopic diffusion for the species in solids is in general negligible with respect to heat conductivity. With the heat diffusion time scale as reference, the equations for a one step chemical alloying reaction read

$$\partial_t T \;-\; \kappa\,\nabla^2 T \;=\; Q\omega\,, \tag{VII.8}$$

$$\partial_t Y \;=\; -\omega\,, \tag{VII.9}$$

where T is the temperature divided by a reference temperature, Y the concentration of the deficient reactant and Q a heat release parameter. Concerning the reaction term quite a number of different models are employed in the literature. They generally contain an Arrhenius term for the temperature dependence and use a first order reaction, i.e.,

$$\omega \;=\; K_0\, Y\, e^{-\frac{E}{T}}\,, \tag{VII.10}$$

where E is a dimensionless activation energy. Since this expression is difficult to treat with analytical approaches it is often replaced by a zero–order mechanism substituting Y in (VII.10) by $\chi(Y) = H(Y)$, the Heaviside function [100]. Although simpler, this expression involves additional modelling, and furthermore it might generate difficulties in the numerical solution due to the discontinuity of χ.

Apart from the reaction realistic processes involve other physical mechanisms such as melting which leads to additional heat release (e.g. [20]) or microscopic diffusion of the reactant into the fine grains constituting the material. SMOOKE and KOSZYKOWSKI [131] employ

$$\omega(T,Y) = \frac{D}{R^2}\,F(Y)\,e^{-\frac{E}{T}} \quad , \qquad F(Y) = \frac{Y^{1/3}}{1-Y^{1/3}} \tag{VII.11}$$

which has been developed by BOOTH [28] considering a material made up of densely packed spheres subject to melting and microscopic diffusion of the reactant. Here, R is the radius of the spheres and D a microscopic diffusion coefficient. Due to the absence of macroscopic species diffusion and the related smoothing property the system (VII.8), (VII.9) is more difficult to treat numerically than the thermodiffusive equations. Nevertheless the same algorithm could be employed.

In our computations we experienced the need to adjust the term F in (VII.11) for Y near one. This is justified, since, after all, (VII.11) can just be a model of limited validity. First, the maximum range of Y is by definition the interval $[0,1]$. However, $F(1)$ cannot be evaluated. For global conservation the physical

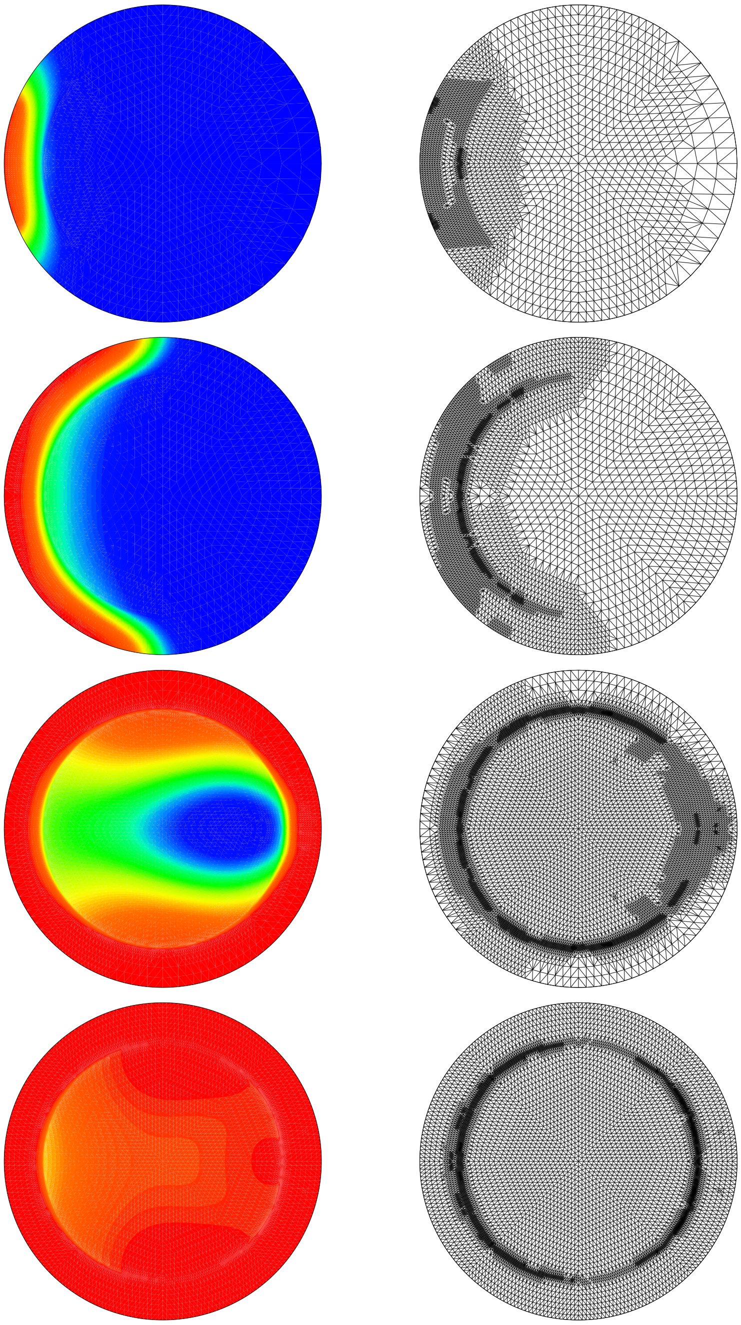


Figure VII.7: Non–uniformly packed solid. Concentration of the reactant and grids at times $t = 0.05,\ 0.065,\ 0.07,\ 0.072$.

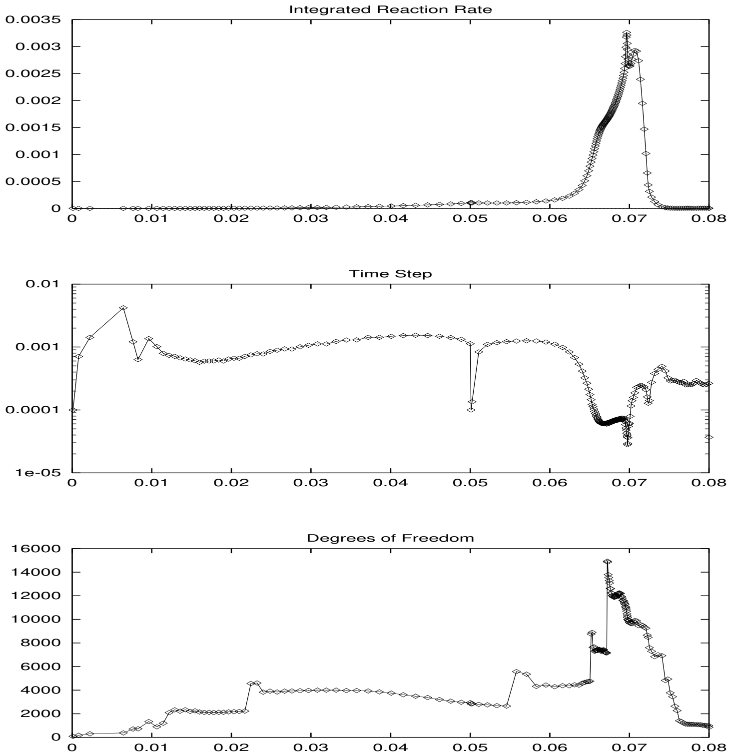


Figure VII.8: Non–uniformly packed solid. Temporal evolution of the total reaction rate $\int \omega d\Omega$, the time step τ, and the number of nodes N for $TOL = 1.5\ 10^{-3}$.

range in fact is $[0, Y_I]$ where Y_I is the maximum value of the initial condition. Hence, choosing $Y(x, 0) = 0.999$ instead of 1 [131] removes this difficulty. Second, the employed spatial approximation does not guarantee that Y remains within a certain interval. Therefore $F(Y)$ has been replaced by the tangent at $Y = 0.999$ and $Y = 0.001$ to the right and to the left of these values, respectively. This prolongation of $\omega(T, Y)$ to arguments Y outside the original range of definition $[0, 1]$ results in a value of ω which brings the solution back to the physically meaningful range in case of overshooting due to finite precision. Such minor adjustments although much smaller than the accuracy of the model may decide on the failure or the success of a computation. In different practical computations we have experienced this strategy to be much more robust than e.g. replacing overshoots by the physically limiting values. This is particularly important when adaptive discretization is used which has less tendency to attenuate fine scale oscillations.

In the following we report on the solution of (VII.8), (VII.9), (VII.11) for a non–uniformly packed solid in cylindrical geometry. We consider the value of D/R^2 being equal to 5800 within a circle of radius $r_1 = 0.0018$ and to increase linearly up to 16 times this value at $r_2 = 0.0024$. This yields a strongly increased flame velocity close to this radius which constitutes the outer border. The reactor is ignited by a Dirichlet condition $T_w = 300 + 24000\,t,\ t < 0.05$, for the wall temperature on a 45 degrees section of the boundary on the left hand side. For $t \geq 0.05$ and all other locations the boundary conditions are homogeneous Neuman conditions. Further parameters are $E = 11000$, $\kappa = 0.0001$, $Q = 2700$. Fig. VII.7 nicely illustrates how the reaction front first propagates along the outer wall before entering the core. The graphs should be related to those of Fig. VII.8 revealing that the reaction phase is rather short compared to the ignition. The small time steps at $t = 0.05$ result from the requirement of exactly attaining this instant where the boundary conditions are modified. Even in this simple geometry the interaction of different fronts can generate a rather complicated pattern in space and time. The reaction rate for example exhibits a peak where both peripherical fronts have merged and propagate into the interior whereas the maximum of the temperature appears at a later time.

§3. 2D: Dopant Diffusion in Silicon

The quality of electronical materials based on silicon substrates is strongly influenced by the quality of doping processes. Impurity atoms of higher or lower chemical valence, such as arsenic, phosphorus, and boron, are introduced under high temperatures (900^oC–1100^oC) into a silicon crystal to change its electrical properties. This is the central process of modern silicon technology. The diffusion mechanism of dopants is a topic of continuing investigations. It cannot be described involving only direct interchange with neighbouring silicon atoms. In order to explain this anomalous behaviour various pair diffusion models have been proposed [56, 64, 75].

We consider a general model for phosphorus diffusion in silicon under extrinsic doping conditions described by GHADERI and HOBLER [64]. At such high concentrations we have to include the charged species and the internal electric field of the crystal, both of which can have profound effects on diffusion. In principle, this leads to a very large number of drift–diffusion–reaction equations: one for each charge state of every species, plus one Poisson equation to describe the internal electrostatic potential. The number of equations can be reduced substantially by making additional equilibrium assumptions concerning the reaction terms. The resulting model turns out to be very interesting for numerical investigation.

§3.1. Diffusion Model under Extrinsic Conditions

Dopant atoms occupy substitutional sites in the silicon crystal lattice, losing (donors such as arsenic and phosphorus) or gaining (acceptors such as boron) at the same time an electron. We denote such substitutional defects by A. Since a diffusion mechanism based only on the direct interchange with neighbouring silicon atoms turns out to be energetically unfavourable [56], native point defects called interstitials (I) and vacancies (V) are taken into account. Interstitials are silicon atoms which are not placed on a lattice site and move through the crystal unconstrained. Vacancies are empty lattice sites. Both can form mobile pairs with dopant atoms, designated by AI and AV. There is no general consensus on the exact nature of the pair mobility mechanism. One way to visualize it is as follows: in the case of AI-pairs, an interstitial and a dopant atom share a lattice site called an interstitialcy. The dopant can now change partners by moving through an intermediate interstitial stage (denoted by A_i in Fig. VII.9). On the other hand, dopants and vacancies exhibit a certain affinity. So, a vacancy near a dopant moves around this dopant quickly, and an occasional interchange between dopant and vacancy leads to a random walk effect (see Fig. VII.9).

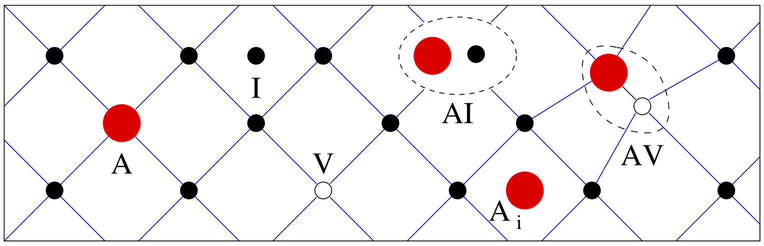


Figure VII.9: Visualization of pair diffusion in a silicon crystal lattice.

In the case of extrinsic doping conditions we consider silicon crystals where the concentration C_A of the dopant is much higher than the intrinsic carrier concentration n_i, i.e., $C_A \gg n_i$. It is well–known that this assumption forbids neglecting the charge states of the species. For the interstitials we consider the charge states

$$I^{(i)}\,, \qquad i=-1,0,+1\,, \tag{VII.12}$$

for the vacancies

$$V^{(j)}\,, \qquad j=-2,-1,0,+1,+2\,. \tag{VII.13}$$

The unpaired dopant on a lattice site has always the fixed charge state $A^{(q)}$, for instance $q=+1$ for phosphorus. Thus, the pairs under consideration are

$$(AI)^{(q+i)} \quad \text{and} \quad (AV)^{(q+j)}\,. \tag{VII.14}$$

Next, we state the set of reactions between the species and formulate the reaction rates in terms of concentrations C_X^z, where $X=I,V,AI,AV,A$, and z denoting the charge state. We consider

- Dopant–defect pairing:

$$A^{(q)}+I^{(i)} \rightleftharpoons (AI)^{(q+i)}\,, \quad R^i_{A,I}=k^i_{A,I}C^q_A C^i_I-\tilde{k}^i_{A,I}C^{q+i}_{AI}\,, \tag{VII.15}$$

$$A^{(q)}+V^{(j)} \rightleftharpoons (AV)^{(q+j)}\,, \quad R^j_{A,V}=k^j_{A,V}C^q_A C^j_V-\tilde{k}^j_{A,V}C^{q+j}_{AV}\,, \tag{VII.16}$$

- Defect recombination:

$$(AI)^{(q+i)}+V^{(j)} \rightleftharpoons A^{(q)}-(i+j)n\,,$$

$$R^{ij}_{AI,V} = k^{(q+i)j}_{AI,V}C^{q+i}_{AI}C^j_V-\tilde{k}^{(q+i)j}_{AI,V}C^q_A\left(\frac{n}{n_i}\right)^{-i-j}\,, \tag{VII.17}$$

$$(AV)^{(q+j)}+I^{(i)} \rightleftharpoons A^{(q)}-(i+j)n\,,$$

$$R^{ji}_{AV,I} = k^{(q+j)i}_{AV,I}C^{q+j}_{AV}C^i_I-\tilde{k}^{(q+j)i}_{AV,I}C^q_A\left(\frac{n}{n_i}\right)^{-i-j}\,, \tag{VII.18}$$

- Frenkel pairs:

$$I^{(i)}+V^{(j)} \rightleftharpoons -(i+j)n\,,$$

$$R^{ij}_{I,V} = k^{ij}_{I,V}C^i_I C^j_V-\tilde{k}^{ij}_{I,V}\left(\frac{n}{n_i}\right)^{-i-j}\,, \tag{VII.19}$$

- Ionization of the defects:

$$I^{(i)} \rightleftharpoons I^{(0)}-in\,, \quad R^i_I=k^i_I C^i_I-\tilde{k}^i_I C^0_I\left(\frac{n}{n_i}\right)^{-i}\,, \tag{VII.20}$$

$$V^{(j)} \rightleftharpoons V^{(0)}-jn\,, \quad R^j_V=k^j_V C^j_V-\tilde{k}^j_I C^0_V\left(\frac{n}{n_i}\right)^{-j}\,, \tag{VII.21}$$

- Ionization of the pairs:

$$
\begin{aligned}
(AI)^{(q+i)} &\rightleftharpoons (AI)^{(q)} - in\,, \\
R^{i}_{AI} &= k^{i}_{AI} C^{q+i}_{AI} - \tilde{k}^{i}_{AI} C^{q}_{AI} \left(\frac{n}{n_i}\right)^{-i}, && \text{(VII.22)} \\
(AV)^{(q+j)} &\rightleftharpoons (AV)^{(q)} - jn\,, \\
R^{j}_{AV} &= k^{j}_{AV} C^{q+j}_{AV} - \tilde{k}^{j}_{AV} C^{q}_{AV} \left(\frac{n}{n_i}\right)^{-j}, && \text{(VII.23)}
\end{aligned}
$$

where the k supplied with the respective indices denote the forward reaction–rate constants, and the $\tilde{k}$ the reverse ones.

We assume that the electrons n and the holes p obey the Boltzmann statistics

$$
n = n_i \exp\left(\frac{e\psi}{k_B T}\right), \quad p = n_i \exp\left(-\frac{e\psi}{k_B T}\right), \qquad \text{(VII.24)}
$$

where ψ denotes the electrostatic potential, k_B is the Boltzmann constant, T the absolute temperature, and e is the elementary charge. Using the thermal voltage defined by $U_T = k_B T/e$, we have the relations

$$
n_i = (np)^{1/2}\,, \quad \psi = U_T \ln\left(\frac{n}{n_i}\right). \qquad \text{(VII.25)}
$$

For each of the charged species we can establish the corresponding balance equation taking into account the reaction terms (VII.15)–(VII.23) and drift–diffusion terms. This yields seventeen coupled partial differential equations of the form

$$
\begin{aligned}
\partial_t C^{\alpha+z}_X + \nabla \cdot J^{\alpha+z}_X &= -R^{\alpha+z}_X\,, \quad X = I, V, AI, AV, A, && \text{(VII.26)} \\
J^{\alpha+z}_X &= -D^{\alpha+z}_X \left(\nabla C^{\alpha+z}_X + (\alpha+z) C^{\alpha+z}_X \nabla \ln\left(\frac{n}{n_i}\right)\right) && \text{(VII.27)}
\end{aligned}
$$

where z represents the possible charge states for the component X, $\alpha = 0$ for $X = I, V$, and $\alpha = q$ for $X = AI, AV, A$. The highly nonlinear term $R^{\alpha+z}_X$ is the sum of recombination and generation terms. Note that according to $J^q_A \equiv 0$ no diffusion term occurs in the equation for C^q_A. The electrostatic potential is computed by an additional Poisson equation

$$
-\nabla^2 \psi = \frac{e}{\varepsilon}\left(-n + p + \sum_{X,z} (\alpha + z) C^{\alpha+z}_X\right), \qquad \text{(VII.28)}
$$

where ε denotes the dielectric constant.

In order to reduce the equations we assume equilibrium concerning the ionization, i.e., we state that

$$R_X^z = 0 \quad \text{for } X = I, V, AI, AV\,. \tag{VII.29}$$

As direct consequence we have

$$C_X^{\alpha+z} = K_X^{\alpha+z}\left(\frac{n}{n_i}\right)^{-z} C_X^{\alpha}\,, \tag{VII.30}$$

where the equilibrium constants $K_X^{\alpha+z}$ are the quotient of reverse and forward reaction–rate constants. Defining total concentrations

$$C_X := \sum_z C_X^{\alpha+z} = \sum_z K_X^{\alpha+z}\left(\frac{n}{n_i}\right)^{-z} C_X^{\alpha}\,. \tag{VII.31}$$

we can substitute C_X^{α} in (VII.30) to get

$$C_X^{\alpha+z} = \frac{K_X^{\alpha+z}\left(\frac{n}{n_i}\right)^{-z}}{\sum_z K_X^{\alpha+z}\left(\frac{n}{n_i}\right)^{-z}} C_X\,, \tag{VII.32}$$

showing that each charged state $C_X^{\alpha+z}$ can directly be computed from the total concentration C_X. Summing up over all charged species in (VII.26), using (VII.32) and (VII.25) yields after some calculation

$$\begin{aligned} \partial_t C_X + \nabla\cdot J_X &= -R_X\,, & \text{(VII.33)}\\ J_X &= -\mathcal{D}_X(n)\left(\nabla C_X + \mathcal{Q}_X(n) C_X \nabla \ln\left(\frac{n}{n_i}\right)\right), & \text{(VII.34)}\\ -U_T\nabla^2 \ln\left(\frac{n}{n_i}\right) &= \frac{e}{\varepsilon}\left(-n + p + \sum_{X,z} \mathcal{Q}_X(n) C_X\right), & \text{(VII.35)} \end{aligned}$$

where

$$\mathcal{D}_X(n) = \frac{\sum_z D_X^{\alpha+z} K_X^{\alpha+z}\left(\frac{n}{n_i}\right)^{-z}}{\sum_z K_X^{\alpha+z}\left(\frac{n}{n_i}\right)^{-z}}\,,\ \mathcal{Q}_X(n) = \frac{\sum_z (\alpha+z) K_X^{\alpha+z}\left(\frac{n}{n_i}\right)^{-z}}{\sum_z K_X^{\alpha+z}\left(\frac{n}{n_i}\right)^{-z}}\,. \tag{VII.36}$$

The total reaction rates $R_X = \sum_z R_X^{\alpha+z}$ are

$$\begin{gathered} R_I = R_{A,I} + R_{AV,I} + R_{I,V}\,,\ R_{AI} = -R_{A,I} + R_{AI,V}\,,\\ R_V = R_{A,V} + R_{AI,V} + R_{I,V}\,,\ R_{AV} = -R_{A,V} + R_{AV,I}\,, \end{gathered}$$

with

$$R_{A,I} = \frac{\sum_i k^i_{A,I}\, K^i_I \left(\frac{n}{n_i}\right)^{-i}}{\sum_i K^i_I \left(\frac{n}{n_i}\right)^{-i}} \left(C_A\, C_I - C^*_I \frac{C^*_A}{C^*_{AI}} C_{AI} \right),$$

$$R_{A,V} = \frac{\sum_j k^j_{A,V}\, K^j_V \left(\frac{n}{n_i}\right)^{-j}}{\sum_j K^j_V \left(\frac{n}{n_i}\right)^{-j}} \left(C_A\, C_V - C^*_V \frac{C^*_A}{C^*_{AV}} C_{AV} \right),$$

$$R_{AI,V} = \frac{\sum_i \sum_j k^{(q+i)j}_{AI,V}\, K^{q+i}_{AI}\, K^j_V \left(\frac{n}{n_i}\right)^{-i-j}}{\sum_i K^{q+i}_{AI} \left(\frac{n}{n_i}\right)^{-i} \sum_j K^j_V \left(\frac{n}{n_i}\right)^{-j}} \left(C_{AI}\, C_V - C^*_V \frac{C^*_{AI}}{C^*_A} C_A \right),$$

$$R_{AV,I} = \frac{\sum_j \sum_i k^{(q+j)i}_{AV,I}\, K^{q+j}_{AV}\, K^i_I \left(\frac{n}{n_i}\right)^{-i-j}}{\sum_j K^{q+j}_{AV} \left(\frac{n}{n_i}\right)^{-j} \sum_i K^i_I \left(\frac{n}{n_i}\right)^{-i}} \left(C_{AV}\, C_I - C^*_I \frac{C^*_{AV}}{C^*_A} C_A \right),$$

$$R_{I,V} = \frac{\sum_i \sum_j k^{ij}_{I,V}\, K^i_I\, K^j_V \left(\frac{n}{n_i}\right)^{-i-j}}{\sum_i K^i_I \left(\frac{n}{n_i}\right)^{-i} \sum_j K^j_V \left(\frac{n}{n_i}\right)^{-j}} \left(C_I\, C_V - C^*_I C^*_V \right).$$

Here, we have substituted the backward reaction–rate coefficients by the forward ones, the equilibrium concentrations C^*_I, C^*_V, and the equilibrium ratios C^*_{AI}/C^*_A, and C^*_{AV}/C^*_A. These expressions are obtained by setting the reaction rates $R_{X,Y}$ equal to zero. All equilibrium concentrations, except C^*_A, depend on n/n_i. This dependence can be determined from (VII.31). Using values for the intrinsic case $n{=}n_i$ we derive

$$C^*_X = C^*_X\,|_{n=n_i}\ \frac{K^{\alpha+z}_X \left(\frac{n}{n_i}\right)^{-z}}{\sum_z K^{\alpha+z}_X}$$

and further for the equilibrium ratios

$$\frac{C^*_A}{C^*_{AX}} = \left.\frac{C^*_A}{C^*_{AX}}\right|_{n=n_i} \frac{K^{\alpha+z}_X \left(\frac{n}{n_i}\right)^{-z}}{\sum_z K^{\alpha+z}_X}, \qquad X = AI, AV\,.$$

Hence, we can replace the general equilibrium values by their intrinsic versions.

The whole model requests the knowledge of an enormous list of parameters. Most of them are essentially unknown or at least controversial. The set of parameters we have used in our simulations was taken from [99]. Table VII.10 summarizes some of them. For the point defects we use for all charge states the intrinsic diffusivities $D_{I|n=n_i}$ and $D_{V|n=n_i}$ which were obtained from experiments via gold diffusion. The remaining forward reaction–rate constants $k^{rs}_{X,Y}$

and $k^s_{A,Y}$ ($X, Y \in \{I, V, AI, AV\}$, and r, s denote the charge states) may be expressed as

$$k^{rs}_{X,Y} = 4\pi r_c (D^r_X + D^s_Y) \exp\left(-\frac{E^{rs}_{X,Y}}{k_B\,T}\right)$$

and

$$k^s_{A,Y} = 4\pi r_c (D_{A|n=n_i} + D^s_Y) \exp\left(-\frac{E^s_{A,Y}}{k_B\,T}\right).$$

Therein, r_c denotes the capture radius set equal to 5Å. All the barrier energies $E^{rs}_{X,Y}$ and $E^s_{X,Y}$ are taken to be zero, except for the Frenkel pair reaction, where we take the value $0.3eV$. The effective intrinsic diffusion coefficient of phosphorus is well–known [57]

$$D_{A|n=n_i} = 3.850 \exp\left(-\frac{3.660}{k_B\,T}\right).$$

In order to complete the discussion concerning the physical parameters, we set

$$\left.\frac{C^*_A}{C^*_{AI}}\right|_{n=n_i} = 3.129 \cdot 10^{16} \exp\left(-\frac{1.680}{k_B\,T}\right)$$

and compute $(C^*_A/C^*_{AV})_{|n=n_i}$ from the relationship

$$D_{A|n=n_i} = \left.\frac{C^*_{AI}}{C^*_A}\right|_{n=n_i} D_{AI|n=n_i} + \left.\frac{C^*_{AV}}{C^*_A}\right|_{n=n_i} D_{AV|n=n_i}\,,$$

the derivation of which can be found in [64]. Here, we set $D_{AX|n=n_i} = D^0_{AX}$, $X = I, V$. The given quantities are valid in our case over a temperature range of $900^oC - 1200^oC$.

We also have to supply initial and boundary conditions for the system (VII.33)–(VII.35). The diffusion process is only one part of the whole semiconductor device fabrication. Thus, these conditions depend strongly on the interaction with other processes. A detailed description elaborated for different situations can be found in [75].

Here we consider the case of an implanted dopant profile which is initially set to a Gaussian curve. Desirable experimentally measured initial distributions are hardly available. Since Gaussian profiles are sometimes not well–fitted to high channeling effects in the lower concentration area [56], Pearson-IV distributions with exponential channeling tail are frequently used. Appropriate initial conditions for the defects are given by

$$C_X(0) = C^*_X(0)\,, \quad X = I, V\,.$$

$\mathcal{K}$	A	B	E [eV]
$D_{I\|n=n_i}$ $[cm^2s^{-1}]$	$2.629 \cdot 10^{11}$	0.000	4.436
$D_{V\|n=n_i}$ $[cm^2s^{-1}]$	$2.639 \cdot 10^{06}$	0.000	4.002
D^0_{AI} $[cm^2s^{-1}]$	$8.570 \cdot 10^{-1}$	0.000	1.720
D^1_{AI} $[cm^2s^{-1}]$	$1.780 \cdot 10^{06}$	0.000	3.340
D^2_{AI} $[cm^2s^{-1}]$	$4.128 \cdot 10^{-3}$	0.000	1.330
D^{-1}_{AV} $[cm^2s^{-1}]$	$6.123 \cdot 10^{03}$	0.000	2.550
D^0_{AV} $[cm^2s^{-1}]$	$5.466 \cdot 10^{05}$	0.000	3.040
D^1_{AV} $[cm^2s^{-1}]$	$7.094 \cdot 10^{09}$	0.000	4.090
D^2_{AV} $[cm^2s^{-1}]$	$1.509 \cdot 10^{00}$	0.000	1.840
D^3_{AV} $[cm^2s^{-1}]$	$1.509 \cdot 10^{00}$	0.000	1.840
K^{-1}_I	$0.754 \cdot 10^{00}$	0.868	0.185
K^0_I	$1.000 \cdot 10^{00}$	0.000	0.000
K^1_I	$1.326 \cdot 10^{00}$	0.868	0.185
K^{-2}_V	$0.569 \cdot 10^{00}$	2.252	0.480
K^{-1}_V	$0.754 \cdot 10^{00}$	0.070	0.015
K^0_V	$1.000 \cdot 10^{00}$	0.000	0.000
K^1_V	$1.326 \cdot 10^{00}$	2.557	0.545
K^2_V	$1.758 \cdot 10^{00}$	4.702	1.002
K^0_{AI}	$1.000 \cdot 10^{00}$	0.000	0.000
K^1_{AI}	$1.995 \cdot 10^{06}$	0.000	1.880
K^2_{AI}	$4.422 \cdot 10^{12}$	0.000	3.020
K^{-1}_{AV}	$8.601 \cdot 10^{13}$	0.000	3.260
K^0_{AV}	$1.000 \cdot 10^{00}$	0.000	0.000
K^1_{AV}	$9.501 \cdot 10^{15}$	0.000	3.780
K^2_{AV}	$7.068 \cdot 10^{03}$	0.000	0.920
K^3_{AV}	$1.317 \cdot 10^{20}$	0.000	4.760
$C^*_{I\|n=n_i}$ $[cm^{-3}]$	$1.132 \cdot 10^{18}$	0.000	1.377
$C^*_{V\|n=n_i}$ $[cm^{-3}]$	$1.642 \cdot 10^{24}$	0.000	2.226

Figure VII.10: Parameters of phosphorus diffusion in silicon.
$$\mathcal{K} = A \exp\Big(\frac{B\,T}{T+636}\Big)\ \ exp\Big(-\frac{E}{k_B\,T}\Big).$$

For the pairs AI and AV we take constant background dopings of order $O(10^8)$, whereas for the initial electrostatic potential the Poisson equation has to be approximated.

In order to define boundary conditions we allow the wafer surface to absorb or supply an arbitrary number of point defects which is usually modelled by flux conditions of the form

$$J_X \cdot \mathrm{n} = h_X(C_X - C_X^*)$$

where $X = I$ or V, and h_X denotes a specific transmission coefficient. Assuming the reaction rate of the neutral point defects to be infinitely fast at the interface, we can use Dirichlet boundary conditions at the wafer surface

$$C_X = C_X^* \,. \tag{VII.37}$$

Note that C_X^* depends strongly on the electrostatic potential, making this condition highly nonlinear. At all other boundaries homogeneous flux conditions are used. For the pairs we simply have homogeneous flux conditions at all boundaries. The Poisson equation is equipped with homogeneous Dirichlet conditions at the bottom of the wafer and with zero flux conditions otherwise.

§3.2. Some Simulation Results for Phosphorus Diffusion

We have performed a series of two–dimensional simulations to test our adaptive strategies for the above pair diffusion model. The computational domain is defined by the rectangle

$$\Omega = \{x = (x_1, x_2) \in \mathbb{R}^2 \,,\; 0 < x_1 < 10^{-3} \,,\; 0 < x_2 < 10^{-4}\,\}$$

where the unit spatial measurement is given in cm. The wafer surface is located at $x_1 = 0$ and the bottom of the wafer is at $x_1 = 10^{-3}$. The initial coarse grid used throughout our simulations is shown in Fig. VII.11. The relatively large expansion of the computational domain guarantees that the solution is not affected by the boundary condition at the bottom.

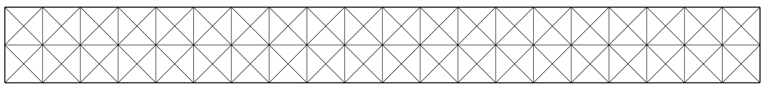


Figure VII.11: Phosphorus diffusion. Coarse grid consisting of 160 triangular elements.

The implanted phosphorus concentration has been set initially to the Gaussian profile

$$C_A(x,0) = C_A^{\max} \exp\left(-\frac{1}{2}\,\frac{f(x-a)}{\sigma^2}\right),$$

where $C_A^{\max} \gg n_i$ is the maximal value of the function, $a = (a_1, a_2)$ determines the position of the profile, σ is the standard deviation and

$$f(x-a) = (x_1 - a_1)^2 + \frac{1}{4}\big(\big||x_2 - a_2| - b\big| + |x_2 - a_2| - b\big)^2 .$$

If $b = 0$, then we have the usual Gaussian profile; for $b > 0$ the maximum extends to a whole line of length b in the x_2–direction. We choose

$$\sigma = 0.027 \cdot 10^{-4}, a_1 = 0.02 \cdot 10^{-4}, a_2 = 0.5 \cdot 10^{-4}, b = 1.0 \cdot 10^{-5} .$$

For the boundary conditions of the point defects Dirichlet conditions (VII.37) are employed (see [101] for a discussion of flux conditions). In the following our standard setting is $C_A^{\max} = 6 \cdot 10^{20}\,\mathrm{cm}^{-3}$ and $T = 900^o C$. All concentrations will be plotted using a logarithmic scale.

In Fig. VII.12 we see some phosphorus profiles and the corresponding dynamic meshes near the wafer surface at different time points. The phosphorus concentration at $t = 30\,\mathrm{min}$ shows its typical "kink and tail" behaviour caused by the anomalous diffusion mechanism. A detailed discussion of this phenomenon can be found in [115]. Steep gradients are well resolved by the dynamic meshes not wasting degrees of freedom.

Fig. VII.13 shows how the phosphorus goes into the silicon. Cuts along the middle axis $x_2 = 5 \cdot 10^{-5}$ are plotted for the concentration C_A and for the total sum $C_S = C_A + C_{AI} + C_{AV}$. We observe that the concentration of the pairs is negligible compared with the extrinsic dopant concentration. Both profiles are nearly identical.

The evolution of the defects is illustrated in Fig. VII.14. At the beginning the total concentration of interstitials increases very fast forced by the decay of AI–pairs which dominate the phosphorus diffusion. Due to the Dirichlet boundary condition a profile change from "concave to convex" takes place around $t = 10^{-5}\,\mathrm{sec}$. During this period of high activities our adaptive algorithm increases drastically the number of grid points in order to achieve the prescribed accuracy (see Fig. VII.16). Analogous effects can be observed for the concentration of neutral point defects, whereas the total concentration of vacancies simply decreases during the process. Note that $C_I^{(0)}$ and $C_V^{(0)}$ are always equal to the intrinsic equlibrium values at the wafer surface.

Fig. VII.15 shows profiles for various initial peak concentrations $C_A^{\max}$ (including the intrinsic case where $C_A^{\max} = 10^{18}\ \mathrm{cm}^{-3}$) at different temperatures. It can be seen clearly that the phosphorus diffusion proceeds faster with higher temperatures and the typical "kink and tail" behaviour vanishes.

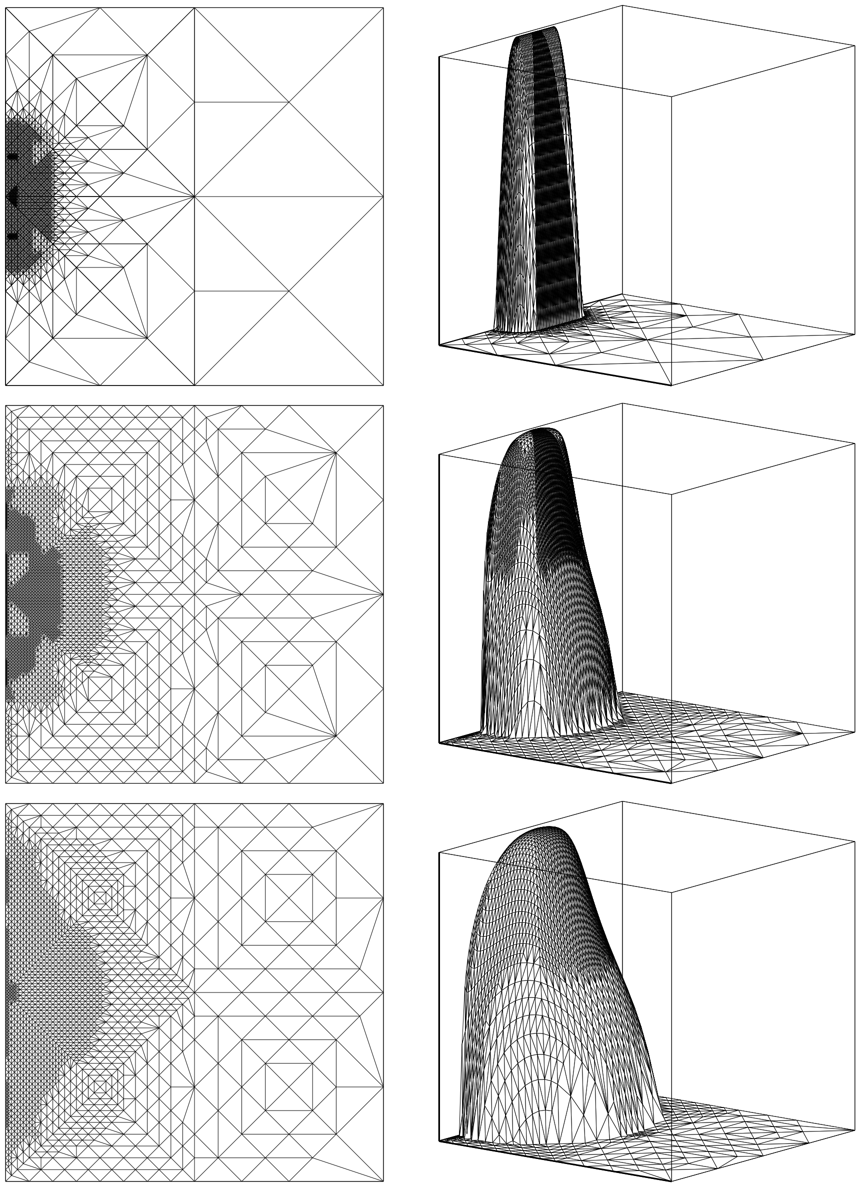


Figure VII.12: Phosphorus diffusion. Evolution of the dynamic meshes and the phosphorus concentration near the wafer surface at $t=0$, $t=3$, and $t=30\,\mathrm{min}$.

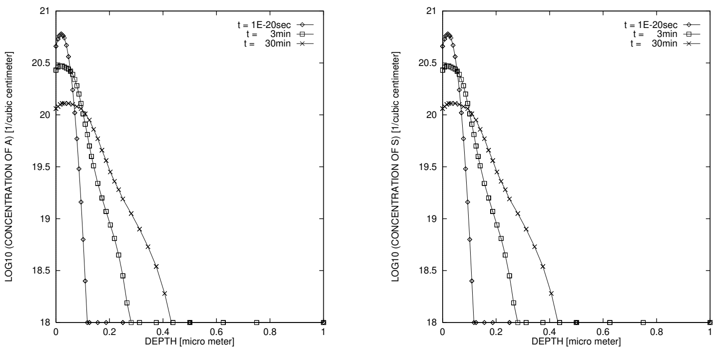


Figure VII.13: Phosphorus diffusion. Cuts through the phosphorus concentration C_A (left) and the total concentration $C_S = C_A + C_{AI} + C_{AV}$ (right) at $t=0$, $t=3$, and $t=30\,\mathrm{min}$.

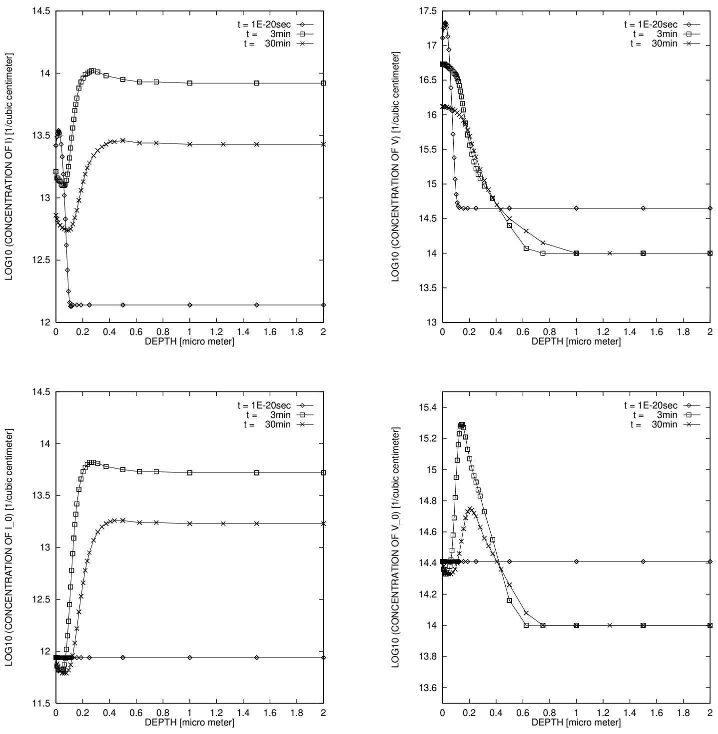


Figure VII.14: Phosphorus diffusion. Cuts through the total and neutral concentration of interstitials (left) and vacancies (right) at $t=0$, $t=3$, and $t=30\,\mathrm{min}$.

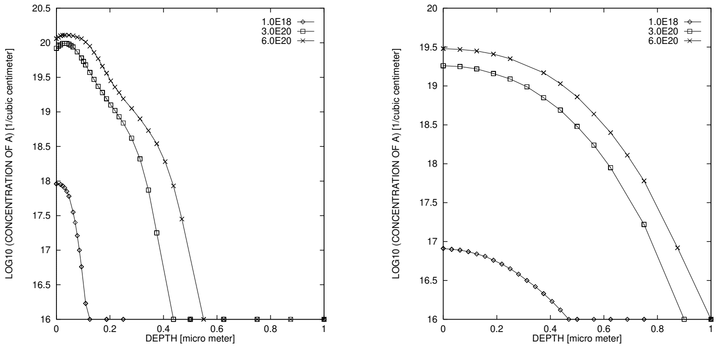


Figure VII.15: Phosphorus diffusion. Phosphorus concentration after 30 min at $900^o C$ (left) and $1100^o C$ (right) for different maximum values $C_A^{\max}$.

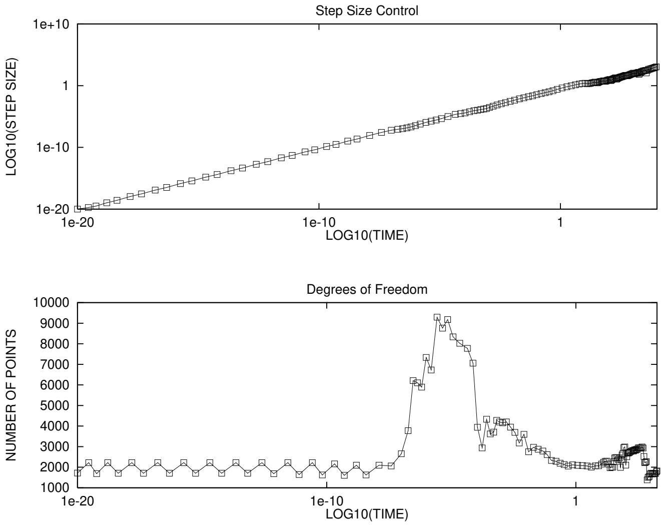


Figure VII.16: Phosphorus diffusion. Evolution of time steps and number of spatial discretization points for TOL=0.02.

In Fig. VII.16 we have plotted the evolution of time steps and number of grid points used by the adaptive algorithm. The chosen time steps range from 10^{-20} up to 300 and increase monotonically in time. The spatial dynamics of the whole system show an irregular behaviour. While approximately $2,000$ points are sufficient to represent the steep initial contributions, about $10,000$ nodes around $t = 10^{-5}$ are necessary to guarantee a relative tolerance $TOL = 0.02$. As explained above this large number of grid points is caused by the sudden shape change of the interstitials concentration.

§4. 3D: Bio–Heat Transfer in Regional Hyperthermia

Hyperthermia, i.e., heating tissue to 42–43$°C$, is a method of cancer therapy. It is normally applied as an additive therapy to enhance the effect of conventional radio- or chemotherapy. The standard way to produce local heating in the human body is the use of electromagnetic waves. We are mainly interested in regional hyperthermia of deep seated tumors. For this type of treatment usually a phased array of antennas surrounding the patient is used (see Fig. VII.17). The distribution of absorbed power within the patient's body can be steered by selecting the amplitudes and phases of the antennas' driving voltages. The space between the body and the antennas is filled by a so-called water bolus to avoid excessive heating of the skin.

From the viewpoint of computational medicine there are different challenges: (i) modelling and calculation of the electromagnetic field and the forced temperature, (ii) optimization of the channel adjustments to achieve favourable interference patterns for a successful cancer therapy, and (iii) visualization of vector fields and temperature distributions on a very complicated geometry. It should be possible to perform all steps of a simulation for each individual patient within a medical planning system [21].

In the following we use the fully adaptive code Kardos to solve a three–dimensional nonlinear heat transfer model within an optimization process to adjust the antennas. The simulation requires the numerical solution in a complex geometry involving a nonlinearity due to the perfusion term and different material properties of the tissues. As a prerequisite we need a three–dimensional geometric model in which the different tissue compartments are represented (see Fig. VII.18). Prior to grid generation, a segmentation of the CT data is performed, i.e., the relevant tissue compartments are defined on each scan. Then, the generation of a patient model consists of three steps: First, the compartment surfaces are extracted from the segmented CT data. For this purpose, Hege et al. have generalized the well–known marching cubes algorithm [93] for non-binary classifications [73]. This method creates a consistent description of the compartment interfaces. They are composed of so-called patches each separating two different compartments. Second, the surfaces are simplified to make them suitable for tetrahedron generation. An algorithm from computer graphics [65] has been extended to avoid intersections and ensure a high quality (i.e. aspect ratio) of the surface triangles. Third, each tissue compartment is

Figure VII.17: Patient lying in a Sigma 60 Applicator of the BSD 2000 Hyperthermia System. The patient is surrounded by 8 antennas emitting radiowaves. A water–filled bolus is placed between patient and antennas.

filled with tetrahedra using an advancing front algorithm. The compartment's surface is composed from the corresponding patches. At the beginning one starts with this surface. Then repeatedly a triangle of the advancing front is selected and a fourth point is searched such that the resulting tetrahedron resembles an equilateral one as much as possible. This procedure is continued until the whole compartment is filled with tetrahedra (see Seebass et al. [129]).

It is a rather difficult task to establish an appropriate physical model for the heat transport in the human body. Several approaches can be found in the literature (see eg. [150, 82]). The basis for our modelling is Pennes' bio–heat–transfer equation which employs a temperature–dependent blood perfusion model. A similar two–dimensional model was studied in [140] for ferromagnetic thermoseed hyperthermia. Finite element solutions for the electromagnetic fields [22] are taken as input data.

The optimization process is based on a specially designed objective function. The aim is to achieve a stationary temperature distribution which avoids "hot spots" (temperature greater than $44°C$) in healthy tissue and "cold spots" (temperature less than $42°C$) in the tumor region.

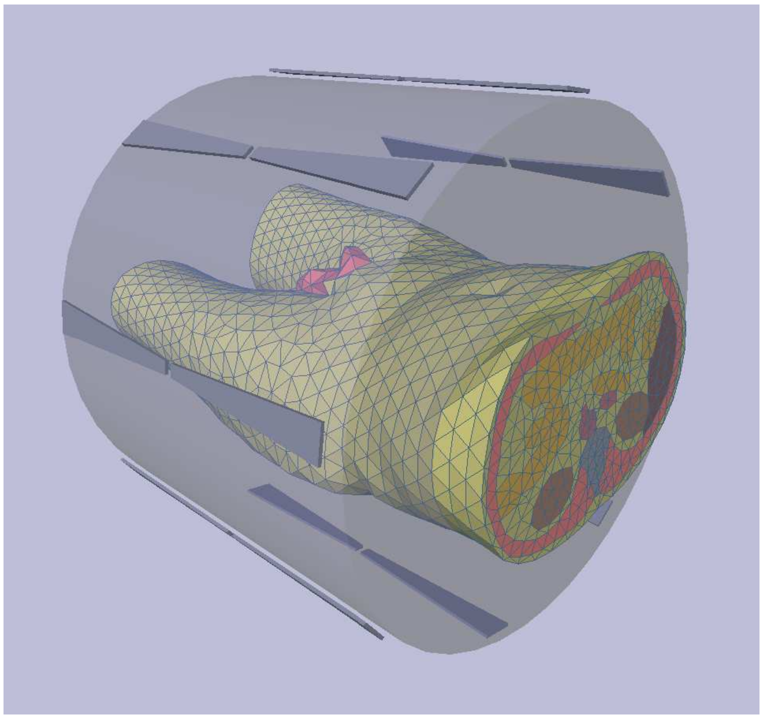


Figure VII.18: Three–dimensional finite element model of the patient's abdomen.

§4.1. Mathematical Modelling and Optimization

The basic model used in our simulation is the instationary bio–heat–transfer equation proposed by Pennes [111]

$$\rho c \frac{\partial T}{\partial t} = \operatorname{div}(\kappa \operatorname{grad} T) - c_b W (T - T_b) + Q_e \,, \tag{VII.38}$$

where ρ is the density, c and c_b are specific heat of tissue and blood, κ is the thermal conductivity; T_b is the blood temperature; and W is the mass flow rate of blood per unit volume of tissue. The power Q_e deposited by an electric field E in a tissue with electric conductivity σ is given by

$$Q_e = \frac{1}{2}\sigma |E|^2 \,. \tag{VII.39}$$

In hyperthermia applicators utilizing electromagnetic waves the antennas normally are grouped into *channels* that can be controlled independently. For such an applicator the total electric field E can be computed by superposition

$$E = \sum_{j=1}^{N_{chan}} a_j \; exp(-i\theta_j) \; E_j, \tag{VII.40}$$

where the channel j has amplitude a_j and phase delay θ_j. E_j is the electric field generated by the antennas of channel j. If complex values z_j are defined as

$$z_j = a_j \; exp(-i\theta_j) \tag{VII.41}$$

the absorbed power Q_e can be expressed as a quadratic function of z_j

$$Q_e = \frac{1}{2}\sigma \sum_{j,k=1}^{N_{chan}} z_j^* E_j^* E_k z_k \; . \tag{VII.42}$$

Besides the differential equation boundary conditions determine the temperature distribution. The heat exchange between body and water bolus can be described by the flux condition

$$\kappa \frac{\partial T}{\partial n} = \beta (T_{bol} - T) \tag{VII.43}$$

where T_{bol} is the bolus temperature and β is the heat transfer coefficient. No heat loss is assumed in remaining regions. We use for our simulations $\beta = 45W/m^2/^oC$ and $T_{bol} = 25^oC$.

Studies that predict temperatures in tissue models usually assume a constant–rate blood perfusion within each tissue. However, several experiments have shown that the response of vasculature in tissues to heat stress is strongly temperature–dependent [132]. When heated up to $41-43^oC$, temperatures that are commonly used in clinical hyperthermia, the blood flow in normal tissues, e.g., skin and muscle, increases significantly. In contrast, the tumor zone often appears to be so vulnerable to heat that the blood flow decreases on heating.

For the temperature dependence of blood perfusion we slightly simplified the curves presented in [140]. For healthy tissue (muscle and fat) we assume sigmoidal curves consisting of a Gaussian profile describing the perfusion increase between $37°C$ and $45°C$ and a plateau for temperatures above $45°C$ (see Fig. VII.19). In the raising part our curve differs from the one used in [140] only slightly. The differences are small compared to the uncertainties of the underlying experimental data [132]. In [140] a decrease of perfusion above 45°C is assumed. This is motivated by the observation that vasculature is destroyed if tissue is heated to such temperatures for about 30 minutes. We do not assume such a decrease of perfusion. With our objective function for optimization, this should not matter, because the objective function guarantees that temperatures in healthy tissue are always below 45°C. The curve assumed for fat tissue takes into account that fat tissue has a smaller capability to increase perfusion than muscle tissue. For tumor tissue a curve is used with the same shape as the curve for tumor core in [140]. We choose slightly different absolute values to make the results comparable with prior studies assuming constant–rate perfusion. The absolute values for blood perfusion are open for discussion, and the capability to increase perfusion also strongly depends on the cardiac state of the individual patient. But in this study we are mainly interested in qualitative effects of temperature–dependent blood flow.

In detail we use the following temperature–dependent blood perfusions:

Temperature–dependent blood perfusion in muscle:

$$W_{muscle} = \begin{cases} 0.45 + 3.55 \exp\left(-\dfrac{(T-45.0)^2}{12.0}\right), & T \leq 45.0 \\ 4.00\,, & T > 45.0 \end{cases} \tag{VII.44}$$

Temperature–dependent blood perfusion in fat:

$$W_{fat} = \begin{cases} 0.36 + 0.36 \exp\left(-\dfrac{(T-45.0)^2}{12.0}\right), & T \leq 45.0 \\ 0.72\,, & T > 45.0 \end{cases} \tag{VII.45}$$

Temperature–dependent blood perfusion in tumor:

$$W_{tumor} = \begin{cases} 0.833\,, & T < 37.0 \\ 0.833 - (T-37.0)^{4.8}/5.438E{+}3\,, & 37.0 \leq T \leq 42.0 \\ 0.416\,, & T > 42.0 \end{cases} \tag{VII.46}$$

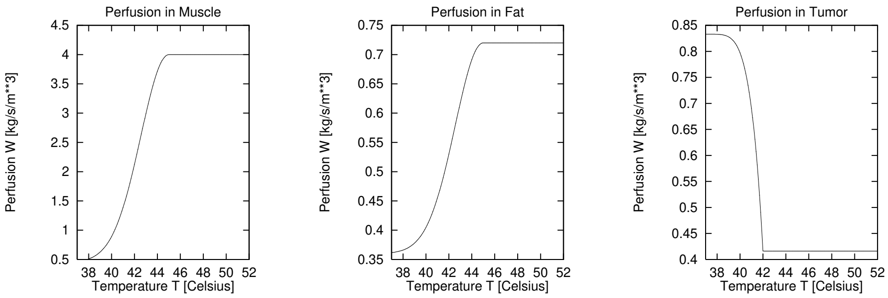


Figure VII.19: Nonlinear models of temperature–dependent blood perfusion for muscle tissue, fat tissue, and tumor.

Tissue	Thermal conductivity κ $[W/m/^\circ C]$	Electric conductivity σ $[1/m/\Omega]$	Density ρ $[kg/m^3]$	Specific heat c $[Ws/kg/^\circ C]$	Mass flow rate W $[kg/s/m^3]$
Fat	0.210	0.04	900	3,500	W_{fat} $(VII.45)$
Tumor	0.642	0.80	1,000	3,500	W_{tumor} $(VII.46)$
Bladder	0.600	0.60	1,000	3,500	5.000
Kidney	0.577	1.00	1,000	3,500	66.670
Liver	0.640	0.60	1,000	3,500	16.670
Muscle	0.642	0.80	1,000	3,500	W_{muscle} $(VII.44)$
Bone	0.436	0.02	1,600	1,000	0.540
Aorta	0.506	0.86	1,000	3,500	83.330
Intestine	0.550	0.60	1,000	3,500	3.333

Table VII.1: Material properties of tissues.

The material properties of the involved tissues are summarized in Tab. VII.1. For blood we take $T_b = 37^oC$ and $c_b = 3500 Ws/kg/^oC$. If a constant–rate perfusion model is applied, we assume mean perfusion values for muscle, $W_{muscle} = 2.3 kg/s/m^3$, and fat, $W_{fat} = 0.54 kg/s/m^3$. The maximum value $W_{tumor} = 0.833 kg/s/m^3$ is taken for tumor tissue.

The goal is to control the amplitudes z_j, $j = 1, \ldots, N_{chan}$, of the independent channels in order to achieve an effective hyperthermia therapy. A favourable temperature distribution is characterized as follows:

- Within the tumor a therapeutic temperature level of $42 - 43^oC$ is reached.
- No larger regions of healthy tissue are heated to above $42 - 43^oC$.
- Temperature in healthy tissue does not exceed certain temperature limits depending on the tissue type.

Taking into account these requirements an objective function is defined for optimization by

$$q = \int\limits_{\substack{x \,\in\, \text{tumor} \\ T \,<\, T_{ther}}} (T_{ther} - T)^2 \, dV \; + \int\limits_{\substack{x \,\notin\, \text{tumor} \\ T \,>\, T_{health}}} (T - T_{health})^2 \, dV \; + \; p \int\limits_{\substack{x \,\notin\, \text{tumor} \\ T \,>\, T_{lim}}} (T - T_{lim})^2 \, dV \,, \tag{VII.47}$$

where a therapeutic level $T_{ther} = 43^oC$ is used, and a temperature $T_{health} = 42^oC$ that should not be exceeded in healthy tissue. The limits T_{lim} are chosen tissue–dependent: $T_{lim} = 42^oC$ for more sensible tissue compartments (bladder, intestine) and $T_{lim} = 44^oC$ otherwise. To ensure high penalization for temperatures exceeding the limits $p = 1000$ is set. Optimization of the temperature distribution means now to choose the amplitudes z_j for each channel in such a way that the resulting temperature field minimizes the objective function q.

The definition of the objective function as an integral of squares guarantees that regions with large deviations from the desired temperatures, i.e., "hot spots" in healthy tissue and "cold spots" in the tumor, contribute large amounts to the objective function. A similar optimization strategy for a phased array hyperthermia system based on a simpler objective function is described in [104]. In contrast to the objective function proposed there, we add the second term which attempts to control excessive heating of healthy tissue. Moreover, the objective function is evaluated not only in a small number of selected points, but for the entire three–dimensional temperature distribution.

Using a piecewise linear finite element solution T_h which represents an approximation of the stationary temperature distribution on an adaptive spatial mesh M_h, and applying an integration formula based only on the vertices x_i (mass lumping), we get an approximation of the objective function (VII.47)

$$\begin{aligned} q_h &= \sum_{i\in M_{h1}} \frac{w_i}{4}(T_{ther} - T_h(x_i))^2 + \sum_{i\in M_{h2}} \frac{w_i}{4}(T_h(x_i) - T_{health})^2 \\ &\quad + p\sum_{i\in M_{h3}} \frac{w_i}{4}(T_h(x_i) - T_{lim})^2 \end{aligned} \tag{VII.48}$$

with

$$\begin{aligned} M_{h1} &= \{i\colon\, x_i \in \text{tumor},\, T_h(x_i) < T_{ther}\}, \\ M_{h2} &= \{i\colon\, x_i \notin \text{tumor},\, T_h(x_i) > T_{health}\}, \\ M_{h3} &= \{i\colon\, x_i \notin \text{tumor},\, T_h(x_i) > T_{lim}\}, \end{aligned}$$

and w_i stands for the volume of all tetrahedra of which x_i is a vertex.

It is now useful to split the temperature. The stationary temperature field T can be computed as sum of the basal temperature T_{bas} determined by $Q_e = 0$ and the temperature increment T_{inc} caused by the hyperthermic application. We easily derive the stationary equations for T_{bas} and T_{inc}

$$\begin{aligned} \operatorname{div}(\kappa \operatorname{grad} T_{bas}) - c_b W[T_{bas}](T_{bas} - T_b) &= 0\,, \\ \kappa\frac{\partial T_{bas}}{\partial n} - \beta(T_{bol} - T_{bas}) &= 0\,, \end{aligned} \tag{VII.49}$$

and

$$\begin{aligned} \operatorname{div}(\kappa \operatorname{grad} T_{inc}) - c_b(\, W[T_{inc} + T_{bas}]T_{inc} & \\ +(\, W[T_{inc} + T_{bas}] - W[T_{bas}]\,)\,(T_{bas} - T_b)\,) + Q_e &= 0\,, \\ \kappa\frac{\partial T_{inc}}{\partial n} + \beta T_{inc} &= 0\,. \end{aligned} \tag{VII.50}$$

This splitting allows one to distinguish clearly between local effects forced by the permanent cooling of the human body at the surface and the heating by the electromagnetic field.

In a next step we derive formulas for the quick calculation of the temperature field for arbitrary amplitudes z_j. Let us first consider the linear model with a constant–rate perfusion in each tissue. Then from (VII.50) it can be seen directly that T_{inc} depends linearly on the distribution of the absorbed power Q_e. Hence, a superposition principle is valid:

$$T_{inc}(\alpha_1 Q_e^{(1)} + \alpha_2 Q_e^{(2)}) = \alpha_1 T_{inc}(Q_e^{(1)}) + \alpha_2 T_{inc}(Q_e^{(2)}) . \qquad \text{(VII.51)}$$

According to the representation (VII.42) we get

$$T_{inc}(Q_e) = \sum_{j,k=1}^{N_{chan}} z_j^* \, T_{inc}(E_j^* E_k) \, z_k \ , \qquad \text{(VII.52)}$$

and finally for the whole stationary temperature distribution

$$T(Z) = T_{bas} + \sum_{j,k=1}^{N_{chan}} z_j^* \, T_{inc}(E_j^* E_k) \, z_k \ , \qquad \text{(VII.53)}$$

where Z is the vector of all z_j. The temperature increments $T_{inc}(E_j^* E_k)$ can be derived from N_{chan}^2 basic calculations combining two channels. Consequently, for an arbitrary set of parameters z_j the objective function can be computed very fast. The same holds for the first and second derivatives of the finite element solution T_h with respect to the parameters z_j.

In the nonlinear case, relation (VII.51) is no longer valid. Nevertheless, we can fix the nonlinear perfusion terms with respect to a given intermediate state Z_n of all amplitudes. Then we utilize representation (VII.52) as an approximation in a neighborhood of Z_n to perform the minimization process. Doing so we get a better Z_{n+1} for which we solve the nonlinear heat equation. The arising perfusion $W(T(Z_{n+1}))$ is once again fixed and the optimization is done. Improving successively the constant–rate model of the perfusion in such a way, we end up with a nearly optimal adjustment of the parameters z_j for the nonlinear model.

To start the optimization we calculate an initial optimized $Z_0^{(0)}$ employing our constant–rate perfusion model. Next we adjust the total power, i.e., we scale the amplitudes of $Z_0^{(0)}$ such that for the nonlinear model the maximum temperature in healthy tissue does not exceed 44^oC. Employing a damped Newton method for the optimization, the whole iteration process reads as in Fig. VII.20. The inner iteration is terminated if the objective function has changed by less than 0.02 within the last 10 iterations. To control the outer iteration we always compute the new stationary temperature $T(Z_{n+1}^{(0)})$ and compare it with the old one. If the difference becomes small enough (less than 0.05^oC), we stop the optimization process.

Choose initial value $Z_0^{(0)}$

for $n = 0, 1, \ldots$

  Calculate stationary temperature $T(Z_n^{(0)})$

  Calculate $W_n := W(T(Z_n^{(0)}))$

  Calculate $T_{\mathrm{inc}}(E_j^* E_k)$, $j, k = 1, \ldots, N_{\mathrm{chan}}$, employing W_n

  for $k = 0, 1, \ldots$

    Calculate $D^i q_h := \left.\dfrac{D^i q_h}{dZ^i}\right|_{Z=Z_n^{(k)}}, \quad i = 1, 2$

    Calculate $\Delta Z := -(D^2 q_h)^{-1} D^1 q_h$

    Find $\alpha_0 \in \{1, \frac{1}{2}, \frac{1}{4}, \ldots\}$ **such that**

      $q_h(Z_n^{(k)} + \alpha_0 \Delta Z) < q_h(Z_n^{(k)}) + \frac{1}{2}\alpha_0 D^1 q_h \Delta Z$

    Define $Z_n^{(k+1)} := Z_n^{(k)} + \alpha_0 \Delta Z$

    Finished?

  Define $Z_{n+1}^{(0)} := Z_n^{(k+1)}$

  Finished?

Figure VII.20: Flow diagram of the whole optimization process.

§4.2. Simulation for Two Individual Patients

We report some data concerning optimization processes for two individual patients. The simulations were done for the Sigma 60 Applicator of the BSD 2000 Hyperthermia System which consists of eight radio frequency antennas grouped in four antenna pairs. Each group can have its own amplitude and phase. Our aim is to control four different complex values z_j.

The patients have different tumor locations. Fig. VII.21 shows specific sagittal and transversal sections of both patients where the contours of bone and tumor are colored black and grey, respectively. It can be seen clearly that the tumor of the second patient is strongly shielded by bones whereas the first tumor is located in a more central position.

For both patients the optimization is completed after five outer iteration steps. The corresponding values of the objective function q_h and the maximum temperature difference are shown in Tab. VII.2. We state that the values of the objective function are reduced by nearly the same factor 2/3.

	n	0	1	2	3	4	5
Patient 1	q_h	1,732	1,458	1,327	1,263	1,229	1,214
	$\|\|\delta T\|\|_\infty$	-	3.5	0.7	0.18	0.085	0.043
Patient 2	q_h	4,264	2,743	2,796	2,813	2,819	2,823
	$\|\|\delta T\|\|_\infty$	-	3.2	0.3	0.15	0.077	0.044

Table VII.2: History of objective function q_h and maximal temperature difference during the optimization process.

In Fig. VII.22 the convergence history of the vector $Z_n^{(0)}$ is presented. For each outer iteration step all complex amplitudes z_j are plotted as vertices of a quadrilateral. We observe that the use of the nonlinear heat transfer model leads to a more uniform adjustment of $|z_j|$ and to a slight reduction of the phase differences. Moreover, the phases of the antenna pairs at the left and right of the patient come successively closer. They can be identified in the diagrams as neighbors of the channel with fixed phase zero. We observe a more symmetric adjustment of the phases.

Let us now compare the optimized temperature distribution based on an adaptively improved spatial grid with the temperature field computed on the coarse grid. Fig. VII.24 shows two cuts through the computational domain of the second patient involving the tumor boundary to give an impression of the local refinement process. The coarse grid contains $7,140$ vertices (degrees of freedom for the finite element solution), while the refined grid has $35,936$ vertices. Starting with the coarse grid two refinement steps are necessary to reach an accuracy of 2%. The corresponding uniform grid would have about $420,000$ degrees of freedom which demonstrates the power of the proposed adaptive method.

Fig. VII.23 illustrates the influence of the adaptive mesh control on the adjustment of the antenna pairs and the objective function. The optimization process based on the coarse grid requests five outer iteration loops and reaches a maximum temperature difference $||\delta T||_\infty = 0.03^oC$ at $q_h = 2,505$. Comparing the final value of the objective function with the value given in Tab. VII.2 for the fine mesh, $q_h = 2,823$, the attained change ranges in the order of 10%. This is also reflected by the adjustment of the applicator. The same difference can be observed for the temperature increment T_{inc} with respect to coarse and fine meshes. The local refinement controlled by a posteriori error estimates leads to a better resolution of the solution in regions with high temperature gradients and material transitions.

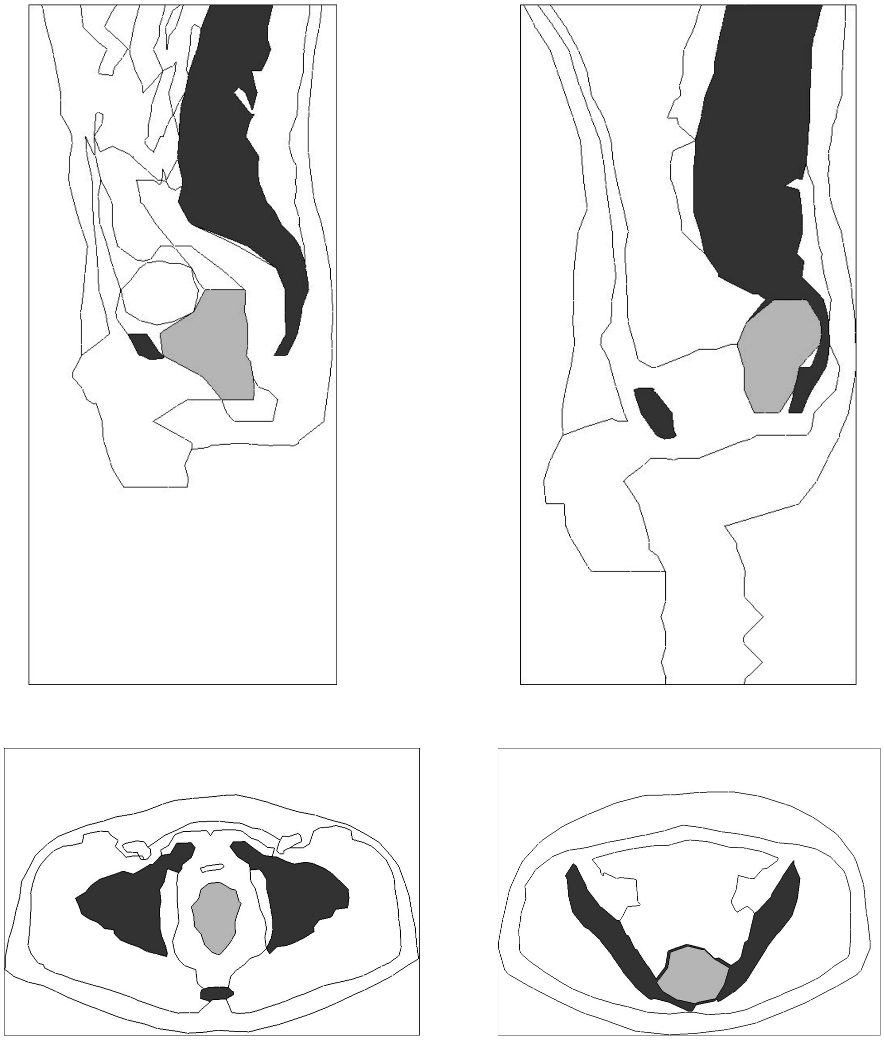


Figure VII.21: Contours of tissue compartments in specific sagittal (top) and transversal (bottom) sections. The location of tumors (grey) with respect to bone (black) is shown for the first patient (left) and the second patient (right).

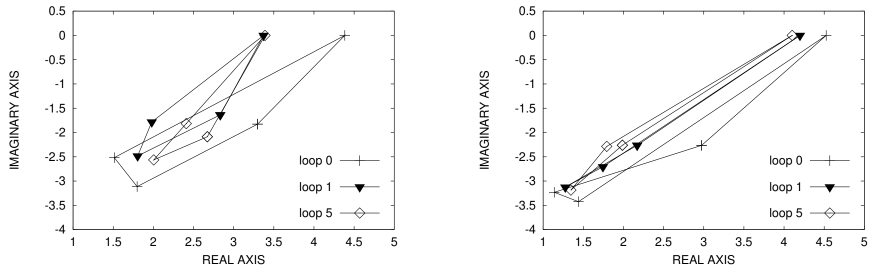


Figure VII.22: Patient 1 (left) and patient 2 (right). Optimization of the four complex amplitudes plotted in a quadrilateral for each outer iteration step.

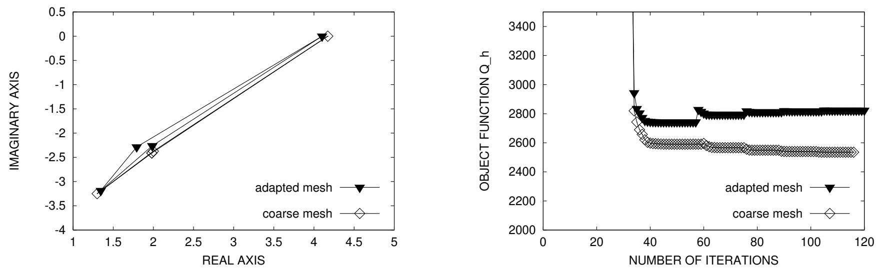


Figure VII.23: Patient 2. Influence of the adapted mesh control on the optimized adjustment of the antenna pairs and the minimization history of the objective function.

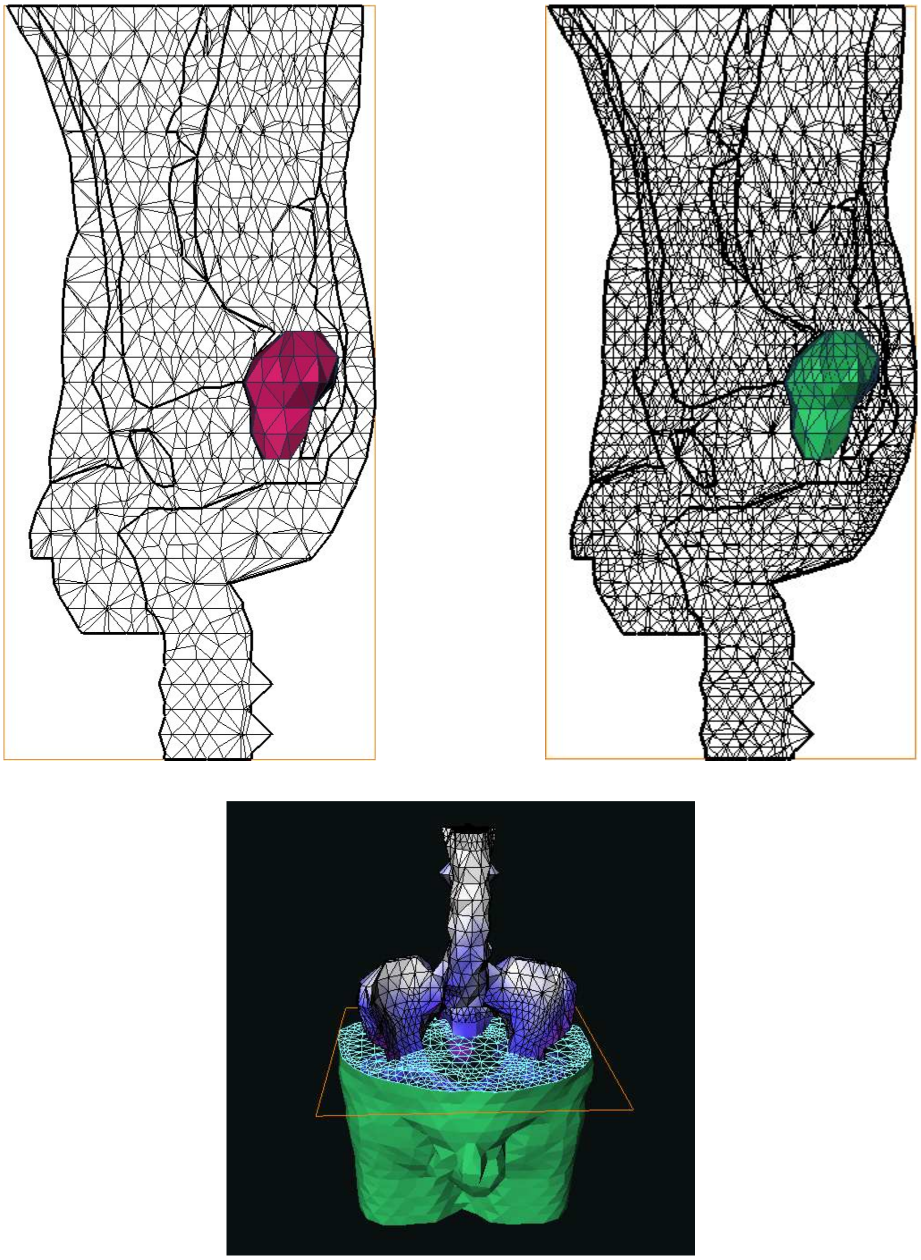


Figure VII.24: Patient 2. Coarse (top left) and refined (top right and bottom) grid with tumor boundary used for the computation of the optimized temperature distribution.

Appendix A: Advanced Tools from Functional Analysis

§1. Gelfand Triple

Let $\mathcal{X}$ be a reflexive Banach space, densely and continuously embedded in a Hilbert space $\mathcal{H}$,

$$\mathcal{X} \hookrightarrow \mathcal{H} .$$

The corresponding inner product will be denoted by $\langle \cdot, \cdot \rangle_{\mathcal{X}}$. The Hahn–Banach separation theorem ([119], Theorem 3.4) shows that $\mathcal{H}$ can be densely embedded into the conjugate dual space $\mathcal{X}'$ of $\mathcal{X}$,

$$\mathcal{H} \hookrightarrow \mathcal{X}^* = \mathcal{X}' .$$

Any such scale

$$\mathcal{X} \hookrightarrow \mathcal{H} \hookrightarrow \mathcal{X}'$$

is called a Gelfand triple, or sometimes, a rigging of $\mathcal{H}$. Defining the norm in $\mathcal{X}'$ by

$$\|x'\|_{\mathcal{X}'} = \sup_{x \in \mathcal{X}} \frac{\langle x', x \rangle}{\|x\|_{\mathcal{X}}}, \qquad x' \in \mathcal{X}' ,$$

giving the dual conjugate pairing of the two spaces $\mathcal{X}'$ and $\mathcal{X}$ by $\langle \cdot, \cdot \rangle$ which is conjugate linear in the second argument, we have

$$C^{-1} \|x\|_{\mathcal{X}'} \leq \|x\|_{\mathcal{H}} \leq C \|x\|_{\mathcal{X}}, \qquad x \in \mathcal{X}$$

for some positive and moderate constant C. Using density and uniform continuity arguments, the inner product $(\cdot, \cdot)$ of $\mathcal{H}$ extends uniquely to a sesquilinear form on $\mathcal{X}' \times \mathcal{X}$,

$$\langle x', x \rangle = \lim_{h \to x'} (h, x)_{\mathcal{H}}, \quad x \in \mathcal{X}.$$

The extension can be considered as new representation of all $x' \in \mathcal{X}'$.

If $\mathcal{X}$ is additionally a separable Hilbert space then $\mathcal{X}'$ is a separable Hilbert space as well. In that case the Riesz representation theorem ([42], Chap. VI.1, Theorem 10) shows that there is an unitary operator

$$\mathcal{R} : \mathcal{X}' \to \mathcal{X} \ ,$$

defined by

$$\langle x, \mathcal{R}x' \rangle_{\mathcal{X}} = \langle x', x \rangle \, , \quad x' \in \mathcal{X}', x \in \mathcal{X}.$$

The inverse of $\mathcal{R}$ is given by the adjoint operator

$$\mathcal{R}^* : \mathcal{X} \to \mathcal{X}' \ ,$$

which satisfies

$$\langle \mathcal{R}^* x, y \rangle = \langle y, x \rangle_{\mathcal{X}} \, , \quad x \in \mathcal{X}, y \in \mathcal{X}.$$

The map $\mathcal{R}$ will be called the Riesz representation map of the Gelfand triple in the case of Hilbert spaces. If we wish, we can therefore identify the Hilbert space $\mathcal{X}$ with its anti–dual $\mathcal{X}'$.

§2. Sesquilinear Forms and Bounded Operators in Hilbert Spaces

Let $\mathcal{X}$ and $\mathcal{Y}$ be two complex Hilbert spaces. A sesquilinear form on $\mathcal{X} \times \mathcal{Y}$ is defined as a mapping

$$(x, y) \to a(x, y) \in \mathbb{C}$$

satisfying for $\lambda_1, \lambda_2 \in \mathbb{C}$ the conditions

$$\begin{array}{rlcl} (i) & a(\lambda_1 x_1 + \lambda_2 x_2, y) & = & \lambda_1 \, a(x_1, y) + \lambda_2 \, a(x_2, y), \\ (ii) & a(x, \lambda_1 y_1 + \lambda_2 y_2) & = & \overline{\lambda_1} \, a(x, y_1) + \overline{\lambda_2} \, a(x, y_2). \end{array}$$

Thus the mapping is a linear form for y fixed and antilinear (or semi–linear) form for x fixed. A sesquilinear form is *continuous* or *bounded* on $\mathcal{X} \times \mathcal{Y}$ if there exists a constant $C > 0$ such that

$$|a(x, y)| \le C \, \|x\|_{\mathcal{X}} \|y\|_{\mathcal{Y}}, \qquad x \in \mathcal{X}, y \in \mathcal{Y}. \tag{A.1}$$

Denoting by $\mathcal{S}(\mathcal{X} \times \mathcal{Y})$, $\mathcal{S}(\mathcal{X})$ if $\mathcal{Y} = \mathcal{X}$, the set of all continuous sesquilinear forms on $\mathcal{X} \times \mathcal{Y}$, this inequality allows us to define the norm $\|a\|_{\mathcal{S}(\mathcal{X} \times \mathcal{Y})}$ of a by the smallest constant C in (A.1). We have the following

Theorem A.1. *There exists an isomorphism of* $\mathcal{S}(\mathcal{X}\times\mathcal{Y})$ *onto* $\mathcal{L}(\mathcal{X},\mathcal{Y})$ *which associates with a sesquilinear form* $a(x,y)$ *the operator* $A\in\mathcal{L}(\mathcal{X},\mathcal{Y})$ *defined by*

$$\langle Ax,y\rangle_{\mathcal{Y}} = a(x,y), \qquad x\in\mathcal{X}, y\in\mathcal{Y}$$

and we have

$$\|A\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} = \|a\|_{\mathcal{S}(\mathcal{X}\times\mathcal{Y})}.$$

Proof. Applying Riesz' theorem for x fixed, the mapping $y\to a(x,y)$ defines $Ax\in\mathcal{Y}$ such that $a(x,y)=\langle\mathcal{R}x,y\rangle_{\mathcal{Y}}, y\in\mathcal{Y}$. The mapping A is linear and continuous. The converse follows immediately. ∎

Defining the adjoint sesquilinear form $a^*\in\mathcal{S}(\mathcal{Y}\times\mathcal{X})$ by

$$a^*(y,x)=\overline{a(x,y)}, \qquad y\in\mathcal{Y}, x\in\mathcal{X},$$

from Riesz' theorem follows that there exists $A^*\in\mathcal{L}(\mathcal{Y},\mathcal{X})$ called the adjoint of A such that

$$a^*(y,x)=\langle A^*y,x\rangle_{\mathcal{X}}, \qquad y\in\mathcal{Y}, x\in\mathcal{X}.$$

Now let $\mathcal{Y}=\mathcal{X}$ and $\mathcal{X}'$ be the anti–dual of $\mathcal{X}$, which we do not identify here with $\mathcal{X}$. We consider continuous sesquilinear forms on $\mathcal{X}\times\mathcal{X}'$ and get the analogue of Theorem A.1.

Theorem A.2. *There exists an isomorphism of* $\mathcal{S}(\mathcal{X})$ *onto* $\mathcal{L}(\mathcal{X},\mathcal{X}')$ *which associates with a sesquilinear form* $a(x,y)$ *the operator* $A\in\mathcal{L}(\mathcal{X},\mathcal{X}')$ *defined by*

$$\langle Ax,y\rangle = a(x,y), \qquad x,y\in\mathcal{X},$$

and we have

$$\|A\|_{\mathcal{L}(\mathcal{X},\mathcal{X}')} = \|a\|_{\mathcal{S}(\mathcal{X})}.$$

To assume that A is an isomorphism of $\mathcal{X}$ onto $\mathcal{X}'$ we have to define an additional property: The sesquilinear form $a(x,y)\in\mathcal{S}(\mathcal{X})$ is said to be $\mathcal{X}$*–elliptic* if it satisfies

$$\Re e\; a(x,x)\geq\alpha\|x\|^2_{\mathcal{X}}, \qquad x\in\mathcal{X}$$

with a real constant $C>0$.

We then have the well–known

Theorem A.3. (Lax–Milgram Theorem). *Let* $a(x,y)\in\mathcal{S}(\mathcal{X})$ $\mathcal{X}$*–elliptic and let* $A\in\mathcal{L}(\mathcal{X},\mathcal{X}')$ *be the associated operator. Then* A *is an isomorphism of* $\mathcal{X}$ *onto* $\mathcal{X}'$.

Proof. See [42], Chap. VI.3, Theorem 7. ∎

§3. Unbounded Operators in Hilbert Spaces

In many situations, e.g., concerning evolution equations $\mathcal{X}$–ellipticity of the underlying sesquilinear form cannot be achieved. Here, the construction of a Gelfand triple (see also §A.1)

$$\mathcal{X} \hookrightarrow \mathcal{H} \hookrightarrow \mathcal{X}'$$

with a second Hilbert space $\mathcal{H}$ is very useful. The sesquilinear form $a(x,y) \in \mathcal{S}(\mathcal{X})$ is said to be *$\mathcal{X}$–coercitive* (with respect to $\mathcal{H}$) if there exist $\kappa_0 \in \mathbb{R}$ and $\alpha > 0$ such that

$$\Re e\, a(x,x) + \kappa_0 \|x\|_{\mathcal{H}}^2 \geq \alpha \|x\|_{\mathcal{X}}^2, \qquad x \in \mathcal{X}.$$

As a direct consequence we observe that the sesquilinear forms $a(x,y)+\kappa(x,y)_{\mathcal{H}}$ are $\mathcal{X}$–elliptic for all $\kappa \geq \kappa_0$. The theorem of Lax–Milgram shows the existence of associated operators $A_\kappa \in \mathcal{L}(\mathcal{X}, \mathcal{X}')$ which are isomorphisms of $\mathcal{X}$ onto $\mathcal{X}'$. That means, if we take $x \in \mathcal{H}$, there is a unique $y \in \mathcal{X}$ such that $A_\kappa y = x \in \mathcal{H}$. Defining the domain of A_κ by

$$\mathcal{D}(A_\kappa) = A_\kappa^{-1}(\mathcal{H}) = \{x \in \mathcal{H} \ : \ A_\kappa x \in \mathcal{H}\}$$

we can consider A_κ restricted to $\mathcal{D}(A_\kappa)$ as an unbounded operator in $\mathcal{H}$. The Lax–Milgram theorem states that this restriction is an isomorphism of $\mathcal{D}(A_\kappa)$ onto $\mathcal{H}$.

Usually, the situation is the following: Given an unbounded operator A with domain $\mathcal{D}(A)$ in a Hilbert space $\mathcal{H}$, one wishes to extend this operator to an isomorphism of $\mathcal{X}$ onto $\mathcal{X}'$ associating with it a sesquilinear form on $\mathcal{X} \times \mathcal{X}$. We do not dwell further on this question here, and refer for a deeper study to KATO [81].

§4. Analytic Semigroups

A linear operator A in a complex Banach space $\mathcal{X}$ is called *sectorial* if it is a closed densely defined operator and there are constants $\phi \in (0, \pi/2)$, $M \geq 0$, and $a \in \mathbb{R}$ such that the sector

$$S_{a,\phi} = \{\lambda \in \mathbb{C} \ : \ \phi \leq |\arg(\lambda - a)| \leq \pi, \ \lambda \neq a\}$$

contains no part of the spectrum of A, and

$$\|(\lambda - A)^{-1}\|_{\mathcal{L}(\mathcal{X})} \leq \frac{M}{|\lambda - a|}, \qquad \lambda \in S_{a,\phi}.$$

Denoting by $\rho(A)$ the resolvent set of A, we get for a sectorial operator $S_{a,\phi} \subset \rho(A)$. The angle of the section $S_{a,\phi}$ is $2\pi - 2\phi > \pi$. Note, the above definition is very general such that many elliptic boundary value problems define sectorial operators.

A family $\{S(t)\}_{t\geq 0}$ of elements $S(t) \in \mathcal{L}(\mathcal{X})$ forms an *analytic semigroup* in $\mathcal{X}$ if

(i) $S(0) = I$, the identity;

(ii) $S(t+s) = S(t)S(s), \quad s \geq 0, t \geq 0,$ the semigroup property;

(iii) $S(t)x \to x$ as $t \to 0^+$ for all $x \in \mathcal{X}$;

(iv) $t \to S(t)x$ is analytic on $0 < t < \infty$ for all $x \in \mathcal{X}$.

The infinitesimal generator G of this semigroup is defined by

$$Gx = \lim_{t \to 0^+} \frac{1}{t}(S(t)x - x)$$

and has a domain containing all $x \in \mathcal{X}$ for which the limit exists in $\mathcal{X}$.

Analytic semigroups are generated by sectorial operators. We have ([74], Theorem 1.3.4)

Theorem A.4. *A is a sectorial operator, iff $-A$ is the infinitesimal generator of an analytic semigroup $\{e^{-At}\}_{t\geq 0}$, where*

$$e^{-At} = \frac{1}{2\pi i}\int_{\Gamma} (\lambda + A)^{-1} e^{\lambda t} d\lambda \,.$$

Here, Γ is a contour to the right of the spectrum $\sigma(-A)$ with $\arg \lambda \to \pm\theta$ as $|\lambda| \to \infty$, for some $\theta \in (\pi/2, \pi)$. The semigroup is differentiable for $t > 0$ and can be extended analytically into a domain $\triangle_\varepsilon = \{z \in \mathbb{C} : |\arg z| < \varepsilon, z \neq 0\}$ containing the positive real axis. If $\Re e\, \sigma(A) > a$, we get for $t > 0$

$$\|e^{-At}\|_{\mathcal{L}(\mathcal{X})} \leq Ce^{-at}\,,$$

and

$$\|\frac{d}{dt}e^{-At}\|_{\mathcal{L}(\mathcal{X})} = \|-Ae^{-At}\|_{\mathcal{L}(\mathcal{X})} \leq \frac{C}{t}e^{-at}$$

for some constant C.

Multiplying the semigroup $\{e^{-At}\}_{t\geq 0}$ by $e^{-\omega t}$ with some appropriate positive ω, we get an uniformly bounded analytic semigroup generated by the operator $-\hat{A}$ with $\hat{A} = A + \omega I$ and $0 \in \rho(\hat{A})$. This translation does not affect the possibility of extending the analytic semigroup into some sector $\triangle_\varepsilon$. Moreover, $\|S(z)\|$ is uniformly bounded in every closed subsector $\triangle_{\varepsilon'} = \{z \in \mathbb{C} : |\arg z| \leq \varepsilon' < \varepsilon\}$.

The property of $-\hat{A}$ to be the infinitesimal generator of a bounded analytic semigroup is equivalent ([110], Theorem 2.5.2) to the existence of constants $\delta \in (0, \pi/2)$ and $M \geq 0$ such that

$$\rho(-\hat{A}) \supset \Sigma_\delta = \{\lambda \in \mathbb{C} : |\arg \lambda| < \frac{\pi}{2} + \delta\} \cup \{0\},$$

and

$$\|(\lambda I + \hat{A})^{-1}\|_{\mathcal{L}(\mathcal{X})} \le \frac{M}{|\lambda|}, \quad \lambda \in \Sigma_\delta .$$

If Σ_δ includes a whole neighborhood of zero, a sometimes stronger inequality with $M_1 > M$ is valid ([43], Chap. XVII.6, Theorem 1)

$$\|(\lambda I + \hat{A})^{-1}\|_{\mathcal{L}(\mathcal{X})} \le \frac{M_1}{1 + |\lambda|}, \quad \lambda \in \Sigma_\delta .$$

Setting for $\alpha > 0$

$$\hat{A}^{-\alpha} = \frac{1}{\Gamma(\alpha)} \int_0^\infty t^{\alpha-1} e^{-\hat{A}t}\, dt,$$

which is a bounded linear one–to–one operator on $\mathcal{X}$, fractional powers of the operator $\hat{A}$ are defined by the inverse of $\hat{A}^{-\alpha}$,

$$\hat{A}^\alpha = (\hat{A}^{-\alpha})^{-1}, \quad \alpha > 0 .$$

Additionally, we set $\hat{A}^\alpha = I$ for $\alpha = 0$. Some simple properties of $\hat{A}^\alpha$ are

(*i*) $\hat{A}^\alpha$ is a closed densely defined operator on $\mathcal{X}$;

(*ii*) $\alpha \ge \beta > 0$ implies $\mathcal{D}(\hat{A}^\alpha) \subset \mathcal{D}(\hat{A}^\beta)$;

(*iii*) $\hat{A}^{\alpha+\beta} = \hat{A}^\alpha \hat{A}^\beta = \hat{A}^\beta \hat{A}^\alpha$ on $\mathcal{D}(\hat{A}^\gamma)$ where $\gamma = \max(\alpha, \beta, \alpha+\beta)$;

(*iv*) $\hat{A}^\alpha e^{-\hat{A}t} = e^{-\hat{A}t} \hat{A}^\alpha$ on $\mathcal{D}(\hat{A}^\alpha), t > 0$.

§5. Vectorial Functions Defined on Real Intervals

Let $\mathcal{V}$ be a Banach space and $\Omega \subset \mathbb{R}^d$ be a bounded open set. Functions dependent on $x \in \Omega$ and $t \in (0,T)$ will be treated as functions of t with values in $\mathcal{V}(\Omega)$ ([43], Chap. XVIII, §1). Then, $L^2(0,T;\mathcal{V})$ stands for the Lebesgue–Bochner space of L^2-summable functions on $(0,T)$ which have values in $\mathcal{V}$. Equipped with the norm

$$\|v\|_{L^2(0,T;\mathcal{V})} = \left(\int_0^T \|v(t)\|_{\mathcal{V}}^2\, dt \right)^{1/2}$$

where dt is the Lebesgue measure on $(0,T)$, the space $L^2(0,T;\mathcal{V})$ becomes a Hilbert space whenever $\mathcal{V}$ is a Hilbert space.

We further define Sobolev spaces

$$H^q(0,T;\mathcal{V}) := \left\{v \mid \partial_t^j v \in L^2(0,T;\mathcal{V}),\, j=0,\dots,q\right\}$$

with the norm

$$\|v\|_{H^q(0,T;\mathcal{V})} = \left(\int_0^T \sum_{j=0}^{q} \|\partial_t^j v(t)\|_{\mathcal{V}}^2 \, dt\right)^{1/2} .$$

If $\mathcal{V}$ is a Hilbert space, so is the space $H^q(0,T;\mathcal{V})$. Here, all partial derivatives $\partial_t^j v$ have to be understood in the distributional sense. Clearly, $H^0(0,T;\mathcal{V}) = L^2(0,T;\mathcal{V})$.

To describe continuity in time we introduce the space $C^0([0,T];\mathcal{V})$ consisting of all functions v which are continuous in t in the norm of $\mathcal{V}$, that is, such that $\|v(\cdot,t+\delta t) - v(\cdot,t)\|_{\mathcal{V}} \to 0$ as $\delta t \to 0$ for all $t \in [0,T]$. $C^0([0,T];\mathcal{V})$ is a Banach space equipped with the norm

$$\|v\|_{C^0([0,T];\mathcal{V})} = \sup_{0\le t\le T} \|v(t)\|_{\mathcal{V}} .$$

The preceding definition can be generalized to functions which, together with all time derivatives of orders $\le q$, are continuous in t in the norm of $\mathcal{V}$. We set

$$C^q([0,T];\mathcal{V}) := \left\{v \mid \partial_t^j v \in C^0([0,T];\mathcal{V}),\, j=0,\dots,q\right\}$$

provided with the norm

$$\|v\|_{C^q([0,T];\mathcal{V})} = \sum_{j=0}^{q} \sup_{0\le t\le T} \|\partial_t^j v(t)\|_{\mathcal{V}} .$$

We also use the shorthand notation $L_t^2(\mathcal{V})$, $H_t^q(\mathcal{V})$, and $C_t^q(\mathcal{V})$ for the above spaces.

Let us recall one important fact.

Theorem A.5. *The space $H^q(0,T;\mathcal{V})$, $q \ge 1$, is continuously embedded in $C^{q-1}([0,T];\mathcal{V})$.*

Proof. For $q=1$, the embedding follows directly from Theorem 3.1 [92], setting $\mathcal{X} = \mathcal{Y} = \mathcal{V}$, $m=1$, $a=0$, and $b=T$. The general result is then obtained by induction with respect to q. ∎

We mention that there are stronger identifications available considering absolutely continuous functions ([19], Theorem 2.2).

Appendix B: Consistency and Stability of Rosenbrock Methods

§1. Order Conditions

The order of consistency describes how rapidly the local error of an integration method tends to zero for a decreasing step size. Formal Taylor expansion of the local error leads to order conditions for the coefficients of the Rosenbrock methods given in (II.18). The procedure is conceptually simple, but needs special care because the resulting expressions become quite lengthy for higher order. The conditions for order $p \leq 4$ are:

$$\begin{aligned}
p=1 \quad &: \quad \textstyle\sum_{i=1}^{s} b_i = 1, \\
p=2 \quad &: \quad \textstyle\sum_{i,k=1}^{s} b_i\beta_{ik} = \frac{1}{2}, \\
p=3 \quad &: \quad \textstyle\sum_{i=1}^{s} b_i\alpha_i^2 = \frac{1}{3}, \\
&\quad\;\; \textstyle\sum_{i,k,l=1}^{s} b_i\beta_{ik}\beta_{kl} = \frac{1}{6}. \\
p=4 \quad &: \quad \textstyle\sum_{i=1}^{s} b_i\alpha_i^3 = \frac{1}{4}, \\
&\quad\;\; \textstyle\sum_{i,k,l=1}^{s} b_i\alpha_{ik}\beta_{kl}\alpha_i^2 = \frac{1}{8}, \\
&\quad\;\; \textstyle\sum_{i,k,l,m=1}^{s} b_i\beta_{ik}\alpha_{kl}\alpha_{km} = \frac{1}{12}, \\
&\quad\;\; \textstyle\sum_{i,k,l,m=1}^{s} b_i\beta_{ik}\beta_{kl}\beta_{lm} = \frac{1}{24}.
\end{aligned} \tag{B.1}$$

For higher order conditions we refer to [72].

§2. The Stability Function

The s–stage Rosenbrock method (II.18) applied to the scalar equation $y' = \lambda y$ yields after one step with $\tau > 0$

$$y_1 = R(\tau\lambda)y_0 \tag{B.2}$$

with the so–called stability function

$$R(z) = 1 + zb^T(I - z\mathcal{B})^{-1}\mathbf{1}.$$

In general $R(z)$ becomes a rational function of the form

$$R(z) = \frac{P(z)}{(1-\gamma z)^s},$$

where P is a polynomial of degree $\leq s$. If the method is of order p, we have

$$e^z - R(z) = Cz^{p+1} + O(z^{p+2}) \quad \text{for } z \to 0,$$

showing that $R(z)$ is a rational approximation to e^z. Properties of stability functions have been extensively studied [105, 149].

We call an integration method *A–stable* if $|R(z)| \leq 1$ for arbitrary $z = \tau\lambda$, $\Re e(\lambda) \leq 0$. Then, the solution process in (B.2) is stable in the sense that the damping property of the solution operator $e^{\lambda t}$ is correctly reflected. Stiff components with $\lambda << 0$ will be damped out much faster if we additionally require $R(\infty) = 0$, the condition of *L–stability*. Rosenbrock methods with good stability properties can be constructed for $p \leq s$ without any difficulties.

§3. The Property "Stiffly Accurate"

PROTHERO and ROBINSON [114] proposed to study the model problem

$$y' = \lambda(y - \psi(t)) + \psi'(t)$$

with the solution

$$y(t+\tau) = e^{\lambda\tau}(y(t) - \psi(t)) + \psi(t+\tau).$$

When solving stiff differential equations we are mainly interested in step sizes τ which are much larger than $|\lambda|^{-1}$. Prothero and Robinson therefore suggested to investigate the error behaviour of an integration method for the above equation when simultaneously $\tau \to 0$ and $\Re e(\lambda\tau) \to -\infty$. In this case we observe $y(t+\tau) \to \psi(t+\tau)$. A Rosenbrock method handles this particular transition to infinite stiffness and yields asymptotically exact results if the parameters satisfy

$$\alpha_{si} + \gamma_{si} = b_i \quad (i = 1, \ldots, s) \qquad \text{and} \qquad \alpha_s = 1. \tag{B.3}$$

This result was shown by HAIRER and WANNER [72] who call such methods *stiffly accurate*. There it is also argued that stiff accuracy is advantageous when solving stiff differential–algebraic equations. It can be shown that (B.3) implies $R(\infty) = 0$ automatically.

Although unbounded differential operators are strongly related to infinite stiffness, the virtue of stiff accuracy for general nonlinear PDEs is not clear. An interesting interpretation is given in [122] which can be adopted to our formulation (II.18) as follows. We easily derive with (B.3)

$$K'_{ns} = F(t_{n+1}, K_{ns}) + F_u(t_n, u_n)(u_{n+1} - K_{ns}),$$

and assuming that the Jacobian is invertible, we get further

$$u_{n+1} = F_u(t_n, u_n)^{-1}(K'_{ns} - F(t_{n+1}, K_{ns})) + K_{ns}.$$

This equation can be interpreted as the result of one modified Newton iteration applied to

$$K'_{ns} - F(t, y) = 0,$$

with starting value K_{ns}. The authors argue now that if K_{ns} is a sufficiently good starting guess and K'_{ns} is close to a true derivative then the Rosenbrock solution u_{n+1} is an approximation of a collocation equation - a property which seems desirable.

Table of Notations

Function Spaces

Ω	domain in $\mathbb{R}^d$
$\mathcal{X}$	Banach space
$C^k(\overline{\Omega}, \mathcal{X})$	k–times continuously differentiable functions $f : \Omega \to \mathcal{X}$, all derivatives have a bounded maximum norm
$C^k(\overline{\Omega})$	$C^k(\overline{\Omega}, \mathbb{R})$
$C_t^k(\mathcal{X})$	$C^k([0,T], \mathcal{X})$
$L^p(\Omega, \mathcal{X})$	Lebesgue–Bochner spaces, Appendix A
$L^p(\Omega)$	$L^p(\Omega, \mathbb{R})$
$L_t^p(\mathcal{X})$	$L^p(0,T;\mathcal{X}) = L^p((0,T), \mathcal{X})$
$H^s(\Omega, \mathcal{X})$	Sobolev spaces, Appendix A
$H^s(\Omega)$	$H^s(\Omega, \mathbb{R})$
$H_t^s(\mathcal{X})$	$H^s(0,T;\mathcal{X}) = H^s((0,T), \mathcal{X})$

Linear Operators

A	linear operator
$D(A)$	domain of A
$\sigma(A)$	spectrum of A
$\rho(A)$	resolvent set of A
A^α	fractional orders of A

Operator Spaces

$\mathcal{X}, \mathcal{Y}$	topological vector spaces
$\mathcal{L}(\mathcal{X}, \mathcal{Y})$	bounded linear operators from $\mathcal{X}$ to $\mathcal{Y}$
$\mathcal{L}(\mathcal{X})$	$\mathcal{L}(\mathcal{X}, \mathcal{X})$
$\mathcal{X}'$	dual space of $\mathcal{X}$

Bibliography

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975

[2] S. Adjerid and J.E. Flaherty, *Second–order finite element approximations and a posteriori error estimation for two–dimensional parabolic systems*, Numer. Math., 53 (1988), pp. 183–198

[3] M. Ainsworth and J. T. Oden, *A unified approach to a posteriori error estimation using element residual methods*, Numer. Math., 65 (1993), pp. 23–50

[4] H. Amann, *Dynamic Theory of Quasilinear Parabolic Equations II. Reaction–Diffusion Systems*, Differential and Integral Equations, 3 (1990), pp. 13–75

[5] H. Amann, *Nonhomogeneous Linear and Quasilinear Elliptic and Parabolic Boundary Value Problems*, in: H.-J. Schmeisser, H. Triebel (eds.), Function Spaces, Differential Operators and Nonlinear Analysis, Teubner–Texte zur Mathematik 133, pp. 9–126, Teubner Stuttgart, Leipzig, 1996

[6] H. Amann, *Linear and Quasilinear Parabolic Problems I. Abstract Linear Theory*, Birkhäuser-Verlag, Basel, Boston, Berlin, 1995

[7] E. Anderson, Z. Bai, C. Bischof et al., *LAPACK User's Guide*, SIAM 1992

[8] I. Babuška and A. K. Aziz, *On the angle condition in the finite element method*, SIAM J. Num. Anal., 13 (1976), pp. 214–226

[9] I. Babuška and A. K. Noor, *Quality assessment and control of finite element solutions*, Technical Note BN–1049 (1986), Univ. of Maryland

[10] I. Babuška and W. C. Rheinboldt, *Error estimates for adaptive finite element computations*, SIAM J. Numer. Anal., 15 (1978), pp. 736–754

[11] I. Babuška and W. C. Rheinboldt, *A posteriori error estimates for the finite element method*, Int. J. Numer. Methods Engrg., 12 (1978), pp. 1597–1615

[12] I. Babuška and B. Szabo, *Trends and new problems in finite element methods*, in: J.R. Whiteman (ed.), The Mathematics of Finite Element and Applications, pp. 1–33, John Wiley & Sons Ltd., England, 1997

[13] M.J. Baines, *Moving Finite Elements*, Clarendon Press, Oxford, 1994

[14] N. Yu. Bakaev, *On the bounds of approximations of holomorphic semigroups*, BIT, 35 (1995), pp. 605–608

[15] E. Bänsch, *Local Mesh Refinement in 2 and 3 Dimensions*, IMPACT Comput. Sci. Engrg., 3 (1991), pp. 181–191

[16] R. E. Bank, *Hierarchical bases and the finite element method*, in: A. Iserles (ed.), Acta Numerica (1996), pp. 1–43

[17] R. E. Bank, *PLTMG: A Software Package for Solving Elliptic Partial Differential Equations - User's Guide 8.0*, SIAM, 1998

[18] R. E. Bank and R. K. Smith, *A posteriori error estimates based on hierarchical bases*, SIAM J. Numer. Anal., 30 (1993), pp. 921–935

[19] V. Barbu, *Nonlinear semigroups and differential equations in Banach spaces*, Noordhoff International Publishing, Leyden, The Netherlands, 1976

[20] A. Bayliss and B.J. Matkowsky, *Fronts, Relaxation Oscillations, and Period Doubling in Solid Fuel Combustion*, J. Comp. Phys., 71 (1987), pp. 147–168

[21] R. Beck, P. Deuflhard, H.-C. Hege, M. Seebass, and D. Stalling, *Numerical Algorithms and Visualization in Medical Treatment Planning*, in H.-C. Hege and K. Polthier (eds.), Visualization and Mathematics, pp. 303-328, Springer-Verlag, Heidelberg, 1997

[22] R. Beck, P. Deuflhard, R. Hiptmair, R.H.W. Hoppe, and B. Wohlmuth, *Adaptive Multilevel Methods for Edge Element Discretizations of Maxwell's Equations*, SC 97–66, Konrad–Zuse–Zentrum für Informationstechnik Berlin, Germany, 1997, to be published in Surveys of Mathematics in Industry

[23] F. Benkaldoun, B. Denet, and B. Larrouturou, *Numerical Investigation of the Extinction Limit of Curved Flames*, Combust. Sci. and Tech., 64 (1989), pp. 187–198

[24] B.A.V. Bennet and M.D. Smooke, *Local rectangular refinement with application to axisymmetric laminar flames*, Comb. Theory and Modeling, 2 (1998), pp. 221–258

[25] M. Berzins, P.J. Capon, and P.K. Jimack, *On spatial adaptivity and interpolation when using the method of lines*, Appl. Numer. Math., 26 (1998), pp. 117–133

[26] J. Bey, *Analyse und Simulation eines Konjugierte–Gradienten–Verfahrens mit einem Multilevel–Präkonditionierer zur Lösung dreidimensionaler, elliptischer Randwertprobleme für massiv parallele Rechner*, Diplomarbeit, RWTH Aaachen, 1991

[27] M. Bieterman and I. Babuška, *An adaptive method of lines with error control for parabolic equations of the reaction–diffusion type*, J. Comput. Phys., 63 (1986), pp. 33–66

[28] F. Booth, *The Theory of Self–Propagating Exothermic Reactions in Solid Systems*, Trans. Faraday Soc., 49 (1953), pp. 272–281

[29] F. A. Bornemann, *An adaptive multilevel approach to parabolic equations. I. General theory and 1D implementation*, IMPACT of Comput. in Sci. and Engrg., 2 (1990), pp. 279–317

[30] F. A. Bornemann, *An adaptive multilevel approach to parabolic equations. II. Variable–order time discretization based on a multiplicative error correction*, IMPACT of Comput. in Sci. and Engrg., 3 (1991), pp. 93–122

[31] F. A. Bornemann, *An adaptive multilevel approach to parabolic equations. III. 2D error estimation and multilevel preconditioning*, IMPACT of Comput. in Sci. and Engrg., 4 (1992), pp. 1–45

[32] F. A. Bornemann and P. Deuflhard, *The cascadic multigrid method for elliptic problems*, Numer. Math., 75 (1996), pp. 135–152

[33] F. A. Bornemann, B. Erdmann, and R. Kornhuber, *Adaptive multilevel methods in three space dimensions*, Int. J. Num. Meth. Engrg., 36 (1993), pp. 3187–3203

[34] F. A. Bornemann, B. Erdmann, and R. Kornhuber, *A posteriori error estimates for elliptic problems in two and three space dimensions*, SIAM J. Numer. Anal., 33 (1996), pp. 1188–1204

[35] J. H. Bramble and S. R. Hilbert, *Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation*, SIAM J. Numer. Anal., 7 (1970), pp. 112–124

[36] D. Braess, *The contraction number of a multigrid method for solving the Poisson equation*, Numer. Math., 37 (1981), pp. 387–404

[37] K.E. Brenan, S.L. Campbell, and L.R. Petzold, *Numerical Solution of Initial–Value Problems in Differential–Algebraic Equations*, North–Holland, New York, 1989

[38] W.B. Bush and F.E. Fendel, *Asymptotic Analysis of Laminar Flame Propagation for General Lewis Number*, Combust. Sci. and Tech., 1 (1970), pp. 421–428

[39] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978

[40] P. CLEMENT, *Approximation by finite element functions using local regularization*, RAIRO Anal. Numér., R–2 (1975), pp. 77–84

[41] M. CROUZEIX, S. LARSSON, AND V. THOMÉE, *Resolvent estimates for elliptic finite element operators in one dimension*, Math. Comp., 63 (1994), pp. 121–140

[42] R. DAUTRAY AND J. L. LIONS, *Mathematical Analysis and Numerical Methods for Science and Technology II. Functional and Variational Methods*, Springer-Verlag, Berlin, Heidelberg, New York, 1988

[43] R. DAUTRAY AND J. L. LIONS, *Mathematical Analysis and Numerical Methods for Science and Technology V. Evolution Problems I*, Springer-Verlag, Berlin, Heidelberg, New York, 1992

[44] K. DEKKER AND J.G. VERWER, *Stability of Runge–Kutta methods for stiff nonlinear differential equations*, North–Holland Elsevier Science Publishers, 1984

[45] B. DENET AND P. HALDENWANG, *Numerical Study of Thermal-Diffusive Instability of Premixed Flames*, Combust. Sci. and Tech., 86 (1992), pp. 199–221

[46] A. DERVIEUX, B. LARROUTUROU, AND R. PEYRET, *On some Adaptive Numerical Approaches of Thin Flame Propagation Problems*, Computers and Fluids, 17 (1989), pp. 39–60

[47] P. DEUFLHARD, *Uniqueness theorems for stiff ODE initial value problems*, in: D.F. Griffiths and G.A. Watson (eds.), Numerical analysis 1989, Proceedings of the 13th Dundee Conference, pp. 74–87, Pitman Research Notes in Mathematics Series 228, Longman Scientific and Technical, 1990

[48] P. DEUFLHARD AND F. BORNEMANN, *Numerische Mathematik II. Integration gewöhnlicher Differentialgleichungen*, de Gruyter Lehrbuch, Berlin, New York, 1994

[49] P. DEUFLHARD, J. LANG, AND U. NOWAK, *Adaptive algorithms in dynamical process simulation*, in: H. Neunzert (ed.), Progress in Industrial Mathematics at ECMI'94, pp. 122–137, Wiley–Teubner, 1996

[50] P. DEUFLHARD, P. LEINEN, AND H. YSERENTANT, *Concepts of an adaptive hierarchical finite element code*, IMPACT of Comput. in Sci. and Engrg., 1 (1989), pp. 3–35

[51] J. DOUGLAS AND T. DUPONT, *Galerkin Methods for Parabolic Equations*, SIAM J. Numer. Anal., 7 (1970), pp. 575–626

[52] I. S. DUFF AND J. K. REID, *Some design features of a sparse matrix code*, ACM Trans. Math. Software, 5 (1979), pp. 18–35

[53] V. EIJKHOUT AND P. VASSILEVSKI, *The role of the strengthened Cauchy–Buniakowskii–Schwarz inequality in multilevel methods*, SIAM Review, 33 (1991), pp. 405–419

[54] B. ERDMANN, J. LANG, AND R. ROITZSCH, *KASKADE Manual, Version 2.0*, TR93–5, Konrad–Zuse–Zentrum für Informationstechnik Berlin, 1993

[55] K. ERIKSSON AND C. JOHNSON, *Adaptive finite element methods for parabolic problems IV: Nonlinear problems*, SIAM J. Numer. Math., 32 (1995), pp. 1729–1749

[56] P.M. FAHEY, P.B. GRIFFIN, AND J.D. PLUMMER, *Point defects and dopant diffusion in silicon*, Rev. Mod. Phys., 61 (1989), pp. 290–383

[57] R.B. FAIR AND J.C.C. TSAI, *A quantitive model for the diffusion of phosphorus in silicon and the emitter dip effect*, J. Electrochem. Soc., 124 (1977), pp. 1107–1118

[58] I. FRIED, *Condition of finite element matrices generated from nonuniform meshes*, AIAA J., 10 (1972), pp. 219–221

[59] J. FRÖHLICH AND J. LANG, *Two–dimensional cascadic finite element computations of combustion problems*, Comp. Meth. Appl. Mech. Eng., 158 (1998), pp. 255–267

[60] J. Fröhlich, J. Lang, and R. Roitzsch, *Selfadaptive finite element computations with smooth time controller and anisotropic refinement*, in: J.A. Desideri, P.Le. Tallec, E. Onate, J. Periaux, E. Stein (eds.), Numerical Methods in Engineering, 523-527, Wiley, New York, 1996

[61] J. Fröhlich and R. Peyret, *A Spectral Algorithm for Low Mach Number Combustion*, Comp. Meth. Appl. Mech. Eng., 90 (1991), pp. 631–642

[62] H. Fujita and T. Kato, *On the Navier–Stokes Initial Value Problem. I*, Arch. Rat. Mech. Anal., 16 (1964), pp. 269–315

[63] D. Fujiwara, *Concrete charakterization of the domains of the fractional powers of some elliptic differential operators of the second order*, Proc. Japan Acad., 43 (1992), pp. 82–86

[64] K. Ghaderi ang G. Hobler, *Simulation of phosphorus diffusion in silicon using pair diffusion model with a reduced number of parameters*, J. Electrochem. Soc., 142 (1995), pp. 1654–1658

[65] M. Garland and P.S. Heckbert, *Surface Simplification Using Quadric Error Metrics*, in Computer Graphics Proceedings, pp. 209-216, Addison Wesley, 1997

[66] G. H. Golub and C. F. van Loan, *Matrix Computations (second edition)*, The John Hopkins University Press, Baltimore and London, 1989

[67] M. E. Go Ong, *Hierarchical Basis Preconditioners for Second Order Elliptic Problems in Three Dimensions*, Ph.D. Thesis, University of California, Los Angeles, 1989

[68] P. Grisvard, *Charaktérisation de quelques espaces d'interpolation*, Arch. Rat. Mech. Anal., 25 (1967), pp. 40–63

[69] K. Gustafsson, *Control of error and convergence in ODE solvers*, PhD thesis, Department of Automatic Control, Lund Institute of Technology, Sweden, 1992

[70] K. Gustafsson, M. Lundh, and G. Söderlind, *A PI stepsize control for the numerical solution of ordinary differential equations*, BIT, 28 (1988), pp. 270–287

[71] E. Hairer, S. P. Norsett, and G. Wanner, *Solving ordinary differential equations I, Nonstiff Problems*, Springer–Verlag, Berlin, Heidelberg, New York, 1987

[72] E. Hairer and G. Wanner, *Solving ordinary differential equations II, Stiff and Differential-Algebraic Problems*, Springer–Verlag, Berlin, Heidelberg, New York, 1991

[73] H.-C. Hege, M. Seebass, D. Stalling, and M. Zöckler, *A Generalized Marching Cubes Algorithm Based On Non-Binary Classifications*, SC 97–05, Konrad–Zuse–Zentrum für Informationstechnik Berlin, Germany, 1997

[74] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Mathematics 840, Springer-Verlag, Berlin, Heidelberg, New York, 1981

[75] A. Höfler and N. Strecker, *On the coupled diffusion of dopants and silicon point defects*, Technical Report 94–11, Integrated Systems Laboratory, Swiss Federal Institute of Technology Zurich, 1994

[76] W. Huang and R.D. Russell, *A high dimensional moving mesh strategy*, Appl. Numer. Math., 26 (1998), pp. 63–76

[77] H. Jarausch, *On an adaptive grid refinement technique for finite element approximations*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 1105–1120

[78] A.K. Kapila, *Reactive–diffusive system with Arrhenius kinetics: dynamics and ignition*, SIAM J. Appl. Math., 39 (1980), pp. 21–36

[79] P. Kaps, S. W. H. Poon, and T. D. Bui, *Rosenbrock methods for stiff ODEs: a comparison of Richardson extrapolation and embedding technique*, Computing 34 (1985), pp. 17–40

[80] P. Kaps and G. Wanner, *A study of Rosenbrock–type methods of high order*, Numer. Math., 38 (1981), pp. 279–298

[81] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, Heidelberg, New York, 2nd ed., 1984

[82] A. KOTTE, J. VAN LEEUWEN, J. DE BREE, J. VAN DER KOIJK, H. CREZEE, AND J. LAGENDIJK, *A description of discrete vessel segments in thermal modelling of tissues*, Phys. Med. Biol., 41 (1996), pp. 865–884

[83] J. LANG, *High–resolution selfadaptive computations on chemical reaction–diffusion problems with internal boundaries*, Chem. Engrg. Sci., 51 (1996), pp. 1055–1070

[84] J. LANG, *Adaptive FEM for reaction–diffusion equations*, Appl. Numer. Math., 26 (1998), pp. 105–116

[85] J. LANG, B. ERDMANN, AND M. SEEBASS, *Impact of nonlinear heat transfer on temperature control in regional hyperthermia*, SC97–73, Konrad–Zuse–Zentrum für Informationstechnik Berlin, 1997 (accepted for publication in IEEE Trans. Biomedical Engrg.)

[86] J. LANG AND W. MERZ, *Numerical simulation of single species dopant diffusion in silicon under extrinsic conditions*, SC 97–47, Konrad–Zuse–Zentrum für Informationstechnik Berlin, 1997

[87] J. LANG AND W. MERZ, *Dynamic mesh design control in semiconductor device simulation*, to be published in Proceedings of Int. Conference on Systems, Signals, Control, and Computers, Durban, 1998

[88] J. LANG AND A. WALTER, *A finite element method adaptive in space and time for nonlinear reaction–diffusion systems*, IMPACT of Comput. in Sci. and Engrg., 4 (1992), pp. 269–314

[89] J. LANG AND A. WALTER, *An adaptive Rothe method for nonlinear reaction–diffusion systems*, Appl. Numer. Math., 13 (1993), pp. 135–146

[90] I. LASIECKA, *Convergence estimates for semidiscrete approximations of nonselfadjoint parabolic equations*, SIAM J. Numer. Anal., 21 (1984), pp. 894–909

[91] M. LENOIR, *Optimal isoparametric finite elements and error estimates for domains involving curved boundaries*, SIAM J. Numer. Anal., 23 (1986), pp. 562–580

[92] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, Springer-Verlag, Berlin, Heidelberg, New York, 1972

[93] W.E. LORENSEN AND H.E. CLINE, *Marching Cubes: A high resolution 3D surface construction algorithm*, Computer Graphics, 21 (1987), pp. 163-169

[94] CH. LUBICH AND A. OSTERMANN, *Runge–Kutta approximation of quasi–linear parabolic equations*, Math. Comp., 64 (1995), pp. 601–627

[95] CH. LUBICH AND A. OSTERMANN, *Linearly implicit time discretization of non–linear parabolic equations*, IMA J. Numer. Anal., 15 (1995), pp. 555–583

[96] CH. LUBICH AND M. ROCHE, *Rosenbrock methods for differential–algebraic systems with solution–dependent singular matrix multiplying the derivative*, Comput., 43 (1990), pp. 325–342

[97] A. LUNARDI, *Analytic semigroups and optimal regularity in parabolic problems*, Birkhäuser-Verlag, Basel, Boston, Berlin, 1997

[98] J. F. MAITRE AND F. MUSY, *The contraction number of a class of two level methods; an exact evaluation for some finite element subspaces and model problems*, Lecture Notes in Mathematics 960, Proceedings of Multigrid Methods, Cologne 1981, pp. 535–544, Springer–Verlag, Heidelberg, 1982

[99] D. MATHIOT AND J.C. PFISTER, *Dopant diffusion in silicon: A consistent view involving nonequilibrium defects*, J. Appl. Phys., 55 (1984), pp. 3518–3530

[100] B.J. MATKOWSKY AND V. VOLPERT, *Spiral Gasless Condensed Phase Combustion*, SIAM J. Appl. Math., 54 (1994), pp. 132–146

[101] W. MERZ, *Analysis und Numerische Berechnung der Diffusion von Fremdatomen in Homogenen Strukturen*, Habilitationschrift, Zentrum Mathematik TU München, 1998

[102] K. MILLER AND R.N. MILLER, *Moving finite elements*, SIAM J. Appl. Math., 18 (1981), pp. 1019–1032

[103] P.K. Moore, *A posteriori error estimation with finite element semi– and fully discrete methods for nonlinear parabolic equations in one space dimension*, SIAM J. Numer. Anal., 31 (1994), pp. 149–169

[104] K.S. Nikita, N.G. Maratos, and N.K. Uzunoglu, *Optimal Steady–State Temperature Distribution for a Phased Array Hyperthermia System*, IEEE Trans. Biomed. Engrg., 40 (1993), pp. 1299–1306

[105] S. P. Norsett and A. Wolfbrandt, *Attainable order of rational approximations to the exponential function with only real poles*, BIT, 17 (1977), pp. 200–208

[106] U. Nowak, *Adaptive Linienmethoden für nichtlineare parabolische Systeme in einer Raumdimension*, Ph.D. Thesis, Free University Berlin, 1993

[107] A. Ostermann and M. Roche, *Runge–Kutta methods for partial differential equations and fractional orders of convergence*, Math. Comp., 59 (1992), pp. 403–420

[108] A. Ostermann and M. Roche, *Rosenbrock methods for partial differential equations and fractional orders of convergence*, SIAM J. Numer. Anal., 30 (1993), pp. 1084–1098

[109] C. Palencia, *On the stability of variable stepsize rational approximations of holomorphic semigroups*, Math. Comp., 62 (1994), pp. 93–103

[110] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, Heidelberg, New York, 1983

[111] H.H. Pennes, *Analysis of tissue and arterial blood temperatures in the resting human forearm*, J. Appl. Phys., 1 (1948), pp. 93–122

[112] N. Peters, *Discussion of Test Problem A*, in N. Peters and J. Warnatz (eds.), Numerical Methods in Laminar Flame Propagation, Notes on numerical fluid mechanics, Vol. 6, pp. 1–14. Vieweg, 1982.

[113] N. Peters and J. Warnatz, *Numerical Methods in Laminar Flame Propagation*, Notes on Numerical Fluid Mechanics, Vol. 6, Vieweg, 1982

[114] A. Prothero and A. Robinson, *On the stability and accuracy of one–step methods for solving stiff systems of ordinary differential equations*, Math. Comp., 28 (1974), pp. 145–162

[115] W.B. Richardson and B.J. Mulvaney, *Nonequilibrium behaviour of charged point defects during phosphorus diffusion in silicon*, J. Appl. Phys., 65 (1988), pp. 2243–2247

[116] M. Roche, *Rosenbrock methods for differential algebraic equations*, Numer. Math., 52 (1988), pp. 45–63

[117] R. Roitzsch, B. Erdmann, and J. Lang, *The Benefits of Modularization: from KASKADE to KARDOS*, Proceedings of the 14th GAMM-Seminar Kiel on Concepts of Numerical Software, 1998

[118] H. H. Rosenbrock, *Some general implicit processes for the numerical solution of differential equations*, Computer J. (1963), pp. 329–331

[119] W. Rudin, *Functional Analysis*, McGraw-Hill Publishing Co., New York, London, 1973

[120] W. Ruppel, *Entwicklung von Simulationsverfahren für die Reaktionstechnik*, manuscript, BASF research, 1993

[121] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving non–symmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869

[122] A. Sandu, J.G. Verwer, J.G. Blom, E.J. Spee, G.R. Carmichael, and F.A. Potra, *Benchmarking stiff ODE solvers for athmospheric chemistry problems II: Rosenbrock solvers*, Atmos. Environ., 31 (1997), pp. 3459–3472

[123] J. M. Sanz–Serna and J. G. Verwer, *Stability and convergence at the PDE/stiff ODE interface*, Appl. Numer. Math., 5 (1989), pp. 117–132

[124] J. M. Sanz–Serna, J. G. Verwer, and W. H. Hundsdorfer, *Convergence and order reduction of Runge–Kutta schemes applied to evolutionary problems in partial differential equations*, Numer. Math., 50 (1986), pp. 405–418

[125] G. SAVARÉ, *A(Θ)–stable approximations of abstract Cauchy problems*, Numer. Math., 65 (1993), 319–335

[126] G. SAVARÉ, *Private communication*, 1998

[127] S. SCHOLZ, *Order barriers for the B–convergence of ROW methods*, Computing, 41 (1989), pp. 219–235

[128] L. R. SCOTT AND S. ZHANG, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493

[129] M. SEEBASS, D. STALLING, J. NADOBNY, P. WUST, R. FELIX, AND P. DEUFLHARD, *Three-Dimensional Finite Element Mesh Generation for Numerical Simulations of Hyperthermia Treatments*, in C. Franconi, G. Arcangeli, and R. Cavaliere (eds.), Proc. 7th Int. Congress on Hyperthermic Oncology, Roma, Vol. 2, pp. 547-548, 1996

[130] L. F. SHAMPINE, *Implementation of Rosenbrock Methods*, ACM Trans. Math. Software, 8 (1982), pp. 93–113

[131] M.D. SMOOKE AND M.L. KOSZYKOWSKI, *Two-Dimensional Fully Adaptive Solutions of Solid-Solid Alloying Reactions*, J. Comp. Phys., 62 (1986), pp. 1–25

[132] C.W. SONG, A. LOKSHINA, J.G. RHEE, M. PATTEN, AND S.H. LEVITT, *Implication of Blood Flow in Hyperthermic Treatment of Tumors*, IEEE Trans. Biomed. Engrg., 31 (1984), pp. 9–16

[133] P. SONNEVELD, *CGS: A fast Lanczos–type solver for solving nonsymmetric linear systems*, J. Sci. Stat. Comput., 10 (1989), pp. 36–52

[134] T. STEIHAUG AND A. WOLFBRANDT, *An attempt to avoid exact Jacobian and nonlinear equations in the numerical solution of stiff ordinary differential equations*, Math. Comp., 33 (1979), pp. 521–534

[135] G. STEINEBACH, *Order–reduction of ROW–methods for DAEs and method of lines applications*, Preprint 1741 (1995), Technische Hochschule Darmstadt, Germany

[136] K. STREHMEL AND R. WEINER, *B–convergence results for linearly implicit one step methods*, BIT, 27 (1987), pp. 264–281

[137] K. STREHMEL AND R. WEINER, *Linear–implizite Runge–Kutta–Methoden und ihre Anwendungen*, Teubner Texte zur Mathematik 127, Teubner Stuttgart, Leipzig, 1992

[138] H. J. STETTER, *Tolerance proportionality in ODE-codes*, in R. März (ed.), Proceedings of the Second Conference on Numerical Treatment of Ordinary Differential Equations, Vol. 32 of Seminarberichte, Humboldt University, Berlin, 1980

[139] V. THOMÉE, *Galerkin Finite Element Methods for Parabolic Problems*, Springer Series in Computational Mathematics, Springer-Verlag, Berlin, Heidelberg, New York, 1997

[140] D.T. TOMPKINS, R. VANDERBY, S.A. KLEIN, W.A. BECKMAN, R.A. STEEVES, D.M. FREY, AND B.R. PALIVAL, *Temperature–dependent versus constant–rate blood perfusion modelling in ferromagnetic thermoseed hyperthermia: results with a model of the human prostate*, Int. J. Hyperthermia, 10 (1994), pp. 517–536

[141] R.A. TROMPERT AND J.G. VERWER, *A static regridding method for two–dimensional parabolic partial differential equations*, Appl. Numer. Math., 8 (1991), pp. 65–90

[142] H. A. VAN DER VORST, *BI–CGSTAB: A fast and smoothly converging variant of BI–CG for the solution of nonsymmetric linear systems*, SIAM J. Sci. Stat., 13 (1992), pp. 631–644

[143] A. VANDE WOUWER, P. SAUCEZ, AND W.E. SCHIESSER, *Some user–oriented comparisons of adaptive grid methods for partial differential equations in one space dimension*, Appl. Numer. Math., 26 (1998), pp. 49–62

[144] R. VERFÜRTH, *A posteriori error estimates and adaptive mesh refinement techniques*, Teubner Scripten zur Numerik, B. G. Teubner, Stuttgart, 1995

[145] R. VERFÜRTH, *Robust a posteriori error estimators for a singularly perturbed reaction–diffusion equation*, Numer. Math., 78 (1998), pp. 479–493

[146] R. VERFÜRTH, *A posteriori error estimates for non-linear problems. $L^r(0,T;L^r)$–error estimates for finite element discretizations of parabolic equations*, Math. Comput., 67 (1998), pp. 1335–1360

[147] J.G. VERWER, *Convergence and order reductions of diagonally implicit Runge–Kutta schemes in the method of lines*, in: D.F. Griffiths, G.A. Watson (eds.), Numerical Analysis, pp. 220–237, Pitman Research Notes in Mathematics, Boston, 1986

[148] J. G. VERWER, E. J. SPEE, J. G. BLOM, AND W. H. HUNDSDORFER, *A second order Rosenbrock method applied to photochemical dispersion problems*, Report MAS–R9717, CWI Amsterdam, The Netherlands, 1997

[149] G. WANNER, *On the choice of γ for singly–implicit RK or Rosenbrock methods*, BIT, 20 (1980), 102–106

[150] S. WEINBAUM AND L.M. JIJI, *A new simplified bio–heat equation for the effect of blood flow on local average tissue temperature*, J. Biomech. Engrg. Trans. ASME, 107 (1985), pp. 131–139

[151] G.B. WHITHAM, *Linear and nonlinear waves*, Wiley–Interscience, New York, 1974

[152] A. WOLFBRANDT, *A study of Rosenbrock processes with respect to order conditions and stiff stability*, Ph.D. Thesis, Chalmers University of Technology, Göteburg, Sweden, 1977

[153] P. ZEGELING, *r–refinement for evolutionary PDEs with finite elements or finite differences*, Appl. Numer. Math., 26 (1998), pp. 97–104

[154] S. ZHANG, *Multilevel Iterative Techniques*, Ph.D. Thesis, Pennsylvania State University, 1988

[155] O. C. ZIENKIEWICZ, D. W. KELLEY, J. DE S. R. GAGO, AND I. BABUŠKA, *Hierarchical finite element approaches, adaptive refinement, and error estimates*, The Mathematics of Finite Elements and Applications, Academic Press, New York, pp. 313–346, 1982